Predicting the Success of Crowdfunding Campaigns on Kickstarter

SERHII LYSIN, University of Twente, The Netherlands

This thesis analyses the ability of machine learning models to predict the success of crowdfunding projects on the biggest reward-based platform - Kickstarter. This paper uses a dataset with more than 150,000 campaigns from 2009 to 2018 to evaluate feature importances and critically review the performance of three tree-based models: Random Forest Classifier, XGBoost and LighGBM. This research identified that duration, subcategory, and fundraising goal are the most significant features that influence the success of projects on Kickstarter. All prediction models were evaluated based on cross-validation accuracy, precision, recall, F1 score, ROC ACU score and log loss. The findings imply that XGBoost and LightGBM performed at the same high level, while the Random Forest Classifier was chosen as the best model for the given task. Case studies validated the models' predictive reliability.

This study contains certain limitations and weaknesses, including problems with biased dataset. Finally, possible future development was discussed, considering the integration of real-time analysis and expanding the utilised dataset.

Additional Key Words and Phrases: crowdfunding, Kickstarter, machine learning, campaign success prediction, Random Forest Classifier, XGBoost, LightGBM.

1 INTRODUCTION

1.1 Overview of Crowdfunding

The crowdfunding method of fundraising significantly changed the way social projects, creative campaigns and startups raise money. According to Mora-Cruz and Palos-Sanchez (2023)[14], crowdfunding platforms have made it easier for founders to reach their potential audience, specifically assisting creators who were struggling to find financial resources using standard methods.

This type of platform provides various types of compensation models such as Reward, Equity, and Lending. Project initiators can select the approach that suits better for their strategy of attracting users or investors.

1.2 Types of Crowdfunding Campaigns

There are four types of crowdfunding campaigns that founders can use depending on their product, industry and other parameters. These four models offer various forms of returns to the backers in exchange for their support (Koch & Cheng, 2016)[12].

Donation-based crowdfunding: This model does not provide financial compensation or return to backers. Most often this type of campaign is used by charity funds and organisations.

Reward-based crowdfunding: Project supporters receive physical items or experiences in return for their investments. This approach is widely used by initiators from creative industries like arts, films and music.

Lending-based crowdfunding: This approach is similar to a loan as people receive their money back with interest. Usually, project founders can set an interest rate by themselves.

Equity-based crowdfunding: It is a similar technique to conventional methods of fundraising, project backers receive equity (stake) of the company or business.

Each crowdfunding model has different goals and target audiences, which influence how projects attract and engage supporters. This paper will concentrate on reward-based platforms, with a particular focus on Kickstarter.

1.3 Reward-based Crowdfunding: A Focus on Kickstarter

Kickstarter is a world-leading reward-based crowdfunding platform, which has a bunch of datasets created already, becoming a perfect example to track the dynamics of modern crowdfunding. Most of the projects on the platform are looking for investments to implement their ideas and concepts. In case the campaign is successful, project creators offer their backers to receive products as soon as they are produced. This method provides such advantages:

Direct Engagement: Project initiators can directly communicate and engage with their backers by answering questions, giving updates and replying to comments.

Market Validation: The platform lets creators test hypotheses and find customers.

Flexible Funding Goals: Kickstarter allows founders to adjust campaign parameters to their needs. Initiators can set project duration, funding goal and other parameters depending on the project‘s concept.

1.4 Importance of Predicting Campaign Success

The ability to predict the success of crowdfunding campaigns is important for both creators and funders. Schraven et al. (2020) [17] noted that accurate forecasts can influence how project founders conduct strategic decisions like selecting campaign parameters. Such analysis can help initiators improve their chances of success by taking the right steps before launching the project to meet the expectations of a target audience. Predictions allow potential backers to reduce the risk of investing in projects that have very little chance of succeeding. It underlines the mutually beneficial existence of such analysis that can help reduce unnecessary risks before and during the campaign (Schraven, van Burg, van Gelderen, & Masurel, 2020)[17].

Precise prediction instruments allow founders to allocate their resources properly - they can invest their time, money or effort in particular activities that maximise their chances of becoming successful. This can include targeting specific countries or optimising customer support on the platform. This approach encourages participation rate and enables a more dynamic environment for creative projects to grow (Etter, Grossglauser, & Thiran, 2013)[6].

2 FUNDAMENTALS OF CROWDFUNDING ON KICKSTARTER

2.1 Mechanics of Kickstarter

The mechanics of Kickstarter, as discussed by Mollick (2014)[13], show how this crowdfunding platform allows a wide variety of project founders, from artists to tech innovators, to raise funds directly from a large audience without relying on traditional financial methods. Wang, Ghosh, and Liu (2024)[18] explain an All-or-nothing approach that Kickstarter is using. According to the authors, it means that the project is considered successful only in case it succeeds in achieving its fundraising goal. In case the project fails to succeed, project backers receive their money back. This approach is beneficial for both project creators and backers as it forces founders to prepare better for the launch of the campaign.

