Title
Kickstarter Project Success Prediction

Team Members:

Bhavana Gangula
016003417
bhavana.gangula@sjsu.edu
Neeharika Yeluri
016680508
neeharika.yeluri@sjsu.edu
Mohamed Shafeeq Usman
015232529
mohamedshafeeq.usman@sjsu.edu

San Jose State University
CMPE 257: Machine Learning
Spring 2023, Prof. Jahan Ghofraniha

Tale of Content

1. Executive Summary

The research examined a dataset of 20,632 Kickstarter projects as of February 1st, 2017 in order to comprehend the Kickstarter market. The objective was to pinpoint the characteristics that give campaigns a better chance of success and to provide insights that may guide future campaign development. Despite the fact that the platform publishes guidance and best practices articles on its site, the survey revealed that more than half of Kickstarter projects fail. This is significant because, in contrast to its rival Indiegogo, Kickstarter uses an "all or nothing" fundraising strategy, which means that if a campaign is unsuccessful, both the project's creators and supporters will be dissatisfied and the project won't be finished in any way.

Based on past data, we used 6 classifiers: Logistic Regression, XGBoost Algorithm, SVC, Random Forest, Decision Tree and KNN to forecast the performance of Kickstarter initiatives. The model's strong accuracy, precision, recall, and F1 score show that it successfully identified successful advertisements while reducing false positives. The research also revealed a number of essential characteristics linked to effective campaigns, such as the project type, financing target, and campaign length. According to the study's findings, the XG Boost model may be a useful tool for gauging the performance of Kickstarter projects. However, since the model is based on previous data, it may not be able to forecast with any degree of accuracy how successful fresh campaigns would be with different characteristics. Further investigation is required to examine the generalizability of the model and the underlying causes of why certain characteristics are linked to campaign effectiveness.

Overall, the research offers insightful information on the elements that contribute to the effectiveness of crowdsourcing and may help build future campaigns that are more successful.

2. Introduction

On the online platform known as Kickstarter, people and organizations can start campaigns to raise money for their innovative projects or ideas. The campaigns are available for a defined amount of time, and the project's authors specify a financial target that must be attained for it to be financed. According to the "all or nothing" financing strategy used by Kickstarter, no money will be given to the project if the funding target is not reached before the conclusion of the campaign. The dataset most likely contains information on each Kickstarter campaign, including the campaign's name, description, fundraising target, project type, campaign location, campaign length, and number of backers who contributed to the project.

The dataset might be explored to learn more about the project kinds that are more likely to succeed on the platform, the most popular project categories, the ideal campaign time, and the regions where campaigns are most successful.

Identifying the success elements of the campaign, such as the caliber of the project pitch, the types of prizes provided to supporters, the number of backers attained, and the social media and marketing tactics used to promote the campaign might also be accomplished by analysing the data. Understanding these characteristics may aid artists in creating future Kickstarter projects that are more successful, hence boosting their chances of doing well on the site.

Dataset Link:

https://www.kaggle.com/datasets/sripaadsrinivasan/kickstarter-campaigns-dataset?resource=download

The dataset contains 20632 rows and 68 columns. There are 39 numerical features and 29 categorical features present in the dataset with no duplicate or null values.

3. Problem statement

The goal of the Kickstarter dataset is to determine the elements that influence the success or failure of projects on the platform for crowdsourcing. Despite the fact that Kickstarter gives creators tips and best practices, more than half of the projects there still fall short of their financial targets. Due to Kickstarter's "all or nothing" fundraising strategy, when a campaign is unsuccessful, both the project's creators and supporters are upset since the project will not be finished in any way. Therefore, knowing the characteristics that make campaigns more likely to succeed is essential to informing campaign design in the future.

The dataset offers a comprehensive source of data on more than 20,000 Kickstarter projects, including campaign names, summaries, fundraising targets, categories, locations, timelines, and supporter details. Investigating this data might provide information on the project categories that are most successful, the ideal campaign length, and the marketing tactics that provide the best funding rates.

The issue statement calls for data analysis to pinpoint the variables that affect campaign success, such as the caliber of the project pitch, the incentives provided to supporters, the total number of backers attained, and the social media and marketing tactics used to spread the word about the campaign. By being aware of these characteristics, designers may go on and create Kickstarter projects that will be more successful.

4. Differentiator/Contribution

There are a few significant differences between the project and the literature study despite the fact that both analyze the Kickstarter dataset to determine the elements that make a campaign successful:

1. Approach: To forecast the success of Kickstarter projects, the project uses machine learning, more particularly working on different classifiers and based on the results providing the best

model. The literature review, in contrast, focuses on examining the descriptive statistics of the dataset to spot patterns and trends.

2. Factors: In addition to the factors examined in the literature study, such as project type and financing status, the project takes a broader range of variables into account, such as fundraising objective, campaign length, and location.

3. Timeframe: While the literature study looks at a wider range of data from 2009 to 2018, the research analysis a particular snapshot of the Kickstarter dataset as of February 1st, 2017.

4. Scope: While the literature review covers a wider range of topics related to crowdfunding, including the role of social networks and the impact of rewards on backer behavior, the project specifically focuses on identifying the characteristics that set campaigns up for a higher rate of success on Kickstarter.

Overall, while the project and the literature review analyze the Kickstarter dataset similarly, the project uses a wider range of variables and a particular machine learning methodology, whereas the literature review adopts a broader approach and covers a wider range of crowdfunding-related topics.

5. Literature Review

Crowdfunding has become a well-liked substitute for conventional fundraising strategies, with websites like Kickstarter allowing producers to solicit money from a large online audience. The success rate of crowdfunding projects is still low, with the majority of initiatives failing to reach their financial targets, despite the success stories. According to research around 40% of crowdfunding efforts on Kickstarter were successful, meaning that a significant fraction of them failed to reach their financial targets [1]. The success of a campaign may be attributed to a variety of variables, according to the authors, including the project pitch's quality, the prizes provided to supporters, the total number of backers reached, and the social media and marketing tactics used to spread the word about the campaign.

