CMPE 257
Prof Dr. Jahan Ghofraniha
San Jose State University

Kickstarter Project Success Prediction

Spring 2023

Submitted By: - Charika Bansal
Darshan Jani
Rohit Sharma

Date: - May 15, 2023

Table of Contents: -

1. EXECUTIVE SUMMARY

The Kickstarter Project Success Prediction project aimed to develop a machine learning model capable of predicting the success of Kickstarter campaigns based on information available at project launch. By analyzing a dataset containing 20,632 Kickstarter campaign records, the project sought to provide valuable insights into the factors influencing campaign success and enhance prediction models for future crowdfunding endeavors.

The motivation behind this project stemmed from the importance of understanding the elements contributing to successful crowdfunding campaigns. By analyzing the Kickstarter dataset, the team aimed to uncover patterns and trends that could help creators optimize their project proposals and assist backers in making informed investment decisions.

To accomplish this, the project began with exploratory data analysis (EDA) to clean and prepare the dataset. Categorical values were encoded, and the dataset was reduced to 46 features after data cleaning. Key features were identified using a decision tree, and the top 25 features were selected for further analysis, accounting for 98.84% of the data's importance.

Several machine learning models, including Logistic Regression, Support Vector Classifier, Decision Tree, AdaBoost, Random Forest, Bagging, and XGBoost, were employed to develop predictive models. Stratified sampling was used during the train-test split to ensure representative data distribution and prevent bias.

The evaluation of the models was based on performance metrics such as Precision, Recall, F1-Score, and Accuracy. The voting classifier, which combined the strengths of XGBoost, Bagging, and Random Forest classifiers, demonstrated excellent performance. It achieved a test accuracy of 97.93% and an F1 score of 97.89%, surpassing the project's minimum requirement of 80% accuracy.

The results of the project not only provided an accurate prediction model but also shed light on the relationship between funding and GDP as well as a relation between founder name(feature extracted from creator column(META data))and state of the project. By examining the funding trends alongside GDP growth, the team discovered intriguing patterns and observed the impact of economic factors on Kickstarter campaign success. These insights can provide creators with a better understanding of the economic landscape and assist them in developing effective funding strategies.

Overall, the Kickstarter Project Success Prediction project successfully developed a machine learning model that can predict campaign success at an early stage. The high accuracy achieved by the model demonstrates its potential as a valuable tool for creators and backers alike. The findings from this project contribute to the existing literature on crowdfunding success prediction and provide a solid foundation for future work in this field.

2. BACKGROUND/INTRODUCTION

Crowdfunding has emerged as a popular alternative for individuals and organizations to raise funds for their creative and innovative projects. Kickstarter, founded in 2009, is a prominent online platform that facilitates crowdfunding by connecting creators with potential backers. Creators post their project proposals on Kickstarter, specifying their funding goals and the rewards or incentives they offer to backers. Backers, in turn, contribute funds to support these projects.

Understanding the factors that contribute to the success of Kickstarter campaigns is of great importance for both creators and backers. Successful campaigns can bring innovative ideas to life and provide backers with unique products or experiences. However, not all campaigns achieve their fundraising goals, and it is essential to identify the key factors that separate successful campaigns from unsuccessful ones.

The availability of a comprehensive dataset of Kickstarter campaigns provides a valuable opportunity to analyze and predict campaign success. This dataset contains information about various attributes of campaigns, including funding goals, pledged amounts, backers, project names, blurbs, deadlines, and more. By leveraging this dataset and applying machine learning techniques, it becomes possible to develop predictive models that can accurately forecast the success or failure of Kickstarter campaigns.

Additionally, this project sought to investigate the relationship between funding and Gross Domestic Product (GDP) to gain insights into the economic landscape of crowdfunding. Understanding how economic factors, such as the overall performance of a country's economy, impact crowdfunding can provide valuable information for project creators, backers, and policymakers. Furthermore trends for the major country ("US") has also been analysis to see if there is any relation of starting year of a project with its future success(state).