2.2 Information to launch a campaign

To launch a project on Kickstarter, creators have to prepare information that will be used for communication with potential backers. This data is needed to explain the project‘s value proposition. These details include:

- **Funding Goal:** Amount of money to be raised during the campaign.
- **Campaign Duration:** The time frame in which the funding objective should be achieved. The project will not be stopped in case it reaches the goal before the deadline.
- **Project Description:** Clear explanations of the campaign, including value proposition.
- **Reward Tiers:** Different reward levels, meaning that backers with distinct investments receive distinct compensations.
- **Project Video**
- **Creator Information:** Details about creators, their experience, personal values and connections.

According to Chen, Jones, Kim, and Schlamp (2024)[3], this information is important to give potential backers all the required information to decide whether to support or not a particular campaign.

3 RESEARCH QUESTION

To address the needs of current Kickstarter platform users, the following research question was formulated:

> *How can machine learning models accurately predict the success of crowdfunding campaigns on Kickstarter based on key campaign variables?*

This central question can be explored through the following sub-questions:

(1) What are the most significant factors influencing the success of Kickstarter campaigns?
(2) How do different machine learning algorithms compare in their ability to predict the outcomes of Kickstarter campaigns?
(3) How can the obtained research results and model predictions be utilized and improved for future applications?

4 DATA PREPARATION AND PRELIMINARY ANALYSIS

4.1 Data Source

To conduct analysis a dataset "Kickstarter Campaigns" by Yash Kantharia found on Kaggle.com was utilised. It contains information on more than 150,000 crowdfunding campaigns launched between 2009 and 2018. This dataset has 20 columns (features) for each campaign.

4.2 Data Cleaning and Deduplication

The initial dataset had 192,553 rows. After removing duplicated cases, the dataset was reduced to 168,468 rows. This step was important to prevent misleading model performance. Considering a large number of duplicates, cross-validation accuracy would be much higher as some duplicates can fall into both training and testing sets.

4.3 Conversion of Non-Numerical Data to Numerical Data

To facilitate the analysis, categorical variables were converted to numerical values:

- **Status:** The campaign status was categorized as Success [1] or Fail [0].
- **Country:** Countries were mapped to numerical identifiers, with each country assigned a unique number.
- **Main Category:** Main categories were converted to numerical values, with each category assigned a unique number.
- **Subcategory:** Subcategories were also converted similarly, with each subcategory assigned a unique number.

4.4 Removal of Dependent Columns

To ensure that there are no dependencies between features and that the feature importance graph is not biased, the following columns were removed:

- Start_Q and End_Q: Features start_month and end_month provide more precise information than starting and ending quarters.
- State and City: These features are directly dependent on the parameter country.
- Launched and Deadline: These parameters have a strong dependence on start_month, end_month and duration.
- Currency: This feature was directly dependent on the country.

4.5 Balancing the Dataset

The initial class distribution of the status variable was unbalanced:

- Success: 94,230 (55.9%)
- Fail: 74,237 (44.06%)

As XGBoost and LightGBM models are known to be overfitting, to address the imbalance, undersampling was conducted by randomly removing a certain amount of successful cases. The dataset was balanced to have an equal number of successful and failed campaigns:

- Success: 74,237
- Fail: 74,237

This balancing was important to ensure that the predictive model would not be biased towards class 1 (successful cases).

4.6 Final Dataset

After all manipulations, the final dataset includes the following parameters (columns):

- Main Category (main_category)
- Subcategory (sub_category)
- Duration of the campaign in days (duration)
- Fundraising goal in USD (goal_usd)
- Country (country)
- Length of the campaign description in words (blurb_length)
- Length of the campaign name in words (name_length)
- Status of the campaign, successful or failed (status)
- Start month of the campaign (start_month)
- End month of the campaign (end_month)

4.7 Correlation Matrix Analysis

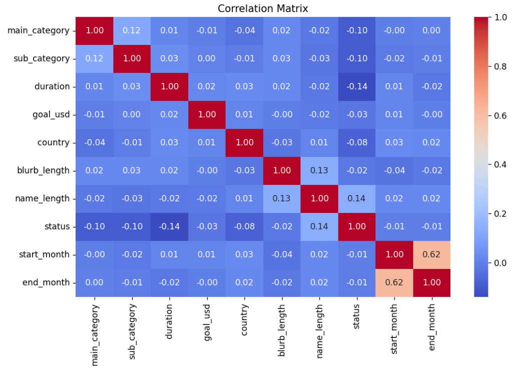

Fig. 1. Correlation Matrix

To ensure that linear dependencies in the dataset are weak, a correlation matrix of all features was built. This figure demonstrates the strength of linear dependencies between every possible pair of parameters. The most important insights are:

- Start Month and End Month (0.62): These features show a high correlation, meaning they are closely related. However, they can not be removed as these parameters might give valuable insights.
- Name Length and Blurb Length (0.14): These variables show a moderate, but acceptable correlation.
- Status: The status variable shows low correlations with most variables, implying that it is not directly dependent on one specific feature but on a combination of parameters.