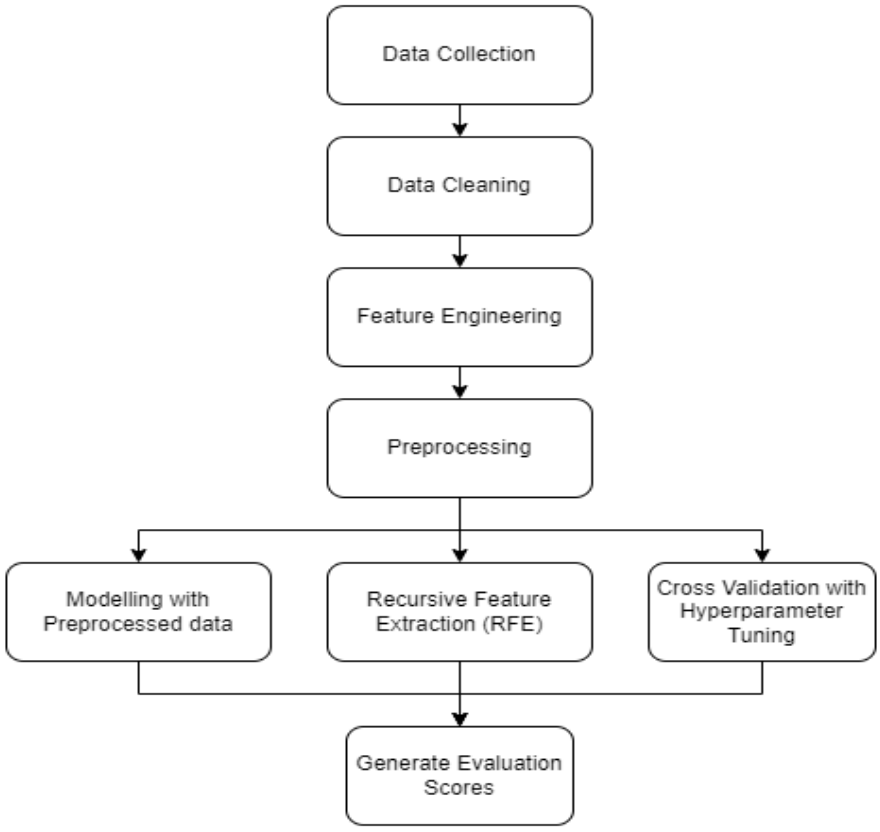
Another research looked at how long a campaign lasts and how successful it is at raising money. Shorter campaigns (30 days or fewer) had a better success rate than lengthier campaigns, according to the authors' analysis of data from Kickstarter projects [2] . This implies that choosing a suitable campaign time is essential to reaching financial targets. The literature has also looked at how social influence affects the success of crowdfunding campaigns. The research discovered a favorable correlation between social media engagement, such as Facebook likes and Twitter followers, and the success of Kickstarter financing initiatives. The research also discovered that campaigns had a better chance of success when there were more updates and comments.

According to a study, some sorts of projects, notably those in the fields of technology and entertainment, are more likely to succeed on Kickstarter than others. The research did point out that success rates varied considerably within each category, underscoring the significance of other elements including the quality of the proposal pitch and incentives provided [3]. Numerous

authors have noted the necessity of providing appealing and distinctive rewards in order to draw backers. The significance of rewards in crowdfunding campaigns has been extensively studied. According to research providing high-quality and distinctive prizes might boost the success of crowdfunding campaigns, especially for less well-known initiatives. Additionally, the authors discovered that backers were more inclined to support campaigns that offered rewards they thought were worthwhile. The effect of campaign updates on the success of crowdfunding has also been studied in the literature, to sum up [4]. According to a study campaign updates that highlighted the project's progress towards its fundraising target and included particular project information were more likely to be funded. Additionally, the study discovered that donors were more likely to support campaigns that offered frequent updates [5].

The literature suggests that a variety of elements, including the quality of the project pitch, the rewards provided to backers, the number of backers attained, the social media and marketing strategies employed to promote the campaign, the proper campaign duration, and the project category, all play a role in the success of crowdfunding campaigns on Kickstarter. Creators may create more successful Kickstarter projects and improve their chances of success on the site by being aware of these criteria.

6. Methodology



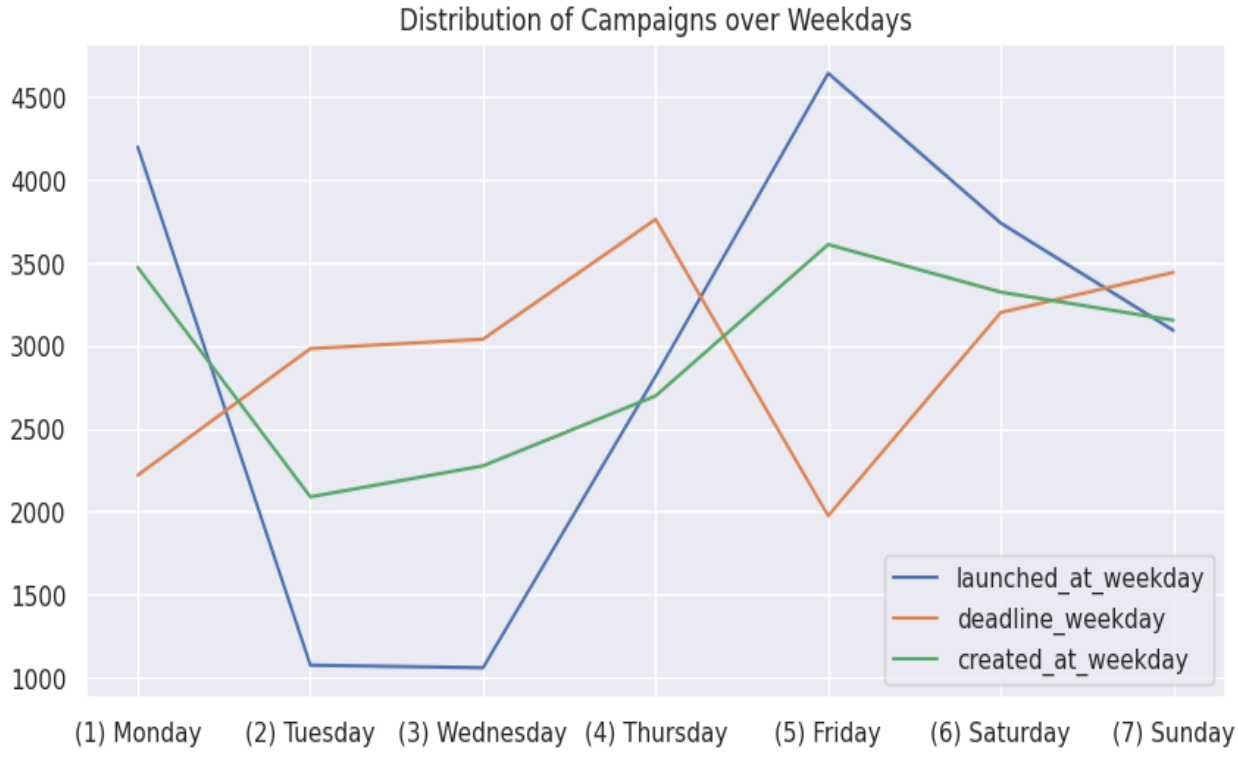
6.1. Exploratory Data Analysis

The Kickstarter dataset was analyzed to identify trends and insights. The dataset consists of over 20,000 records, each representing a campaign on Kickstarter platform. The data includes information such as the category of the campaign, the country in which it was launched, the funding goal, and the campaign's outcome (successful, failed, canceled, live, or suspended).

