Clemens Günther Hofer, BSc

Factors of Success in Crowdfunding Campaigns

Master's Thesis

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Software Engineering and Management

submitted to

Graz University of Technology

Supervisor

Dipl.-Ing. Dr.techn. Simon Walk, BSc

Institute of Interactive Systems and Data Science
Head: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Graz, May 2018

This document is set in Palatino, compiled with pdfLaTeX2e and Biber.

The LaTeX template from Karl Voit is based on KOMA script and can be found online: https://github.com/novoid/LaTeX-KOMA-template

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date	Signature

Abstract

Crowdfunding represents a new, but important source of financing. Numerous crowdfunding platforms have emerged in recent years. Although crowdfunding is used by many people and even by well-known companies to finance their project ideas, very little is known about the factors that lead to success of crowdfunding campaigns. This is reflected by the fact that only 40% of the campaigns on one of the most popular platforms, "*Kickstarter*", reach their initial goal. To improve our understanding of success in Crowdfunding, we collect information about campaigns finished between 2009 and 2016, as well as roughly 39,000 recently launched campaigns which additionally include time series for each campaign. We use the empirical data to evaluate the anatomy and mechanics of Kickstarter, for example when most money is pledged during the funding period. Based on the information from our analysis, we build classification and regression models, which we use to not only predict if a campaign will reach its goal, but also how much money the campaign will collect. The results show that we are able to accurately predict if a campaign will receive enough money to reach the initial goal as well as how much money will be collected by a campaign in the end. We found that the first hours of the funding period are essential for the success of a campaign.

Kurzfassung

Crowdfunding stellt eine neue, aber wichtige Finanzierungsquelle dar. Zahlreiche Plattformen sind in den letzten Jahren entstanden. Obwohl Crowdfunding von vielen Menschen und sogar bekannten Unternehmen genutzt wird, um ihre Projekt-Ideen zu finanzieren, ist sehr wenig über die Erfolgsfaktoren bei Crowdfunding-Kampagnen bekannt. Das zeigt sich in der Tatsache, dass nur 40% der der Kampagnen auf einer der bekanntesten Plattformen, "*Kickstarter*", das gesetzte Finanzierungsziel erreichen. Um die Materie um Crowdfunding besser zu verstehen, sammeln wir Informationen über bereits vergangene Kampagnen zwischen 2009 und 2016, sowie ca. 39.000 kürzlich veröffentlichte Kampagnen, welche zusätzlich Zeitreihen für jede einzelne Kampagne beinhalten. Die daraus resultierenden empirischen Daten benutzen wir, um die Anatomie und die Mechaniken hinter Kickstarter zu erforschen, zum Beispiel wann das meiste Geld während der Finanzierungsperiode gesammelt wird. Basierend auf diesen Informationen haben wir verschiedene Klassifikations- und Regressionmodelle erstellt, welche nicht nur vorhersagen, ob eine Kampagne ihr Finanzierungsziel erreichen wird, sondern auch wie viel Geld die Kampagne letztendlich erreichen wird. Die Ergebnisse zeigen, dass wir relativ genau vorhersagen können, ob eine Kampagne genug Geld sammeln wird, um ihr Finanzierungsziel zu erreichen und außerdem wie viel Geld insgesamt am Ende eingesammelt wird. Wir haben herausgefunden, dass die ersten Stunden der Finanzierungsperiode essentiell für den Erfolg einer Kampagne sind.

Contents

List of Figures

List of Tables

1. Introduction

Crowdfunding is a new source of financing with increasing popularity over the last years. One of the most popular platforms in this area is Kickstarter[1], which was founded in 2009. Since then, over 14 million people backed crowdfunding campaigns spending a total of roughly USD 3.5 billions. As the number of campaigns and the amount of money pledged still has risen since then, crowdfunding is not only used by enthusiasts or experts any more. Even popular companies, for example Cannonical, the company behind the Linux distribution Ubuntu[2], discovered crowdfunding as an opportunity apart from the conventional sources of financing. Other companies established only because of a successful crowdfunding campaign, for example Pebble, which created the an e-paper smartwatch.

In general, crowdfunding is used as financing source for various kinds of projects such as technical products, social projects or art projects. One advantage of crowdfunding over traditional funding channels is that the campaign creators receive direct feedback from the crowd if the idea behind the campaign has the potential to succeed in the market. Nevertheless, only roughly 40% of the campaigns on Kickstarter are funded successfully.

In our work, we investigate the factors that are important for the success of a campaign. For this reason, we examine the anatomy and mechanics of Kickstarter. This includes information about where campaigns are launched, how large the campaigns are or when most money is pledged. This information represents the foundation that allows us to investigate the factors that influence the success of certain projects as well as why some campaigns are more successful than others.

[1] **www.kickstarter.com**.
[2] **www.ubuntu.com**.

Furthermore, we want to predict the success of a campaign. More precisely, we want to predict if a campaign is likely to reach its goal and, in particular, how much money a campaign will collect in the end. Apart from that, we investigate how much data we need to make an accurate prediction.

To answer these questions we build a crawler to collect data from Kickstarter. For our work we crawl finished and live campaigns. For the empirical analysis we use data of finished campaigns from 2009 to 2016, which help us to characterize the anatomy of Kickstarter.

The data of the live campaigns include frequent updates, representing time series, which allow us to investigate the development of each campaign. Then, we create datasets with different combinations of features, including several different time series, information about the creator, descriptions, comments, updates and many more. With these datasets we conduct different classification and regression experiments.

The classification experiments aim to predict if a campaign will reach its funding goal. We build models with Naive Bayes, Support Vector Machines and Logistic Regression. Finally we compare the results of the experiments with existing work to determine how well our models perform.

The goal of the regression experiments is to predict how much money a campaign will collect. Hence, we conduct experiments with Linear Regression, Ridge Regression and Kernel Ridge Regression with polynomial kernel. To the best of our knowledge there is no similar published work where regression analyses are applied on crowdfunding campaign data.

The evaluation of the empirical data provides novel insights into the anatomy of Kickstarter. We provide statistics on different aspects of crowdfunding campaigns. We investigate how high the funding goals are in general and how much money campaigns collect in the end. In this context we provide information when most money is pledged during the funding period of a campaign and we examine the gap between successful and failed campaigns. We also compare campaigns of different categories with each other to investigate if certain findings apply to campaigns in all categories or if they behave differently. Furthermore, we present statistics on how different aspects have developed over the years.

For the classification and regression experiments we crawl datasets which consist of roughly 39,000 live campaigns, including time series, and roughly 287,000 finished campaigns. The results of the classification experiments demonstrate which features and how much data is needed to make an accurate prediction.

We believe that our work helps to improve the understanding of crowdfunding and, in particular, of Kickstarter. The classification and regression experiments show that it is possible to accurately predict the success and the amount of money collected by a campaign only a few hours after the start of a campaign. We think that the findings in our work could provide assistance for potential campaign creators to improve the chances of success.

1.1. Contributions

The work consisted of different work packages. The first step was the creation of a web crawler for live and finished campaigns on Kickstarter. The web crawling itself is still in progress and comprises more than 300,000 campaigns which is more than in the most other published work. Then we made an empirical analysis on the dataset for which we mainly used the data of finished campaigns. This analysis gives new insights on crowdfunding and in particular on Kickstarter. Finally, the main task of the work was the prediction of campaign outcomes wherefore we conducted different classification and regression experiments.

These mentioned work packages were mainly fulfilled by the author of this thesis. The design and implementation of the web crawler and also the creation of the statistics were also carried out by the author of this thesis. The interpretation of the empirical analysis were conducted in collaboration with Simon Walk. The feature engineering and the prediction experiments were also in my responsibility.

2. Related Work

Crowdfunding represents a new, but important source of financing. This is evident in the rising number of newly launched platforms over the course of the last years. Researchers have analyzed various aspects of and around the crowdfunding processes. However, most of the existing work relates to larger and well-established platforms such as Kickstarter[1] and Indiegogo[2].

In general, literature on crowdfunding can be divided into research on anatomy of crowdfunding platforms and the investigation of factors that make a campaign successful.

Specifically, we discuss related work that examines the anatomy of crowdfunding such as the different types of crowdfunding, typical topics of campaigns, or motivations for creators and backers to become active on such platforms.

We will then discuss work that analyzes economical aspects of crowdfunding campaigns to provide assistance to founders when creating their own crowdfunding campaign.

Finally we discuss work that inspired the research that is presented in this thesis. In particular, we will look at research articles where datasets were used to predict the success of a campaign. We conclude this section by highlighting our contributions to the research field of crowdfunding.

[1] **www.kickstarter.com**.
[2] **www.indiegogo.com**.

2.1. Anatomy of Crowdfunding

One of the most popular papers in this area was published by Mollick [Mol14]. The author built a model with basic features of the campaign such as funding goal, number of backers, or category. Then he calculated the correlations between the features of the model. One of the main findings was that the goal and the success rate have a negative correlation, meaning campaigns with higher funding goals have lower success rates. Another finding was that a large social network, for example, on Facebook[3], of the creator has a positive impact on the success of the campaign.

Another empirical study by Mollick and Kuppuswamy [MK14] focused on the outcomes of crowdfunding campaigns. They asked creators of campaigns in the categories "*technology*", "*video games*" and "*design*" what happened with the project after the end of the crowdfunding campaign. They found that over 90% of the organizations behind successful projects still exist. The average age of the campaign creators was 35 with 70% of the them disagreeing that most of their backers are family and friends. The reasons for failure, from the viewpoint of the creators, mainly laid in inadequate marketing, wrong target audience, and bad campaign descriptions. The study showed that less than 10% of the creators are discouraged and stop the project after a failed campaign with over 60% of them still working on the initial idea behind the project.

Greenberg and Gerber [GG14] concentrated on campaigns that failed on their first attempt in their analysis. It revealed that only roughly 2% of those which failed were relaunched. After the relaunch, 43% of these campaigns were successful, which is close to the average success rate on Kickstarter. The average time between the end and the relaunch of the same project was around 37 days. They also found that creators often reduce the funding goal and adapt the pitch video.

Zvilichovsky et al. [ZIB15] investigated reciprocity on Kickstarter. In the context of this analysis reciprocity describes the willingness of a user to fund other campaigns if his/her own campaign has been funded and vice versa. In order to examine the user relations, they crawled 78,061

[3] **www.facebook.com**.

campaigns, including profile and backing-history of the involved users. They distinguished between direct and indirect reciprocity. Direct reciprocity refers to the situation when two specific users back the campaigns of each other, while indirect reciprocity describes that campaigns from users are more likely to be funded due to the fact that these users backed other campaigns, although they did not back campaigns of their own backers. They reported that both forms of reciprocity exist on Kickstarter. Users that backed other campaigns had a success rate of 61.8% for their own campaigns, while users without backing history only reached a success rate of 48.6%. A look at the total numbers revealed that campaigns of users with a backing history had an average goal of about USD 8,000 and reached on average USD 13,500 in comparison to non-backing users with an average goal of USD 5,100 and USD 6,900 money raised.

Agrawal et al. [ACG11] investigated where backers geographically come from and how different groups of backers behave depending on their location. They analyzed the platform "*Sellaband*[4]", which is a crowdfunding platform for musicians who look for funding in order to finance a record. The average distance between the artist and the investor was about 3,000 miles, with local investors exhibiting different behaviors than others. Specifically local investors were more likely to invest the first USD 20,000. Additionally, they found that "*Family & Friends*" investors are disproportionately local. Agrawal et al. [ACG15] came to similar results.

Giudici et al. [GGR13] analyzed the connections of crowdfunding campaigns to Facebook in order to determine the territorial and individual social capital. They define social capital as the goodwill that is gained from the structure and content of social relations. For their study they investigated 11 Italian crowdfunding platforms and tried to identify if the social capital of a campaign-creator has an impact on the success. For this reason, they introduced variables to determine the individual and the territorial social capital and performed a probit regression. The geo social capital consists of statistical factors in an area such as "*social meetings*", "*money given to associations*" or "*participation in political meetings*". Individual social capital measures the number of Facebook contacts (in logarithm) of a creator. The results show that the individual social capital had a positive impact on

[4] **www.sellaband.com**.

the success, while the influence of the territorial social capital was not significant.

Gerber et al. [GHK12] made an analysis to investigate why people are motivated to create and fund campaigns on crowdfunding platforms. They conducted semi-structured interviews with 30 project backer and creator pairs. In the interviews the authors asked questions about the background of the people, the motivation to participate in crowdfunding, and the influence of crowdfunding for their work. The results show that the main motivation for creators were to raise funds, to receive validation for their abilities, to replicate successful experience of others, and to expand awareness of projects through social media. The main motivations for backers were the rewards themselves, the support of creators, and the contribution to a creative community.

In the work of Crosetto and Regner [CR14], they tried to identify the determinants of campaign success by examining the dynamics of crowdfunding. They analyzed 2,711 campaigns on the platform "*Startnext*[5]", which is the biggest crowdfunding platform in Germany. The campaign duration on this platform is split into a starting and a funding phase. In the starting phase a previously defined number of fans has to be collected before the campaign advances into the funding phase, where the crowdfunding itself takes place. According to Mollick [Mol14] the size of the funding goal was negatively correlated with success, while the presence of videos, images, and the number of blog entries on the campaign site positively correlated. Another result of the study shows that more money was pledged right at the beginning and towards the end of a campaign. There was a substantial gap between failed and successful campaigns in terms of collected funds. In general, there were only few campaigns that barely missed their goal, but a lot of campaigns that rose very little to no money or reached their goal without excessive overfunding.

Xu et al. [Xu+14] considered the importance of campaign updates for the success on Kickstarter. Campaign updates are a way to keep potential backers in the loop about the proceedings of a campaign, both during and after the crowdfunding period. Therefore, updates are means of communication for the campaign creators. The authors found that campaigns without any

[5] **www.startnext.com**.

updates had a success rate of 32.6%, while campaigns with updates had a success rate of 58.7%. They used Latent Dirichlet Allocation (LDA) in order to find out about the topics of updates. They divided updates into six different categories: "*social promotion*", "*progress report*", "*new content*", "*reminder*", "*answer question*", "*new reward*" and "*appreciation*". Additionally, they split the campaign duration into three phases and claim that the distribution of these update categories changed during the campaign. Updates of the biggest category "*social promotion*" occurred regularly at the beginning and at the end of a campaign. Progress reports and new content appeared more often at the beginning while reminders were posted frequently towards the end of the campaign.

Xiaoyu et al. [Xia+17] examined factors that lead to overfunding on "*Crowdcube*[6]". Crowdcube is a British crowdfunding platform which uses equity crowdfunding in contrast to Kickstarter or Indiegogo. In equity crowdfunding creators offer shares or securities of their company on crowdfunding platforms. Due to this form of crowdfunding the factor of risk comes into existence for the investor. They found that the relation between money raised and the funding were higher for campaigns with higher targets. Consequently this also applied to the number of investors. Another finding was that the sharing of pitch videos on LinkedIn[7] had a positive impact on the success, while there was no such relation for videos shared on Facebook.

The paper from Trindade et al. [TSC17] investigated determinants of success of campaigns in the areas of music and sports. The authors examined a Portuguese platform called "*PPL*[8]" and their dataset included 171 campaigns from 2011 to 2016. The authors used linear regression in order to determine differences between determinants of success of campaigns from the two categories music and sports. The results show that the connection to Facebook and the number of images had a positive impact on campaigns in the sport category, but not on the success of campaigns in the music category. In contrast, the success of campaigns in the music category was positively influenced by the views of the campaign page and negatively influenced by higher funding goals.

[6] **www.crowdcube.com**.
[7] **www.linkedin.com**.
[8] **www.ppl.com.pt**.

Kuppuswamy and Bayus [KB15] investigated the dynamics of backers on the crowdfunding platform Kickstarter. The dataset for the empirical study included campaigns from May 2009 to February 2012, but only campaigns with at least one backer were considered in the study. They found that the average goal per campaign was around USD 9,000, but received only roughly USD 2,100, with an average amount pledged per backer around USD 70. A key finding of the study was that most of the successful campaigns reached their goal by a small margin, while failed campaigns normally missed their goal clearly. The authors also claim that campaigns received more pledges on weekdays rather than on weekends. They also mentioned that campaigns attract more backers right after the start and towards the end of the campaign. The effect was maybe be caused by the contribution of "*Family & Friends*". Another potential explanation for these two effects is the search function of Kickstarter. It offers the options "*Recently launched*" and "*Ending soon*", which made those campaigns more visible to potential backers. The authors also found that campaign updates are posted more often in early and late stages of the campaign.

Lin et al. [LLC16] investigated the role of rewards on Kickstarter. Therefore, they analyzed a dataset of roughly 3,000 campaigns. The statistical analysis showed that the number of rewards, limited rewards and rewards that are added close to the end of the campaign had a positive correlation with the success.

Also in 2016, Hong et al. [HYK16] investigated different factors of success on Kickstarter. They examined 755 campaigns that were completed in 2015. The authors used multiple regression to calculate correlations between the features. They found that prior funding experience of the creator increased the chances of success. The number of comments and updates on the campaign page were also positively related to campaign success.

Inbar and Barzilay [IB14] investigated the impact of communities on Kickstarter. For this reason, they collected data of 6.6 million users and about 160,000 completed campaigns. For their analysis the authors identified three different types of communities: "*project-centered*", "*category-centered*" and "*platform-centered*". The user membership between those categories is hierarchic and variable. Users start their backing history in a project-centered community when they back their first campaign. This type of community

focuses on one specific campaign or campaigns from a specific owner. The authors stated that 95% of the users on Kickstarter created their account to back one specific campaign. Category-centered communities are centered around different categories or topics. Members of this group are normally formerly members of a project-centered community became interested in other campaigns of the same topic. Platform-centered communities consist of members, who back campaigns across different categories and topics. Additionally, the authors identified "*category-diverged*" users as another type. These are users which were category-centered before and began to broaden their interest to other categories. They claimed that a herding-effect occurs once a campaign has raised about 40% of its goal. Herding, in this context, refers to people being influenced by the the decisions of other people, for example if many people back a certain campaign, other people are more likely to back the same campaign, too. In order to investigate the impact of these different types of communities on success, the authors applied a logistic regression with basic campaign features, creator attributes (e.g. Facebook friends), and the ratio of the introduced community types. The results show that a higher ratio of category-centered backers had a positive effect on the success of a campaign, while a higher number of category-diverged users even had a negative influence on the success.

2.2. Economy of Crowdfunding

Belleflamme et al. [BLS14] observed that it is important to reach the right audience for a crowdfunding campaign. Hence, they compared crowdfunding with pre-ordering and profit-sharing. Pre-ordering is a form of reward-based crowdfunding. They claimed that different campaigns are better suited for one of those types than others. They suggested that pre-ordering schemes are better if the funding needs of the creator are lower and profit-sharing schemes are better if those needs are higher. Furthermore, a profit-sharing scheme was preferred if there was an information asymmetry between the entrepreneur and the customers, for example, in an early stage of product development.

Koch and Siering [KS15] provided an empirical study about the factors that

lead to the success of a campaign. They hypothesized that different factors such as media richness, campaign updates, goal, or creator experience have a positive impact on the success of a campaign. Therefore, they created variables that describe these attributes in the campaign and calculated the correlations between them. They found that on the one hand the provision of video material, the presence of campaign updates, the creator experience, and the funding reciprocity had a positive impact on the campaign success and on the other hand that the funding goal amount had a negative impact on the campaign success.

The impact of positive language patterns is the topic of the work from Lins et al. [LFL16]. The focus of their work was impression management. Impression management is a process where people try to alter the image others have of them. Therefore, they tried to connect the number of backers and the language that is used with the help of logistic regression. Their results show that success was influenced by words and phrases which are related to innovativeness and positiveness. In this work the authors stated that too much emphasis on innovativeness could also have a negative influence on the success of a campaign.

The paper from Belleflamme et al. [BLS13] deals with individual crowdfunding practices, meaning crowdfunding campaigns that do not use an established third-party platform such as Kickstarter or Indiegogo. They pointed out that crowdfunding, apart from these platforms, provides entrepreneurs with the freedom to perfectly tailor the campaign towards their needs. Therefore, every campaign differs from another, which forced Belleflamme et al. [BLS13] to collect the data for this study by hand. Hence, their dataset consisted of only 69 campaigns. The average funding goal of these campaigns was roughly 150,000 €, which is significantly higher than the average goal on Kickstarter. Additionally, they found out that campaigns from nonprofit organizations had a higher chance reaching their funding goal than profit orientated companies.

The target audience of the work from Xiao et al. [Xia+14] are potential campaign creators. They investigated three different aspects of a crowdfunding campaign, namely the campaign homepage, the communication of the creator, and the reward scheme. They found that the presence of a video on the campaign homepage had a positive effect on the success of a

campaign. The communication in form of campaign updates and comments on the campaign page was also important. Regarding the reward scheme they found that a higher backing price in a campaign's reward stream had a positive impact and a higher number of different rewards had a negative influence on the campaign success.

Beier and Wagner [BW14] discussed the timing in crowdfunding campaigns. They analyzed a Swiss crowdfunding platform called "*100-days.net*". In contrast to most other platforms the campaign duration is fixed at 100 days. The empirical study analyzed 492 campaigns on the platform to identify the best starting day for a campaign and when the most money is pledged. They found that Mondays and Tuesdays were the best weekdays to start a campaign. The biggest increase in backers was on Mondays, but most money was pledged on Fridays. The fixed campaign duration on this platform leads to the fact that the weekday of the campaign end is always two days after the weekday of the start. If we consider that most campaigns on Kickstarter have a campaign duration of 30 days, we can see that the same weekday constellation appears here, which makes the results comparable.