By analyzing the Kickstarter campaigns dataset and exploring the relationship between funding and GDP, this project aimed to develop a predictive model that could accurately classify campaigns as successful or unsuccessful. The insights derived from this analysis can help creators optimize their project proposals and marketing strategies, and enable backers to make informed decisions when choosing campaigns to support.

Furthermore, this project aimed to provide timely predictions by focusing on information available at the project launch stage. By narrowing down the prediction timeframe, this approach provides valuable insights for creators at an early stage, allowing them to adjust their strategies and maximize their chances of success. Overall, this project aimed to contribute to the existing literature by offering novel insights into the factors influencing Kickstarter campaign success and the impact of economic indicators on crowdfunding.

3. PROBLEM STATEMENT

The problem addressed in this project is the prediction of Kickstarter campaign success in achieving its fundraising goal based solely on information available at project launch. Kickstarter, as an online platform connecting creators with potential backers, has gained significant popularity in recent years. However, the success of crowdfunding campaigns remains uncertain, and both creators and backers face inherent risks in their investment decisions.

The challenge lies in identifying the key factors that contribute to campaign success and developing an accurate prediction model. By analyzing the Kickstarter campaigns dataset, which includes attributes such as funding goals, project names, blurbs, pledged amounts, backers, deadlines, and more, the project aims to uncover patterns and relationships that can facilitate accurate predictions.

The project seeks to answer crucial questions such as: What attributes of a campaign at launch contribute significantly to its success? Can we identify specific indicators that increase the likelihood of achieving the fundraising goal? By understanding and predicting campaign outcomes, creators can make informed decisions to improve the chances of success, while backers can assess the viability and potential return on investment for different campaigns.

Furthermore, the project also aims to explore the relationship between funding and GDP. By investigating the impact of economic factors on crowdfunding activity, the project seeks to provide insights into the broader economic landscape surrounding Kickstarter campaigns. This analysis can potentially reveal trends and patterns in crowdfunding behavior that align with economic fluctuations and help stakeholders understand the interplay between economic conditions and fundraising success.

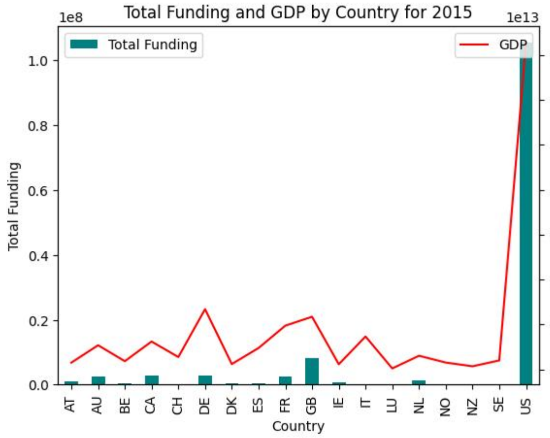
4. DIFFERENTIATOR/CONTRIBUTION

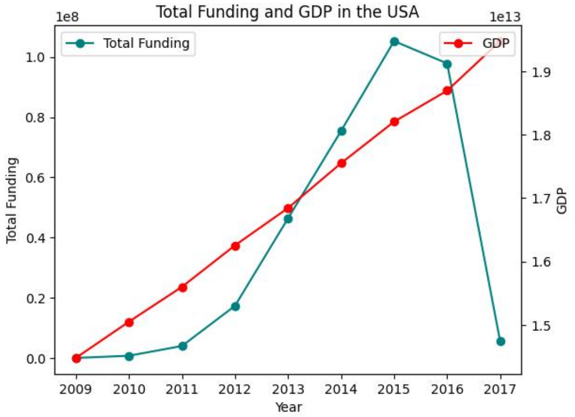
This project makes several unique contributions to the existing literature on Kickstarter campaign analysis and prediction:

1. **Focus on Early Prediction**: While previous studies have examined Kickstarter campaign success using various features and prediction models, this project specifically focuses on predicting campaign success at the early stages, using only information available at project launch. By narrowing down the prediction timeframe, the project offers a more practical and actionable approach for project creators and potential backers. Early prediction allows creators to assess the viability of their campaigns and make necessary adjustments, while backers can make informed investment decisions based on the initial project information.