The first step in data exploration was to check the data's shape and information. The dataset consists of 20632 rows and 68 columns. The columns represent various attributes such as category, country, currency, deadline, launched_at, created_at, state, and others. There were many missing values in friends, is_starred, is_backing, and permissions columns of the dataset so we removed them.

The data was then explored to identify the distribution of the attributes. The histograms showed that the majority of the campaigns had a funding goal not more than $20,000. The distribution of the campaign duration, in days, showed that most campaigns lasted between 20 and 30 days. The distribution of the campaign's success rates showed that the majority of the campaigns failed to reach their funding goals.

The distribution of campaigns was also explored over the weekdays and hours. The campaigns were most launched on Tuesday, and the deadline was mostly on Friday. The distribution of campaigns was also analyzed over UTC Day by hour, where campaigns were most launched between 12 pm and 4 pm UTC.



The dataset was then grouped by country, and the success and failure rates of campaigns in each country were analyzed. The results showed that the campaigns launched in the US had the highest success rate, followed by the UK and Canada.

Recommendations based on above analysis:

The analysis of the Kickstarter dataset has provided several insights that can be used to improve the chances of success for future campaigns.

- **Launch the campaign on Tuesday:**

  Based on the analysis, campaigns launched on Tuesday had the highest success rates. Therefore, it is recommended to launch campaigns on Tuesdays.

- **Set a realistic funding goal:**

  The majority of the campaigns failed to reach their funding goals. It is recommended to set realistic funding goals based on the campaign's category, duration, and location.

- **Target countries with higher success rates:**

  The success rates of campaigns vary by country. It is recommended to target countries with higher success rates such as the US, UK, and Canada.

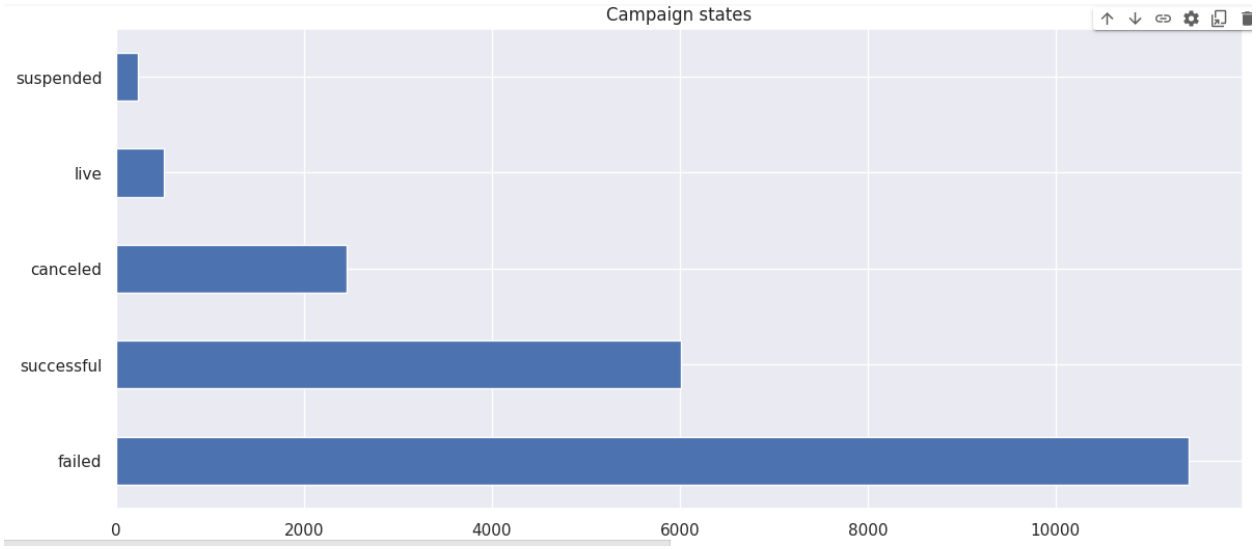
- **Optimize the campaign duration:**

  The analysis showed that most campaigns lasted between 20 and 30 days. It is recommended to optimize the campaign duration based on the category, funding goal, and location of the campaign.

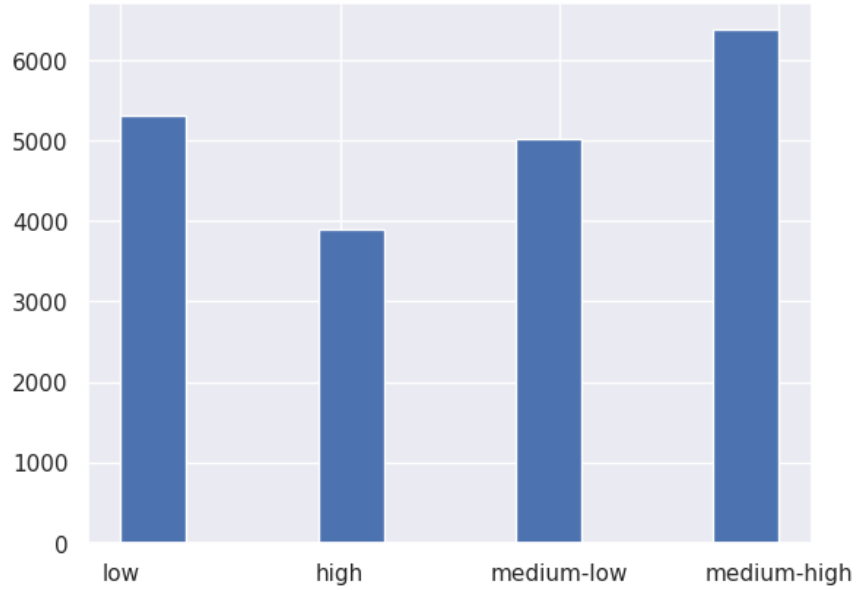
- **Launch the campaign during the optimal hours:**

  Based on the analysis, campaigns were most launched between 12 pm and 4 pm UTC. It is recommended to launch the campaign during the optimal hours to increase the chances of success.