Another analysis of 100-days.net in 2015 was conducted by Beier et al. [BWC15]. They tried to examine the impact of different marketing strategies in crowdfunding campaigns. Therefore, they distinguished between "*on-page*" and "*off-page*" marketing. On-page-marketing describes the different measures that are used on the crowdfunding platform itself, while off-page-marketing is about all other means of marketing that are related to a certain campaign, but not on the platform itself (e.g. a separate Website or Facebook). They built a model with basic campaign features and social-media related features such as linkage to Facebook and Twitter[9] or the presence of a Website. The results of the regression analysis show that the richness of media usage was related to the success of a campaign.

[9] **www.twitter.com**.

2.3. Prediction of Crowdfunding Campaigns

Only a limited number of researchers have published work on predicting the success of crowdfunding campaigns [EGT13; KC16; Rao+14; Gre+13; LRR16].

Etter et al. [EGT13] tried to connect the success of a campaign with the social network behind the creator. They crawled campaigns from Kickstarter and additionally all tweets from Twitter that mention the campaign. Then they extracted features such as the number of tweets and the number of different users that tweeted about a campaign. They tried to predict whether a certain campaign reaches its goal using support vector machines with RBF-Kernel or not and reached an accuracy of 76% already 4 hours after the campaign was launched. Li et al. [LRR16] also used data from Kickstarter and social networks (Twitter and Facebook). They built a model for predicting the success of crowdfunding camapaigns with log-logistic regression. They included the first 3 days of campaign progress in their dataset and achieved the best results with an accuracy of 90.03%. In contrast to Etter et al. [EGT13] and Li et al. [LRR16] we only use data that is available on Kickstarter for the prediction experiments in our work.

Mitra and Gilbert [MG14] analyzed the language that is used in crowdfunding campaigns. They found out that some phrases have an impact on the success of a campaign, both positively and negatively. They claimed that the language that is used within a campaign has an impact on the chances of success. Zhou et al. [Zho+16] shared Mitra and Gilbert [MG14] interests and focused on the language in campaign descriptions on Kickstarter. The dataset contained all campaigns on Kickstarter from 2009 to November 2014, whereby canceled, purged, and suspended campaigns were not taken into account. Additionally, only campaigns with a goal which is at least USD 100 and at most USD 1,000,000 and with a description of a minimum of 100 words were considered. In order to investigate the impact of the campaign description on the success, they built two different models: the first model only contained basic campaign properties (e.g. funding goal, duration), while the second model additionally contained features such as the readability and tone of the description, and the experience and expertise of the campaign creator. They used decision trees to predict if a campaign is

likely to reach its goal and to compare the accuracy of both models. While the basic model reached an accuracy (F1 score) of 69%, the extended model had an accuracy of 73%. The results show that the introduced features had a positive impact on the prediction accuracy. In our work we also will follow this approach to filter out certain campaigns from our dataset. We will mainly focus on quantitative features in our prediction and will not further investigate the language that is used.

Koch and Cheng [KC16] analyzed both, quantitative and qualitative factors in their work. They used qualitative factors, such as the video tone quality, the presence of a prototype or the risk of delay. To test their hypothesis they built different logistic regression models, one with only quantitative factors and three including qualitative factors, in order to compare the results. The logistic regression used a binary label (funding successful or not). The model with only quantitative factors reached an accuracy of roughly 79%, while the combined models reached an accuracy of 83%. So the inclusion of qualitative factors had a positive effect on the prediction result. The work that is presented in this thesis follows a similar approach. In addition to the classification experiments we will also use regression. This means that we will not only predict if a certain campaign will be successful or not, but also how much money will be pledged.

Rao et al. [Rao+14] analyzed features that change during the duration of a campaign, for example money, pledged. They found that success can be predicted with an accuracy of 83% on average after 15% of the campaign duration was passed. They used decision trees for their prediction and claimed that there are 3 major turning points in a campaign, namely at 10%, at 45% and at 85% of the campaign duration, where major changes happen. The first turning point was explained by "*seed money*", which is money from people that are closely related to the creator or from the creator itself. The last turning point was explained by a "*deadline effect*". The deadline effect possibly occurs towards the end of a campaign and means that people are more likely to pledge a campaign if the end of the funding period is in sight.

Greenberg et al. [Gre+13] have provided a tool for campaign creators to get feedback on their campaigns. They used a pre-scraped dataset with campaigns that finished between June and November 2012. For the prediction

they used features such as the goal, the category, the number of rewards, but also linkages to social networks (e.g. Facebook or Twitter). The authors used different algorithms for the prediction such as support vector machines and decision trees. Their best prediction used logistic model trees and reached an accuracy of roughly 68%. In contrast to this work we will not provide a tool for campaign creators, but we will examine factors that make the campaign successful. The dataset that is used for our predictions is from 2017. This means that there are 5 years between the dataset that is used in Greenberg et al. [Gre+13] and our dataset.

Stadler et al. [Sta+15] compared the two biggest crowdfunding platforms Kickstarter and Indiegogo with each other and tried to find factors that lead to success of the campaigns on these platforms. One of the main differences between those platforms is the funding-model. While Kickstarter only provides fixed-funding, Indiegogo also allows flexible-funding. Fixed-funding follows the principle "*all or nothing*", which means that the creator only gets the money if the funding goal is reached. Flexible-funding allows the creator to collect the money, even if not enough money was pledged to reach the goal. The dataset consisted of 12,825 Kickstarter campaigns and 13,288 Indiegogo campaigns. The average funding goal in the dataset was USD 22,410.62 for Kickstarter campaigns and USD 98,217.85 for campaigns on Indiegogo. The success rate was 43.13% on Kickstarter and 27.04% on Indiegogo. They hypothesized that campaigns which reach a quarter of their goal in an early stage are likely to succeed. They used t-tests and Bonferroni post-hoc-test to prove their hypothesis. Another finding was that the early phase of a campaign is more important in context of fixed-funding in comparison to flexible funding.

The work from Althoff et al. [ADJ14] is about altruistic requests, which can be described as someone asking for a favor without physical rewards. They analyzed "*Random Acts of Pizza*" from Reddit, where people can post a request in order to receive a pizza from another user on an altruistic basis. The reasons why people post such a request are quite different: some claim to have a hard time and others just refer to "*being fed*", for example. In their paper they analyzed textual features of the request, for example, politeness, reciprocity, sentiment or length as well as social features such as the status of the requesters and their similarity to the users. Then they built a logistic regression model with these features and predicted success. For

the evaluation of their model receiver operation characteristic (ROC) was used. The best model reached a performance of 0.669, which is significantly above the baseline.

Chung and Lee [CL15] followed an approach similar to ours. They collected a dataset from Kickstarter and tried to predict the success of campaigns. In order to do this, they divided the available data into three categories: static data, which normally does not change during the campaign, temporal data, including information about pledge amounts and backers, and twitter linkages. They used Naive Bayes, Random Forests and AdaBoost M1 for the prediction. AdaBoost M1 achieved the best accuracy with 76.4%. We have to mention that the data was not filtered, which might have helped to improve the prediction. The second approach in the paper was to predict the money pledged for a campaign. Therefore, they extended their model with 3 classes representing a range of the money that was pledged, which means that they transformed a regression problem into a classification problem. The best result had an accuracy of 77.2%. As distinct to Chung and Lee [CL15], in our work, we want to predict the amount that is pledged with regression and not with classification.

3. Materials & Methods

In this chapter we describe the materials and methods used for our analysis. In Section 3.1, we start with the data acquisition, where we introduce a custom-built web crawler for Kickstarter.

We describe the dataset that we have crawled in Section 3.2 and provide a first empirical characterization.

The main goal of our experiments is the prediction and identification of features which define success in Kickstarter campaigns. In order to achieve this, we describe all methods used in our prediction experiments in Section 3.3, followed by details about the feature engineering in Section 3.4.

3.1. Data Acquisition

We build a web crawler for Kickstarter in the programming language "*Python*". As Python does not provide all the features we need, we made use of some freely available libraries such as "*urllib*[1]" and "*Beautiful Soup*[2]". We used the library "*urllib*" to make HTTP-requests to the URL "*https://www.kickstarter.com/*" and its sub-URLs. The library "*Beautiful Soup*" parses HTML documents and provides a convenient way for the user to access the hierarchy of the DOM-tree and to search for certain tags. Our crawler uses both hierarchy and CSS classes to identify HTML elements that contain desired information.

Our internal representation of a campaign comprises a main part with **(i) general information** that is always visible and different tabs with **(ii) additional information** of the campaign. The main part contains the pitch

[1] **https://docs.python.org/3/library/urllib.html**.
[2] **https://www.crummy.com/software/BeautifulSoup/bs4/doc/**.

video, the creator of the campaign, the amount of money that has been pledged, the funding goal, the number of backers, the days until the end of the campaign, the category, and the location of the campaign. The creator information is saved as new record at every update, in order to get history of the campaigns created and backed. The additional information is available for each campaign, and is split into grouped views (tabs) called "*Campaign*", "*FAQ*", "*Updates*", "*Comments*" and "*Community*".

The campaign tab contains the description of the campaign and the different rewards including their price, their expected delivery, the number of backers of the reward, and if the reward is limited to a certain number of backers. In the FAQ tab users can ask creators questions about the campaign. The creator has the option to make the question visible for all users afterwards. The updates tab can be used for communication by the creator. The creator can post updates about the project's progress during and after the campaign in order to keep backers in the loop about the proceedings of the campaign. Additionally, users have the option to comment and like an update.

Users can also engage in discussions or make general remarks about the campaign in the comments tab. The community tab provides information about the composition of the backer community. It shows the top ten countries and cities backers come from. In addition, it shows how many new and returning backers were attracted to the campaign. New backers are backers that did not back a campaign before on Kickstarter. Returning backers already backed a campaign in the past.

On Kickstarter every campaign has a unique URL through which it can be reached. Therefore, our crawler basically needs the URL in order to crawl the desired information. The data acquisition itself is split up into two parts. First, we wanted to crawl as many finished campaigns as possible in order to create statistics to examine the anatomy of Kickstarter and to track changes over the years. Secondly, we want to track live campaigns during the whole campaign duration, which means that we want to have snapshots of the campaign at different points in time. Therefore, we distinguish between a "*Live Campaigns Dataset*" and a "*Finished Campaigns Dataset*".

Unfortunately, there is no complete listing of finished campaigns on Kickstarter itself, but they are still reachable through the campaign URL. For this reason, we had to find lists of old campaign URLs. The first source for

URLs of finished campaigns was "*kicktraq*[3]". We were able to crawl roughly 200,000 campaign URLs from this source, but they locked their archive due to excessive usage. The second source was "*webrobots*[4]", from which we extracted the URLs from the provided CSV files. After we had extracted the URLs from both sources, we merged them together and eliminated duplicate URLs from the dataset.

The data acquisition of live campaigns only uses data from Kickstarter itself. In the first step we used the search function on Kickstarter and searched for campaigns that have recently been launched and added them to our list. In the second step we looped through the list and crawl the information of the campaign. We only crawled the basic information and the campaign tab of live campaigns in order to get regular updates. Additionally, we looked roughly every 30 minutes for new campaigns in order to get a snapshot of the campaign as soon as possible. With this approach we get 4 - 5 campaign updates per campaign on a single day, if we consider that there are roughly 4,500 live campaigns at the moment.

Due to changes of the Kickstarter interface, we had to adapt the crawler several times. As we did not notice these changes in first place, we were missing some data. We expected that we would encounter problems like these and, for this reason, we stored the compressed source code of the HTML document in the database. With this source code we were able to restore the missing data.

After a campaign finished its funding period, we crawled the campaign tab once more and additionally all other available tabs. The crawling of all tabs within the campaign duration would need too many HTTP requests and would take therefore too much time. Additionally, most information from these tabs is labeled with a timestamp. Therefore, it was sufficient to crawl them altogether after the campaign ended. As the structures of live and finished campaigns are different, we also had to build separate crawlers for both cases.

[3] **www.kicktraq.com**.
[4] **https://webrobots.io/kickstarter-datasets/**.

3.2. Dataset

In this section we describe the dataset that we acquired with the means that are explained in Section 3.1. In general, there are two different datasets which are quite similar in structure, but different regarding the start of the campaigns that are contained in the datasets.

The first dataset, we refer to it as "*Finished Campaigns Dataset*", contains campaigns that had already been finished before the start of our research. We crawled finished campaigns from 2009 to 2016, but without any claim of comprehensiveness. Finished campaigns are normally still reachable under their URL after the end of the campaign, but creators can request to delete their campaign from Kickstarter.

The second dataset consists of campaigns that we tracked from its inception to the end. The target of this dataset is to collect time-series of every campaign. We regularly looked for new campaigns on Kickstarter and tracked them until the end of their funding period. Therefore, we can investigate how the number of backers or the money pledged evolves over time. We are also able to detect changes from the creator, for example, if the description or the category were changed. We refer to this dataset as "*Live Campaigns Dataset*". A detailed overview of all stored attributes can be found in Appendix A.

3.2.1. Finished Campaigns Dataset

Characterization

An overview of the "*Finished Campaigns Dataset*" can be seen in Table 3.1. We used this dataset to collect new insights on Kickstarter. Specifically, we provide some figures to investigate where campaigns come from, how many of them are funded successfully and which factors have an impact on success. Basically we consider a campaign successful if the funding goal has been reached. We also investigate how these features have developed over the years.

Table 3.1.: Short description of "*Finished Campaigns Dataset*"

Crawling period	2009 - 2016
Number of campaigns	286,966
-successful	114,362
-failed	146,106
-canceled / suspended	26,498
Success rate	39.9%
Money pledged overall USD	USD 2,154 M
Backers overall	23.4 M

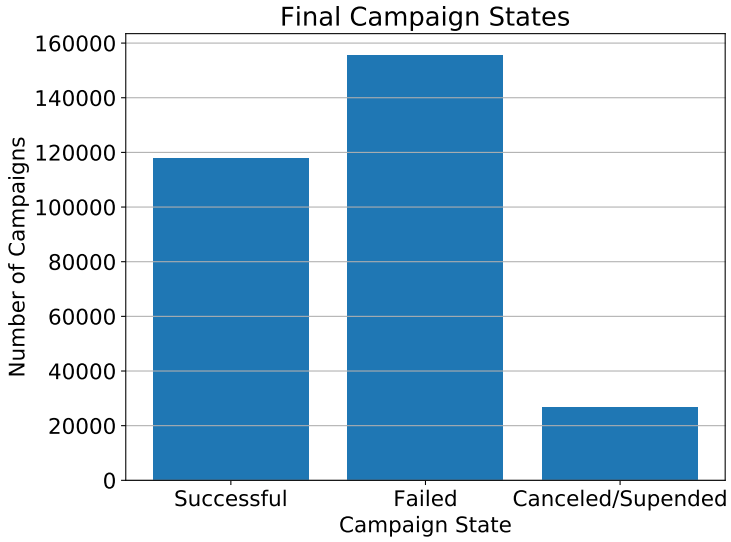
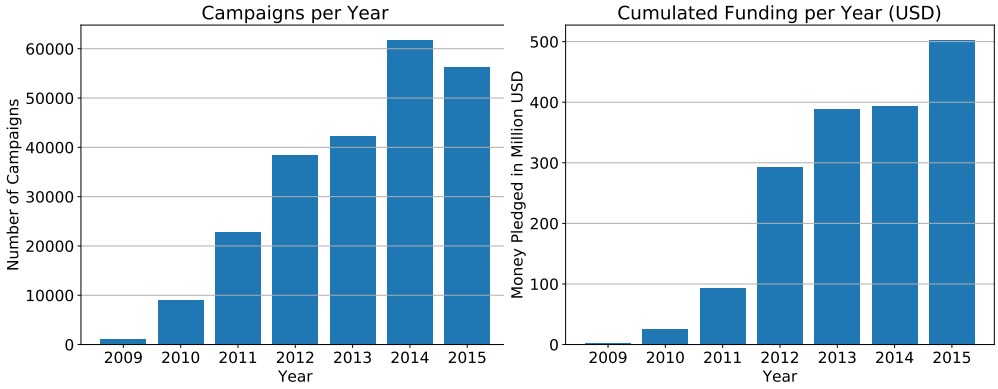


Figure 3.1.: **Final Campaign States of the Finished Campaigns Dataset.** In this figure we can see the distribution of the final campaign states. We see that 39.9% of the campaigns reached their goal. There is also a considerable number of canceled and suspended campaigns.

Final Campaign States

The distribution of the final campaign states can be seen in Figure 3.1. The figure shows that only roughly 40% of the campaigns on Kickstarter reach their funding goal. The number of canceled or suspended campaigns is also considerable. We pooled canceled and suspended campaigns together, because it is sometimes ambiguous why a certain campaign terminated before its dedicated end. The difference if a campaign is canceled or suspended depends on who indicated the end. While canceled campaigns are ended by the creator himself/herself, some campaigns are also terminated by Kickstarter, often due to legal reasons.

Finding 1: Only roughly 40% of Kickstarter campaigns are funded successfully. Roughly 9% of all campaigns were canceled or suspended.



(a) Number of Campaigns from 2009 to 2015

(b) Cumulated Funding from 2009 to 2015

Figure 3.2.: **Number of Campaigns and Money Pledged per Year.** In Fig. 3.2a we can see the number of campaigns from 2009 to 2015. We see that the number of campaigns increased from 2009 until 2014. In 2015 the number decreased again. Fig. 3.2b shows how much money was pledged to successful campaigns from 2009 to 2015. We can see that the number increased from one year to another, even if there were fewer campaigns in 2015 than in 2014.

Campaigns per Year

In Figure 3.2 we can see the development of the size of Kickstarter from 2009 until 2015. Figure 3.2a shows that the number of campaigns increased every year until 2014. In 2015 there were slightly fewer campaigns launched, but still more than in 2013. Figure 3.2b shows the cumulated funding in U.S. dollar for every year. The numbers only include successful campaigns in U.S. dollar, which means that the numbers represent money that has actually been paid. We can see that the trend is similar to the number of campaigns, but interestingly even if the number of campaigns in 2015 was lower than in the year before, more money has been collected.

Finding 2: The number of campaigns grew steadily from 2009 to 2014, but decreased in 2015. Nevertheless, the money that was pledged to successful campaigns increased from one year to another.

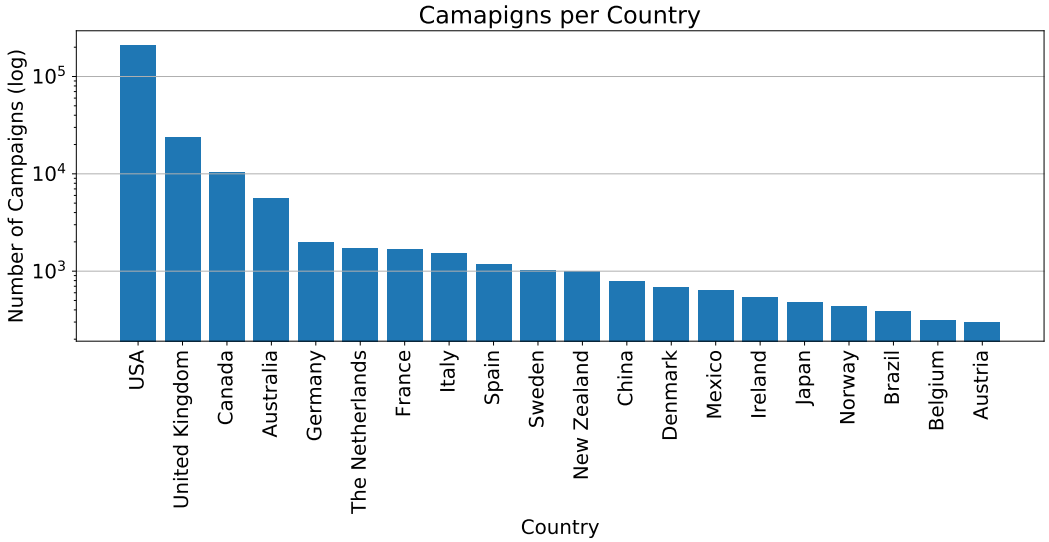


Figure 3.3.: **Distribution of Countries where Campaigns Come from.** On the x-axis we can see the 20 countries where campaigns come from and the number of campaigns on the logarithmic y-axis. We can see that by far most campaigns come from the United States. Apart from that, there is a focus on English-speaking countries and countries in Middle Europe.

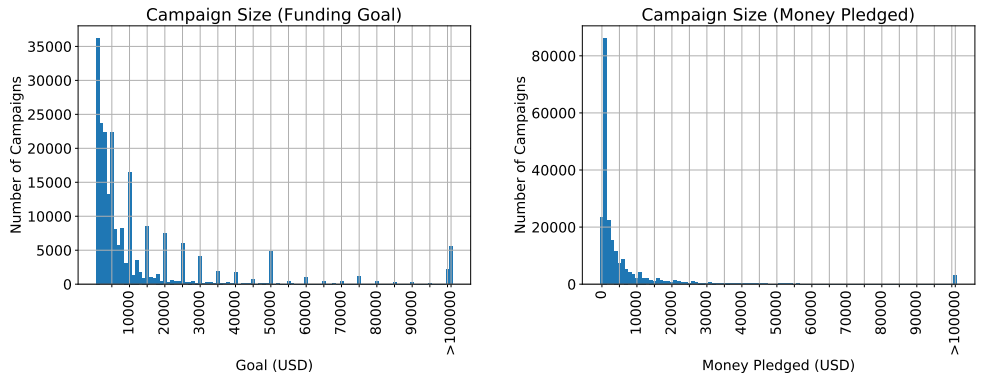
Countries of Origin

In Figure 3.3 we can see the distribution of countries where the campaigns in the dataset come from. By far, the most campaigns come from the United States, which can be explained by the fact that Kickstarter has its origin there. There is also a focus on other English-speaking countries, such as Canada or Australia, followed by countries from Middle Europe.

Finding 3: Most campaigns have their origin in the USA. Apart from that, there is a focus on English-speaking countries and countries in Middle Europe.

