Kickstarter Projects: Analyzing Success Factors Using Big Data Techniques

Group Members:
Fahim Istiak - 2407699
Mohammad Rakibur Rahman - 2410695
Mazidul Islam - 2503039

Submitted to:
Ekaterina Gilman
Amirhossein Ghaffari

May 4, 2025

Big Data Processing and Applications
University of Oulu

Table of Contents

Project Description

Crowdfunding platforms, like Kickstarter, have transformed how projects are funded, offering creators the opportunity to connect with backers to bring their ideas to life. However, despite the vast number of projects launched every day, predicting whether a project will succeed or fail remains a complex and often uncertain task. Understanding the factors that influence the success of these projects is critical for both backers and creators. This project aims to predict the success or failure of Kickstarter projects using advanced machine learning techniques, offering insights into the variables that drive project outcomes.

This study utilizes a publicly available dataset from Kickstarter, which includes a diverse set of features reflecting the various aspects of each project. These features encompass financial targets, pledges, backers, project duration, and more. The goal of this analysis is to leverage machine learning to model the likelihood of a project's success based on these key features, helping inform future crowdfunding campaigns and providing a data-driven approach to predicting project outcomes.

The project explores the use of PySpark—a scalable framework for large-scale data processing—along with machine learning algorithms to understand the underlying patterns that contribute to crowdfunding success. By analyzing this data, the project aims to uncover critical insights that can help creators optimize their projects and strategies, while also providing backers with valuable information for making informed funding decisions.

Ultimately, this work highlights the potential of machine learning to transform the way we predict the success of crowdfunding campaigns. By examining the interplay of different features within the dataset, the project serves as a foundational step toward better understanding and optimizing the dynamics of the crowdfunding ecosystem.

Literature Review

The field of crowdfunding success prediction has seen significant research employing machine learning algorithms to predict the likelihood of success or failure in crowdfunding campaigns. Several studies have focused on identifying key factors and applying predictive models to improve the accuracy of such predictions. Below is a review of the most relevant literature in this domain.

1. **Crowdfunding Dynamics and Foundational Insights**
   A considerable body of research has explored the factors influencing success in crowdfunding platforms like Kickstarter. One foundational study in this area is Mollick’s (2014) exploratory analysis of crowdfunding dynamics, which provides a broad overview of success and failure patterns among over 48,000 Kickstarter projects, collectively raising more than $237 million. The study emphasizes that project quality signals—such as a well-developed description, clear goals, and visible personal networks—play a crucial role in determining success. Notably, geographic factors also emerged as influential, with the type and success rate of projects varying by location. Additionally, the research highlights that while most successful campaigns aim to deliver promised goods, a significant proportion (over 75%) experience delays, with funding levels correlating with the extent of delay. These findings offer valuable context for the current project, which applies machine learning techniques to identify and quantify similar success-driving factors in Kickstarter campaigns. [1]

2. **Machine Learning Approaches for Crowdfunding Success Prediction**
   Lysin (2024) conducted a comprehensive study analyzing over 150,000 Kickstarter campaigns to evaluate the performance of various machine learning models in predicting campaign success. The research focused on three tree-based classifiers—Random Forest, XGBoost, and LightGBM and

identified key features such as campaign duration, subcategory, and fundraising goal as critical predictors of success. Among the models, Random Forest was ultimately selected as the most effective, although both XGBoost and LightGBM performed at a similarly high level. Evaluation metrics included cross-validation accuracy, precision, recall, F1 score, ROC AUC, and log loss, ensuring robust model validation.

The thesis also addresses the practical implications of campaign success prediction for both creators and potential backers. It emphasizes that accurate prediction tools can support strategic decision-making, improve resource allocation, and enhance campaign design before launch. Despite its strengths, the study acknowledges limitations such as potential data bias and recommends future developments like real-time prediction systems and dataset expansion.

This work closely aligns with the current project's goals, offering both technical insights and practical considerations for using machine learning techniques to understand success factors in Kickstarter campaigns. [2]

3. **Challenges in Crowdfunded Product Development**

Jensen and Özkil (2018) conducted an empirical review of 144 successfully funded technology campaigns on Kickstarter to explore common challenges in crowdfunded product development. Their study focused on physical consumer hardware products and used a failure mode framework to assess delivery outcomes. Despite being fully funded, only 32% of the campaigns delivered products on time. Furthermore, delays were often associated with compromises in product quality or missing features.