5 DEVELOPING PREDICTIVE MODELS AND EVALUATING THEIR ACCURACY

This research paper applied three three-based prediction models for conducting analysis: Random Forest Classifier, XGBoost model, and LightGBM model. A study by Greenberg, Hariharan, Gerber, and Pardo (2013)[7] showed that tree-based models performed the best in predicting crowdfunding project success, specifically Random Forests and Logistic Model Trees. Oduro, Yu, and Huang (2022)[15] in their research stated that gradient-boosting models performed with one of the lowest test error rates, only 5-6%. That is why the decision was made to use the above-mentioned prediction models.

5.1 Random Forest Classifier

Leo Breiman (2001)[1] explained that Random Forests are a combination of tree predictors where each tree relies on the values of a random vector that is independently sampled and has an identical distribution for all trees in the forest.

Random Forest Classifier is a particular application of the Random Forest approach which is used for classification tasks, specifically with categorical target features.

5.1.1 Upsides and Downsides. As Etter et al. (2013)[6] stated, one of the biggest upsides of the Random Forest Classifier model is reducing possible overfitting and improving the model's generalisation ability by building decision trees with random subsets of training data. Moreover, the mentioned approach makes them resistant to noise in the data and outliers. On the other side, the authors mention that this tree-based prediction model does not perform well on imbalanced datasets and can be computationally more expensive compared to simpler models.

5.1.2 Grid Search for Hyperparameter Tuning. The grid search was conducted for the Random Forest Classifier to find the best combination of hyperparameters that perform the best. To prevent model overfitting and ensure model stability, such hyperparameters were selected:

- Number of estimators (`n_estimators`)
- Maximum depth of the trees (`max_depth`)
- Minimum samples required to split an internal node (`min_samples_split`)
- Minimum samples required to be at a leaf node (`min_samples_leaf`)
- Number of features to consider for looking for the best split (`max_features`)

The results of the search depict that the combination `max_depth = 20`, `max_features = None`, `min_samples_leaf = 4`, `min_samples_split = 10`, and `n_estimators = 200` performs the best with a cross-validation accuracy of 77.01%.

Metric	Value
Cross-Validated Accuracy	77.01%
Overall Accuracy	88.27%
Precision	92.00%
Recall	83.83%
F1 Score	0.88
ROC AUC Score	96.24%
Log Loss	0.31

Table 1. Random Forest Classifier Performance Metrics

5.1.3 Evaluating Accuracy. After optimisation by Grid Search, the Random Forest Classifier was evaluated by the following metrics: Overall accuracy, cross-validation accuracy, Precision, Recall, F1-score, ROC AUC curve, and Log loss.

- Overall accuracy (77.07%) and cross-validation accuracy (88.25%) indicate strong predictive performance.
- Precision (92%) and Recall (84%) demonstrate a high ability to identify positive cases, with only 16% of successful cases missed.
- F1-score (0.88) stated a reasonable ratio between Recall and Precision.
- The ROC Curve depicts that the model can differentiate between classes on a good level.
- Log loss (0.31) demonstrates that the prediction model is moderately reliable and makes well-calibrated forecasts.

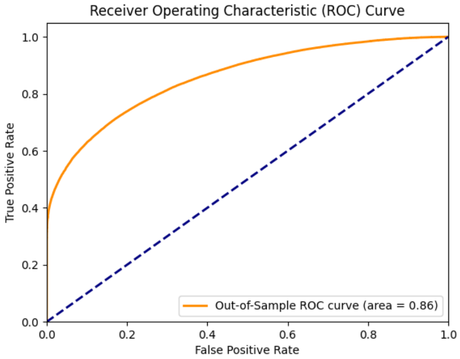

Fig. 2. Out-of-sample Random Forest Classifier - ROC Curve

5.2 XGBoost model

XGBoost is an easy-to-scale portable tree-boosting model that is often used by data scientists for a variety of machine-learning tasks. Moreover, the authors demonstrated that the XGBoost is faster than other models used in experiments (Chen & Guestrin, 2016)[4].

5.2.1 Upsides and Downsides. Experiments by Chen and Guestrin (2016)[4] showed that the XGBoost model was faster compared to other models. The authors stated that the major advantage of the model is the ability to support various weighted classifications. Gu, Cao, Wang, Yu, and Qing (2022)[8] stated that XGBoost has moderately high computing performance and a strong ability to deal with non-linear datasets. However, the model is very complex and directly dependent on its hyperparameters. It means that without precise tuning of hyperparameters, the model can give low-quality predictions.

5.2.2 Grid Search for Hyperparameter Tuning. Considering the conclusions of the above-mentioned research, a grid search was conducted for the XGBoost model to avoid low-quality performance. The optimisation procedure was carried out based on hyperparameters, including the most important ones:[8]:

- The maximum depth of the decision tree (`max_depth`)
- How much to adjust the model with each new tree (`learning_rate`)
- The number of trees to fit in the model (`n_estimators`)
- The portion of the training data used to build each tree (`subsample`)
- The portion of features used to build each tree (`colsample_bytree`)

The outcome of the grid search shows that the combination `colsample_bytree = 0.8`, `learning_rate = 0.1`, `max_depth = 7`, `n_estimators = 200`, and `subsample = 0.9` achieves the best results with a cross-validation accuracy of 78.37%.