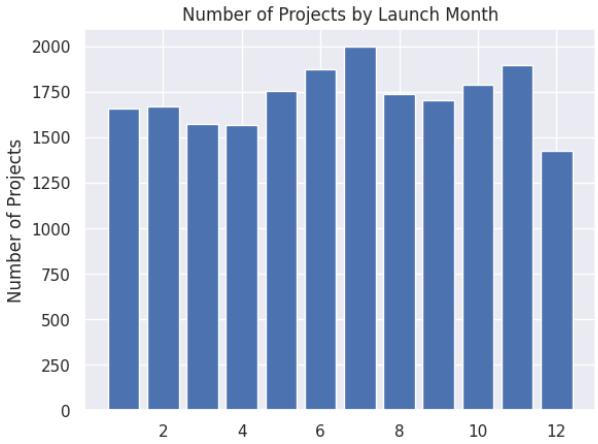
We converted the 'launched_at' and 'deadline' columns to pandas datetime objects to enable easier manipulation of dates and added new columns for launch month, launch year, and goal class to help in our analysis. Then, first examined the overall success rate of Kickstarter projects, finding that approximately 29.17 % of campaigns were successful while 55.33% failed, 11.92 % of campaigns were canceled and 3.58 % were others. We also found that the most successful projects were in the Music and Film & Video categories, while the least successful were in the Journalism and Crafts categories.

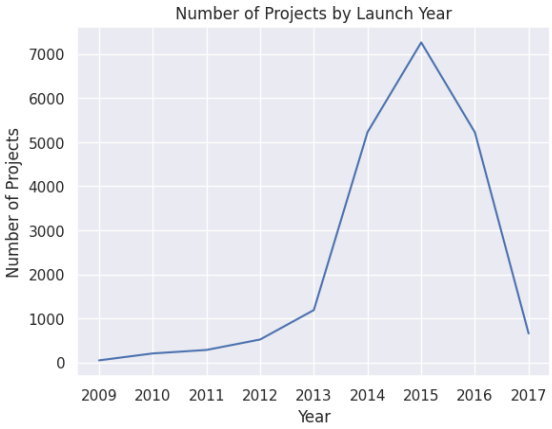


We also investigated the distribution of funding goals, finding that most projects had goals less than $10,000, with a significant number of projects having goals less than $1,000. We categorized the funding goals into four classes and found that most projects fell into the low and medium-high categories.

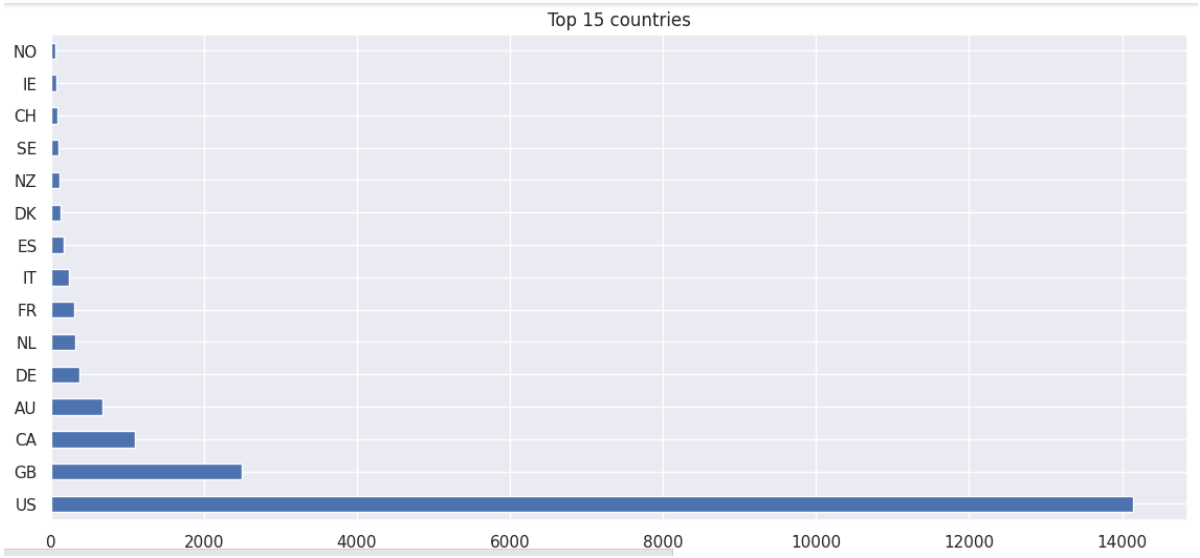


We examined the number of projects launched per month and year and found a steady increase in projects from 2009 to 2015, with a decrease in recent years. We also found that most projects were launched in June and July.





Finally, we looked at the top 15 countries with the most Kickstarter projects, finding that the United States had the most projects by far, followed by the United Kingdom and Canada.

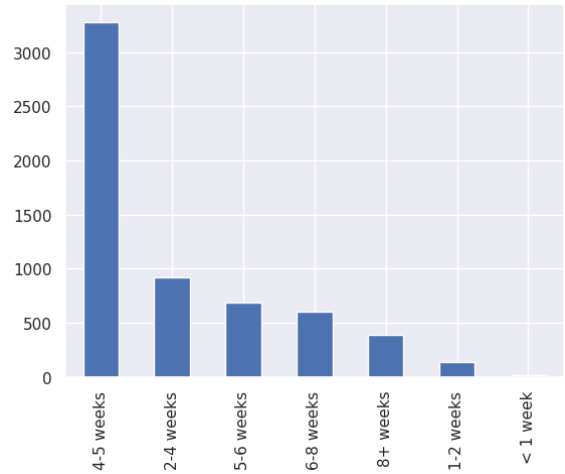


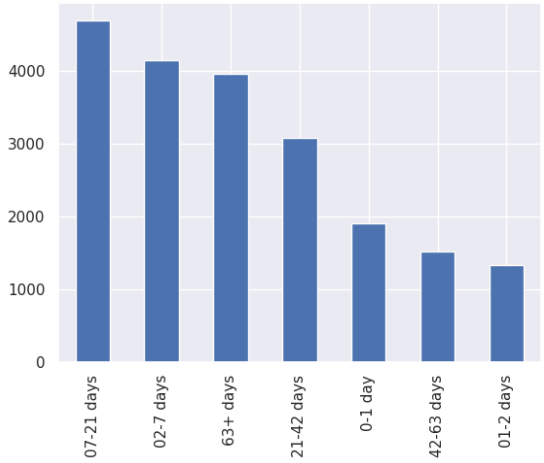
Text column analysis:

Then we cleaned the data by dropping rows with missing values in the 'blurb' column. We then calculated the title and description word counts for each state and created word clouds for the 'name' and 'blurb' columns to visualize the most commonly used words in the campaign titles and descriptions. From the 'name' word cloud, we can see that the most commonly used words are 'cancelled,' 'app,' 'project,' 'smart,' and 'music.'