Campaign Size

Figure 3.4 illustrates the campaign sizes on Kickstarter. We only consider campaigns with U.S. dollar as currency. The x-axis shows the different money amounts in buckets of USD 1,000 and y-axis shows the number of campaigns. In Figure 3.4a we can see that most Kickstarter campaigns



(a) Campaign Size Measured by Funding Goal

(b) Campaign Size Measured by Money Pledged

Figure 3.4.: **Campaign Size Measured in Funding Goal and Money Pledged.** Fig. 3.4a measures the campaign size by the funding goal. We see that most campaigns have a funding goal of USD 10,000 or less. 3.4b illustrates the campaign size depending on the money pledged. There are a lot of campaigns that pledge few or nothing. There are also quite a few campaigns that pledge USD 100,000 or more.

have a funding goal of less than USD 10,000. There are also some "*spikes*" in the figure. Those spikes show that creators tend to use round funding goals which are dividable by 5,000. In Figure 3.4b we measure the campaign size with the money that is pledged. We can see that 48.7% pledge less than USD 1.000, but roughly 1.5% of the campaigns pledge more than USD 100,000.

Finding 4: Most campaigns have a funding goal of USD 10,000 or less. Nearly half of the campaigns pledge less than USD 1.000.

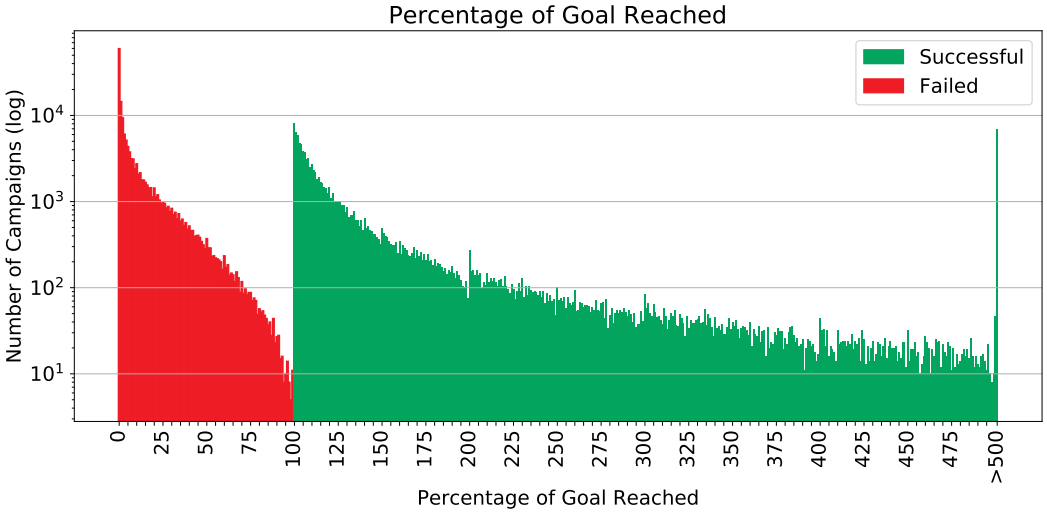


Figure 3.5.: **Distribution of Campaigns with Relation between Goal and Money Pledged.** The figure illustrates the relation between the goal and the money pledged of the campaigns. On the x-axis we can see how much percent of the funding goal were reached and on the logarithmic y-axis we have the number of campaigns for the corresponding percentage. There is an extensive gap between successful and failed campaigns. We can also see that only a few campaigns miss their goal slightly.

Fraction of Goal Reached

Figure 3.5 indicates the money pledged of campaigns compared to their funding goals. The relations on the x-axis are expressed as a percentage; successful campaigns are indicated in green, campaigns that failed are colored in red. Overall, the chart demonstrates an extensive gap between successful and failed campaigns. There are very few campaigns that only slightly missed their goals.

Finding 5: There are very few campaigns that only slightly miss their goals.

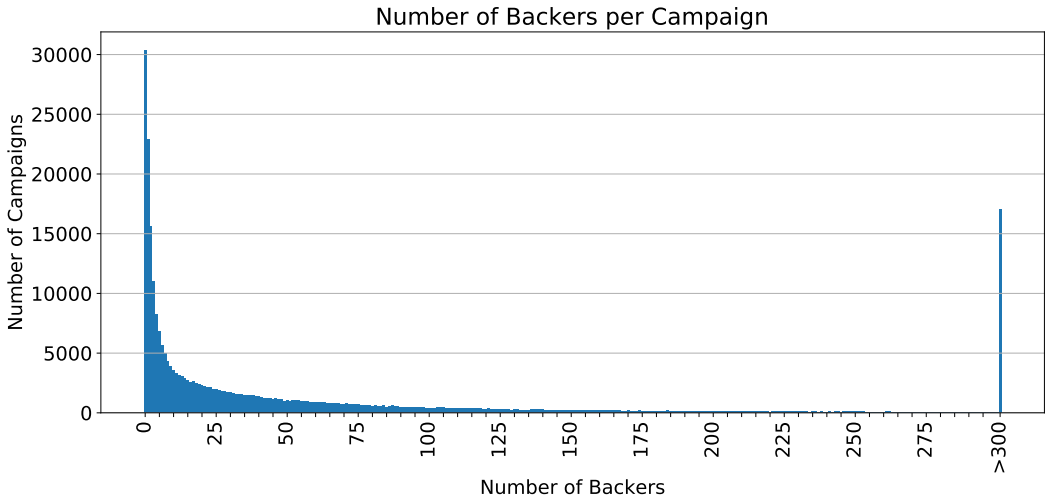


Figure 3.6.: **Number of Backers.** On the x-axis we can see the the number of backers and on the y-axis the corresponding number of campaigns. Most campaigns have fewer than 50 backers, but there are also some campaigns with more than 300 backers.

Number of Backers

The Figure 3.6 shows the distribution of campaigns compared with their number of backers. On the x-axis we can see the number of backers and on the y-axis the number of campaigns. Most campaigns have fewer than 50 backers, but there are also some campaigns which have more than 300 backers. Once again, we can see that there is a small number of very successful campaigns.

Finding 6: Most campaigns have fewer than 50 backers, but there are also campaigns with more than 300 backers.

Figure 3.7 shows the mean amount of money backers pledge per campaign. Successful campaigns are colored in green and failed campaigns in red. We can see that most people pledge amounts less than USD 100 and tend to pledge round amounts, which can be seen through the spikes in the chart. The average amount pledged for successful campaigns is USD 89.23 in comparison to USD 59.47 for failed campaigns. Therefore, we can see that

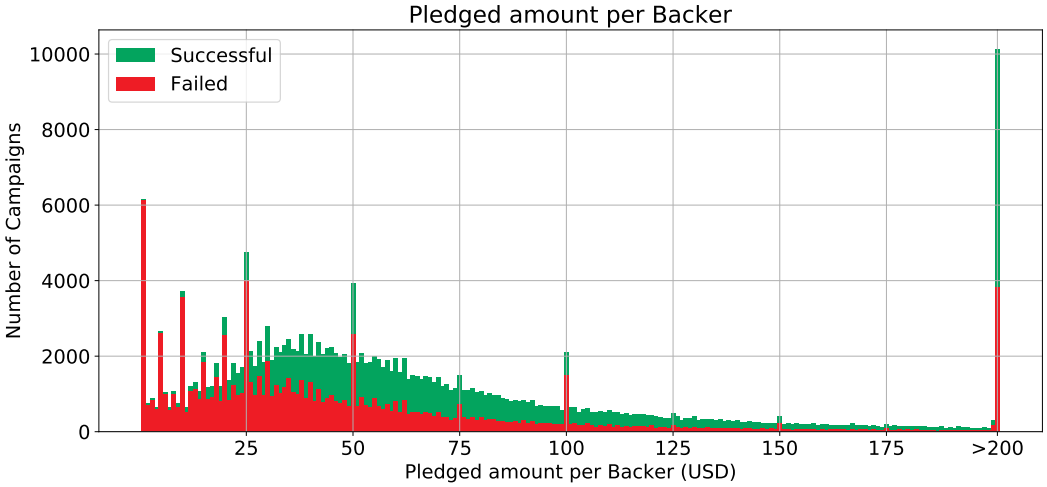


Figure 3.7.: **Money Pledged per Backer.** On the x-axis we have the mean amount pledged by each backer and on the y-axis we have the corresponding number of campaigns, whereby successful campaigns are indicated in green and failed campaigns in red. We can see that the success rate tends to be higher if the backers pledge higher amounts on average.

the success rate is higher for campaigns where the backers pledge more on average.

Finding 7: Backers tend to pledge round amounts. The average amount pledged is roughly USD 90 for successful campaigns compared to roughly USD 60 for failed campaigns. Therefore, the success rate tends to be higher if backers pledge more on average.

Categories

Figure 3.8 provides an overview over the 15 main categories for campaigns. While the x-axis indicates the category, the stacked bars in green and red illustrate the number of campaigns for the corresponding category. The segmentation of the bar indicates the number of successful and failed campaigns. The blue bars correspond to the amount that was pledged for campaigns in each category, whereby only successful campaigns are considered here. Consequently, we can see from the relation of the stacked and the blue bar, that there are categories with a larger number of campaigns, but relatively little money pledged and vice versa. If we compare the categories

"*art*" and "*design*", we can see that "*art*" has slightly more campaigns, but campaigns of "*design*" received far more money. We can also see that the success rates differ across the categories. While "*theater*" has twice as many successful campaigns as failed ones, only a fifth of the campaigns in the category "*journalism*" reach their goal. The category with the most campaigns is "*film & video*" with roughly 49.000 campaigns. The category with the fewest campaigns is "*dance*" with only roughly 3.100 campaigns. If we have a look on the success rates, we can see that the category "*theater*" reaches the highest success rate with 65.8%. The category "*technology*" in comparison only has 24.7% successful campaigns. The most popular category, according to the money pledged, is "*games*" with USD 487.4 million pledged. The campaigns in the category "*crafts*" only managed to pledge USD 6.24 million in comparison.

Finding 8: There are considerable differences regarding the success rate, the number of campaigns and the money pledged across the categories. There are categories with a larger number of campaigns, but relatively little money pledged and vice versa. The category with the highest success rate is "*theater*" with a success rate of 65.8%. Compared to that, only 24.7% of the campaigns in the category "*technology*" reach their goal. Campaigns in the category "*games*" collected USD 487.4 million in comparison to USD 6.24 million for campaigns in the category "*dance*".

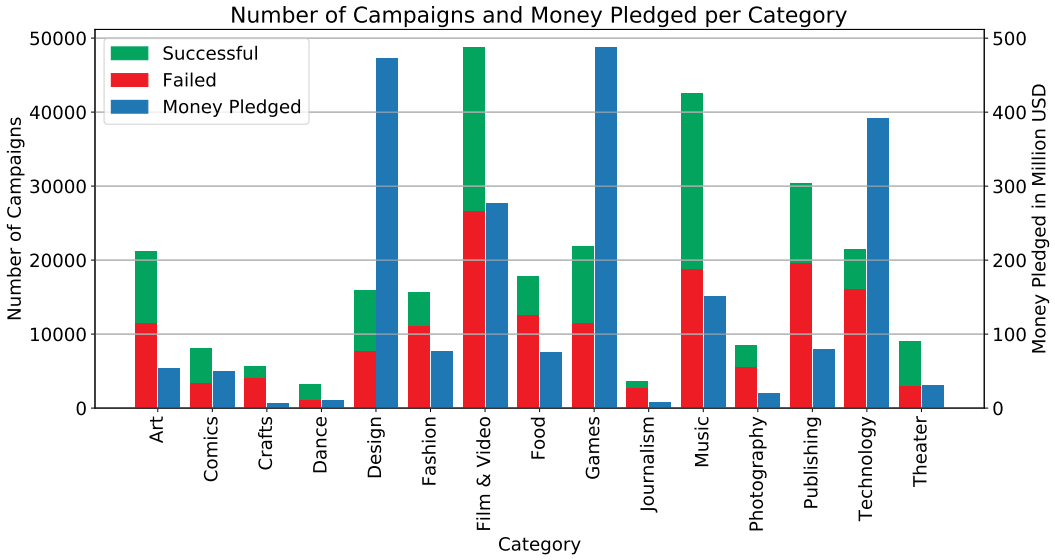


Figure 3.8.: **Money Pledged and Number of Campaigns per Category.** On the x-axis we see the 15 main categories for campaigns on Kickstarter. The stacked bar in green and red indicates the number campaigns for each category, split up in successful and failed campaigns. The blue bars correspond to the money that was pledged for campaigns in each category. We see that there are considerable differences regarding the success rate, the number of campaigns and the money pledged across the categories.

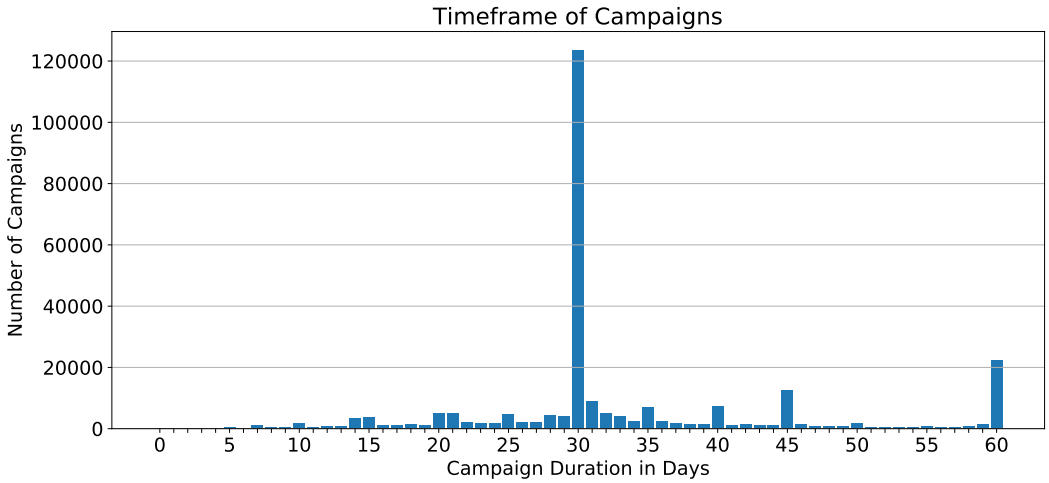


Figure 3.9.: **Distribution of Campaign Durations.** The x-axis represents the possible durations for a campaign and y-axis shows the corresponding number of campaigns. We can see that most Kickstarter campaigns have a duration of 30 days.

Campaign Duration

Figure 3.9 illustrates the number of campaigns compared to their campaign duration. We can see that 45.9% of the campaigns have a duration of 30 days, which is likely caused by the fact that Kickstarter recommends this duration to campaign creators. There are almost no campaigns with a duration of less than 15 days. Another spike with 8.3% can be seen at 60 days, which is the longest possible campaign duration on Kickstarter today.

Finding 9: Roughly 46% of the Kickstarter campaigns have a duration of 30 days.

Figure 3.10 shows the development of the mean campaign duration over time. We can see that from 2009 to 2012 creators tended to choose a longer campaign duration than in the following years. The mean duration in 2010 of over roughly 70 days is not even possible anymore, as Kickstarter limited the maximum duration from 90 to of 60 days in 2011[5]. From the middle of 2012 the mean duration has remained roughly 30 days. Interestingly the mean duration of failed campaigns is generally higher.

[5]**https://www.kickstarter.com/blog/shortening-the-maximum-project-length**.

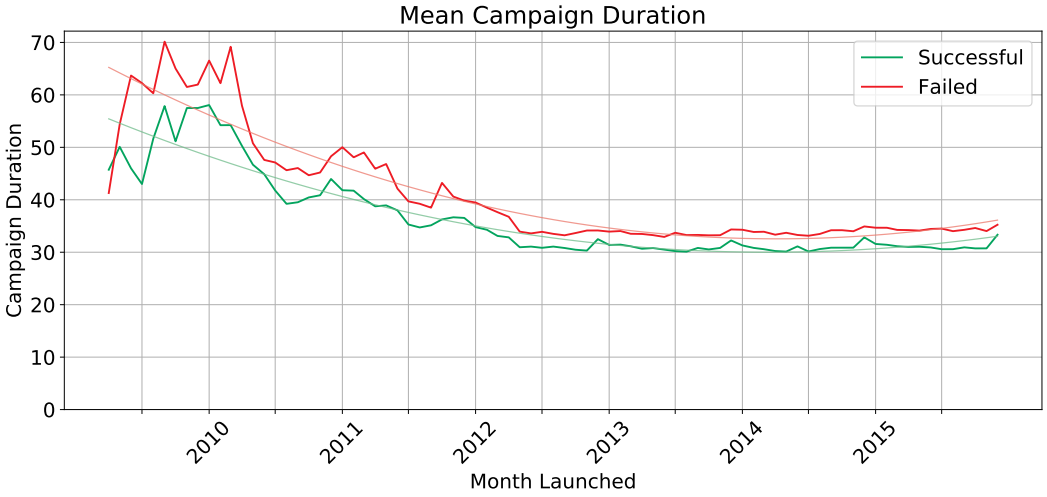


Figure 3.10.: **Mean Campaign Duration Over Time.** We can see the launch month on the x-axis and the corresponding mean campaign duration on the y-axis. The mean campaign duration decreased from roughly 60 days in 2009 to roughly 30 days in the middle of 2012. From then on the mean campaign duration has remained stable. Failed campaigns generally have longer campaign durations.

Finding 10: The mean campaign duration decreased from roughly 60 days to roughly 30 days from 2009 to 2013 and has remained stable since then. Failed campaigns have a longer campaign duration on average.

In Figure 3.11 we compare the success rate for each category with the duration of the campaign. For most categories we find that both campaigns with a shorter or longer duration have a lower success rate. We can see that the duration where the highest success rate is reached differs across the categories. The success rate in the category *"design"* is nearly linear and decreases slightly for longer durations. The categories *"fasion"*, *"journalism"* and *"publishing"* in comparison have a quite low success rate for long campaign durations in comparison to shorter ones. Generally, we can observe that campaigns with a duration between 15 and 30 days reach the highest success rates.

Finding 11: Campaigns with a duration between 15 and 30 days have the highest success rates.

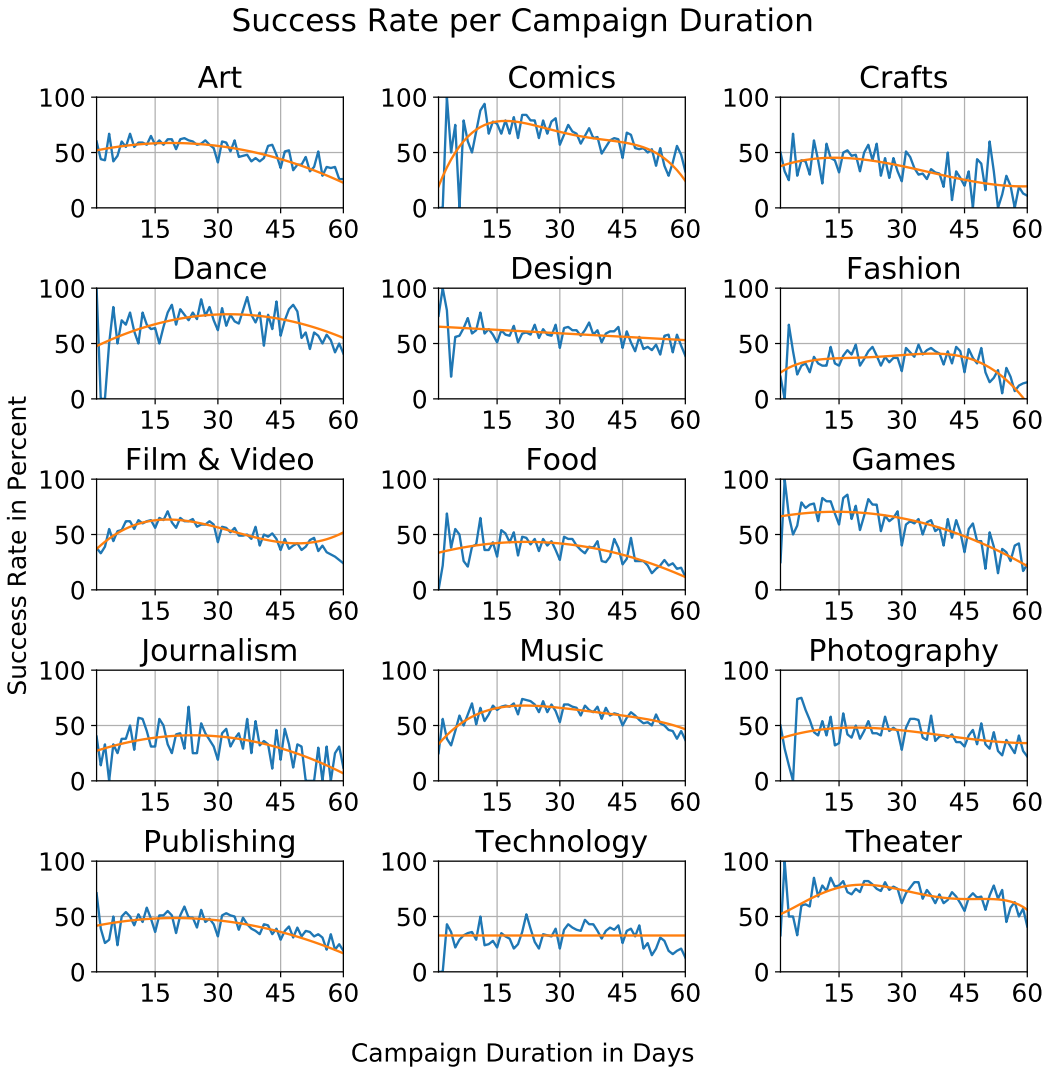


Figure 3.11.: **Success Rate by Campaign Duration per Category.** The x-axis illustrates the campaign durations in days and the y-axis the success rate for each duration. We see that most categories have lower success rates for longer or short campaign durations.