Metric	Value
Cross-Validated Accuracy	78.37%
Overall Accuracy	80.31%
Precision	83.67%
Recall	75.32%
F1 Score	0.79
ROC AUC Score	89.42%
Log Loss	0.40

Table 2. XGBoost Model Performance Metrics

5.2.3 Evaluating Accuracy. Considering the conducted Grid Search for the hyperparameters of the XGBoost model, results were evaluated by the subsequent metrics: Overall accuracy, cross-validation accuracy, Precision, Recall, F1-score, ROC AUC curve, and Log loss.

- Overall accuracy (80.31%) and cross-validation accuracy (78.37%) demonstrate the ability to conduct successful predictions on both seen and unseen data.
- Precision (83.67%) and Recall (75.32%) indicate an ability to successfully identify positive cases, with a relatively small proportion of false positive scenarios.
- F1-score (0.79) is a bit lower than for the Random Forest Classifier, but it nevertheless indicates a well-balanced Precision and Recall.
- The ROC Curve depicts that the model can accurately distinguish between successful and failed cases.
- Log loss (0.40) displays that this prediction model is less accurate than the Random Forest Classifier, and the reason should be investigated.

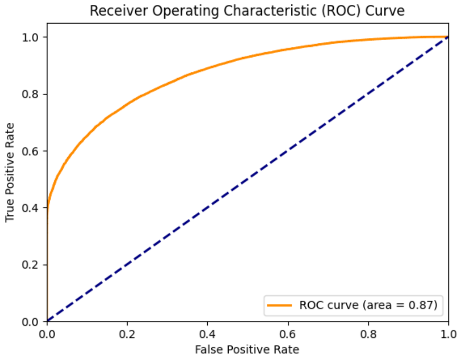

Fig. 3. Out-of-sample XGBoost Model - ROC Curve

5.3 LightGBM model

Experiments by Ke et al. (2017)[11] showed that LightGBM is a gradient-boosting decision tree (GBDT) model that has a faster process of training by up to 20 times while maintaining relatively the same accuracy compared to conventional GBDTs.

5.3.1 Upsides and Downsides. According to research by Ke et al. (2017)[11], LightGBM is considered to be the fastest while maintaining almost the same accuracy as baselines. Moreover, the authors note that LightGBM can significantly outperform XGBoost in aspects of processing speed and memory usage.

A study by Hajihosseinlou, Maghsoudi, and Ghezelbash (2023)[9] represents that the main drawback of the LightGBM model is similar to the XGBoost model as it is sensitive to overfitting.

5.3.2 Grid Search for Hyperparameter Tuning. As Hajihosseinlou, Maghsoudi, and Ghezelbash (2023)[9] stated, the LightGBM model is very sensitive to overfitting and to avoid it, the Grid Search was implemented for the hyperparameters. As this gradient-boosting model is complex, the following hyperparameters were chosen for the analysis:

- The maximum depth of the decision tree (`max_depth`)
- The maximum number of leaves in one tree (`num_leaves`)
- The minimum improvement in model accuracy needed to split a leaf node further (`min_split_gain`)
- The amount by which the model's updates are reduced to avoid overfitting (`learning_rate`)
- The number of trees to fit (`n_estimators`)
- The fraction of samples used to build each tree (`subsample`)
- The fraction of features used to build each tree (`colsample_bytree`)

The result of the grid search shows that the mix `colsample_bytree = 0.8`, `learning_rate = 0.1`, `max_depth = 10`, `min_split_gain = 0.2`, `n_estimators = 200`, `num_leaves = 63`, `subsample = 0.8` accomplishes the best outcomes with a cross-validation accuracy of 78.37%.

Metric	Value
Cross-Validated Accuracy	78.37%
Overall Accuracy	80.41%
Precision	83.86%
Recall	75.30%
F1 Score	0.79
ROC AUC Score	89.44%
Log Loss	0.40

Table 3. LightGBM Model Performance Metrics

5.3.3 Evaluating Accuracy. Evaluating results of the Grid Search and selected hyperparameters for the LightGBM model, results were evaluated by the same metrics as other models: Overall accuracy, cross-validation accuracy, Precision, Recall, F1-score, ROC AUC curve, and Log loss.

- Overall accuracy (80.41%) and cross-validation accuracy (78.37%) show almost the same performance as the XGBoost model with a bit higher ability to predict already seen data.
- Precision (83.68%) and Recall (75.31%) demonstrate the same ability to successfully identify positive cases as the XGBoost model.
- F1-score (0.79) is identical to the above-mentioned model and indicates a good ratio between Precision and Recall.
- The ROC Curve illustrates the excellent ability of the model to predict both possible outcomes of the crowdfunding projects.
- Log loss (0.40) states that the model is on the same level of accuracy as XGBoost.