From the 'blurb' word cloud, we can see that the most commonly used words are 'help,' make,' 'new,' app,' use' and 'world' etc..

We extracted the number of days between the creation and launch dates, launch and deadline dates, and launch and state change dates. We then binned the 'launch_to_deadline_days' column into different time intervals and plotted the distribution of the successful campaigns in each bin. We observed that the majority of successful campaigns had a duration of 4-5 weeks.





We also binned the 'create_to_launch_days' column into different time intervals and plotted the distribution of all campaigns in each bin. We observed that the majority of campaigns were launched within 7-21 days of their creation.

6.2. Data Preprocessing

- The next step involves feature engineering, starts by dropping unnecessary columns from the dataset, such as 'id', 'photo', 'name', 'blurb', 'pledged', 'state', 'slug', 'disable_communication', 'currency', 'currency_symbol', 'currency_trailing_code', 'urls', 'source_url'.
- It also converts date and time columns ('created_at', 'launched_at', and 'state_changed_at') to pandas datetime format.
- The pipeline then calculates the duration of each campaign's creation period ('create_to_launch'), launch period ('launch_to_deadline'), and period from launch to state change ('launch_to_state_change').
- The duration of each period is converted to the number of days using timedelta and total_seconds functions.
- It also creates two new columns, 'SuccessfulBool' and 'USorGB'. 'SuccessfulBool' is a binary column indicating whether the campaign was successful (1) or not (0), based on whether 'state_changed_at' is greater than or equal to 'deadline'.
- 'USorGB' is a binary column indicating whether the campaign was based in the US or UK (1) or not (0).
- The pipeline then creates another binary column 'TOPCOUNTRY', indicating whether the campaign was based in one of the top 15 countries in terms of the number of campaigns launched.

- It also creates two more binary columns, 'LaunchedTuesday' and 'DeadlineWeekend'. 'LaunchedTuesday' is a binary column indicating whether the campaign was launched on a Tuesday (1) or not (0).
- 'DeadlineWeekend' is a binary column indicating whether the campaign's deadline was on a weekend day (Saturday or Sunday) (1) or not (0). It then drops more unnecessary columns, such as 'deadline', 'state_changed_at', 'created_at', 'launched_at', 'creator', 'location', and 'profile'. It then takes variables for categorical columns, such as 'deadline_weekday', 'state_changed_at_weekday', 'created_at_weekday', 'launched_at_weekday', ‘category', and 'goal_class'.

6.3. Feature Selection

After preprocessing the data, the next step is feature selection. Feature selection is important because it helps to identify the most relevant features for predicting the target variable, which in this case is the success or failure of a Kickstarter campaign. Here, Recursive Feature Elimination (RFE) is used to select the top 25 features for prediction. RFE is a wrapper method that works by recursively removing features from the dataset and fitting a model to the remaining features until the desired number of features is reached. The RFE class from the sklearn.feature_selection module is used to perform the feature selection.

The top 25 features selected by RFE are then used to subset the training and testing sets for use in model training and evaluation.

6.4. Model Training and Evaluation

After selecting the features, the next step is to train and evaluate different machine learning algorithms to predict the success of Kickstarter campaigns. We used 6 algorithms: Logistic Regression, Naive Bayes, Support Vector Machines (SVM), XGBoost, Random Forest, and Decision Tree to determine which is best for the features that are extracted from the dataset.

Each algorithm is trained on the selected features and evaluated using four metrics: accuracy, precision, recall, and F1-score.

The LogisticRegression class from the sklearn.linear_model module is used to train and evaluate the Logistic Regression algorithm. The GaussianNB class from the sklearn.naive_bayes module is used to train and evaluate the Naive Bayes algorithm. The SVC class from the sklearn.svm module is used to train and evaluate the SVM algorithm. The XGBClassifier class from the xgboost module is used to train and evaluate the XGBoost algorithm. The RandomForestClassifier class from the sklearn.ensemble module is used to train and evaluate the Random Forest algorithm. The DecisionTreeClassifier class from the sklearn.tree module is used to train and evaluate the Decision Tree algorithm.

7. Implementation and Results

7.1. Implementation

7.1.1. Data Splitting

The first step in our analysis was to split the data into training and testing sets. We used a test size of 0.2, which means that 20% of the data was used for testing and the remaining 80% for training.

7.1.2. Initial Model Training

We initially trained the following classification algorithms on our training set: Logistic Regression, XGBoost, SVC, Gradient Boost, Random Forest, Decision Tree, KNN, and Naive Bayes. These models were trained on the raw data, meaning no feature selection was performed at this stage.

7.1.3. Feature Extraction with RFE

Next, we used Recursive Feature Elimination (RFE) to extract the most relevant features from the dataset and re-ran the same models on this new reduced dataset. We calculated accuracy, precision, recall, and F1 scores for all the models using both the original dataset and the feature-extracted dataset to see if there was any improvement in the model performance.

7.1.4. Model Selection with Hyperparameter Tuning

Hyperparameter tuning is the process of finding the best set of hyperparameters for a machine learning model that maximizes its performance on a given dataset. In this project, we used GridSearchCV to perform hyperparameter tuning for each of the models in our dictionary.

We defined a dictionary params that specifies the hyperparameters to tune for each model. For each model, we performed a grid search using 5-fold cross-validation and computed the best hyperparameters using the best_params_ attribute of the GridSearchCV object.

We stored the best models obtained from hyperparameter tuning in a dictionary called best_models. We will use these models in the cross-validation step to compare their performance.

7.1.5. Cross-Validation

Cross-validation is a technique used to evaluate the performance of a machine learning model on a given dataset. In this project, we used 5-fold cross-validation to evaluate the performance of each model.

For each model, the best hyperparameters are identified using grid search, and the model's performance is evaluated using cross-validation.

The best hyperparameters for logistic regression are identified as 'C': 0.1 and 'penalty': 'l2'. The cross-validation scores for logistic regression range from 0.842 to 0.933, with a mean cross-validation score of 0.871.

The best hyperparameters for the decision tree are identified as 'max_depth': 5 and 'min_samples_leaf': 10. The cross-validation scores for decision trees range from 0.975 to 0.978, with a mean cross-validation score of 0.979.