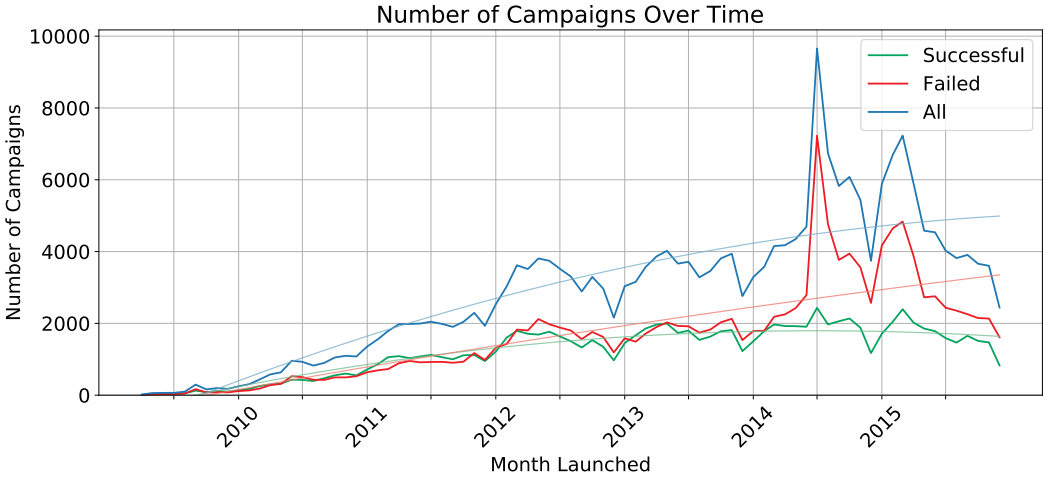


Figure 3.12.: **Number of Campaigns Over Time.** On the x-axis we see the launch month of the campaigns and on the y-axis the corresponding number of campaigns. The green line indicates successful campaigns and the red line deals with the failed campaigns. The blue line is the sum of both and therefore represents the number of all campaigns. We see that the number of campaigns increased from 2009 to the middle of 2014. The number of campaigns went down at the beginning of 2015, but then increased again within the the first half of the year. Towards the end of 2015 the number of campaigns decreased again.

Campaigns Over Time

In Figure 3.12 we see the month launched on the x-axis and number of campaigns on the y-axis. While the green line represents successful campaigns, the red line indicates failed campaigns. The blue line is the sum of both. We only considered campaigns until the end of 2015, as we were not able to get a considerable number of campaigns after that year. The number of campaigns that were launched on Kickstarter rose steadily until 2014. The number of newly launched campaigns suddenly spiked in 2014, but fell towards the end of that year. The decrease by the end of 2015 might be caused by missing campaigns in our dataset. Generally, there is an increase in the middle of the year and a decrease at the turns of the year. We can see that the two spikes in 2014 and 2015 were mainly caused by failed campaigns. The number of successful campaigns remained steady in the period from 2012 to 2015.

Finding 12: The number of campaigns increased steadily from 2009 to 2014, but went down at the beginning of 2015. Generally, there is an increase in the middle of the year and a decrease at the turn of the year.

Figure 3.13 shows how many campaigns were launched per month in each category. The chart shows that different categories developed differently over time. Successful campaigns are indicated in green, while unsuccessful ones are red. The peak in 2014, which can be seen in 3.12, is not visible in all categories. We can see that some categories, such as "*dance*", "*comics*", "*film & video*" or "*theater*", exhibit cycles regarding the launch of campaigns, which means that more campaigns are launched in the middle of the year than at the turn of the year. We can also see that some categories have grown over the years, for example, "*comics*" or "*games*". The categories "*film & video*" and "*music*" had fewer launches in 2015 than in the year before. We can see that general trends on Kickstarter do not apply to all campaign categories.

Finding 13: The number of campaign launches is higher in the middle of the year than at the turn of the year for campaigns in the cateogories "*dance*", "*comics*", "*film & video*" and "*theater*". While the category "*comics*" has grown over the years, the categories "*film & video*" and "*music*" had fewer launches in 2015 than in the years before.

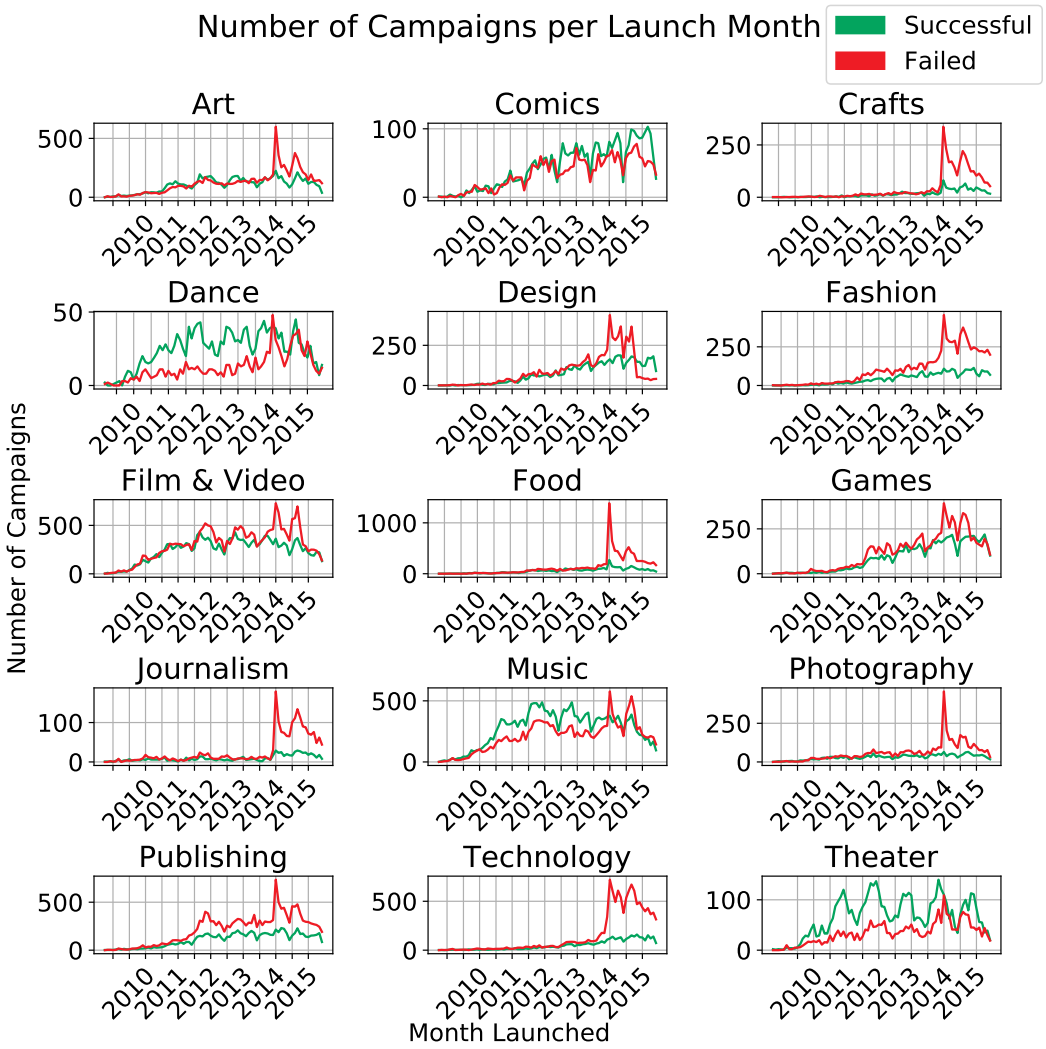


Figure 3.13.: **Number of Campaigns per Launch Month for each Category.** On the x-axis we can see the launch month and the number of the campaigns that were launched in that month on the y-axis. The categories "*dance*", "*comics*", "*film & video*" or "*theater*" exhibit cycles regarding the launch of campaigns. The number of new launched campaigns jumped up in 2014 in some categories, for example, "*art*", "*crafts*", "*design*" and "*photography*".

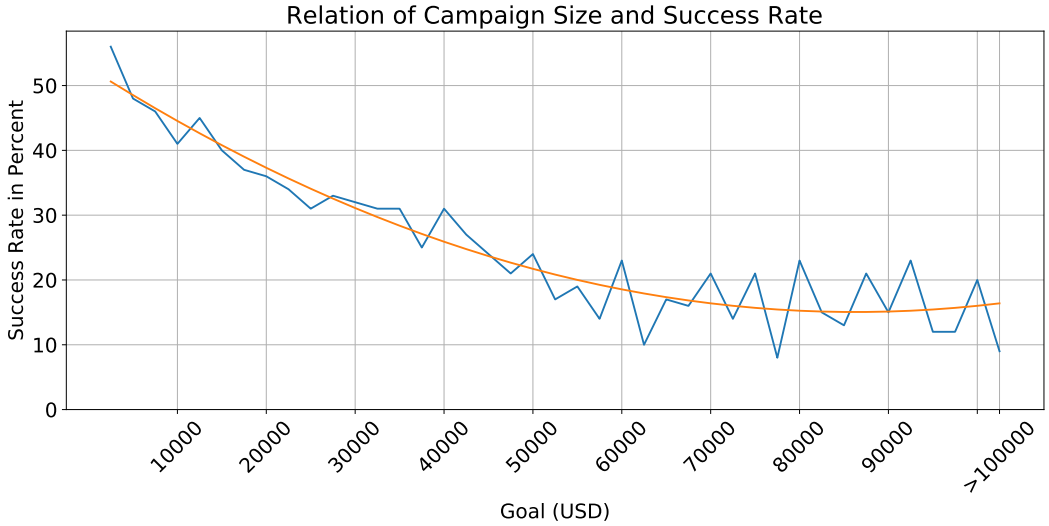


Figure 3.14.: **Relation between Success Rate and Campaign Size.** On the x-axis we can see the funding goal in U.S. dollar and the success rate of campaigns with the corresponding goal on the y-axis. We can see that the success rate decreases the higher the funding goal is.

Success Rate and Campaign Size

Figure 3.14 shows the relation between the funding goal on the x-axis and the success rate on the y-axis. The success rate for campaigns with a funding goal of USD 10,000 is roughly 40%. This rate decreases to roughly 30% for campaigns with a goal of USD 30,000 and to roughly 20% if the funding goal is USD 60,000. Generally, the success rate decreases the higher the funding goal is.

Finding 14: The success rate decreases the higher the funding goal is. Campaigns with a funding goal of USD 10,000 have a success rate of roughly 40%. The success rate decreases to roughly 30% for funding goals of USD 30,000, respectively to 20% for funding goals of USD 60,000.

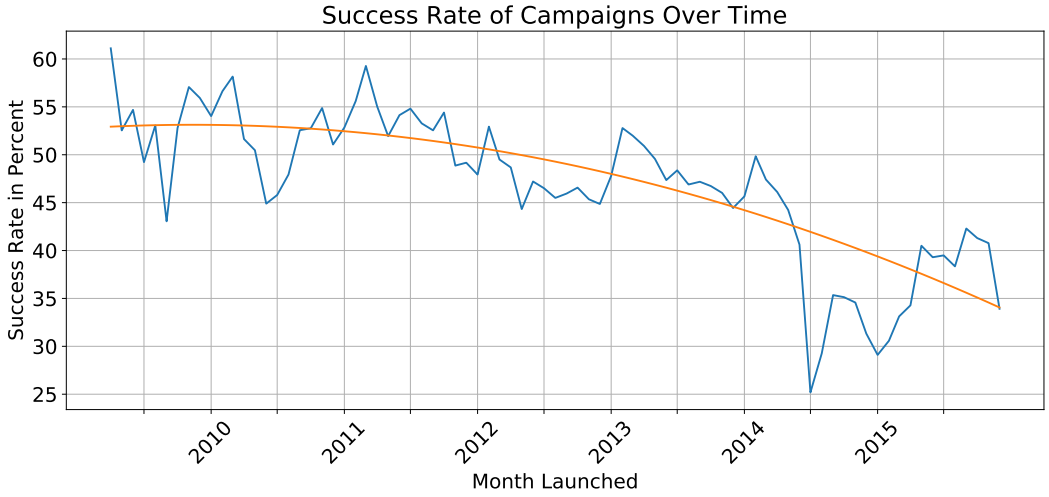


Figure 3.15.: **Success Rate Over Time.** On the x-axis we can see the launch month of campaigns and the corresponding success rate on the y-axis. We can see that the success rate varied between 45% and 60% from 2009 to 2013. After a drop in 2014 the success rate recovered to roughly 40% by 2015.

Development of Success Rate Over Time

In Figure 3.15 we can see how the success rate of campaigns has developed over the years. The mean success rate in the years from 2009 to 2013 fluctuated between 45% and 60%. After a drop in 2014, the success rate rose again, but only reached just above 40%.

Finding 15: The success rate of Kickstarter campaigns decreased from roughly 60% in 2009 to 35% in 2015.

Figure 3.16 shows the development of the success rate for each category. We can see that there are different trends throughout the categories. For example, the categories "*film & video*" and "*games*" have a relative stable success rate in the considered period of time. While the success rate of "*design*" campaigns was higher in 2015 than in the years before, the success rate of most of the other categories decreased towards 2015.

Finding 16: While the success rate of the category "*design*" increased from 2010 to 2015, the success rates of the categories "*dance*", "*food*" and "*photography*" decreased during the considered period.

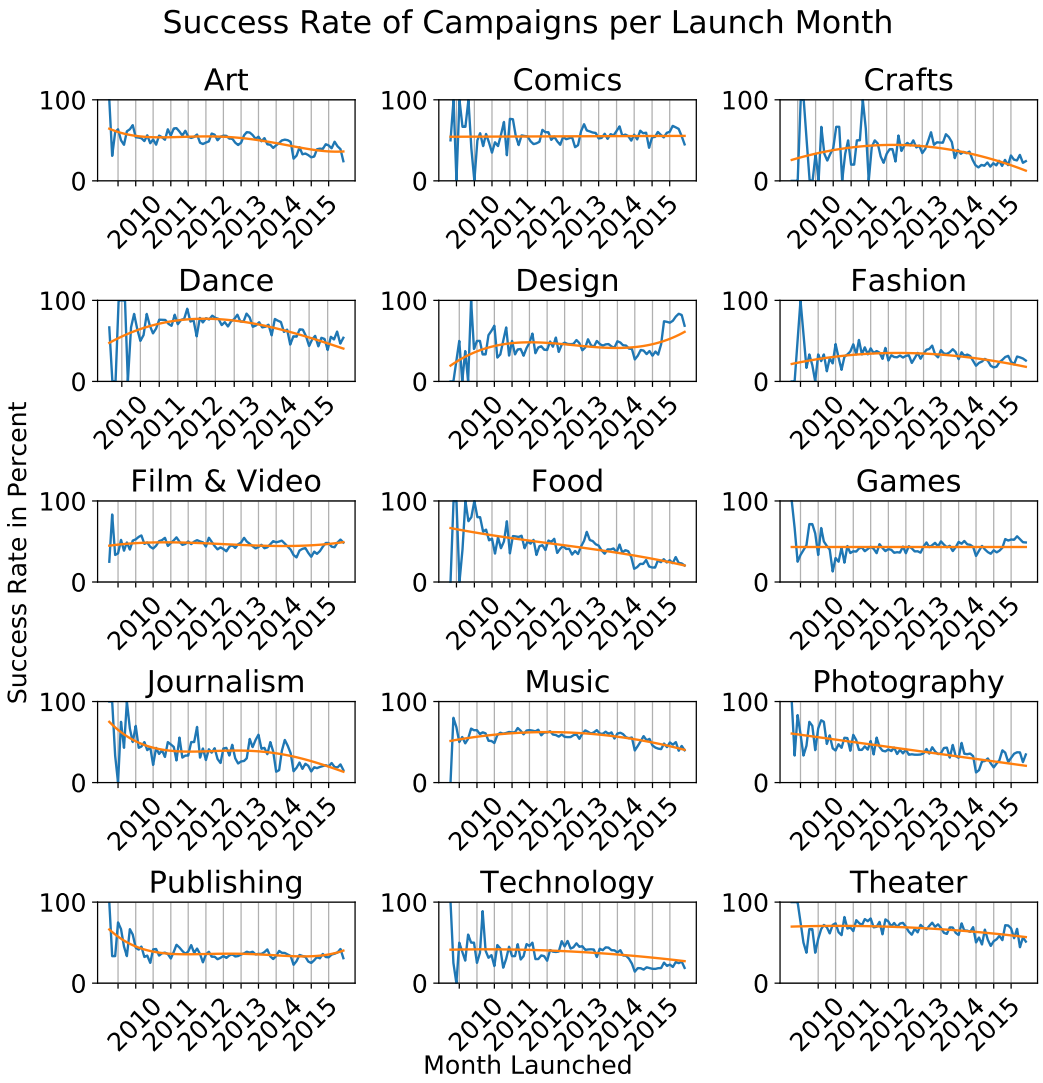
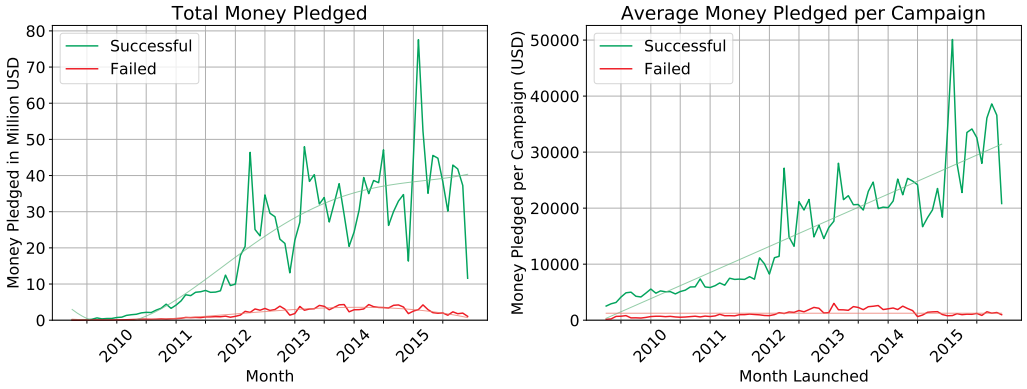


Figure 3.16.: **Success Rate Over Time per Category.** The figure shows how the success rate developed in each category over time. The x-axis illustrates the launch month of campaigns and the y-axis shows the corresponding success rate. Generally, the success rate decreased in most categories from 2009 to 2015.



(a) Total Money Pledged Over Time

(b) Average Money Pledged per Campaign Over Time

Figure 3.17.: **Money Pledged Overall and per Campaign Over Time.** On the x-axis we can see the month launched and the y-axis indicates the money pledged. Fig. 3.17a shows the total money pledged over time. We can see that more money is pledged in the middle of the year than at the turn of the year. Fig. 3.17b shows the average money pledged per campaign. While successful campaigns follow a straight upward trend, failed campaigns remain steady.

Money Pledged Over Time

Figure 3.17 shows the money pledged both overall and per campaign over time. Successful campaigns are indicated in green and failed campaigns are red. In Figure 3.17a we can see that the money pledged jumped up in 2013 and then periodically increased in the middle of the year and decreased at the turn of the year. Figure 3.17b shows the total number of money pledged divided by the number of campaigns. It seems that the mean money pledged for successful campaigns rises over the years. Both figures show that there is a large difference between successful and failed campaigns, because significantly more money was pledged for successful campaigns than for failed campaigns. The money pledged for unsuccessful campaigns remained steady on both charts during the considered period.

Finding 17: More money is pledged in the middle of the year than at the turn of the year. The average money pledged per campaign rose from roughly USD 1 million in 2009 to a maximum of USD 80 million in 2015.

In Figure 3.18 we can see how much money was pledged for successful campaigns in each category. The x-axis indicates the month the campaigns were launched in and the y-axis indicates the money that has been pledged to successful campaigns, which means only money that has been actually collected is considered. The chart is an indicator of how the popularity of each category developed over time in comparison to the others. Some categories follow a straight upward trend, such as the categories "*crafts*", "*fashion*" or "*technology*". Other categories had an upward trend from 2009 until 2013 and then decreased again. The category "*design*" generally follows a slight upward trend, but has some significant peaks in between.

Finding 18: The categories "*crafts*", "*fashion*" and "*technology*" followed a straight upward trend regarding money pledged from 2009 to 2015.