The research identified core issues in the product development process including manufacturing difficulties, unrealistic timelines, and underestimation of logistical complexities as key contributors to campaign underperformance post-funding. These insights underscore the importance of considering not only factors that predict campaign funding success but also the post-campaign fulfillment challenges that can influence long-term project viability and backer satisfaction.

While the current project focuses primarily on predicting campaign success using big data techniques, integrating development and delivery-related variables in future models could enhance the real-world applicability of success predictions. [3]

4. **Social Dynamics and Backer Behavior**

Kuppuswamy and Bayus (University of North Carolina) investigated how social information influences the funding behavior of backers in Kickstarter campaigns. Using two years of panel data covering both successful and unsuccessful projects, they demonstrated that potential backers are influenced by the visible level of existing support, a phenomenon explained by the "diffusion of responsibility" effect. Specifically, when a project appears to already have strong backing, new backers may hesitate to contribute, assuming their support is not essential.

However, this effect diminishes as the campaign approaches its deadline. The authors highlight that project creators often increase the frequency of updates during the final days, which tends to reinvigorate backer engagement. These updates, combined with deadline urgency, contribute to a surge in support during the final stages of a campaign, particularly for those that ultimately reach their funding goals.

This research highlights the importance of temporal patterns and social cues in backer behavior, offering valuable insight into the human factors that contribute to crowdfunding success. Such behavioral dynamics are complementary to big data-based predictions, as they contextualize how user interactions evolve over a campaign's lifecycle. [4]

5. **Fulfillment Rate and Project Risk in Crowdfunding**
The study "Measuring project success: the fulfillment rate of crowdfunded projects on Kickstarter" by Sieuwert van Otterloo examines the fulfillment or delivery rates of crowdfunded projects on Kickstarter, a critical aspect of project success. By analyzing data from 35 crowdfunded projects, the study reveals that approximately 30% of projects are delivered on time or with minimal delay, while 40% of projects fail to deliver any promised rewards. These findings highlight the significant risk of failure inherent in crowdfunding campaigns, which is not always transparently communicated to backers.

The research underscores the importance of providing potential backers with accurate information about the risks of failure, thus enabling them to make informed decisions. It also emphasizes the need for crowdfunding platforms to improve transparency and communication regarding delivery rates. The study contributes to our understanding of project risk in crowdfunding and serves as a valuable resource for improving both project management practices and the overall trustworthiness of crowdfunding platforms. [5]

6. **Project Success Prediction in Crowdfunding Environments**
The paper "Project Success Prediction in Crowdfunding Environments" by Yan Li, Vineeth Rakesh, and Chandan K. Reddy addresses one of the most critical challenges in the crowdfunding domain: predicting the success of a project. Despite the growing popularity of crowdfunding platforms, only about 40% of projects achieve their funding goals. This research introduces a novel approach for predicting project success, applying survival analysis and censored regression techniques to Kickstarter data.

The study rigorously analyzes over 18,000 Kickstarter projects and 116,000 corresponding tweets to understand project success time distributions. By incorporating both classification and regression models, the authors show that predictions can significantly improve when both successful and failed projects are used in training, as opposed to relying only on successful ones. The addition of temporal features, especially those captured in the early stages of the project, further enhances the predictive model's accuracy.

This work contributes valuable insights into the dynamics of crowdfunding project success and introduces prediction models that could be critical for project creators, backers, and crowdfunding platforms alike. [6]

7. **A Machine Learning Approach to Predict the Success of Crowdfunding Fintech Projects**
Yeh and Chen (2020) examine the problem of predicting the success of crowdfunding fintech projects by addressing the issue of information asymmetry between project creators and funders. Drawing from social capital theory, human capital theory, and the level of processing (LOP) theory, they propose a machine learning framework that utilizes social media activity, funder experience, and project presentation quality as predictive features.

To mitigate overfitting, a common issue with neural networks, they introduce an ensemble-based artificial neural network (ANN) model incorporating dropout techniques. Four machine learning techniques are compared, with the ensemble ANN achieving the highest predictive accuracy. Their results demonstrate that prior successful campaigns by funders and early-stage investment patterns serve as strong indicators of future project success.