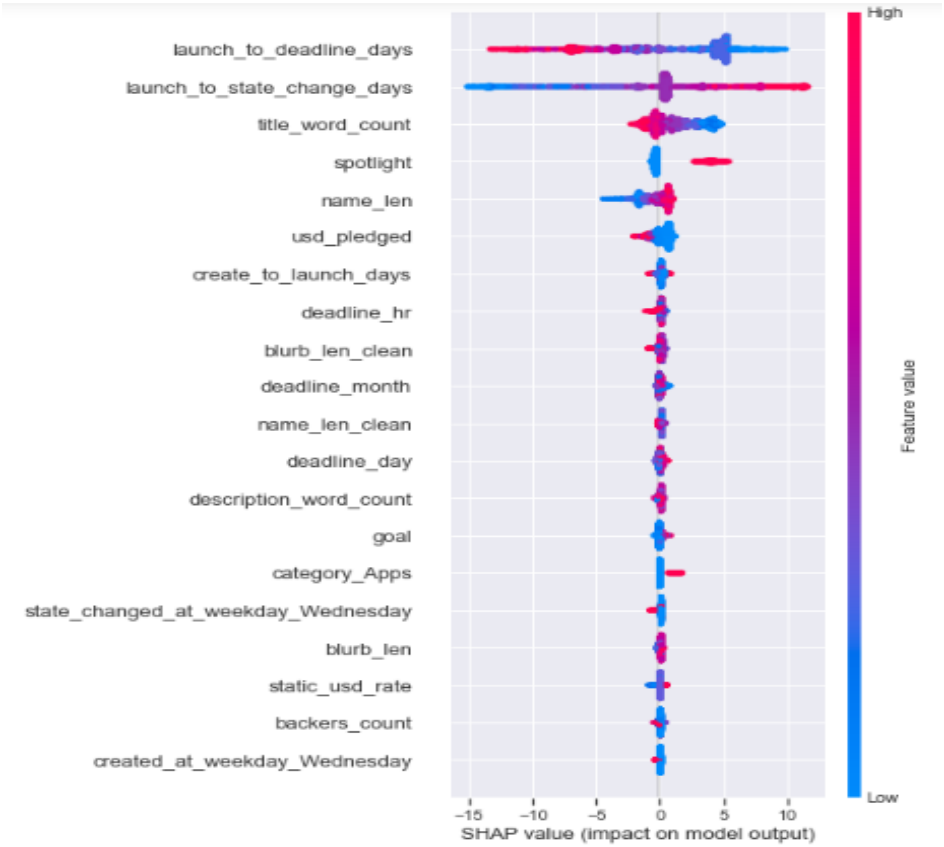
The best hyperparameters for random forest are identified as 'max_depth': 7, 'min_samples_split': 2, and 'n_estimators': 50. The cross-validation scores for random forest range from 0.905 to 0.951, with a mean cross-validation score of 0.931.

The best hyperparameters for XGBoost are identified as 'learning_rate': 0.1, 'max_depth': 5, and 'n_estimators': 500. The cross-validation scores for XGBoost range from 0.982 to 0.995, with a mean cross-validation score of 0.991.

Based on these results, XGBoost appears to be the best model, with the highest mean cross-validation score of 0.991. Therefore, XGBoost is selected as the best model for the given dataset.

7.1.6. SHAP Plots

SHAP plots are a way to visualize the contribution of each feature to the model output for a particular data point. They help to understand how the model made its prediction for that specific data point by showing the impact of each feature on the final output value. The SHAP values indicate the extent to which each feature contributed positively or negatively to the predicted outcome. The color of the dots in the plot represents the value of the feature for that data point, where red dots represent high values and blue dots represent low values.



Here, the shap_values object contains the SHAP values for each feature in the test set. The summary_plot() function is used to plot the feature importance for all observations in the test set. The plot shows the most important features at the top of the plot and the least important features at the bottom. Each dot represents a data point, and its position on the x-axis represents the impact of that feature on the prediction for that data point. The y-axis shows the feature name.

7.2. Results

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.8494	0.8494	1.0	0.9186
XGBoost Algorithm	0.9925	0.9915	0.9996	0.9956
SVC	0.8497	0.8496	1.0	0.9187
Random Forest	0.9754	0.9730	0.9987	0.9857
Decision Tree	0.9855	0.9902	0.9927	0.9915
KNN	0.8604	0.8813	0.9656	0.9215

Table 1. Initial Results

The initial results shown in Table 1 reveals that the XGBoost algorithm had the highest accuracy (0.9925), indicating that it correctly classified 99.25% of the test data. This was followed by Decision Tree (0.9855) and Random Forest (0.9754), which achieved high accuracy. The high accuracy scores of these models suggest that they were able to correctly classify a large proportion of the test data and are promising candidates for further analysis.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Logistic Regression	0.9943	0.9934	1.0	0.9967

XGBoost Algorithm	0.9922	0.9918	0.9990	0.9954
SVC	0.9834	0.9808	1.0	0.9903
Random Forest	0.9839	0.9817	0.9996	0.9906
Decision Tree	0.9890	0.9921	0.9949	0.9935
KNN	0.9812	0.9784	1.0	0.9891

Table 2. Results After Performing Feature Extraction Using RFE

Table 2 shows the results of the models after running them with Feature Selected data using RFE. We can observe that Logistic Regression gives the best results with an accuracy of 0.9943, followed by XGBoost giving similar accuracy of 0.9922. Unlike previous results, rest of the models show good performance with ranging around 98% accuracy, Decision Tree having accuracy close to 99%.

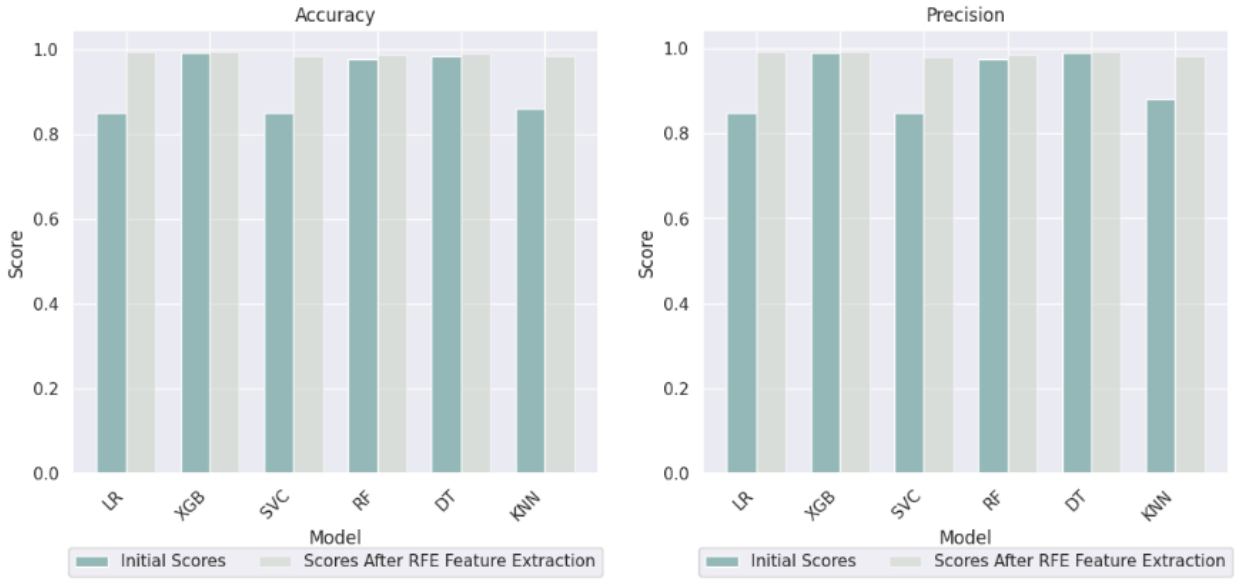


Fig. x Accuracy and Precision Comparison pre and post RFE.