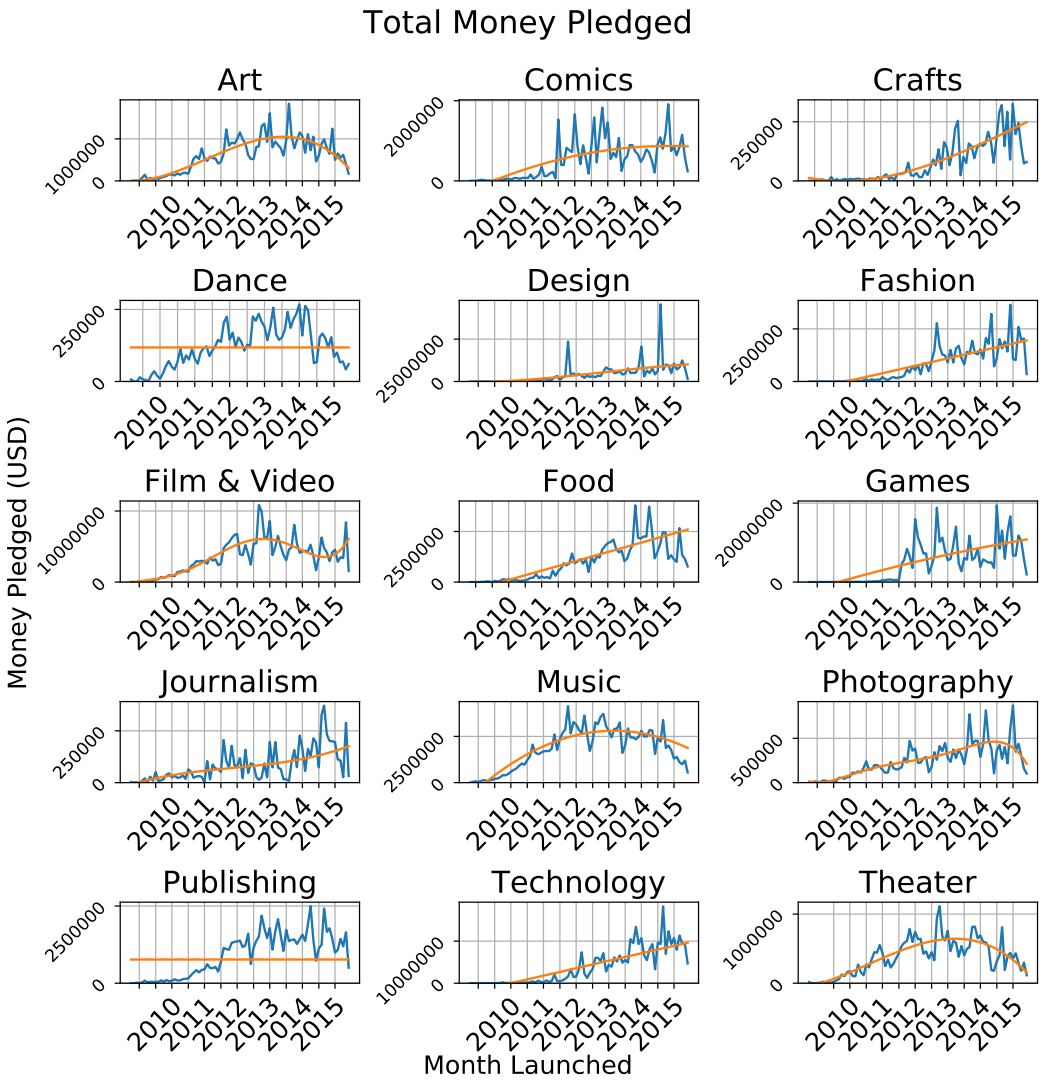


Figure 3.18.: **Total Money Pledged Over Time.** The figure shows the money pledged for each category over time. The x-axis indicates the launch month and the y-axis the amount of money that has been pledged. The categories "*crafts*", "*fashion*" and "*technology*" followed a straight upward trend. Other categories pledged less money in 2015 than in the years before.

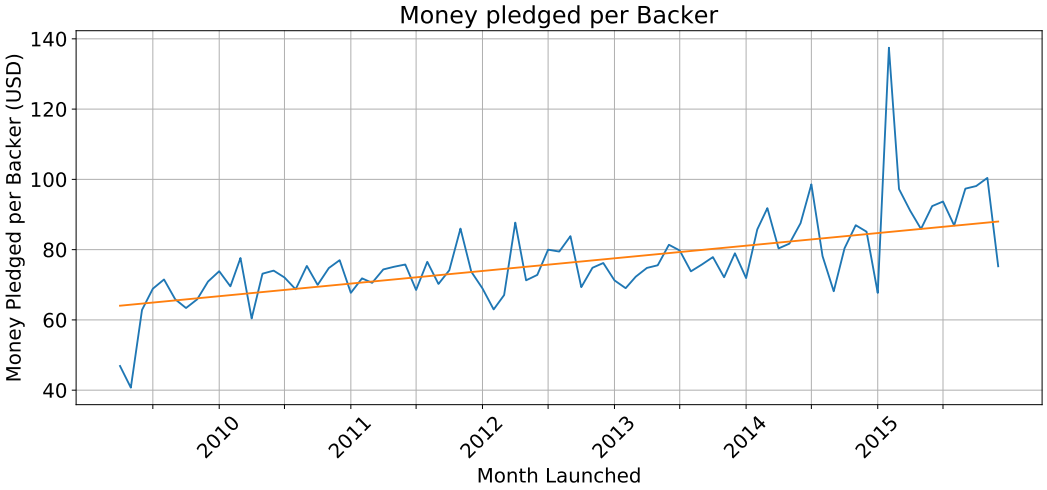


Figure 3.19.: **Money Pledged per Backer Over Time.** We can see the month on the x-axis and at the y-axis the mean amount each backer pledged in this month. While backers pledged USD 60 on average in 2009, they pledged on averge USD 90 in 2015.

Money Pledged per Backer

In Figure 3.19 we can see the month on the x-axis and on the y-axis the mean amount of money each backer pledged in this month. While a backer pledged roughly USD 70 per campaign in 2010, this number rose to approximately USD 90 in 2015.

Finding 19: The amount pledged per backer rose from roughly USD 70 in 2009 to USD 90 in 2015.

Table 3.2.: List of Findings in the *"Finished Campaigns Dataset"*

Finding 1:	Only roughly 40% of Kickstarter campaigns are funded successfully. Roughly 9% of all campaigns were canceled or suspended.
Finding 2:	The number of campaigns grew steadily from 2009 to 2014, but decreased in 2015. Nevertheless, the money that was pledged to successful campaigns increased from one year to another.
Finding 3:	Most campaigns have their origin in the USA. Apart from that, there is a focus on English-speaking countries and countries in Middle Europe.
Finding 4:	Most campaigns have a funding goal of USD 10,000 or less. Nearly half of the campaigns pledge less than USD 1.000.
Finding 5:	There are very few campaigns that only slightly miss their goals.
Finding 6:	Most campaigns have fewer than 50 backers, but there are also campaigns with more than 300 backers.
Finding 7:	Backers tend to pledge round amounts. The average amount pledged is roughly USD 90 for successful campaigns compared to roughly USD 60 for failed campaigns. Therefore, the success rate tends to be higher if backers pledge more on average.
Finding 8:	There are considerable differences regarding the success rate, the number of campaigns and the money pledged across the categories. There are categories with a larger number of campaigns, but relatively little money pledged and vice versa. The category with the highest success rate is "*theater*" with a success rate of 65.8%. Compared to that, only 24.7% of the campaigns in the category "*technology*" reach their goal. Campaigns in the category "*games*" collected USD 487.4 million in comparison to USD 6.24 million for campaigns in the category "*dance*".

Finding 9:	Roughly 46% of the Kickstarter campaigns have a duration of 30 days.
Finding 10:	The mean campaign duration decreased from roughly 60 days to roughly 30 days from 2009 to 2013 and has remained stable since then. Failed campaigns have a longer campaign duration on average.
Finding 11:	Campaigns with a duration between 15 and 30 days have the highest success rates.
Finding 12:	The number of campaigns increased steadily from 2009 to 2014, but went down at the beginning of 2015. Generally, there is an increase in the middle of the year and a decrease at the turn of the year.
Finding 13:	The number of campaign launches is higher in the middle of the year than at the turn of the year for campaigns in the cateogories "*dance*", "*comics*", "*film & video*" and "*theater*". While the category "*comics*" has grown over the years, the categories "*film & video*" and "*music*" had fewer launches in 2015 than in the years before.
Finding 14:	The success rate decreases the higher the funding goal is. Campaigns with a funding goal of USD 10,000 have a success rate of roughly 40%. The success rate decreases to roughly 30% for funding goals of USD 30,000, respectively to 20% for funding goals of USD 60,000.
Finding 15:	The success rate of Kickstarter campaigns decreased from roughly 60% in 2009 to 35% in 2015.
Finding 16:	While the success rate of the category "*design*" increased from 2010 to 2015, the success rates of the categories "*dance*", "*food*" and "*photography*" decreased during the considered period.
Finding 17:	More money is pledged in the middle of the year than at the turn of the year. The average money pledged per campaign rose from roughly USD 1 million in 2009 to a maximum of USD 80 million in 2015.

Finding 18:	The categories "*crafts*", "*fashion*" and "*technology*" followed a straight upward trend regarding money pledged from 2009 to 2015.
Finding 19:	The amount pledged per backer rose from roughly USD 70 in 2009 to USD 90 in 2015.

3.2.2. Live Campaigns Dataset

Characterization

A short characterization of the "*Live Campaigns Dataset*" can be seen in Table 3.3. We crawled campaigns between April 22, 2017 and December 1, 2017. The "*Live Campaigns Dataset*" contains roughly 39,000 campaigns with a success-rate of 38.1%. The canceled and suspended campaigns are pooled, as we could not always determine why a campaign terminated before its actual end. The dataset contains 126 snapshots for every campaign on average.

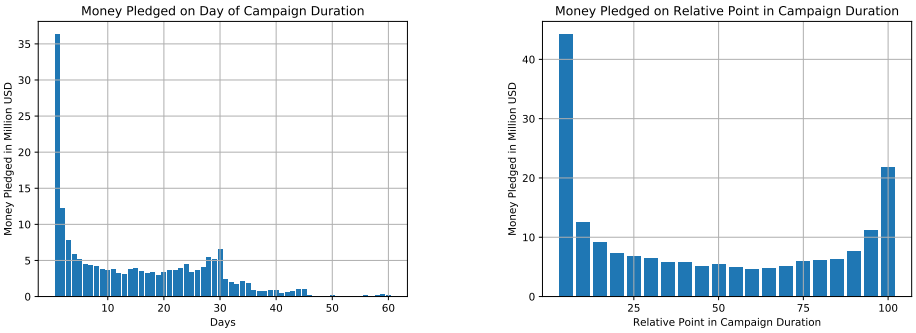
The number "*money pledged overall*" only considers campaigns that collect pledges in U.S. dollars. The number would be slightly higher if we would have considered other currencies with their corresponding exchange rates. Both numbers only consider successful campaigns, that means USD 291.0 million dollars were actually given to campaign creators to finance their campaign.

Money Pledged

The different time series for every campaign allows us to investigate the typical evolution of Kickstarter campaigns. In Figure 3.20 we analyze how much money is pledged on which point in time of the campaign duration.

Table 3.3.: Short description of "*Live Campaigns Dataset*"

Crawling period	April 22, 2017 - December 1, 2017
Number of campaigns -successful -failed -canceled / suspended	38,762 14,779 19,480 4,503
Mean number of snapshots	126
Success rate	38.1%
Money pledged overall USD	USD 291.0 M
Backers overall	4.1 M



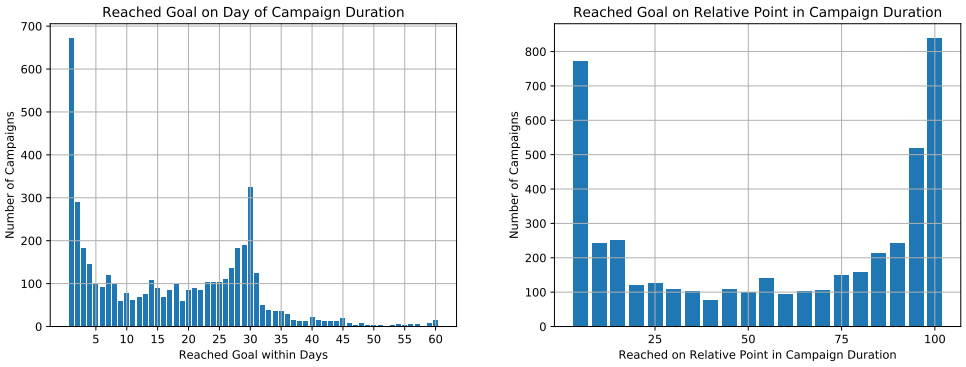
(a) Distribution of the pledged money depending on the day of the campaign duration

(b) Distribution of the pledged money depending on the relative campaign duration

Figure 3.20.: **Pledged Money during Duration.** In Fig. 3.20a we illustrate the days of the campaign duration on the x-axis and the money pledged on the respective day of the duration on the y-axis. We can see that most money was pledged on the first two days and on day 30th day of the duration. In Fig. 3.20b we made the y-axis relative in order to compare campaigns with different durations. We see that most money is pledged at the beginning and the end of the campaign.

In Figure 3.20a we illustrate the day of the campaign duration on the x-axis and the money pledged for all campaigns on the y-axis. We can see that on the first day of the campaign duration by far the most money is pledged with roughly USD 36.3 million, which is 19.4% of the money pledged overall. Another peak appears at the 30th day of the duration with roughly USD 6 million pledged. Figure 3.20b shows the relative point in time of the campaign duration on the x-axis and the money pledged for all campaigns on the y-axis. We can see that there is a peak at the first 5% of the duration with USD 44.2 million pledged. Another peak appears at the last 5% of the campaign duration with USD 21.9 million. We conclude that most money is pledged at the beginning and the end of a campaign. We argue that the second peak in Figure 3.20a is caused by the fact, that by far most Kickstarter campaigns have a duration of 30 days, which can be seen in Figure 3.9.

Finding 1: Most money is pledged at the beginning and towards the end of a Kickstarter campaign.



(a) Day on which successful campaigns reach their goal

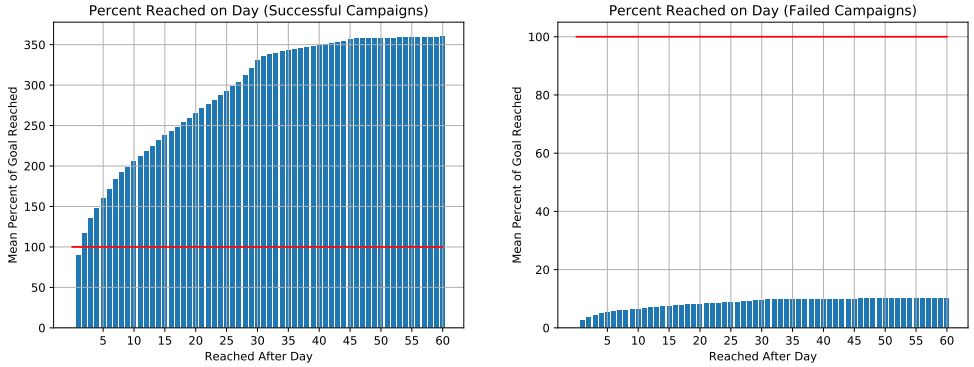
(b) Relative point in time at which campaigns reach their goal

Figure 3.21.: **Point in Time Campaigns Reach their Goal.** In Fig. 3.21a we illustrate the days of the campaign duration on the x-axis and on the number of campaigns the y-axis that became successful on that day. We see that 14.6% of the campaigns become successful on the first day. Another peak appears around the 30th day of the campaign duration. Fig. 3.21b shows the relative campaign duration on the x-axis and the number of campaigns that reached their goal at that relative point on the y-axis. Almost half of the successful campaigns reach their goal within the first 5% and the last 10% of their duration.

Reached Goal on Day

Figure 3.21 draws a similar picture as in Figure 3.20. Figure 3.21a illustrates the duration in days on the x-axis and the number of campaigns that reached their goal on that day of the duration on the y-axis. We can see that 14.6% of successful campaigns reached their goal on the first day. The other peak at day 30 illustrates that 7.1% of the successful campaigns reach their goal on that day. Figure 3.21b shows successful campaigns and when they reached their goal in relation to the campaign duration. We see that 16.8% of successful campaigns reach their goal within the first 5% of their duration. Even more campaigns, namely 18.3%, reached their goal within the last 5% of the duration. Apart from those mentioned periods relatively few campaigns reached their goal in the middle of the duration.

Finding 2: Most campaigns reach their goal at the beginning or towards the end of their duration.



(a) Percent of Goal reached on Day of Campaign Duration (Successful Campaigns)

(b) Percent of Goal reached on Day of Campaign Duration (Failed Campaigns)

Figure 3.22.: **Average Percent of Goal Reached on Day.** The figures illustrate the relation of money pledged and goal in percent on the y-axis. The x-axis shows the day of the campaign duration. Successful campaigns are illustrated in Fig. 3.22a and failed campaigns in Fig. 3.22b. We can see that successful campaigns reached over 100% on average after 3 days. Failed campaigns barely reach 10% of their goal on average through the whole duration.

Percent of Goal Reached on Day

In Figure 3.22 we illustrate how many percent of the goal were pledged on a certain day, both for successful and failed campaigns. The red line in both charts indicates 100%, which is the border for a successful funding. Figure 3.22a shows that successful campaigns pledged on average over 100% of their goal after 3 days. After 30 days successful campaigns reached on average over 300% of their goal on average. In contrast to that, the failed campaigns in Figure 3.22b barely reached 10% on average at the end of the funding period. We can see that successful campaigns exceed their goals threefold on average, while failed campaigns are quite far away from their goal on average.

Finding 3: Successful campaigns reach their goal on average quite early and often exceed their goals by considerable amounts, while failed campaigns are on average at around 10% of their goal by the end of the campaign.

Table 3.4.: List of Findings in the "*Live Campaigns Dataset*"

Finding 1:	Most money is pledged at the beginning and towards the end of a Kickstarter campaign.
Finding 2:	Most campaigns reach their goal at the beginning or towards the end of their duration.
Finding 3:	Successful campaigns reach their goal on average quite early and often exceed their goals by considerable amounts, while failed campaigns are on average at around 10% of their goal by the end of the campaign.

3.3. Predicting Success

We did both classification and regression in this thesis. The classification experiments aim at predicting whether a campaign will reach its funding goal or not. As there is already some work which predicts only the success of the campaign, we extended this task by introducing regression experiments to predict the amount of money a campaign will collect.

3.3.1. Predicting Campaign Outcomes

For the classification we applied different algorithms to our dataset. As we want to predict whether campaigns will reach their goal or not, we can represent it as a binary classification problem. In order to make the different classification experiments comparable, an accurate evaluation is essential. Therefore, we used 5-fold cross-validation, to verify the accuracy of our models. According to Brink et al. [BRF16] cross-validation provides a good way to compare different results with each other and is robust in terms of overfitted models.

In a k-fold cross-validation the whole dataset is split into k chunks. Next, one of those chunks is selected as test set, while the other chunks are used as training set. After that, another chunk is selected as test set and the others as training set. This procedure is repeated until every chunk was used as test set once.

For our work we used 5 folds for the validation, which means that 80% of our data are used as training set and 20% are used as test set to verify the model. Additionally, we stratified and shuffled the samples in order to minimize biases such as unbalanced training sets with unusual big number of successful campaigns. We chose a baseline for the samples of 0.5 with balanced train and test classes, which means that each test and training set contains 50% successful and 50% failed campaigns, which is different to the average success rate on Kickstarter.

It is essential for the evaluation to get meaningful key figures about the accuracy of the model. As we perform binary classifications "*accuracy*" (see Equation 3.1), "*precision*" (see Equation 3.2), "*recall*" (see Equation 3.3) and

"*F1 score*" (see Equation 3.4) are the means of choice according to Géron [Gér17]. While precision states the fraction of the selected elements that are relevant, recall is the proportion of relevant documents that are successfully retrieved. As both metrics represent a fraction, the best value is 1 and the worst possible value is 0. The F1 score is a combination of both figures and is calculated from the harmonic mean of both figures.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + false\ positives + true\ negatives + false\ negatives} \tag{3.1}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{3.2}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{3.3}$$

$$F1\ score = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.4}$$

Naive Bayes

Naive Bayes is a simple and fast algorithm for classification problems in terms of supervised learning, which is why we used it for our predictions. Supervised learning is an approach where we already know the target value for a set of features. In the first step we train the predictor with a training set, which includes the features and their target value. In the second step we evaluate the predictor with the test set. As we know the empirical target values of the test samples, we are able to calculate the error by comparing the empirical and the predicted values with each other.

Naive Bayes can be used with different distribution functions, such as Gaussian, Bernoulli or Multinomial distribution. Generally, the distribution function, which describes the data best, is used. According to Kubat [Kub15] the Naive Bayes algorithm assumes that the trained features are independent

from each other. The algorithm basically counts how often a certain feature appears in combination with a certain class and calculates the probability for this. After that the algorithm estimates the relative frequency of each class in the training set and calculates the conditional probability for each of them. In the end, the class with the highest probability is chosen.

Example

In this example our dataset consists of samples with 2 features and 2 target classes. The features are "*long*" and "*yellow*" and the target classes are "*banana*" and "*other fruit*". For reasons of simplification each feature is either 0 or 1. Table 3.5 provides an overview of the dataset.

Table 3.5.: Example Dataset for Naive Bayes

class	long	yellow	total
banana	4	3	5
other fruit	3	4	10

Then we calculate the possibility for each feature in the class to be present and the probability for the class overall. The probabilities can be seen in Table 3.6.

Table 3.6.: Probabilites for Example Dataset for Naive Bayes

class	long	yellow	total
banana	0.8	0.6	0.33
other fruit	0.3	0.4	0.67

If we now want to predict whether a "*long*" and "*yellow*" fruit is a banana or not, we have to calculate the possibility for each class for a long and yellow fruit. For this reason, we multiply the possibilities for each feature and for the class overall for both classes.

$$P_{banana} = 0.8 * 0.6 * 0.33 = 0.16$$

$$P_{otherfruit} = 0.3 * 0.4 * 0.66 = 0.08$$

As we can see the possibility for a long and yellow fruit is 0.16 for the class "*banana*" and 0.08 for the class "*other fruit*". Therefore, the predictor will chose "*banana*" as target class.

Support Vector Machine

Support Vector Machines (SVM) also use the approach of supervised learning. They can be used both for classification and for regression. In our case, we use them for classification. Compared to Naive Bayes, they require a lot more computation time. In Support Vector Machines each feature is represented as a dimension, that means if you have samples with 2 features, each sample can be expressed as a point in a plane that belongs to a class. Support Vector Machines try to find an optimal decision boundary to separate those classes. More specifically Support Vector Machines try to maximize the separation margin between those classes. Support Vector Machines can be used with different kernels, for example, linear, polynomial or radial basis functions. The choice of the kernel depends on the distribution of the data. According to Hearst et al. [Hea+98] the area of applications is quite broad for Support Vector Machines. They can be used for text categorization, face detection, handwriting recognition, and many other applications.