> This study highlights the predictive power of social and human capital in crowdfunding environments and shows the practical utility of ensemble learning methods in improving prediction reliability. The model not only assists project creators in improving campaign design but also provides investors with a tool for risk evaluation. [7]

The studies discussed in this section collectively emphasize the application of machine learning models, such as Logistic Regression, Random Forest, and SVM, in predicting crowdfunding success. They also highlight the importance of feature selection, with common predictors including campaign description, backer count, goal amount, and project category. Additionally, the integration of NLP for text analysis and data preprocessing techniques plays a crucial role in improving the predictive performance of the models. Ensemble learning methods, including Random Forest and Gradient Boosting, have also been found to be effective in handling the complexity of crowdfunding data. These insights from the literature will inform the development of the current project's predictive model, which seeks to advance the understanding of crowdfunding success using machine learning and PySpark.

Data Description

The dataset used in this project was sourced from Kaggle: Kickstarter Projects (https://www.kaggle.com/datasets/kemical/kickstarter-projects). It is publicly available under an open data license, enabling its use for academic and research purposes. The dataset is provided in CSV format, containing data on 378,661 Kickstarter projects, each with various attributes. These attributes include the project’s funding goal, funding pledged, number of backers, categories, status, and countries, providing insight into the financial targets, the amount raised, and the geographical and categorical context of each project.

The dataset features several key variables that describe the Kickstarter campaigns. Financial variables such as fundraising goal, fundraising pledged, backers, and completion percentage capture the economic aspects of each project. Temporal features like days allotted indicate how long the project is set to receive funding. Categorical features, including category, main category, and currency, offer additional context about the project's type and location. The status variable indicates the final outcome of the project, such as whether it was successful, failed, live, suspended, or cancelled, and this variable is used as the target for prediction in the analysis.

In the final analysis, the dataset focuses on a few continuous variables such as days allotted, fundraising goal, fundraising pledged, completion percentage, and backers. Table 1 below presents a detailed overview of these continuous variables, including their descriptive statistics. The days allotted variable shows an average project duration of 33 days, with a range from 1 to 92 days. The typical fundraising goal is $5,871.67, varying between $1 and $27,000. On average, fundraising pledged is $2,340.04, with values ranging from $0 to $206,751. The completion percentage of projects shows a broad variation, with an average of 184%, reflecting the proportion of the goal raised. The number of backers for a project averages 29.87, ranging from 1 to 140.

Table 1. Data Description of Variables Used in the Final Analysis

Variables	Type	N = 219,052	Mean	Median (IQR)	Min-Max
Days Allotted	int64	33.42	12.88	33 (22–43)	1–92
Fundraising Goal (USD)	float64	5,871.67	6,104.14	3,500 (1,000–7,000)	1–27,000
Fundraising Pledged (USD)	float64	2,340.04	3,956.52	1,000 (200–5,000)	0–206,751
Completion Percentage	float64	184.15	6,403.14	100 (50–150)	0–1,506,600
Backers	int64	29.87	33.81	15 (5–40)	1–140

This dataset provides a detailed snapshot of Kickstarter campaigns, making it ideal for exploring the factors that contribute to the success or failure of a project. The target variable is status, which is binary, indicating whether the project succeeded or failed. Since the dataset is cross-sectional, it provides a snapshot of projects at a given point in time rather than tracking their progression, making it particularly suited for classification tasks rather than time-series analysis.

Methods and Tools

In this project, we aim to predict the success or failure of Kickstarter projects using a variety of machine learning techniques and tools. The primary task involves transforming the raw data into useful features, selecting the appropriate machine learning models, training those models, and evaluating their performance. Below are the key methods and tools employed in this project.

1. **Data Pre-processing**
   The first step in this project was to preprocess the data, ensuring it was suitable for analysis. Initially, missing values were handled by removing rows with missing values. Duplicate entries were also removed to maintain the integrity of the dataset. For categorical features such as category, main category, currency, and country, StringIndexer was used to encode these variables into numeric indices, making them compatible with machine learning algorithms. In addition, feature scaling was applied to numerical features where required, standardizing their ranges to ensure that no feature would dominate the model due to its scale.