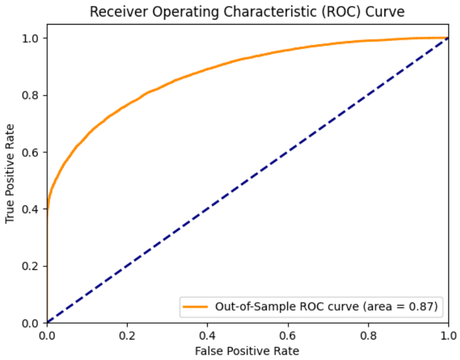

Fig. 4. In-sample ROC Curve - LightGBM Model

5.4 Comparison of Models

To check if trained prediction models are not biased and received high accuracies, especially ROC ACU scores, are truthful, histograms for each feature were built, including splitting features in successful and failed projects.

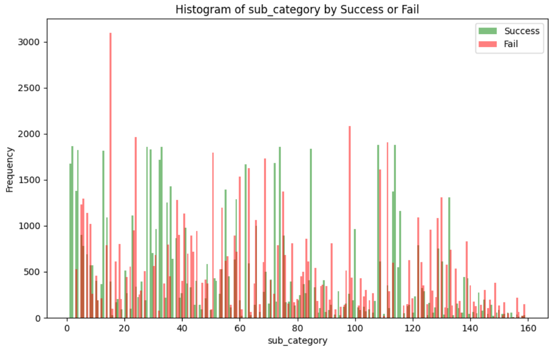

Fig. 5. Histogram for the feature "sub_category"

There were no specific trends visible in any graphs except the histogram for sub_category. Figure 5 depicts that 100% of crowdfunding campaigns launched in subcategories 1, 2, 25, 62, 84, 115, and 116 are successful. While projects of subcategories 18, 24, and 144 failed to achieve fundraising goals. It can be explained by the limitation of the dataset, meaning that the mentioned subcategories had examples only of one type of projects (failed/successful).

Big log loss values might be explained by the existence of outliers in the dataset for such features as goal_usd. The maximum possible value for this parameter is 129,033,257, while the 75th percentile is 15,000 and the 25th percentile is 1,500. IQR is 13,500, thus all values of goal_usd above 35,250 (15,000+1.5*13,500) are considered outliers. The total number of outliers for this feature is 15,741. The Random Forest Classifier has better log loss due to its ability to work with outliers and noise in data.

Based on the evaluation of accuracy, the Random Forest Classifier showed better results than both the XGBoost and LightGBM models based on several factors. The Random Forest classifier demonstrated an overall accuracy of 88.27%, while the same parameter for XGBoost was 80.31% and 80.41% for LightGBM. Precision and recall indicators showed a better ability of the Random Forest Classifier to recognize positive instances and actual positives.

Nearly every parameter for LightGBM and XGBoost was the same or extremely comparable as was shown in the F1 score scenario. With Random Forest receiving the maximum score of 0.88, it was indicated that precision and recall were well-balanced. For XGBoost and LightGBM, the score was a bit lower - 0.79.

Moreover, the ROC AUC score was the highest for the Random Forest Classifier (0.96). At the same time, XGBoost and LightGBM have similar ROC AUC scores of 0.89, meaning all three models show their ability to differentiate between both classes. The log loss was relatively high for all models - Random Forest (0.31) XGBoost and LightGBM (0.40). Such results of analysis lead to the conclusion that the Random Forest Classifier is better for predicting the success of crowdfunding projects on Kickstarter.

6 IMPORTANT FEATURES FOR SUCCESS

6.1 Findings from Correlation Analysis

The correlation matrix (Figure 1) indicated that duration (-0.15), subcategory (-0.10), and main category (-0.09) had the strongest negative correlations with status. Contrarily, name length showed a positive correlation (0.14) with status. Other features exhibited relatively weaker correlations with status. Although correlation analysis provides a useful overview, it is limited to linear relationships and does not account for interactions between features.

6.2 Model-Based Feature Importance

To obtain a more comprehensive understanding of feature importance, a Random Forest Classifier was used as it is known for its ability to compute feature importances. While other models are better for the prediction of outputs.

6.2.1 Steps.

1. **Data Splitting:** The dataset was split into two sets: training with 80% of the data and testing with 20% of the data.
2. **Model Training:** A Random Forest Classifier was initialised, the best hyperparameters were found, and the model was trained.
3. **Feature Importance Extraction:** The feature importance graph (Figure 6) was created, and scores were extracted from the trained model. These scores represent the contribution of each feature to the model's ability to predict the success of crowdfunding campaigns.
4. **Analysis:** Features were sorted based on their impact on predicting campaign success.