Example

In this example we generate a dataset with 100 samples and 2 classes. The features are "*long*" and "*yellow*" and the target classes are "*banana*" and "*other fruit*". The features are represented as decimals.

In Figure 3.23 we can see that each sample of the dataset is represented as point in the chart. The blue points represent the samples of the class "*banana*" and the red ones represent "*other fruits*". Then we train a Support Vector Machine with linear kernel on this samples. The result of this procedure is the decision boundary that separates the blue and the red samples.

The decision boundary represents the function with the maximal margin between the samples of both classes. New samples will be classified depending on their location to the decision boundary. While samples in the blue area will be classified as "*banana*", samples in the red area will be classified as "*other fruit*".

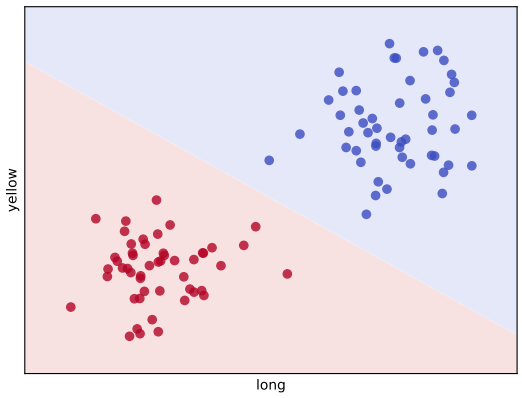


Figure 3.23.: Example of Support Vector Machine

Logistic Regression

Logistic regression is another supervised learning classifier and is mainly used for binary classification. Beforehand we used the "*MinMaxScaler*" from scikit-learn to scale the different features to adjust each feature to the same range. According to Géron [Gér17] Logistic Regression estimates the probability of a sample belonging to a particular class. It calculates a weighted sum of the input features and outputs the logistic of this result. The advantage of this algorithm is that it does not only classify the samples, it also shows the importance of the different features in form of coefficients.

Example

The dataset of this example only consists of the numbers from -5 to 5 as feature and a target value, which is 1 if the feature is positive and 0 otherwise. In Figure 3.24 we can see each sample represented as blue dot. The red line in the figure represents the logistic decision function. We can see that the boundary, if a sample is classified as 0 or 1, is located between 0 and 1 on the x-axis.

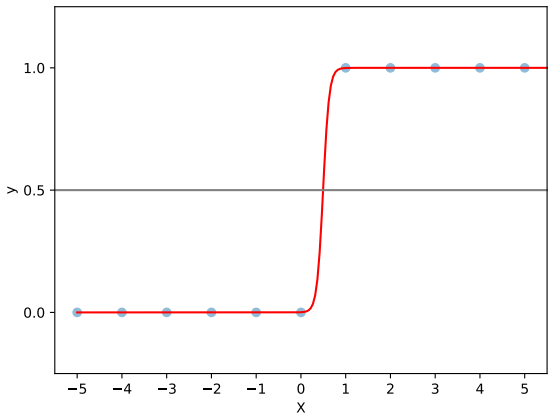


Figure 3.24.: Example of Logistic Regression

3.3.2. Predicting Pledge Amounts

In order to be able to predict how much money a campaign will collect, we made use of regression. For this task, we preprocessed our datasets as follows. First, we performed an outlier detection to remove outliers from our dataset. According to Barnett, Lewis, et al. [B+94] an outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data. Therefore, outliers can lead to a less accurate model.

In our work we used "*One-Class SVM*" for the outlier detection. It works similar to the conventional Support Vector Machine, with the difference that it is an approach of unsupervised learning and can therefore be used for the detection of outliers. We used radial basis function as kernel and filtered 5% of the samples. As filter criterion we used the goal and the ratio of money pledged and goal.

Next, we scaled our features to the same range with the help of "*MinMaxScaler*". This step helps us to compare the coefficients of the features with each other. To determine the best configuration of our regression experiments, we conducted a grid search over all configuration parameters of the regression.

Similar to the classification tasks, we also used k-fold cross-validation with 5 folds.

For the calculation of the accuracy of the prediction we use two different metrics: "*root mean squared error (RMSE)*" (See Equation 3.5) and "*mean absolute error (MAE)*" (See Equation 3.6). The predicted values are represented by the vector $\hat{Y}$ and Y is the vector of the empirical values from the training set.

$$root\ mean\ squared\ error(RMSE) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2} \tag{3.5}$$

$$mean\ absolute\ error(MAE) = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{3.6}$$

Linear Regression

Linear Regression can be used for various regression problems. We consider that every sample in the dataset is represented as point in a cloud where the features of the sample determine the position. According to Raschka [Ras15] the algorithm calculates a linear function which has the smallest distance to all samples in the cloud.

The problem with this approach is that the result of the algorithm is basically a linear function and therefore negative values are possible in the prediction. In contrast to this, the target values in our dataset are always positive. Apart from that, we have also seen that the distribution of our target variable is not linear but likely log normal.

Example

The dataset of this example consists of 1 feature and 1 target class. As feature we have the height of the person and the target value is the shoe size. When we train the linear regression with the samples, we get a line as result which is shown in Figure 3.25. The points in the figure show the different samples which we used for training. The result shows that a person with

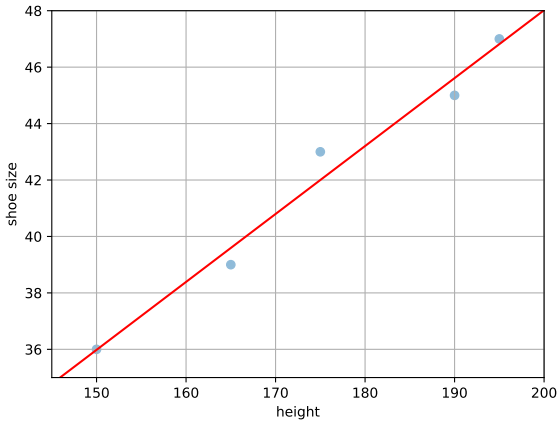


Figure 3.25.: Example of Linear Regression

a height of 175 cm is likely to have shoe size 42 according to our training samples.

Ridge Regression

According to Géron [Gér17] Ridge Regression is a regularized version of Linear Regression and is therefore also a linear model. The algorithm is optimized to keep the model weights as small as possible. Hence, it is less prone to overfitting.

As we found that the distribution of the data is not linear, we decided to perform a polynomial regression, which is possible with "*Kernel Ridge Regression*". This version of Ridge Regression can be used with different kernels such as radial basis functions, sigmoid or polynomial. This is possible due to the usage of the "*kernel trick*". This operation transforms the feature space into a higher dimension where the data is linearly separable.

3.4. Feature Engineering

In the following sections we will introduce three different models that are used for the prediction experiments. These models follow different approaches to identify the factors that make a campaign successful. As Kickstarter follows the "*all or nothing*" principle, we define success as reaching the funding goal. Additionally, we decided to remove noise from our data.

To that end, we introduced following filter criteria to minimize noise in our dataset:

- only campaigns with a minimum of 10 backers
- only campaigns with a goal of minimum USD 1,000
- no USD 0-pledged-campaigns

We argue that campaigns with a goal less than USD 1,000 and less than 10 backers are not really "*crowdfunding*" campaigns. Further, we want to remove campaigns with little to no information (e.g. USD 0-pledged-campaigns).

The number of campaigns that are included in each of the following datasets can vary from each other. This is because of missing time series in the data. Time series could be missing when the crawler discovers new campaigns too late or due to connection problems. Generally, we only considered campaigns with complete time series.

3.4.1. Static Features Dataset

The "*Static Features Dataset*" mainly contains data that is available right from the start of the campaign. Additionally, we added the number of "*Frequently Asked Questions*" and "*comments*" to the dataset. There are no time or money dependent features included. All features are listed in Table 3.7.

We considered the length of the campaign description and the text of the "*Risk and Challenges*" section. Also, the usage of images within the description, the presence and length of the pitch video are included in the dataset.

Table 3.7.: Features of Static Features Dataset

Feature	Description
duration	campaign duration
pitch video length	length of pitch video on the campaign site
sum comments	number of comments
sum faqs	number of FAQs
sum pledge amounts	number of different rewards
num project creator	number of campaigns the creator started before
num images	number of images in the campaign description
len description	length (in characters) of the campaign description
len risks	length (in characters) of the *"Risk & Challenges"* section
success	goal reached (target variable for classification)
money pledged overall	money pledged at the end of the campaign (target variable for regression)

3.4.2. Extended Dataset

The *"Extended Dataset"* extends the *"Static Features Dataset"* with time-series information and additional attributes about the rewards. The feature *"count changes"* indicates the number of changes that were made by the creator. This includes features that can only be manipulated by the creator, for example, the description, the number of rewards or the pitch video. We were able to use this feature as we collected 4 - 5 snapshots of a campaign per day that can be compared over the whole campaign duration. Other time-dependent features are available after 6, 12, 24 and 48 hours.

We decided to use different snapshots to investigate if there is an improvement in accuracy if we use time information from a longer period, for example, if the dataset with features aggregated up to 48 hours performs better than the one after 6 hours.

Time dependent information includes money pledged, backers, and comments after a certain time. An overview of the features can be seen in Table 3.8.

Table 3.8.: Features of Extended Dataset

Feature	**Description**
goal	funding goal
pledged 6/12/24/48 h	money pledged after 6/12/24/48 hours
backers 6/12/24/48 h	number of backers after 6/12/24/48 hours
sum comments 6/12/24/48 h	number of comments after 6/12/24/48 hours
duration	campaign duration
pitch video length	length of pitch video on the campaign site
sum faqs	number of FAQs
sum pledge amounts	number of different rewards
min pledge amounts	price of cheapest reward
max pledge amounts	price of most expensive reward
mean pledge amounts	average price of rewards
num project creator	number of campaigns the creator started before
num images	number of images in the campaign description
len description	length (in characters) of the campaign description
len risks	length (in characters) of the "*Risk & Challenges*" section
count changes	number of changes made in the campaign tab during the duration
success	goal reached (target variable for classification)
money pledged overall	money pledged at the end of the campaign (target variable for regression)

3.4.3. Extended Dataset without Pledged

The "*Extended Dataset without Pledged*" is similar to the "*Extended Dataset*" with the difference that there is no information about the amount pledged included in the dataset. The numbers after 6 and 12 hours are fixed in the data-set to include information about the progression of the campaign. We also limited the time dependent information to 12 hours as many campaigns reach their goal quite early (see Figure 3.21a).

An overview of the features can be seen in Table 3.9.

Table 3.9.: Features of Extended Dataset without Pledged

Feature	**Description**
goal	funding goal
backers6h	number of backers after 6 hours
backers12h	number of backers after 12 hours
sum comments hours6	number of comments after 6 hours
sum comments hours12	number of comments after 12 hours
duration	campaign duration
pitch video length	length of pitch video on the campaign site
sum faqs	number of FAQs
sum pledge amounts	number of different rewards
min pledge amounts	price of cheapest reward
max pledge amounts	price of most expensive reward
mean pledge amounts	average price of rewards
num project creator	number of campaigns the creator started before
num images	number of images in the campaign description
len description	length (in characters) of the campaign description
len risks	length (in characters) of the "*Risk & Challenges*" section
count changes	number of changes made in the campaign tab during the duration
success	goal reached (target variable for classification)
money pledged overall	money pledged at the end of the campaign (target variable for regression)

4. Results & Evaluation

In this chapter we describe the results of the classification and regression experiments outlined in Section 3.3. The configuration parameters for each model are the results of a grid search. To determine the best configuration we used "*accuracy*" for the classification experiments and "R^2" for the regression experiments.

4.1. Classification

In the following section we provide the results of the classification models introduced in Section 3.3.1 with the datasets introduced in Section 3.4. The number of campaigns for the "*Static Dataset*" is 8,488. The "*Extended Dataset*" includes 5,134 and the "*Extended Dataset without Pledged*" 5,136 campaigns.

4.1.1. Naive Bayes

Since Naive Bayes can be used with different distributions, we conducted the experiments with Gaussian and Multinomial distributions. The Naive Bayes model with Gaussian distribution only allows one parameter to adjust, which is the prior probability of the classes. As we balanced our classes and created the train- and test-splits using a stratified shuffle split approach, the baseline for Naive Bayes is 0.5, which means the number of successful and failed campaigns corresponds to each other. We decided not to adjust the prior probabilities, because the probability for each class would be 0.5 in the model by default, which is basically what we want.

Table 4.1 illustrates the results of the Naive Bayes classification with Gaussian distribution. We can see that the results of all datasets are similar, but always outperform the baseline of 0.5. The numbers for precision, recall and F1 score differ from each other across all datasets of this model.

We expected that the datasets which contain time series would perform considerably better, but we can see that the "*Static Dataset*" has the same accuracy as the "*Extended Dataset without Pledged*" with an accuracy of 0.59. All other datasets reached an accuracy of 0.58.

Table 4.1.: Results of Classification with Naive Bayes (Gaussian Distribution)

Dataset	Precision	Recall	F1 Score	Accuracy	Baseline Accuracy
Static Dataset	0.71	0.59	0.52	0.59	0.50
Ext. Dataset 6 h	0.59	0.58	0.56	0.58	0.50
Ext. Dataset 12 h	0.74	0.58	0.50	0.58	0.50
Ext. Dataset 24 h	0.75	0.58	0.50	0.58	0.50
Ext. Dataset 48 h	0.74	0.58	0.50	0.58	0.50
Ext. Dataset w/o Pled.	0.72	0.59	0.53	0.59	0.50

Another approach was Naive Bayes with Multinomial distribution. To find the best configuration, we conducted a grid search. It can happen that there are no probabilities available in the model for certain features, because they were not covered in the training set. Therefore, we need to smooth the probabilities with an additive Laplace smoothing, which is represented by the parameter "*alpha*". As possible values we chose the numbers from 0.001 to 10 in logarithmic steps, which means that every value is the previous value multiplied by 10. As a result of the grid search, we set the parameter alpha to 1. For the same reasons as mentioned before in the Naive Bayes models with Gaussian distribution, we did not set any prior class probabilities.

The results of Naive Bayes with Multinomial distribution in Table 4.2 generally show an improvement compared to the results of the Naive Bayes with Gaussian distribution. The accuracy of the "*Static Dataset*" and the "*Extended Dataset without Pledged*" slightly improved from 0.59 to 0.61.

The models with the "*Extended Dataset*" improved even more. The experiment with the "*Extended Dataset 48 h*" reached an accuracy of 0.75, which

is an improvement of 0.17 to the model with Gaussian distribution. This is also the best result we achieved with Naive Bayes models. The other models of the "*Extended Dataset*" improved by 0.11 to 0.16. We can also see that the numbers for precision, recall and F1 score are roughly the same compared to each other within the same dataset. This stands in contrast to the models with Gaussian distribution, where the numbers for these metrics fluctuate more within the same dataset.

Table 4.2.: Results of Classification with Naive Bayes (Multinomial Distribution)

Dataset	Precision	Recall	F1 Score	Accuracy	Baseline Accuracy
Static Dataset	0.72	0.61	0.56	0.61	0.50
Ext. Dataset 6 h	0.70	0.69	0.69	0.69	0.50
Ext. Dataset 12 h	0.73	0.73	0.73	0.73	0.50
Ext. Dataset 24 h	0.75	0.74	0.74	0.74	0.50
Ext. Dataset 48 h	0.77	0.75	0.75	0.75	0.50
Ext. Dataset w/o Pled.	0.65	0.61	0.59	0.61	0.50

Generally, we can say that all of our Naive Bayes models outperform the baseline of 0.5. The results for the models with Gaussian distribution are similar across all datasets with accuracies of 0.58 and 0.59. The models with Multinomial distribution improved our results and achieved an accuracy of 0.75 when trained on the "*Extended Dataset 48 h*". Additionally, we can see that the results of the "*Extended Dataset*" improve if we include more time series.

4.1.2. Support Vector Machine

Support Vector Machines with linear kernel were another approach in our work to predict the success of campaigns. Additionally, we set the kernel coefficient "*gamma*" to 0.0001 and the tolerance for the stopping criterion to 0.001. We determined both parameters with the help of grid search, in which we chose the values from 0.00001 to 100 in logarithmic step as possible values.

Table 4.3 shows the results of the Support Vector Machine experiments. The Support Vector Machine models further improved the accuracy by 0.02

to 0.13 for the same dataset, compared to the Naive Bayes models with Multinomial distribution.

The model with the "*Static Dataset*" reaches an accuracy of 0.67, compared to 0.61 with the best Naive Bayes model. The models with the "*Extended Dataset*" also perform better by 0.02 to 0.05. In general, we can see that the accuracy increases the more information is included in the dataset. The accuracy of the "*Extended Dataset 6 h*" is 0.72 and goes up to 0.80 for the "*Extended Dataset 48 h*".

The biggest improvement is achieved by the model with the "*Extended Dataset without Pledged*" with an accuracy of 0.74, compared to 0.61 with the best Naive Bayes model with the corresponding dataset. Generally, we can see that the datasets which include time information perform better than the "*Static Dataset*".

Table 4.3.: Results of Classification with Support Vector Machine

Dataset	Precision	Recall	F1 Score	Accuracy	Baseline Accuracy
Static Dataset	0.69	0.67	0.66	0.67	0.50
Ext. Dataset 6 h	0.73	0.72	0.72	0.72	0.50
Ext. Dataset 12 h	0.75	0.75	0.75	0.75	0.50
Ext. Dataset 24 h	0.78	0.78	0.78	0.78	0.50
Ext. Dataset 48 h	0.80	0.80	0.80	0.80	0.50
Ext. Dataset w/o Pled.	0.74	0.74	0.74	0.74	0.50

The results of the Support Vector Machine with linear kernel are generally better than the results of the Naive Bayes models. Also, the values for precision, recall, F1 score, and accuracy are quite similar to each other for the same dataset. We can see that even the "*Static Dataset*" outperforms the baseline of 0.5 by 0.17. We achieve the best result with the "*Extended Dataset 48 h*" with an accuracy of 0.80.

4.1.3. Logistic Regression

Finally, we also applied Logistic Regression to our datasets. As penalization norm we decided to use "*L2*". The values for the tolerance of the stopping

criterion "*tol*" was set to 0.0001. This means we stop if the error of our model on the training set, within 2 periods, is smaller than "*tol*". The inverse of the regularization strength "*C*" was set to 1. We also set the parameter "*fit_intercept*" to "*True*". This configuration was the result of the grid search we conducted, where we tried the values from 0.00001 to 100 for the parameters "*tol*" and "*C*". For the penalty we tried the "*L1*" and the "*L2*" norm and for "*fit_intercept*", which is represented as a boolean, we tried both possible values.

The results of the Logistic Regression models, which we can see in Table 4.4, are similar to the results of the Support Vector Machine models. The "*Static Dataset*" reaches an accuracy of 0.67. Precision is 0.69, recall is 0.67 and the F1 score is 0.66. These numbers are the same as with the corresponding Support Vector Machine model with the same dataset.

The models with the "*Extended Dataset*" also perform similarly to the Support Vector Machine models. We can see that the accuracy increases if more information is included in the dataset. The accuracy of the "*Extended Dataset 6 h*" is 0.72 and goes up to 0.80 for the "*Extended Dataset 48 h*".

The "*Extended Dataset without Pledged*" reaches an accuracy of 0.74. The accuracy is the same as for the corresponding Support Vector Machine model and 0.13 better than the best Naive Bayes model with the same dataset.

One advantage of Logistic Regression in comparison to Naive Bayes and Support Vector Machines is that the algorithm not only classifies the samples, but also computes coefficients. These coefficients show us the influence of each feature for the prediction. The coefficients for all Logistic Regression models are shown in Appendix B.1.

We can see that the feature "*sum comments*" and "*num project creator*" are the highest positive coefficients for the "*Static Dataset*" model. The feature "*pitch video length*" also has a small positive coefficient in this dataset. We can also see that the coefficient of "*sum comments 6 hours*" is negative and "*sum comments 12 hours*" is positive in the "*Extended Dataset withouth Pledged*".

Table 4.4.: Results of Classification with Logistic Regression

Dataset	Precision	Recall	F1 Score	Accuracy	Baseline Accuracy
Static Dataset	0.69	0.67	0.66	0.67	0.50
Ext. Dataset 6 h	0.72	0.72	0.72	0.72	0.50
Ext. Dataset 12 h	0.76	0.76	0.76	0.76	0.50
Ext. Dataset 24 h	0.78	0.78	0.78	0.78	0.50
Ext. Dataset 48 h	0.80	0.80	0.80	0.80	0.50
Ext. Dataset w/o Pled.	0.74	0.74	0.74	0.74	0.50

4.2. Regression

In the following section we provide the results of the regression models introduced in Section 3.3.2 with the datasets introduced in Section 3.4. The "*Static Dataset*" includes 13,355 campaigns. The "*Extended Dataset*" comprises of 7,658 campaigns and the "*Extended Dataset without Pledged*" includes 7,661 campaigns. The numbers of campaigns differ from the numbers from Section 4.1, as we considered all campaigns without introducing a baseline depending on the successful funding. Therefore, the relation of successful and failed campaigns corresponds to the relation in the whole dataset.

4.2.1. Linear Regression

First, we applied Linear Regression models to our datasets. As we can adjust the two parameters "*fit_intercept*" and "*normalize*", which are both boolean values, we performed a grid search with the possible values. As a result, we set the parameter "*fit_intercept*" to "*True*" and "*normalize*" to "*False*".

The results of the Linear Regression models, which we can see in Table 4.5, are considerably different depending on the chosen dataset. The "*Static Dataset*" reaches an R^2 of 0.36. The RMSE is USD 132,212.87 and the MAE is USD 29,963.80.