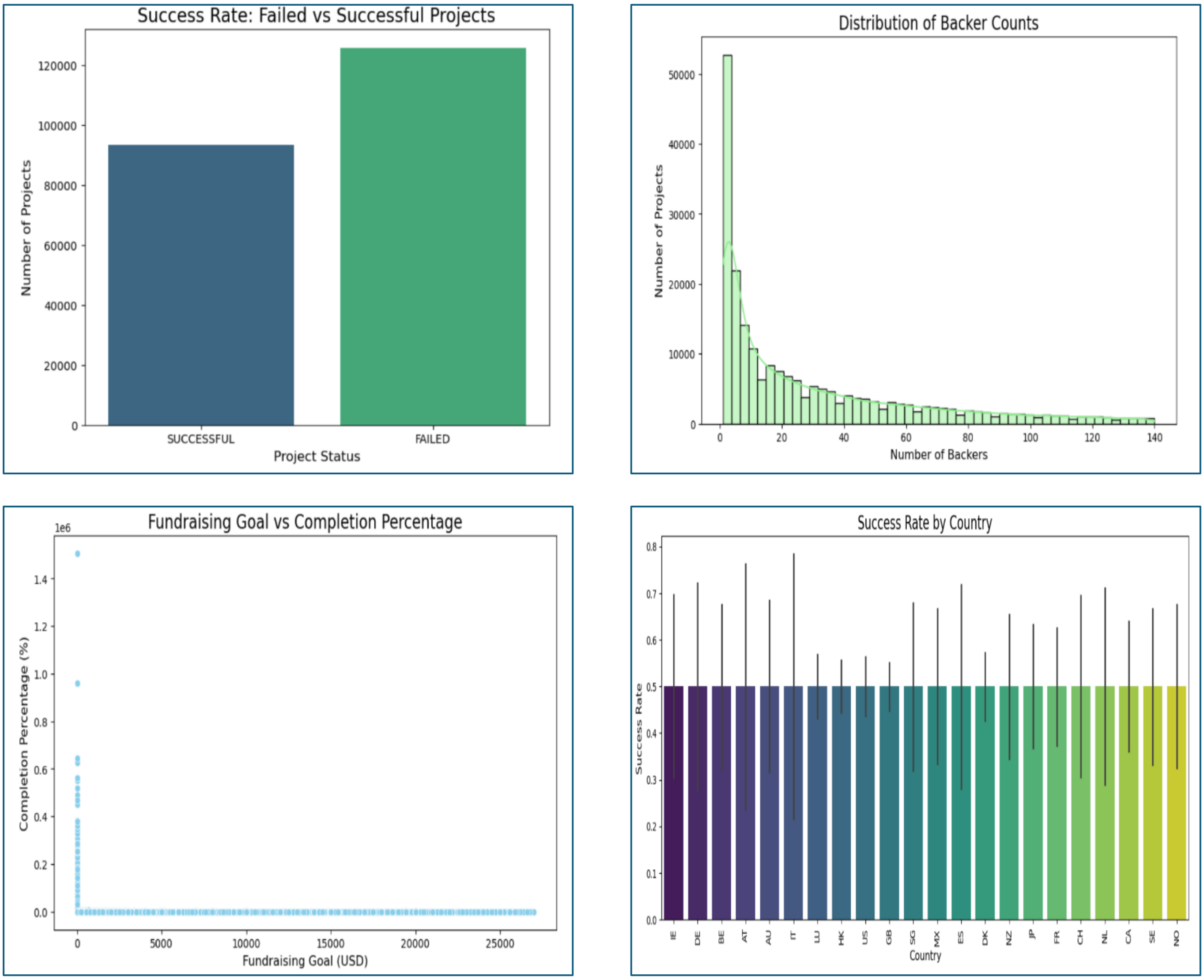

Figure 1: Data visualization

2. **Exploratory Data Analysis (EDA)**
   Before diving into modeling, we performed Exploratory Data Analysis (EDA) to gain insights into the dataset. The primary objective of EDA was to better understand the distribution of the data, identify any patterns, correlations, or anomalies, and determine which features might be important for the prediction task. During the EDA phase, we explored the distribution of key continuous variables such as fundraising goal, fundraising pledged, backers, and completion percentage. We also examined the relationships between status (the target variable) and other features using visualizations like histograms, scatter plots, and box plots. Additionally, we identified any potential outliers in the data, which were subsequently handled through outlier removal. Correlation matrices were used to uncover relationships between numerical variables, and categorical features were analyzed to see how they influenced the outcome variable.