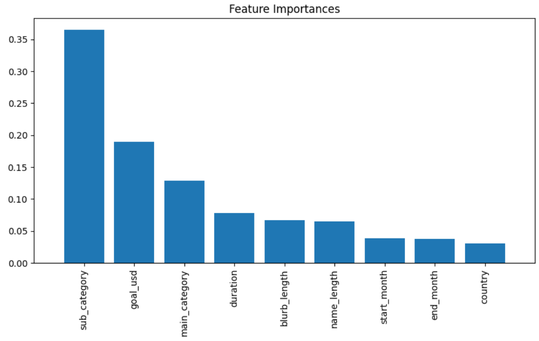

Fig. 6. Feature Importance Graph

6.2.2 Findings from Model-Based Feature Importance. Figure 6 illustrates that sub_category (0.366) is the most influential feature, indicating that specific subcategories within the main categories greatly influence campaign success. As it was discussed in section 5.4, the assumption is that the dataset did not have examples for both successful and failed projects in several sub_categories, which caused this feature to be the most significant.

According to the histogram, goal_usd (0.19) is the second most important parameter, suggesting the importance of setting reasonable funding goals. Duration (0.078) and main_category (0.129) are also influential, underscoring the significance of appropriate time frames. Other features such as blurb_length, name_length, start_month, and end_month were found to have less influence according to the graph.

6.3 Analysis of Other Research Papers

To find the most crucial factors for the success of crowdfunding projects scientific papers were examined including "KickPredict: Predicting Kickstarter Success" (2013)[3], "The Dynamics of Crowdfunding: An Exploratory Study" (2013)[13], "Using Language to Predict Kickstarter Success" (2016)[16], "A Long-Term Study of a Crowdfunding Platform: Predicting Project Success and Fundraising Amount" (2015)[5], "Kickstarter Crowdfunding: How the Predictors of Success Vary by Project Category" (2012)[2], "Success Prediction using Random Forest, CatBoost, XGBoost, and AdaBoost for Kickstarter Campaigns" (2022)[10], and "Prediction of Crowdfunding Project Success with Deep Learning" (2018)[19].

6.3.1 Funding Goal. Selected studies highlight the significance of the funding goal due to its effect on project supporters' perspectives and campaign dynamics.

- *KickPredict* (2013) [3] found that lower goals appear more achievable, leading to increased contributions.

- *Using Language to Predict Kickstarter Success* (2016) [16] emphasised that it is easier to persuade potential backers to support a project with a clear, attainable goal.
- *A Long-Term Study of a Crowdfunding Platform* (2015) [5] demonstrated a strong correlation between reasonable goals and success rates.
- *Kickstarter Crowdfunding* (2012) [2] noted the importance of aligning funding objectives with project potential.
- Recent data analysis in machine learning and deep learning studies (2022, 2018)[19] supports these findings.

6.3.2 Project Duration. The perfect duration balances enough time for contributors to engage while creating a sense of urgency.

- *A Long-Term Study of a Crowdfunding Platform* (2015) [5] highlights that shorter projects cultivate a sense of urgency.
- *Kickstarter Crowdfunding* (2012) [2] found that setting a perfect duration requires analysing the behaviour of the target audience to align with the project's concept.
- *Success Prediction using Machine Learning* (2022)[10] determined optimal time frames through data analysis, emphasising the duration's impact on backer engagement.

6.3.3 Social Networks and Networking Influence. The creator's network is essential for the project's trust and publicity.

- *The Dynamics of Crowdfunding* (2013) [13] underlined that projects of founders with strong personal networks are associated with higher success rates.
- *Long-Term Study of a Crowdfunding Platform* (2015) [5] discovered that the social networking efforts of founders are vital for project advertisement.

6.3.4 Additional Factors. Research papers mentioned other factors, but their influence was not as significant as in the previously mentioned features. These include project updates, FAQs, the number of reward tiers, project descriptions and effective use of language, geographical location, and the creator's track record of backing and creating projects.

6.4 Comparison of Model Training Results and Analysis of Other Research Papers

To identify the most crucial aspects of the success of crowdfunding campaigns theoretical and practical methods were applied. As a part of the practical approach, correlation matrix analysis was conducted and a feature importance histogram was extracted from the trained Random Forest Classifier. These methods showed that parameters sub_category, duration, funding goal, blurb_length, and main_category are the most impactful on campaign dynamics.

The theoretical approach consisted of analysing seven scientific papers on the topic of predicting the success of crowdfunding projects. This procedure indicated that a campaign's ability to succeed is greatly influenced by realistic goals, duration, project description, location and social connections of founders.

Overall, there are a lot of intersections in the results of theoretical and practical methods, giving a clear understanding of the most important parameters for campaign success.

7 TESTING MODELS

As the dataset contains data from 2009 to 2018, it was decided to test trained models on modern projects that are currently in the process of fundraising. All selected campaigns reached their goal by date information was scraped. Data for projects such as main_category, start_month, and country were found on kicktraq.com.

To convert categorical variables such as name, category, start month, and other relevant parameters to numerical type, `convert_cat_to_num()` function was developed. This processed data is then converted into a data frame.

Afterwards, a pre-trained model and a scaler were loaded. Using this model, the class to which the project belongs was predicted: Class [1] indicates success, and Class [0] indicates failure. The model also provides the probability of the project being successful.