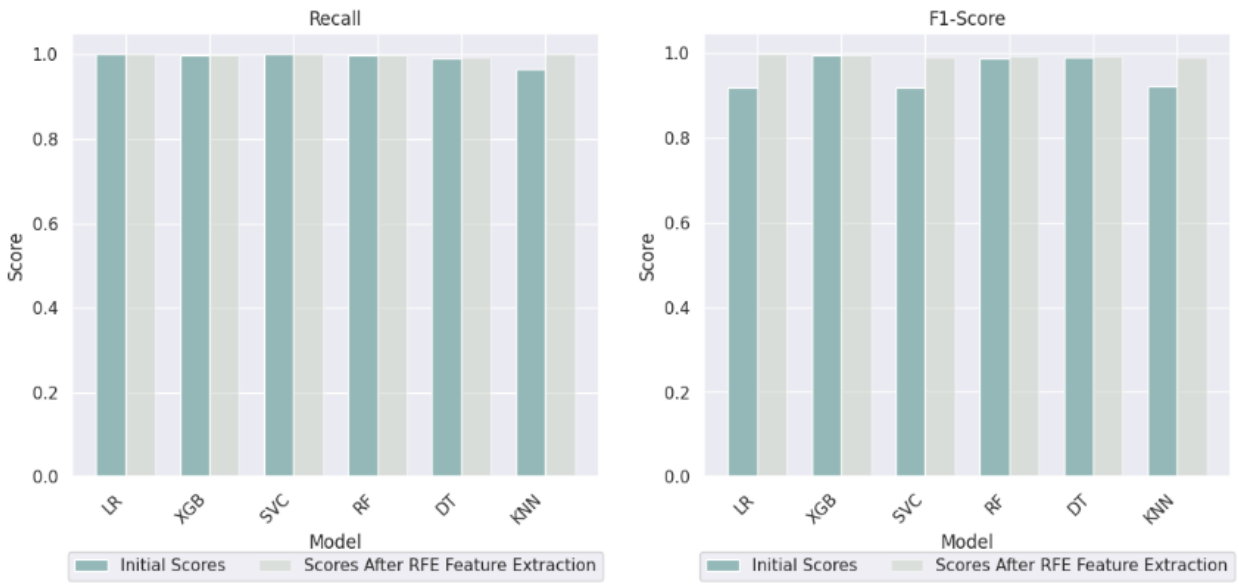


Fig. x Recall and F1-Score Comparison pre and post RFE.

The figures x1 and x2 demonstrates the positive impact of feature selection using RFE on the performance of the models. Several models, which initially performed reasonably well, showed a significant improvement in their accuracy after the feature selection process. For instance, Logistic Regression, SVC, and KNN showed a remarkable increase in their accuracy from about 85% to 99.43%, 98.34%, and 98.12%, respectively. Even XGBoost, which had the highest accuracy initially, maintained a similar level of performance after feature selection. Moreover, other models that had accuracy around 98% also saw an increase of approximately 1% in their accuracy.

It's worth noting that none of the models showed a decrease in performance after the feature selection, indicating that they were all well-fitted with the feature-selected data. Overall, the results demonstrate the effectiveness of RFE in enhancing the performance of the models.

Model	**Accuracy**
Logistic Regression	0.9315
XGBoost Algorithm	0.9949
Random Forest	0.9339

Decision Tree	0.9922
KNN	0.8705

Table 3. Accuracy Results after Hyperparameter Tuning and Cross Validation

Based on the findings in Table 3, the XGBoost Algorithm achieved the highest accuracy score of 0.9949, closely followed by Decision Tree with an accuracy of 0.9922. Random Forest and Logistic Regression both achieved decent results, with accuracy scores of around 93%. However, KNN had the lowest accuracy score of 0.8705.