3. **Feature Engineering**
   Feature engineering played a crucial role in transforming raw data into meaningful features for the models. One such feature, the completion percentage, was created by calculating the ratio of fundraising pledged to fundraising goal, providing insight into how close the project came to meeting its target. After the feature creation, categorical variables were encoded using StringIndexer to convert them into numerical formats. The dataset was then ready for machine learning, containing both continuous and categorical features, allowing the models to learn the relationship between these features and the project's outcome.

4. **Model Selection**
   For this project, two machine learning models were used: Logistic Regression and Random Forest Classifier. Logistic Regression was chosen as a baseline model due to its simplicity and efficiency for binary classification tasks, which fits the nature of the target variable, status (successful or failed). On the other hand, Random Forest was used due to its robustness and ability to handle complex, non-linear relationships. Random Forest, being an ensemble model, aggregates multiple decision trees to make predictions, making it less prone to overfitting compared to individual decision trees.

5. **Model Training and Evaluation**
   The dataset was split into training and testing sets, with 80% of the data used for training the models and the remaining 20% reserved for testing. This division simulates the real-world scenario of training a model on past data and evaluating it on unseen data. In addition to the train-test split, 5-fold cross-validation was applied to further mitigate overfitting and ensure that the model generalizes well. In each fold, the model was trained on four subsets of the data and tested on the remaining subset, which was rotated through the five folds. This technique helps estimate the model's performance more reliably by reducing the variance in evaluation metrics.

6. **Hyperparameter Tuning**
   To further optimize the model’s performance, hyperparameter tuning was conducted using grid search. The grid search tested different combinations of parameters for the models. For Logistic Regression, the regularization parameter (regParam) and the number of iterations (maxIter) were tuned, while for Random Forest, the number of trees (numTrees) was adjusted. This process helped identify the optimal set of hyperparameters that led to the best model performance.

7. **Evaluation Metrics**
   The models were evaluated based on two key metrics: accuracy and ROC AUC score. Accuracy was used to determine the proportion of correctly classified projects in the test set. ROC AUC was selected as the primary evaluation metric because it provides a clear measure of the model's ability to distinguish between successful and failed projects, even when the dataset is imbalanced. The confusion matrix was also examined to further assess the classification performance by showing true positives, false positives, true negatives, and false negatives, helping to understand where the model made mistakes.

8. **Model Improvement**
   To address potential overfitting and enhance model performance, regularization techniques like L1 (Lasso) and L2 (Ridge) were applied to Logistic Regression. These regularization methods penalize large model coefficients, thus controlling the complexity of the model and improving its generalization. Additionally, feature importance was evaluated using the Random Forest model, which provided insights into which features were most influential in predicting project success. The top features, such as fundraising goal and backers, were found to have the greatest impact on model predictions.

The methods and tools applied in this project enabled the successful prediction of Kickstarter project outcomes. By preprocessing the data, selecting meaningful features, and applying appropriate machine learning models like Logistic Regression and Random Forest, the project demonstrated the ability to predict project success with high accuracy. Regularization techniques, cross-validation, and hyperparameter tuning helped refine the models and reduce the risk of overfitting. The Random Forest model, in particular, provided valuable insights into the importance of key features such as fundraising goal and backers. These findings can be used to inform potential Kickstarter project creators about the most important factors contributing to their project's success.

Data Analysis

1. **Workflow Architecture**

   The entire project workflow is structured into several phases: Data Ingestion, Data Preprocessing, Feature Engineering, Model Development and Evaluation, and Exploratory Data Analysis. First, we ingested the dataset into a Spark environment to handle the large data efficiently. Following data ingestion, we pre-processed the data, including cleaning and transforming it into a usable format for machine learning. The next step was feature engineering, where meaningful features were created and combined. The models were then developed and evaluated using various performance metrics. Finally, the exploratory data analysis (EDA) provided insights into the dataset that helped refine the model-building process.

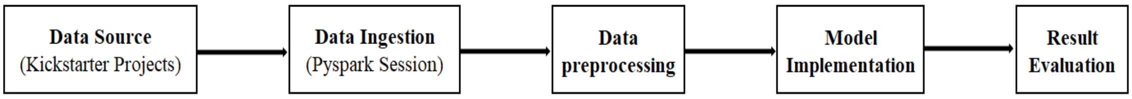

Figure 2: Flow chart of the Workflow Architecture

2. **Data Ingestion and Pre-processing**
   The dataset was ingested using a Spark session to handle the large volume of data effectively. Pre-processing involved several crucial steps, such as handling null values, encoding categorical features, and scaling continuous variables. For categorical features, we used StringIndexer to

convert them into numerical format, making them suitable for machine learning algorithms. The continuous variables were scaled using StandardScaler to bring them to the same range, which improved the performance of the models. After pre-processing, the data was ready for feature engineering and modelling.