7.1 GC6 - The first carbon fibre smart mini cordless rotary pen

- **RandomForest:** Predicted probability for success is 100%, classified as successful (Class 1).
- **XGBoost:** Predicted probability for success is 98%, classified as successful (Class 1).
- **LightGBM:** Predicted probability for success is 98.9%, classified as successful (Class 1).

This project raised $121,248 while a goal was $7,683 in 40 days, meaning it was successful and all models forecasted correctly, however, they showed excessively precise outcomes.

The campaign was launched in the sub_category "Product Design" with id 62. As depicted in Figure 5, all projects in this category were successful, which causes bias for all three prediction models. To validate this parameters like main_category and goal_usd were significantly adjusted, but the probability of success remained unchanged.

7.2 mui Board Gen 2

- **RandomForest:** Predicted probability for success is 68.1%, classified as successful (Class 1).
- **XGBoost:** Predicted probability for success is 56%, classified as successful (Class 1).
- **LightGBM:** Predicted probability for success is 52.9%, classified as successful (Class 1).

This campaign raised $166,210 of $10,000, implying predictions of all trained models were correct. However, the predicted probability was relatively low for the XGBoost and LightGBM models. After comparing the data of the project with histograms, there was no specific trend detected causing the prediction score to be low.

7.3 The Better with Bitters Experience

- **RandomForest:** Predicted probability for success is 35.5%, classified as not successful (Class 0).
- **XGBoost:** Predicted probability for success is 29.6%, classified as not successful (Class 0).
- **LightGBM:** Predicted probability for success is 32.5%, classified as not successful (Class 0).

This crowdfunding campaign was forecasted to fail by all three models despite raising $131,719 out of a goal of $10,000. To investigate the prediction scores, the project's input was compared to histograms to identify trends.

"The Better with Bitters Experience" was launched in the main category "food" with ID 14. Figure 8 demonstrates that the number of failed projects launched in this category is twice that of successful ones. To confirm this hypothesis, when the main category was changed, the prediction score grew greatly, predicting the project's success.

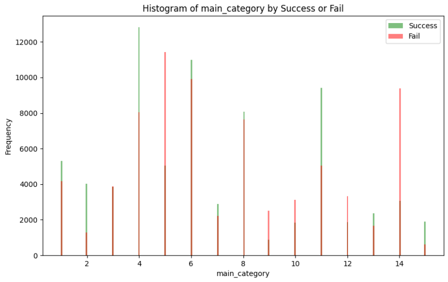

Fig. 7. main_category Feature Histogram

7.4 Summary

	Random Forest	XGBoost	LightGBM
GC6	100% [1]	98% [1]	98.9% [1]
mui Board Gen 2	68.1% [1]	56% [1]	52.9% [1]
The Better with Bitters	35.5% [0]	29.6% [0]	32.5% [0]

Table 4. Prediction by LightGBM, XGboost and Random Forest Classifier

Three projects are not enough to validate the efficiency of the trained models, while the chosen scenarios indicated some existing biases caused by the used dataset.

Table 4 indicates that XGBoost, Light GBM models, and Random Forest Classifier performed similarly. They correctly forecasted outcomes for the GC6 and MUI Board Gen 2 projects, with the Random Forest Classifier being more accurate. The Better with Bitters project was mistakenly predicted to fail due to bias towards the project's main category.

To make more precise judgments, models should be tested and analysed on a broader dataset to detect all biases.

8 POSSIBLE FUTURE APPLICATIONS

Prediction tools and analysis to understand the possibility of success of crowdfunding campaigns can have various target audiences, including project creators, potential backers, and investors. Conducted analysis can help to comprehend global trends in projects, identify which industries receive more investments and paying users, and allow founders to pivot accordingly.

Moreover, a user interface (UI) can be developed, integrating useful pieces of advice into the prediction model. This approach would enable project initiators to not only understand their chances of success but also receive appropriate instructions to increase the probability of achieving their financial goals.

Furthermore, the dataset used for training purposes can be improved by adding new features that might influence the success of crowdfunding campaigns. The number of projects in the dataset can be expanded as well.

Overall, the model can become even more practical if it scrapes data from Kickstarter in real-time and gives valuable insights to founders based on that. The only limitations are computational resources and the legal aspect of collecting data.

9 CONCLUSION

XGBoost and LightGBM models trained fast and performed at relatively the same level on all tests, while the Random Forest Classifier had better accuracy with a longer training time. Cross-validation accuracy for all models was 77-78%, meaning a high ability to accurately predict the success of campaigns on unseen data. Based on received outputs in conditions with limited data, the Random Forest Classifier was selected as the most appropriate model to predict the probability of success of crowdfunding projects.

The list of the most important features received from the Random Forest Classifier was similar to the theoretical analysis of other research papers highlighting the influence of such parameters as funding goal, campaign duration and project description.

After all tests and analysis, it was realised that some unbalanced features and outliers caused models to be biased towards certain sub_categories, goal_usd and main_categories. It means that the dataset quality plays a crucial role in the ability of the predictive models to accurately forecast the results of crowdfunding campaigns.