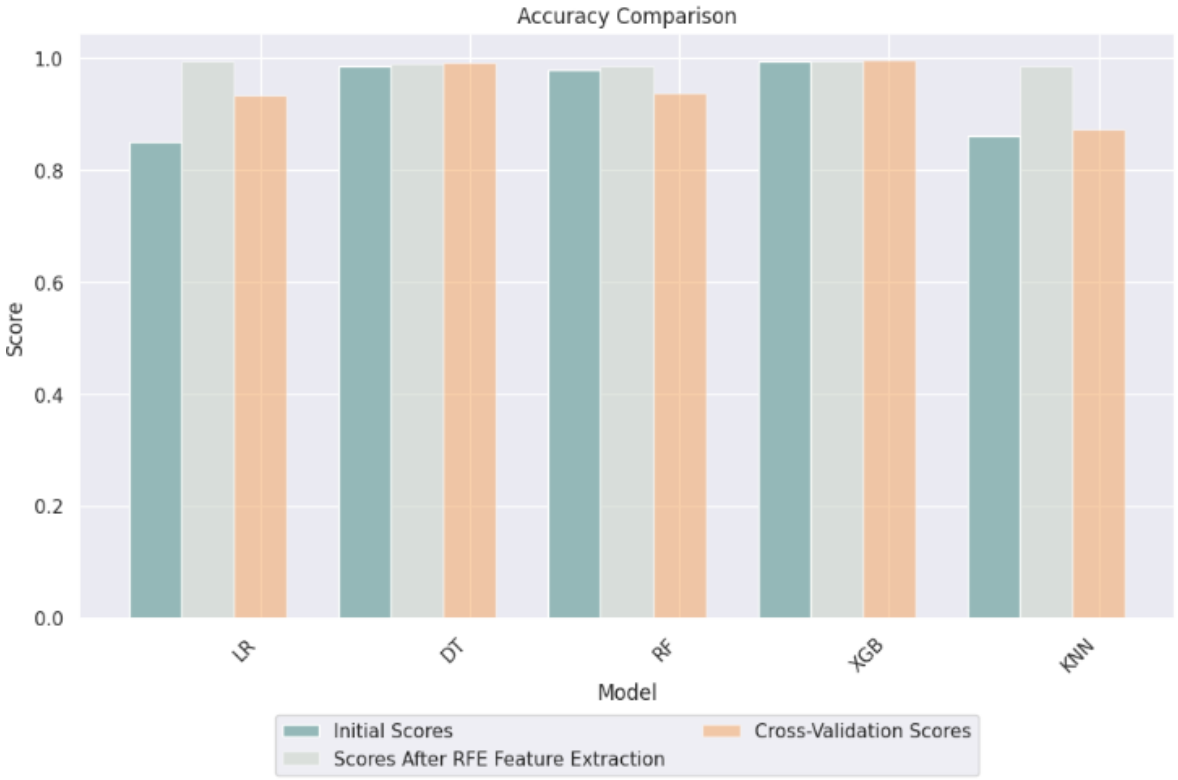


Fig. x Accuracy Comparison of All Models Processed

We can note that while some models performed better after cross-validation, others showed a decline in their accuracy scores. Specifically, the accuracy of Logistic Regression increased by around 8%, while XGBoost's accuracy remained similar and KNN's accuracy remained around 87%. On the other hand, Random Forest and Decision Tree showed a decline in their accuracy scores by 4% and 1%, respectively.

Furthermore, when comparing the cross-validation results with the results after feature selection, XGBoost and Decision Tree performed similarly, while the other models had significantly lower accuracy scores. Specifically, Logistic Regression had a 6% lower accuracy score, Random Forest had a 5% lower accuracy score, and KNN had an 11% lower accuracy score after feature selection.

7.3. Contribution

In our project, each team member made significant contributions to different aspects of the work, ensuring a comprehensive and well-rounded analysis. The following is a detailed elaboration of each team member's contributions:

Bhavana: Worked in the exploratory data analysis (EDA), carefully examined the dataset, conducted statistical analyses, and visualized the data to gain insights into its characteristics.
Neeharika: Both on the preprocessing and post-processing.This involved tasks such as handling missing values, feature selection, and encoding categorical variables and SHAP plots for examining the impact of selected features. Experimented with different hyperparameters using grid search, and evaluated the model's performance using cross-validation. Also, actively participated in the EDA phase, collaborating with Bhavana to explore the dataset further.
Shafeeq: Primary focus was on building and fine-tuning classifier models for our project. Applied various techniques to train and optimize the models. Experimented with different classifier architectures, tuning hyperparameters, and implementing ensemble methods to improve performance with RFE. Also, contributed to model evaluation, employing appropriate metrics and conducting rigorous testing to assess the models' performance.

By dividing the work among the team members in this way, we were able to maximize our efficiency and ensure a comprehensive analysis of the dataset. Each member's unique skills and expertise complemented the others, resulting in a well-rounded project.

8. Conclusion

In conclusion, the results of the study indicate that using machine learning algorithms can accurately classify the given dataset. The XGBoost algorithm performed the best in terms of accuracy before feature selection, with an accuracy score of 0.9925. However, after performing feature selection using RFE, Logistic Regression gave the best results with an accuracy of 0.9943, followed closely by XGBoost with an accuracy score of 0.9922. The other models also showed improved performance after feature selection, with all models showing an increase in accuracy, indicating the effectiveness of RFE in enhancing model performance.

Furthermore, after performing hyperparameter tuning and cross-validation, XGBoost Algorithm achieved the highest accuracy score of 0.9949, followed by Decision Tree with an accuracy of 0.9922. Random Forest and Logistic Regression both achieved decent results, with accuracy scores of around 93%, while KNN had the lowest accuracy score of 0.8705.

After experimenting with different techniques to optimize the model performance, we can conclude that using feature selection with Recursive Feature Elimination (RFE) generated

notable improvements in accuracy across all the models, with some models exhibiting a substantial increase in accuracy. Although XGBoost with cross-validation produced the highest accuracy, the cross-validation approach led to a decline in accuracy for some models

It is important to note that while some models showed improvement in their accuracy scores after cross-validation, others showed a decline. The findings of this project demonstrate the potential of machine learning algorithms in accurately classifying the given dataset, and the importance of using appropriate techniques like feature selection and hyperparameter tuning to improve model performance.

9. Future work

Finally, although this study identified a number of crucial characteristics linked to effective Kickstarter campaigns, it did not investigate the underlying causes of these characteristics' significance. Future studies could look into the psychological and societal elements that influence the success of a campaign, such as the influence of emotional appeals, social identity, and social proof. This would provide a more nuanced knowledge of the elements that contribute to the success of crowdsourcing and might guide the creation of future tactics that are more successful.

For the category column, there are notable missing values that can be addressed in future work. As our analysis revealed a high correlation between category and the target variable, imputation methods could be applied to fill in those missing values. By doing so, we could increase the amount of available data, which may improve the accuracy of our models. This could be an area for future exploration and further investigation.

Appendix A

A copy of the link to where the code resides:

Notebook:
https://colab.research.google.com/drive/1wr9g0F--8Aq6bozfOnno1m08i-hEFWMA?usp=sharing

Github: https://github.com/MdShafeeqU/Kickstarter-Success-Prediction

Appendix B: References

[1] Etter, V., Grossglauser, M., & Thiran, P. (2013). Launch hard or go home! Predicting the success of kickstarter campaigns. COSN 2013 - Proceedings of the 2013 Conference on Online Social Networks, 177–182. https://doi.org/10.1145/2512938.2512957

[2] Kindler, A., Golosovsky, M., & Solomon, S. (2019). Early prediction of the outcome of Kickstarter campaigns: is the success due to virality? Palgrave Communications, 5(1). https://doi.org/10.1057/s41599-019-0261-6

[3] Colistra, R., & Duvall, K. (2017). Show Me the Money: Importance of Crowdfunding Factors on Backers' Decisions to Financially Support Kickstarter Campaigns. Social Media and Society, 3(4). https://doi.org/10.1177/2056305117736942

[4] Robertson, E. N., & Wooster, R. B. (n.d.). Crowdfunding as a Social Movement: The Determinants of Success in Kickstarter Campaigns. http://ssrn.com/abstract=2631320Electroniccopyavailableat:http://ssrn.com/abstract=2631320

[5] Surya Engineering College, & Institute of Electrical and Electronics Engineers. (n.d.). Proceedings of the 3rd International Conference on Computing Methodologies and Communication (ICCMC 2019) : 27-29, March 2019

[6]https://github.com/rileypredum/Kickstarter-Campaign-Success-Prediction/blob/master/kickstarter.ipynb

[7]https://towardsdatascience.com/kickstarter-projects-walk-through-simple-data-exploration-in-python-c2302a997789