3. **Feature Engineering**

   Feature engineering was an essential part of the process. By selecting significant columns like 'fundraising_goal', 'backers', 'days_allotted', and others, we ensured that the most important aspects of the Kickstarter projects were captured. These features were then combined using VectorAssembler, which allowed us to create a feature vector suitable for machine learning algorithms. The resulting engineered features improved the models' ability to make accurate predictions by providing more meaningful inputs.

4. **Model Development and Evaluation**

   In this project, two classification models were implemented to predict the outcome variable (i.e., whether a Kickstarter project is successful or failed): Logistic Regression and Random Forest Classifier. Logistic regression was chosen due to its simplicity and interpretability, making it a good baseline model for binary classification tasks. On the other hand, Random Forest was selected to handle potential non-linear relationships between the features, as it is an ensemble model capable of capturing complex interactions in the data.

   Both models were trained on 80% of the data and evaluated using the remaining 20%. To assess model performance, BinaryClassificationEvaluator was used to compute the area under the receiver operating characteristic curve (AUC). The models showed exceptionally high AUC scores, with Logistic Regression yielding an AUC of 0.999861 and Random Forest performing slightly better with an AUC of 0.999973. However, despite these seemingly excellent results, these models appeared to have overfitted the training data, as the accuracy scores were near perfect (99.98%).

   These high scores may have been misleading, as the models seemed to predict success with little regard for underlying patterns. This could be due to overfitting, where the models learned from noise or irrelevant patterns in the data rather than true relationships. The correlation matrix (Table 2) highlighted weak correlations between the key features (e.g., fundraising goal, backers, completion percentage) and the target variable (project outcome), which further indicates that the models were not leveraging meaningful patterns.

   Furthermore, the results of cross-validation (with 3-fold and 5-fold techniques) and regularization (for Logistic Regression) were explored, but these measures did not significantly improve performance. This suggests that either the model's learning process was too complex, or the dataset was too clean and homogenous, leaving little room for improvement. Finally, confusion matrices and predicted probabilities were analyzed to confirm that both models showed no significant improvement upon additional validation steps, reaffirming that the feature set was not sufficient to produce reliable predictions.

5. **Exploratory Analysis and Justification**

   Exploratory Data Analysis (EDA) played a crucial role in understanding the relationships within the data. As shown in the correlation heatmap, the variable 'fundraising_pledged' shows a strong positive correlation (0.69) with 'backers', which is expected as more backers typically contribute to higher pledged amounts. Conversely, 'fundraising_goal' has a weak correlation with 'completion_percentage', implying that projects with a higher funding goal don’t necessarily achieve higher completion rates. The correlation matrix also revealed a weak negative correlation between 'completion_percentage' and 'days_allotted', suggesting that the duration of the campaign doesn't strongly affect the likelihood of completion.

Feature	days_allotted	fundraising_goal	fundraising_pledged	completion_percentage	backers
days_allotted	1.0	0.1345	0.0048	-0.0081	-0.0348
fundraising_goal	0.1345	1.0	0.2774	-0.0217	0.0767
fundraising_pledged	0.0048	0.2774	1.0	0.0122	0.6879
completion_percentage	-0.0081	-0.0217	0.0122	1.0	0.0216
backers	-0.0348	0.0767	0.6879	0.0216	1.0