The "*Extented Dataset 6 h*" achieves an R^2 of 0.71 and a RMSE of USD 52,373.45 and a MAE of USD 16,260.42. If we include time series up to 48 hours the R^2

goes up to 0.89. The numbers for RMSE and MAE decrease to USD 28,622.33 and USD 10,212.00, respectively.

The result of the "*Extended Dataset without Pledged*" is between the "*Extended Dataset*" and the "*Static Dataset*" with an R^2 of 0.62 and a RMSE of USD 52,967.09 and a MAE of USD 17,948.46.

The coefficients for the Linear Regression models are shown in Appendix B.2. The coefficients of the features "*sum comments*" and "*sum faqs*" are rather high for the "*Static Dataset*". The coefficients of the "*Extended Dataset without Pledged*" show that the coefficient of "*sum comments 6 hours*" is negative and "*sum comments 12 hours*" is positive, which is similar to the coefficients of the Logistic Regression model. The features "*pledged 6/12/24/48 hours*" and "*backers 6/12/24/48 hours*" in the "*Extended Dataset*" are mainly positive.

Table 4.5.: Results of Linear Regression

Dataset	RMSE	MAE	R^2
Static Dataset	132,212.87	29,963.80	0.36
Ext. Dataset 6 h	52,373.45	16,260.42	0.71
Ext. Dataset 12 h	33,172.64	12,138.96	0.85
Ext. Dataset 24 h	30,281.87	11,022.45	0.86
Ext. Dataset 48 h	28,622.33	10,212.00	0.89
Ext. Dataset w/o Pled.	52,967.09	17,948.46	0.62

4.2.2. Ridge Regression

Ridge Regression is another linear regression model to which we applied our datasets. Just as for the Linear Regression models, we set the parameters "*fit_intercept*" to "*True*" and "*normalize*" to "*False*". The parameter "*alpha*", which stands for the strength of the regularization, was set to 1 and "*tol*", which represents the convergence criteria, to 0.001. The values for both numbers are based on a grid search, where we chose the values from 0.00001 to 100 in logarithmic steps as possible values.

Table 4.6 shows the results of the Ridge Regression experiments. Generally, the results are similar to the results of the Linear Regression. The R^2 of the "*Static Dataset*" is 0.34 and therefore lower by 0.02. The numbers for RMSE and MAE are also a bit higher with USD 133,451.66 and USD 30,045.96.

The R^2 of the "*Extended Dataset 6 h*" is a bit lower compared to the corresponding Linear Regression model with 0.67. Similar to the Linear Regression models the R^2 goes up if more time series are added to the dataset. We achieve the best result with the "*Extended Dataset 48 h*" with an R^2 of 0.89, which is the same as for the corresponding Linear Regression model. The RMSE of USD 27,933.87 is roughly USD 700 lower than the RMSE of the Linear Regression model. The MAE of USD 10,182.79 is roughly the same.

The R^2 of the "*Extended Dataset without Pledged*" is 0.65 and therefore shows an improvement of 0.03 to the Linear Regression model with the same dataset. The RMSE is USD 51,070.13 and the MAE is USD 17,533.26 for this model.

The coefficients of the Ridge Regression models, which are stated in Appendix B.3, are similar to the coefficients of the Linear Regression model. The features "*sum comments*" and "*sum faqs*" in the "*Static Dataset*" have positive coefficients and therefore indicate if a campaign is likely to collect more money in total. According to the coefficients of the Logistic Regression and the Linear Regression models we can see that the coefficient of "*sum comments 6 hours*" is negative and "*sum comments 12 hours*" is positive in the "*Extended Dataset without Pledged*".

Table 4.6.: Results of Ridge Regression

Dataset	RMSE	MAE	R^2
Static Dataset	133,451.66	30,045.96	0.34
Ext. Dataset 6 h	51,730.02	16,273.05	0.67
Ext. Dataset 12 h	33,611.06	12,102.66	0.83
Ext. Dataset 24 h	31,836.76	11,086.85	0.86
Ext. Dataset 48 h	27,933.87	10,182.79	0.89
Ext. Dataset w/o Pled.	51,070.13	17,533.26	0.65

4.2.3. Kernel Ridge Regression

Apart from the linear regression models we also conducted Kernel Ridge Regression experiments with a polynomial kernel of second order. We set the parameter "*alpha*", which is responsible for reducing the variance of estimates in this context, to 1 and the zero-coefficient "$coef_0$" also to 1. The values for both parameters are based on a grid search, where we chose the values from 0.1 to 10.0 in steps of 0.1 as possible values.

The results in Table 4.7 of the Kernel Ridge Regression with polynomial kernel vary across the different datasets. The "*Static Dataset*" achieves an R^2 of -0.21, which indicates that the model does not really fit to the data. The RMSE with USD 175,592.71 is roughly USD 40,000 higher than the RMSE of the linear models.

The "*Extended Dataset 6 h*" reaches an R^2 of 0.47 and a RMSE of USD 23,967.66 and a MAE of 10,331.54. In contrast to the linear models, the "*Extended Dataset without Pledged*" outperforms the "*Extended Dataset 6 h*" with an R^2 of 0.56. The RMSE of the "*Extended Dataset without Pledged*" is USD 21,824.06 and the MAE is USD 9,593.24. Compared to the Ridge Regression model with the same dataset the RMSE is roughly USD 30,000 lower and the MAE is roughly USD 8,000 lower, but also the R^2 is lower by 0.09.

According to the results of the linear models the results for the "*Extended Dataset*" improve if more time series are added to the dataset. The R^2 rises from 0.47 for the "*Extended Dataset 6 h*", over 0.63 and 0.71 for the "*Extended Dataset 12 h*" and the "*Extended Dataset 24 h*" to a maximum of 0.73 for the "*Extended Dataset 48 h*". Although the R^2 of this model is 0.16 lower than the

R^2 of the corresponding model with Ridge Regression, the RMSE is roughly USD 10,000 and the MAE is roughly USD 3,000 lower with USD 17,141.00 and USD 7,355.30.

Table 4.7.: Results of Kernel Ridge Regression with Polynomial Kernel of Second Order

Dataset	RMSE	MAE	R^2
Static Dataset	175,592.71	27,943.98	-0.21
Ext. Dataset 6 h	23,967.66	10,331.54	0.47
Ext. Dataset 12 h	20,308.36	8,714.38	0.63
Ext. Dataset 24 h	17,833.20	7,946.33	0.71
Ext. Dataset 48 h	17,141.00	7,355.30	0.73
Ext. Dataset w/o Pled.	21,824.06	9,593.24	0.56

5. Discussion

In this work we performed both classification and regression experiments with different datasets. The target of the classification experiments was to predict whether a campaign will reach its funding goal or not. The regression experiments aimed at predicting the exact amount a campaign will collect. To achieve these targets we built datasets with different sets of features and time series and applied well-known machine learning algorithms on them.

5.1. Classification

We first applied Naive Bayes models with Gaussian and Multinomial distribution to our datasets. These distributions describe how the features in the dataset are distributed, for example, when we use Naive Bayes with Gaussian distribution, we assume that the features are normally distributed.

When looking at the results of the Naive Bayes models, we noticed that the figures for precision and recall show variations within the same dataset. This especially applies to the models with Gaussian distribution. We discovered that the success rate among the predicted values is lower than 10%, although the predicted success rate should be 50%, according to the baseline. Hence, the precision of the prediction of successful campaigns is high, but the recall is quite low. For the failed campaigns it is the opposite.

As a second approach, we applied our datasets to Support Vector Machine models with linear kernel. The results show that the Support Vector Machine models improved the results by 0.03 to 0.13 depending on the used dataset. One explanation for this observation is that the improvement lays in the functionality of both algorithms. Naive Bayes basically counts the

probabilities for each feature and delivers the class with the highest probability. The features themselves are considered isolated from each other. Our dataset likely contains statistically dependent features, for example, the money pledged depends on the number of backers. If there are no backers, there is no money that can be pledged.

Etter et al. [EGT13] also used Support Vector Machines to predict whether a campaign reaches its funding goal or not. They included 4 hours of time series from the start of the campaign to their dataset and achieved an accuracy of 0.76. If we compare that to our model with the "*Extended Dataset 6 h*" with an accuracy of 0.72, we can see that the accuracy of our model is slightly worse, but in contrast to Etter et al. [EGT13], who included data from social networks, our dataset only contains data which is available on Kickstarter.

Logistic Regression is an algorithm which is intended for binary classification problems and is widely used. In addition to the classification, Logistic Regression provides information about the influence of each feature on the outcome of the prediction. Hence, we decided to use this algorithm in our work. The results of the Logistic Regression and Support Vector Machine models are similar. The values for precision, recall and F1 score only vary marginally. We assume that better results could only be achieved with further feature engineering and/or more data.

The coefficients of the Logistic Regression model of the "*Extended Dataset*" show that the goal has considerable negative influence on the success, meaning that the higher the goal, the lower the chances to reach the goal of the campaign. The time series features "*pledged 6/12/24/48 hours*" and "*backers 6/12/24/48 hours*" are also good indicators for the prediction, which is hardly surprising as they are most likely related to the success of a campaign. The feature "*pledged 6/12/24/48 hours*" can indicate how much money will be pledged overall, which is basically the definition of success in combination with the goal. As backers tend to pledge amounts lower than USD 100 (see Figure 3.7), there is also an adequate number of backers needed for a certain amount of money.

The coefficients of the "*Extended Dataset without Pledged*" confirm the observation that a higher goal reduces the chances of reaching the goal. We can also see that the coefficient of "*sum comments 6 hours*" is negative and

"*sum comments 12 hours*" is positive, which is contradictory. We explain this observation with users who start a discussion on the campaign right after the start. Reasons for these discussions could be the demonstration of misconception of the campaign or the aim of drawing attention to the creator that their idea is not new or even legally dubious.

Li et al. [LRR16] also used Logistic Regression models to predict the success of crowdfunding campaigns on Kickstarter and reached an accuracy of 0.90. In comparison to our models, they included information of the campaigns from Twitter and time series of the first 72 hours. Our best Logistic Regression Model reaches an accuracy of 0.80 after 48 hours. We hypothesize that we would reach a similar accuracy with our features after 72 hours of each campaign have passed. However, additional experiments would be required to confirm this claim.

Greenberg et al. [Gre+13] used information from the platform itself, but also linkages to social networks. However, the used dataset does not include any temporal data, similar to our "*Static Dataset*". They achieved an accuracy of 0.68, which is comparable with the results of the "*Static Dataset*" with an accuracy of 0.67.

Chung and Lee [CL15] also used data which is directly available on Kickstarter. They achieved their best result with AdaBoost M1 and reached an accuracy of 0.764. Their dataset only comprises static features, but includes features about the used language and the user. Therefore, the results are hardly comparable with our work. Most of our models that include time information outperform the models introduced in this work, but the accuracy of the models with the "*Static Dataset*" is lower by roughly 0.09.

Summarizing, we see that the inclusion of time series improves the accuracy of all models. Additionally, we are able to predict the success of a campaign 48 hours after the start with an accuracy of 0.80. Therefore, we conclude that the first hours and days are essential for the success of a campaign. Our results are comparable to the ones of other published work.

5.2. Regression

Apart from the classification experiments, where we predicted if campaigns will reach their goal, we also performed regression experiments to predict how much money campaigns will collect. To tackle this problem, we build different models with both linear and polynomial regressions. To our knowledge, no scientific work has been published by now where money pledged is predicted with regression models.

Linear Regression is one of the most well studied regression algorithms. The Linear Regression model with the "*Extended Dataset 48 h*" reaches an R^2 of 0.89, which is relatively high and indicates that the model fits quite good to the data. The MAE of roughly USD 10,000 is quite high, if we consider that many campaigns have a goal of USD 10,000 or less. We assume that this is caused by campaigns with rather high goals, for example, greater than USD 50,000. We note that the range of goals is pretty high (see Figure 3.4a). There are many campaigns with a goal of USD 10,000 or less, but also a considerable number of campaigns with a goal that goes into the hundred thousands or even millions. Additionally, we note that we filtered out campaigns with a goal less than USD 1,000 and campaigns with zero money pledged. We can observe that the datasets that include time series generally reach a higher R^2 and lower errors.

As explained in Section 3.3.2, Ridge Regression is a version of Linear Regression that is less prone to overfitting. Thus, it potentially improves the results. In fact, the results are similar to the results of Linear Regression. Only the model with "*Extended Dataset without Pledged*" exhibits a noticeable improvement of the R^2, namely 0.65 compared to 0.62 with the Linear Regression. The similar results suggest that overfitting is not a problem in the linear regression models.

The coefficients of the Linear Regression and Ridge Regression models only show marginal differences compared to each other. The time series features "*pledged 6/12/24/48 hours*", "*backers 6/12/24/48 hours*" and "*comments 6/12/24/48 hours*" are mainly positive and are also the highest coefficients in the "*Extended Dataset*". We claim that they are good indicators to predict if a campaign will collect more money overall. The time series features "*backers 6/12 hours*" and "*comments 6/12 hours*" in the "*Extended Dataset*

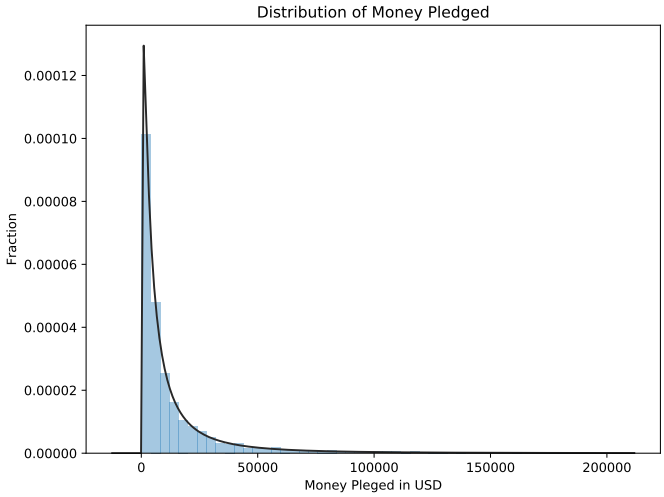


Figure 5.1.: **Distribution of Dependent Variable "*Money Pledged*".** On the x-axis we can see the money pledge in USD and on the y-axis the fraction of each bucket in the whole dataset. Additionally we fitted the log-normal distribution to the data (solid black line).

without Pledged" also have the highest coefficients in the model and therefore support this observation.

Figure 5.1 illustrates that the distribution of the dependent variable "*money pledged*" is not linear, but likely log-norm distributed. Therefore, we claim that the distribution of the dependent variable can be approximated with a polynomial function. As there are many polynomial regression algorithms, we conducted different experiments. As a result, we achieved the best results with Kernel Ridge Regression with a polynomial kernel of second order.

The "*Extended Dataset 48 h*" delivered the best result with an R^2 of 0.73 and a RMSE of USD 17,141.00 and a MAE of USD 7,355.30. Although the R^2 is lower by 0.16 than the R^2 of the model with Ridge Regression, the RMSE is roughly USD 10,000 and the MAE is roughly USD 3,000 lower. The model with the "*Extended Dataset without Pledged*" also has a RMSE of USD 21,824.06, which is also better than the RMSE of all Linear Regression models. We conclude that the time information is essential to make a decent prediction, especially for the polynomial models.

We mentioned that R^2 and RMSE are both lower for the model with polynomial kernel compared to the linear regression models. We explain this observation by the imbalance of samples in the dataset. The target value of most samples is between USD 1,000 and USD 50,000, but there are also few samples where the target values is over USD 1,000,000. We argue that the polynomial model delivers better results for campaigns with high funding goals, but has problems predicting lower funding goals. The influence of campaigns with lower funding goals on the metrics RMSE and MAE is much smaller, but the variance of the prediction, which is considered in the R^2, increases.

We observe that the results for the "*Extended Dataset*" and the "*Static Dataset*" show considerable differences. Therefore, we conclude that the inclusion of time information is important for the quality of the regression results. We also find that the time series feature "*pledged 6/12/24/48 hours*" improves the results for all models.

6. Conclusions & Future Work

In this work we examined the main factors that lead to success of crowdfunding campaigns on Kickstarter. We crawled live campaigns with frequent updates as well as already finished campaigns. On the one hand, we used these datasets to build models and apply different classification and regression algorithms in order to predict the success of the campaigns. On the other hand we used them to examine the empirical data with statistical evaluations.

We found that many campaigns receive only few dollars and that there are also very successful campaigns that attract multiple amounts of money of the initial goal. We also found that only few campaigns miss their goal by a small margin, meaning that if a campaign collects a considerable fraction of the goal, it is very likely that the campaign will reach its goal.

Apart from that, campaigns with a lower goal have higher chances of success, which can be seen in Figure 3.14. Figure 3.17a illustrates that more money is pledged in the middle of the year than at the turns of the year. This information could be used by campaign creators in order to optimize the point of time to start a campaign.

Most money is pledged at the beginning and towards the end of a campaign, which can be seen in Figure 3.20. Figure 3.22 shows that campaigns that pledge a considerable amount of money right after the start are likely to exceed their goal. Campaigns that are missing smaller amounts to their goal are also likely to reach the goal towards the end of the funding period.

The classification and regression experiments showed that the inclusion of time-based information of the first hours helps to accurately predict the outcome of a campaign. The best classification models have an accuracy of 0.72 after 6 hours and go up to an accuracy of 0.80 after 48 hours. The

best linear regression model achieved an R^2 of 0.89 and a RMSE of USD 27,933.87, the best polynomial model reached an R^2 of 0.73 and a RMSE of USD 17,141.00. These results suggest that the first few hours of a campaign are important for the success of a campaign.

The findings of the work could be useful for campaign creators and for potential backers. Creators can make use of the information to optimize their campaign and improve the chances of reaching their goal, respectively maximize the money to collect. The provided information in this thesis can also provide assistance for potential backers to identify potentially successful campaigns early.

For future work, we plan to examine the comments of a campaign and analyze the frequency and the network of users behind the comments to further improve the prediction models. This could also be adapted to the updates that are posted by the creator. We also plan to investigate the language that is used in the campaign, for example, in the description or in the comments. We want to examine if the language of the users has a considerable impact on the success of a campaign.

Additionally, we would like to further generalize our results by conducting similar analysis on other crowdfunding platforms and compare the differences and similarities between those platforms. This could provide assistance for potential creators to choose the best platform for their campaign.

Appendix

Appendix A.

Database Structure

A.1. Entity-Relationship Model of Database

Figure A.1 illustrates the entity-relation model of the database.

Appendix A. Database Structure

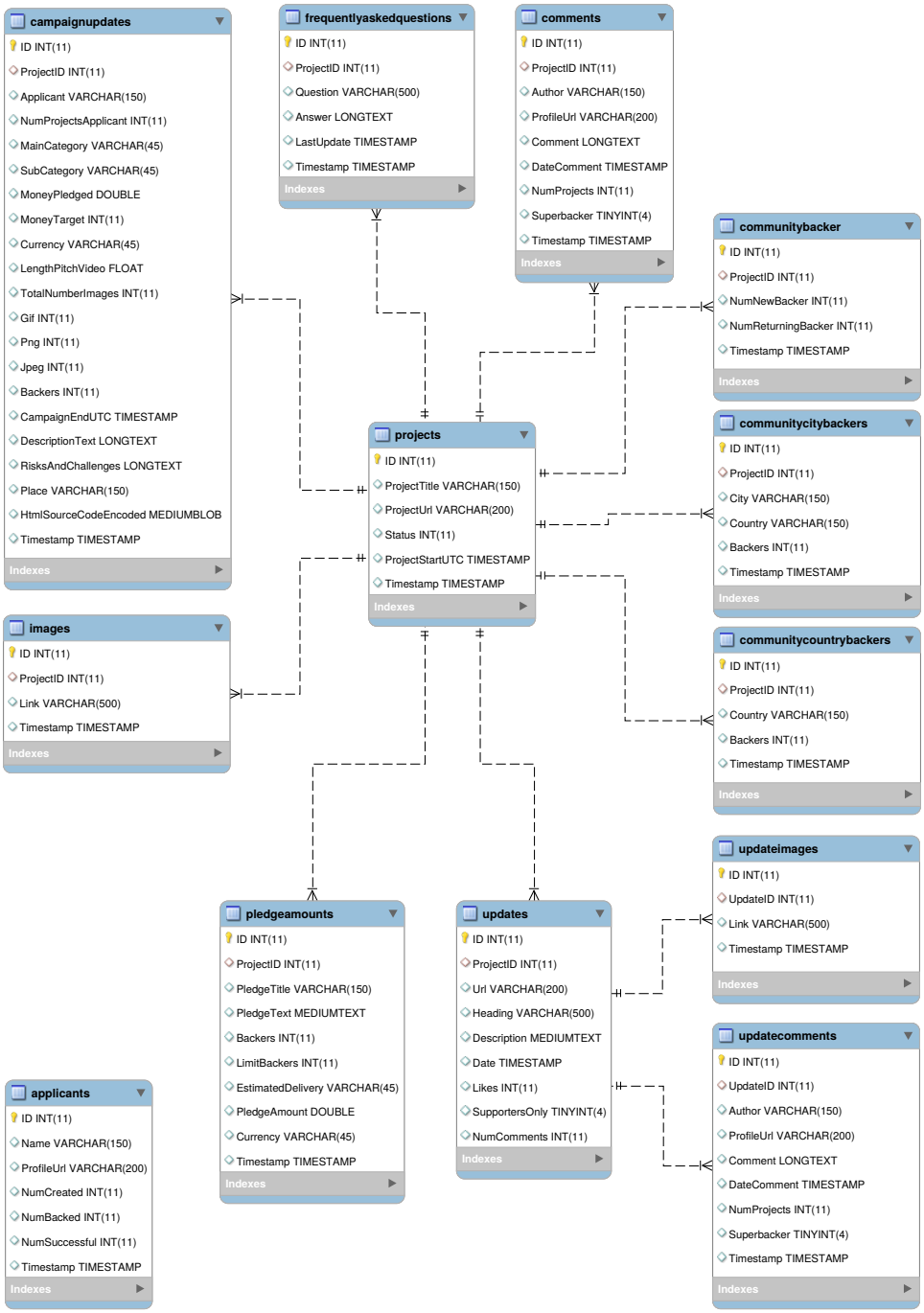


Figure A.1.: Entity–Relationship Model of the Database

A.2. Data of Live Campaigns

The features that are stated in the following tables are periodically crawled during the funding duration of the campaign.