For the next steps, the dataset should be adjusted properly and investigated if it is not biased. As used models are trained on datasets with the latest projects in 2018, they should be tested on modern cases to see if trends on Kickstarter did not change in the previous 6 years.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Jörg Osterrieder, for his guidance in resolving issues that arose during the practical aspects of this thesis. Moreover, I want to thank Maksym Dryhval for providing constructive feedback.

A APPENDIX

During the preparation of this work, the author used ChatGPT and Grammarly in order to correct mistakes in the text, paraphrase, find synonyms and refine the Python code. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of the work.

REFERENCES

[1] L Breiman. 2001. Random Forests. *Machine Learning* 45 (Oct. 2001), 5–32. https://doi.org/10.1023/A:1010950718922
[2] Salvador Briggman. [n. d.]. Kickstarter Crowdfunding: How the Predictors of Success Vary by Project. ([n. d.]).
[3] Kevin Chen, Brock Jones, Isaac Kim, and Brooklyn Schlamp. [n. d.]. KickPredict: Predicting Kickstarter Success. ([n. d.]).

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794. https://doi.org/10.1145/2939672.2939785 arXiv:1603.02754 [cs].

[5] Jinwook Chung and Kyumin Lee. 2015. 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 (HT '15)*. Association for Computing Machinery, New York, NY, USA, 211–220. https://doi.org/10.1145/2700171.2791045

[6] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. 2013. Launch hard or go home! predicting the success of kickstarter campaigns. In *Proceedings of the first ACM conference on Online social networks (COSN '13)*. Association for Computing Machinery, New York, NY, USA, 177–182. https://doi.org/10.1145/2512938.2512957

[7] Michael D. Greenberg, Bryan Pardo, Karthic Hariharan, and Elizabeth Gerber. 2013. Crowdfunding support tools: predicting success & failure. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*. ACM, Paris France, 1815–1820. https://doi.org/10.1145/2468356.2468682

[8] Zhongyuan Gu, Miaocong Cao, Chunguang Wang, Na Yu, and Hongyu Qing. 2022. Research on Mining Maximum Subsidence Prediction Based on Genetic Algorithm Combined with XGBoost Model. *Sustainability* 14, 16 (Jan. 2022), 10421. https://doi.org/10.3390/su141610421 Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

[9] Mahsa Hajihosseinlou, Abbas Maghsoudi, and Reza Ghezelbash. 2023. A Novel Scheme for Mapping of MVT-Type Pb–Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm. *Natural Resources Research* 32, 6 (Dec. 2023), 2417–2438. https://doi.org/10.1007/s11053-023-10249-6

[10] Siddharth Jhaveri, Ishan Khedkar, Yash Kantharia, and Shree Jaswal. 2019. Success Prediction using Random Forest, CatBoost, XGBoost and AdaBoost for Kickstarter Campaigns. In *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, Erode, India, 1170–1173. https://doi.org/10.1109/ICCMC.2019.8819828

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. [n. d.]. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. ([n. d.]).

[12] Jascha-Alexander Koch and Qian Cheng. 2016. The Role of Qualitative Success Factors in the Analysis of Crowdfunding Success: Evidence from Kickstarter. https://papers.ssrn.com/abstract=2808428

[13] Ethan Mollick. 2014. The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing* 29, 1 (Jan. 2014), 1–16. https://doi.org/10.1016/j.jbusvent.2013.06.005

[14] Alexandra Mora-Cruz and Pedro R. Palos-Sanchez. 2023. Crowdfunding platforms: a systematic literature review and a bibliometric analysis. *International Entrepreneurship and Management Journal* 19, 3 (Sept. 2023), 1257–1288. https://doi.org/10.1007/s11365-023-00856-3

[15] Michael Safo Oduro, Han Yu, and Hong Huang. 2022. Predicting the Entrepreneurial Success of Crowdfunding Campaigns Using Model-Based Machine Learning Methods. *International Journal of Crowd Science* 6, 1 (April 2022), 7–16. https://doi.org/10.26599/IJCS.2022.9100003

[16] Kartik Sawhney, Caelin Tran, and Ramon Tuason. [n. d.]. Using Language to Predict Kickstarter Success. ([n. d.]).

[17] Etienne Schraven, Elco van Burg, Marco Gelderen, and Enno Masurel. 2020. Predictions of Crowdfunding Campaign Success: The Influence of First Impressions on Accuracy and Positivity. *Journal of Risk and Financial Management* 13 (Dec. 2020), 331. https://doi.org/10.3390/jrfm13120331

[18] Peng Wang, Bikram Ghosh, and Yong Liu. 2024. Marketing strategies in reward-based crowdfunding: The role of demand uncertainties. *International Journal of Research in Marketing* (March 2024). https://doi.org/10.1016/j.ijresmar.2024.03.001

[19] Pi-Fen Yu, Fu-Ming Huang, Chuan Yang, Yu-Hsin Liu, Zi-Yi Li, and Cheng-Hung Tsai. 2018. Prediction of Crowdfunding Project Success with Deep Learning. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. IEEE, Xi'an, 1–8. https://doi.org/10.1109/ICEBE.2018.00012