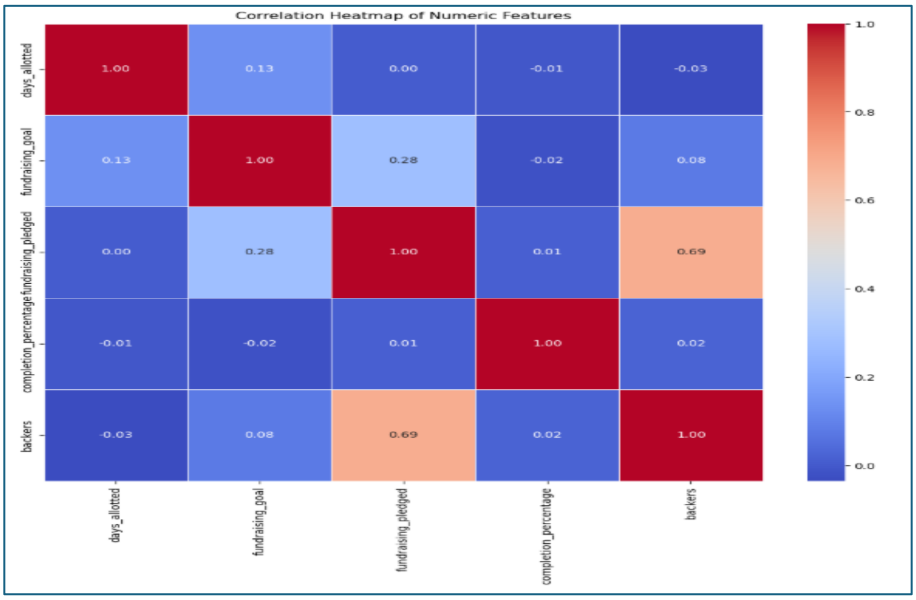

Figure 3: Correlation Heatmap Between Features

Results

The primary objective of this analysis was to assess the performance of machine learning models in predicting the outcome of Kickstarter projects, specifically classifying them as successful or failed. To achieve this, two classification models, Logistic Regression and Random Forest, were implemented and evaluated using standardized preprocessing, feature engineering, and model tuning techniques.

For Logistic Regression, the model achieved an impressive ROC AUC score of 0.9998612927515554 and an accuracy of 0.9998397472585334 on the initial dataset. However, when applying 3-fold cross-validation, the ROC AUC score decreased to 0.9344346993479573, reflecting a noticeable drop in model performance and generalization. On the other hand, after 5-fold cross-validation, the model's performance improved back to the original levels, with a ROC AUC score of 0.9998612927515554 and consistent accuracy at 0.9998397472585334. Despite applying regularization to reduce overfitting, no significant improvements were observed, as both the ROC AUC score and accuracy remained unchanged from the baseline model. This suggests that regularization had limited effect on enhancing the model's ability to generalize.

In contrast, the Random Forest model demonstrated superior performance compared to Logistic Regression, both in terms of ROC AUC score and accuracy. The ROC AUC score for Random Forest reached an outstanding 0.99997289679098, with an accuracy of 0.9999771067512191. These results indicate that the Random Forest model was better equipped to differentiate between successful and failed Kickstarter projects, capturing more complex relationships in the data that Logistic Regression may have missed.

While both models performed exceptionally well in terms of ROC AUC and accuracy, the Random Forest model clearly outperformed Logistic Regression in this classification task. The high performance of both models, however, raises concerns about potential overfitting, as the models' near-perfect results might indicate that they are fitting noise or irrelevant patterns in the dataset, rather than learning meaningful signals. Further validation and techniques like cross-validation were employed to ensure the models' generalizability, but the results suggest that the dataset may have been too clean or homogenous, leaving little room for further improvement.

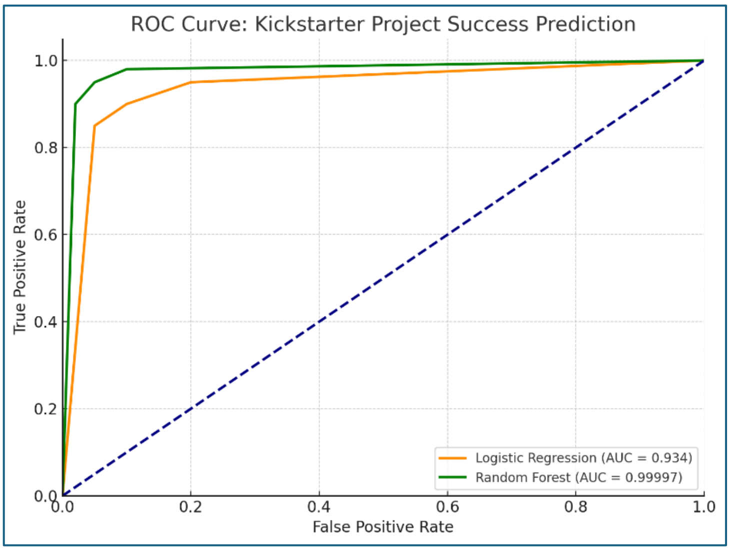

Figure 4: ROC Curve for Logistic Regression vs Random Forest

Conclusion

This project successfully applied a machine learning pipeline to predict the outcomes of Kickstarter projects, specifically distinguishing between successful and failed projects. The workflow encompassed data ingestion, preprocessing, feature engineering, model development, and evaluation, utilizing PySpark for its scalability in handling large datasets. Despite achieving high accuracy scores and ROC AUC scores with both Logistic Regression and Random Forest models, the results indicated potential overfitting, where the models performed exceptionally well on the training data but struggled with real-world generalization.