Table A.1.: Attributes of the Database Table "*projects*"

projects	the campaign itself - main entity
ID	unique id
ProjectTitle	title of a campaign
ProjectUrl	unique url of the campaign
Status	status of the campaign (live, successful, failed)
ProjectStartUTC	start of the campaign in UTC
Timestamp	timestamp

Table A.2.: Attributes of the Database Table "*campaignupdates*"

campaignupdates	snapshots of the campaign tab
ID	unique id of a campaign update
ProjectID	foreign key (project)
Applicant	creator
NumProjectsApplicant	number of campaigns from the creator
MainCategory	main category of the campaign
SubCategory	sub category of the campaign
MoneyPledged	money pledged at the moment
MoneyTarget	funding goal
Currency	currency
LengthPitchVideo	length of pitch video
TotalNumberImages	number of images in the description
Gif	number of gifs in the description
Png	number of pngs in the description
Jpeg	number of gifs in the description
Backers	number of backers at the moment
CampaignEndUTC	end of funding period in UTC
DescriptionText	text of campaign description
RisksAndChallenges	text of „Risk and Challenges"-section
Place	geographical place of the campaign
HtmlSourceCodeEncoded	compressed html source code
Timestamp	timestamp

Table A.3.: Attributes of the Database Table "*images*"

images	images within the campaign description
ID	unique id
ProjectID	foreign key (project)
Link	link to image
Timestamp	timestamp

Table A.4.: Attributes of the Database Table *"pledgeamounts"*

pledgeamounts	different available pledge amount of a campaign
ID	unique id
ProjectID	foreign key (project)
PledgeTitle	title of reward
PledgeText	description text of reward
Backers	backers that chose this reward
LimitBackers	maximum number of backers for this reward
EstimatedDelivery	estimated delivery of reward
PledgeAmount	price of reward
Currency	currency of reward
Timestamp	timestamp

A.3. Data of Finished Campaigns

The features that are stated in the following tables are crawled only once after the campaign has reached the end of the funding duration.

Table A.5.: Attributes of the Database Table *"comments"*

comments	comments of a campaign in the comments tab
ID	unique id
ProjectID	foreign key (project)
Author	username of the comment author
ProfileUrl	profile URL of the author
Comment	comment text
DateComment	date and time of the comment
NumProjects	number of campaigns the comment author created
Superbacker	indicates if the comment author is a „superbacker"
Timestamp	timestamp

Table A.6.: Attributes of the Database Table "*communitybacker*"

communitybacker	number of new and returning backers per campaign
ID	unique id
ProjectID	foreign key (project)
NumNewBacker	number of first-time backers
NumReturningBacker	number of returning backers
Timestamp	timestamp

Table A.7.: Attributes of the Database Table "*communitycitybackers*"

communitycitybackers	cities where most backers of a campaign come from of a campaign
ID	unique id
ProjectID	foreign key (project)
City	city of the community tab
Country	country where the city belongs
Backers	backer per city
Timestamp	timestamp

Table A.8.: Attributes of the Database Table "*communitycountrybackers*"

communitycountrybackers	countries where most backers of a campaign come from
ID	unique id
ProjectID	foreign key (project)
Country	country of the community
Backers	backer per country
Timestamp	timestamp

Table A.9.: Attributes of the Database Table "*frequentlyaskedquestions*"

frequentlyaskedquestions	FAQs from the FAQ tab
ID	unique id
ProjectID	foreign key (project)
Question	question text
Answer	answer text
LastUpdate	last update of question
Timestamp	timestamp

Table A.10.: Attributes of the Database Table "*updates*"

updates	updates from the creator from the updates tab
ID	unique id
ProjectID	foreign key (project)
Url	unique URL of update
Heading	heading of update
Description	description text of update
Date	date and time of the update
Likes	number of likes for the update
SupportersOnly	indicates if the update is only visible for backers
NumComments	number of comments the update

Table A.11.: Attributes of the Database Table "*updatecomments*"

updatecomments	comments on comment of the update tab
ID	unique id
UpdateID	foreign key (update)
Author	author of the update comment
ProfileUrl	profile URL of the author
Comment	comment text
DateComment	date and time of the comment
NumProjects	number of campaigns the comment author created
Superbacker	indicates if the comment author is a „superbacker“
Timestamp	timestamp

Table A.12.: Attributes of the Database Table "*updateimages*"

updateimages	images within an update description
ID	unique id
UpdateID	foreign key (update)
Link	link to image
Timestamp	timestamp

Appendix B.

Coefficients of Models

B.1. Coefficients of Logistic Regression

Table B.1.: Coefficients of Static Dataset of Classification with Logistic Regression

Feature	**Coefficient**
duration	-1.36
pitch video length	2.8
sum comments	4,271.86
sum faqs	2.72
sum pledge amounts	3.48
num project creator	7.61
num images	-3.42
len description	0.78
len risks	-1.33

Table B.2.: Coefficients of Extended Dataset of Classification with Logistic Regression

Feature	Coef. 6 h	Coef. 12 h	Coef. 24 h	Coef. 48 h
goal	-1,695.64	-2,829.29	-3,698.50	-4,791.32
pledged 6 h	495.08	55.25	149.85	280.40
pledged 12 h		575.02	-517.93	-741.36
pledged 24 h			1,143.43	390.91
pledged 48 h				955.93
backers 6 h	292.02	286.52	266.58	246.41
backers 12 h		161.98	-630.03	-202.30
backers 24 h			664.41	-1,180.86
backers 48 h				1,368.73
sum comments 6 h	246.29	-424.11	41.19	124.07
sum comments 12 h		335.19	-173.58	-81.54
sum comments 24 h			22.73	-104.64
sum comments 48 h				-164.87
duration	-0.46	-0.37	-0.39	-0.41
pitch video length	16.10	16.81	15.99	16.23
sum faqs	12.51	9.88	8.06	6.28
sum pledge amounts	2.90	2.83	2.70	2.70
min pledge amounts	3.87	0.00	-0.89	-2.86
max pledge amounts	2.65	14.73	18.46	7.26
mean pledge amounts	46.81	64.90	71.88	98.74
num project creator	4.05	1.23	0.97	1.28
num images	-1.44	-2.59	-3.15	-3.10
len description	0.57	0.78	0.90	0.65
len risks	-1.72	-1.45	-1.32	-1.21
count changes	-2.31	-2.82	-2.99	-3.33

Table B.3.: Coefficients of Extended Dataset without Pledged of Classification with Logistic Regression

Feature	**Coefficient**
goal	-1751.51
backers6h	336.91
backers12h	352.25
sum comments hours6	-357.88
sum comments hours12	326.63
duration	-0.61
pitch video length	15.45
sum faqs	11.52
sum pledge amounts	2.75
min pledge amounts	2.96
max pledge amounts	-14.19
mean pledge amounts	83.64
num project creator	1.42
num images	-2.09
len description	0.58
len risks	-1.65
count changes	-2.8

B.2. Coefficients of Linear Regression

Table B.4.: Coefficients of Static Dataset with Linear Regression

Feature	Coefficient
duration	16,412.65
pitch video length	26,4565.57
sum comments	7,375,892.93
sum faqs	1,317,643.42
sum pledge amounts	-6,553.75
num project creator	22,070.87
num images	11,6469.24
len description	18,612.95
len risks	-78,369.25

Table B.5.: Coefficients of Extended Dataset with Linear Regression

Feature	Coef. 6 h	Coef. 12 h	Coef. 24 h	Coef. 48 h
goal	643,456.56	344,719.47	262,859.08	226,658.07
pledged 6 h	4,886,105.07	-45,425.77	1,251,638.80	1,196,061.78
pledged 12 h		4,283,191.80	-1,246,735.27	209,777.98
pledged 24 h			4,845,971.62	-1,583,999.31
pledged 48 h				4,623,905.02
backers 6 h	-1,313,232.99	717,839.82	144,714.17	320,959.96
backers 12 h		-107,695.87	-1,993,054.14	-1,208,720.54
backers 24 h			1,604,662.96	336,018.33
backers 48 h				413,221.48
sum comments 6 h	2,008,073.69	-3,884,610.83	-1,395,899.44	-183,177.61
sum comments 12 h		3,847,061.00	1,517,660.22	1,721,687.35
sum comments 24 h			189,885.32	-2,065,810.17
sum comments 48 h				1,141,711.81
duration	1,022.89	9,742.51	11,627.66	12,434.17
pitch video length	55,830.10	38,386.38	35,603.73	31,188.34
sum faqs	347,647.66	161,409.05	103,297.76	87,880.77
sum pledge amounts	-3,626.52	-3,236.19	-2,447.66	-5,348.11
min pledge amounts	180,687.45	42,792.75	7,611.50	-22,719.37

max pledge amounts	-25,818.77	32,091.17	40,178.70	43,741.06
mean pledge amounts	-13,582.04	-16,354.79	-30,001.33	-32,822.89
num project creator	31,259.45	-9,543.36	-12,265.13	-11,979.58
num images	25,315.43	5,784.11	1,840.16	5,340.45
len description	8,564.61	-1,235.16	-1,874.07	-334.18
len risks	-39,237.26	-37,569.59	-26,542.72	-21,132.30
count changes	33,200.06	22,963.94	19,098.63	20,053.45

Table B.6.: Coefficients of Extended Dataset without Pledged with Linear Regression

Feature	**Coefficient**
goal	699,986.67
backers6h	1,693,584.56
backers12h	1,755,498.86
sum comments hours6	-6,191,000.52
sum comments hours12	7,515,970.37
duration	4,576.83
pitch video length	72,562.74
sum faqs	369,088.39
sum pledge amounts	5,537.80
min pledge amounts	213,172.63
max pledge amounts	-167,575.66
mean pledge amounts	174,894.96
num project creator	-30,074.07
num images	7,717.78
len description	-1,020.58
len risks	-47,590.29
count changes	27,565.47

B.3. Coefficients of Ridge Regression

Table B.7.: Coefficients of Static Dataset with Ridge Regression

Feature	Coefficient
duration	16,270.94
pitch video length	260,506.41
sum comments	7,215,306.65
sum faqs	1,316,762.66
sum pledge amounts	-7,005.89
num project creator	19,786.84
num images	117,431.19
len description	19,305.02
len risks	-78,849.73

Table B.8.: Coefficients of Extended Dataset with Ridge Regression

Feature	Coef. 6 h	Coef. 12 h	Coef. 24 h	Coef. 48 h
goal	646,320.26	342,826.93	262,898.08	227,057.75
pledged 6 h	5,093,143.17	-31,907.27	1,292,466.21	1,198,023.25
pledged 12 h		4,267,109.24	-1,368,380.46	197,091.03
pledged 24 h			4,939,417.49	-1,649,891.47
pledged 48 h				4,698,355.15
backers 6 h	-1,445,083.46	676,391.17	139,203.14	335,054.88
backers 12 h		-81,278.75	-1,967,392.99	-1,218,546.90
backers 24 h			1,575,776.59	310,283.90
backers 48 h				438,977.19
sum comments 6 h	1,934,691.63	-3,904,666.42	-1,282,291.57	-82,216.65
sum comments 12 h		3,876,806.53	1,473,130.74	1,564,671.26
sum comments 24 h			163,841.18	-1,708,355.00
sum comments 48 h				790,523.41
duration	880.98	9,774.40	11,719.46	12,333.64
pitch video length	56,170.71	39,267.83	34,082.94	30,132.58
sum faqs	352,442.43	161,649.25	103,101.01	86,530.35
sum pledge amounts	-4,254.53	-3,340.47	-2,390.07	-5,556.14
min pledge amounts	177,397.67	34,774.48	4,441.65	-22,114.63

max pledge amounts	-25,516.14	30,887.88	42,595.32	44,956.06
mean pledge amounts	-15,199.21	-14,713.51	-32,262.44	-34,284.59
num project creator	32,419.80	-10,381.27	-12,230.18	-12,133.39
num images	24,630.07	6,136.30	1,737.32	5,237.06
len description	8,404.97	-1,111.21	-1,886.62	-431.67
len risks	-38,867.71	-37,809.48	-26,580.64	-20,914.58
count changes	33,377.25	22,840.89	18,906.01	20,161.39

Table B.9.: Coefficients of Extended Dataset without Pledged with Ridge Regression

Feature	**Coefficient**
goal	70,0345.75
backers6h	1,700,087.82
backers12h	1,796,491.72
sum comments hours6	-6,345,941.17
sum comments hours12	7,571,190.46
duration	4,490.57
pitch video length	73,388.76
sum faqs	368,272.23
sum pledge amounts	5,337.25
min pledge amounts	242,110.43
max pledge amounts	-168,588.79
mean pledge amounts	174,925.63
num project creator	-29,511.59
num images	7,362.42
len description	-792.41
len risks	-47,053.42
count changes	27,815.80

Bibliography

[ACG11] Ajay Agrawal, Christian Catalini, and Avi Goldfarb. *The geography of crowdfunding*. Tech. rep. National bureau of economic research, 2011 (cit. on p. 6).

[ACG15] Ajay Agrawal, Christian Catalini, and Avi Goldfarb. "Crowdfunding: Geography, social networks, and the timing of investment decisions." In: *Journal of Economics & Management Strategy* 24.2 (2015), pp. 253–274 (cit. on p. 6).

[ADJ14] Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. "How to Ask for a Favor: A Case Study on the Success of Altruistic Requests." In: *ICWSM*. 2014 (cit. on p. 15).

[B+94] Vic Barnett, Toby Lewis, et al. *Outliers in statistical data*. Vol. 3. 1. Wiley New York, 1994 (cit. on p. 57).

[BLS13] Paul Belleflamme, Thomas Lambert, and Armin Schwienbacher. "Individual crowdfunding practices." In: *Venture Capital* 15.4 (2013), pp. 313–333 (cit. on p. 11).

[BLS14] Paul Belleflamme, Thomas Lambert, and Armin Schwienbacher. "Crowdfunding: Tapping the right crowd." In: *Journal of business venturing* 29.5 (2014), pp. 585–609 (cit. on p. 10).

[BRF16] Henrik Brink, Joseph Richards, and Mark Fetherolf. *Real-World Machine Learning*. 1st. Greenwich, CT, USA: Manning Publications Co., 2016. ISBN: 1617291927, 9781617291920 (cit. on p. 52).

[BW14] Michael Beier and Kerstin Wagner. "Das richtige Timing beim Crowdfunding-Eine explorative Studie zu Projekten auf 100-Days. Net (How Timing Affects Success in Crowdfunding-An Exploratory Study from Switzerland)." In: (2014) (cit. on p. 12).

[BWC15] Michael Beier, Kerstin Wagner, and HTW Chur. "Erfolg Bei Crowdfunding-Kampagnen: Eine Perspektive Aus E-Commerce Und Social Media." In: *Schweizerisches Institut für Entrepreneurship (SIFE) Kompetenzfeld „Digitale Strategien* (2015), pp. 1–16 (cit. on p. 12).

[CL15] Jinwook Chung and Kyumin Lee. "A long-term study of a crowdfunding platform: Predicting project success and fundraising amount." In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM. 2015, pp. 211–220 (cit. on pp. 16, 76).

[CR14] Paolo Crosetto and Tobias Regner. *Crowdfunding: Determinants of success and funding dynamics*. Tech. rep. Jena Economic Research Papers, 2014 (cit. on p. 7).

[EGT13] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. "Launch hard or go home!: predicting the success of kickstarter campaigns." In: *Proceedings of the first ACM conference on Online social networks*. ACM. 2013, pp. 177–182 (cit. on pp. 13, 75).

[Gér17] Aurélien Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. 2017 (cit. on pp. 53, 56, 59).

[GG14] Michael D Greenberg and Elizabeth M Gerber. "Learning to fail: experiencing public failure online through crowdfunding." In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2014, pp. 581–590 (cit. on p. 5).

[GGR13] Giancarlo Giudici, Massimiliano Guerini, and Cristina Rossi Lamastra. "Why crowdfunding projects can succeed: the role of proponents' individual and territorial social capital." In: (2013) (cit. on p. 6).

[GHK12] Elizabeth M Gerber, Julie S Hui, and Pei-Yi Kuo. "Crowdfunding: Why people are motivated to post and fund projects on crowdfunding platforms." In: *Proceedings of the International Workshop on Design, Influence, and Social Technologies: Techniques, Impacts and Ethics*. Vol. 2. 2012, p. 11 (cit. on p. 7).

[Gre+13] Michael D Greenberg, Bryan Pardo, Karthic Hariharan, and Elizabeth Gerber. "Crowdfunding support tools: predicting success & failure." In: *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2013, pp. 1815–1820 (cit. on pp. 13–15, 76).

[Hea+98] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. "Support vector machines." In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28 (cit. on p. 55).

[HYK16] Meng Hong, Sung-Byung Yang, and Taekyung Kim. "Successful crowdfunding: the effects of founder and project factors." In: *Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World*. ACM. 2016, p. 8 (cit. on p. 9).

[IB14] Yael Inbar and Ohad Barzilay. "Community impact on crowdfunding performance." In: (2014) (cit. on p. 9).

[KB15] Venkat Kuppuswamy and Barry L Bayus. "Crowdfunding creative ideas: The dynamics of project backers in Kickstarter." In: (2015) (cit. on p. 9).

[KC16] Jascha-Alexander Koch and Qian Cheng. "The Role of Qualitative Success Factors in the Analysis of Crowdfunding Success: Evidence from Kickstarter." In: (2016) (cit. on pp. 13, 14).

[KS15] Jascha-Alexander Koch and Michael Siering. "Crowdfunding success factors: the characteristics of successfully funded projects on crowdfunding platforms." In: (2015) (cit. on p. 10).

[Kub15] Miroslav Kubat. *An Introduction to Machine Learning*. Springer, 2015 (cit. on p. 53).

[LFL16] Elmar Lins, Kaja J Fietkiewicz, and Eva Lutz. "How to Convince the Crowd: An Impression Management Approach." In: *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE. 2016, pp. 3505–3514 (cit. on p. 11).

[LLC16] Yusan Lin, Wang-Chien Lee, and Chung-Chou H Chang. "Analysis of rewards on reward-based crowdfunding platforms." In: *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*. IEEE. 2016, pp. 501–504 (cit. on p. 9).

[LRR16] Yan Li, Vineeth Rakesh, and Chandan K Reddy. "Project success prediction in crowdfunding environments." In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM. 2016, pp. 247–256 (cit. on pp. 13, 76).

[MG14] Tanushree Mitra and Eric Gilbert. "The language that gets people to give: Phrases that predict success on kickstarter." In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM. 2014, pp. 49–61 (cit. on p. 13).

[MK14] Ethan Mollick and Venkat Kuppuswamy. "After the campaign: Outcomes of crowdfunding." In: (2014) (cit. on p. 5).

[Mol14] Ethan Mollick. "The dynamics of crowdfunding: An exploratory study." In: *Journal of business venturing* 29.1 (2014), pp. 1–16 (cit. on pp. 5, 7).

[Rao+14] Huaming Rao, Anbang Xu, Xiao Yang, and Wai-Tat Fu. "Emerging Dynamics in Crowdfunding Campaigns." In: *SBP*. Springer. 2014, pp. 333–340 (cit. on pp. 13, 14).

[Ras15] Sebastian Raschka. *Python machine learning*. Packt Publishing Ltd, 2015 (cit. on p. 58).

[Sta+15] Marius Stadler, Ferdinand Thies, Michael Wessel, Alexander Benlian, et al. "Erfolg von Crowdfunding-Kampagnen frühzeitig erkennen: Erfolgsprädiktoren auf Kickstarter und Indiegogo." In: *Wirtschaftsinformatik*. 2015, pp. 1237–1251 (cit. on p. 15).

[TSC17] Graça Trindade, Tiago Miguel Teixeira Campos Silva, and Maria da Conceição Santos. "Determinants of the crowdfunding campaign success in the areas of music and sports." In: *Information Systems and Technologies (CISTI), 2017 12th Iberian Conference on*. IEEE. 2017, pp. 1–6 (cit. on p. 8).

[Xia+14] Shengsheng Xiao, Xue Tan, Ming Dong, and Jiayin Qi. "How to design your project in the online crowdfunding market? Evidence from Kickstarter." In: (2014) (cit. on p. 11).

[Xia+17] Ma Xiaoyu, Yang Mingru, Li Yanjun, and Zhang Jihong. "Signaling factors in overfunding: An empirical study based on Crowdcube." In: *Service Systems and Service Management (ICSSSM), 2017 International Conference on*. IEEE. 2017, pp. 1–6 (cit. on p. 8).

[Xu+14] Anbang Xu, Xiao Yang, Huaming Rao, Wai-Tat Fu, Shih-Wen Huang, and Brian P Bailey. "Show me the money!: An analysis of project updates during crowdfunding campaigns." In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2014, pp. 591–600 (cit. on p. 7).

[Zho+16] Mi Jamie Zhou, Baozhou Lu, Weiguo Patrick Fan, and G Alan Wang. "Project description and crowdfunding success: an exploratory study." In: *Information Systems Frontiers* (2016), pp. 1–16 (cit. on p. 13).

[ZIB15] David Zvilichovsky, Yael Inbar, and Ohad Barzilay. "Playing both sides of the market: Success and reciprocity on crowdfunding platforms." In: (2015) (cit. on p. 5).