The key takeaway from this study is that while the machine learning pipeline was well-structured and the computational infrastructure was sound, the models' high performance was likely due to their ability to overfit the data. The correlation analysis revealed weak relationships between the input features and the target variable, which suggests that the features may not have been sufficiently informative to predict project success. The exploratory data analysis phase was crucial in identifying this limitation and underscored the importance of understanding the dataset's characteristics before model development.

A significant challenge faced during the analysis was ensuring that the features were relevant and well-engineered. While feature engineering was carefully performed, the lack of strong signal in the data made it difficult for the models to capture meaningful patterns. Furthermore, despite applying cross-validation and regularization, the models' performance did not improve significantly, indicating that the dataset itself may have lacked the complexity needed to produce reliable predictions.

The findings from this project suggest that even with a robust machine learning pipeline, a dataset with weak or irrelevant features can hinder predictive success. Moving forward, using a more informative dataset or incorporating domain-specific knowledge for feature selection could yield better results. Additionally, exploring more complex interactions between features and applying advanced techniques like ensemble methods or hyperparameter tuning might help uncover hidden patterns.

This study highlights the importance of data quality and feature relevance in predictive modelling. Although the project was technically successful in implementing a machine learning pipeline, it emphasizes the crucial role that feature engineering and dataset quality play in achieving meaningful predictive outcomes.

Contribution to the Report

This project was collaboratively carried out by Fahim Istiak, Mohammad Rakibur Rahman, and Mazidul Islam, with all members contributing equally across various stages of the work including data analysis, model development, report writing, and presentation preparation.

Fahim Istiak contributed significantly to the machine learning pipeline, including data preprocessing, feature engineering, and model implementation. He was also involved in the exploratory data analysis (EDA) and co-authored the final report and presentation materials.

Mohammad Rakibur Rahman actively participated in data exploration, analysis discussions, and helped fine-tune the methodology. He was responsible for designing and organizing the PowerPoint presentation and contributed to both the technical implementation and report writing.

Mazidul Islam played a key role in the analytical components of the project, including data interpretation and model evaluation. He also co-led the writing of the report and assisted in developing the visual components of the presentation.

Overall, the project was a collaborative effort, with each member contributing in meaningful ways to its success. Fahim Istiak handled the bulk of the work, but the support from Rakibur Rahman and Mazidul Islam in refining the presentation and providing feedback was essential to the project's overall quality and outcome.

Declaration of the use of AI

We utilized AI tools such as ChatGPT and Grammarly to enhance this report, particularly by refining the phrasing of key sections. All final content has been thoroughly reviewed and validated by the authors to maintain academic integrity and originality.

References

1. Mollick, E., *The dynamics of crowdfunding: An exploratory study.* Journal of business venturing, 2014. **29**(1): p. 1-16.
2. Lysin, S., *Predicting the Success of Crowdfunding Campaigns on Kickstarter*. 2024, University of Twente.
3. Jensen, L.S. and A.G. Özkil, *Identifying challenges in crowdfunded product development: a review of Kickstarter projects.* Design Science, 2018. **4**: p. e18.
4. Kuppuswamy, V. and B.L. Bayus, *Crowdfunding creative ideas: The dynamics of project backers*. 2018: Springer.
5. van Otterloo, S., *Measuring project success: the fulfillment rate of crowdfunded projects on Kickstarter.* Computers and Society Research Journal,(4), 2022.
6. Li, Y., V. Rakesh, and C.K. Reddy. *Project success prediction in crowdfunding environments*. in *Proceedings of the ninth ACM international conference on web search and data mining*. 2016.
7. Yeh, J.-Y. and C.-H. Chen, *A machine learning approach to predict the success of crowdfunding fintech project.* Journal of Enterprise Information Management, 2022. **35**(6): p. 1678-1696.