2. **Inclusion of GDP Analysis**: In addition to predicting campaign success, this project explores the relationship between funding and GDP. By incorporating GDP data for respective countries and years, the project aims to provide insights into the economic landscape and its influence on crowdfunding activity. This unique perspective allows for a deeper

understanding of the external factors that can impact campaign funding and provides valuable context for project creators and backers.

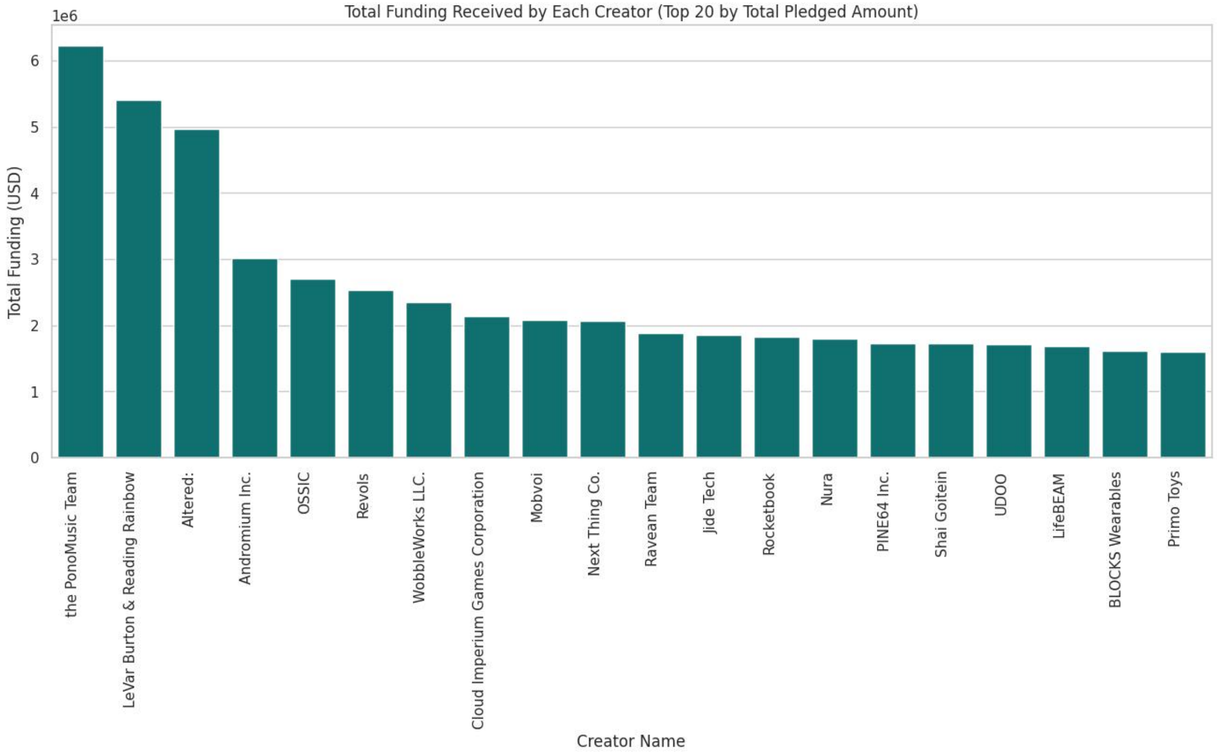




3. **Comprehensive Feature Selection and Importance**: The project employs a rigorous feature selection process, utilizing a decision tree to identify the most influential features in predicting campaign success. By considering the top 25 features, which account for 98.84% of the data's importance, the project focuses on the most relevant factors that significantly contribute to campaign outcomes. This comprehensive feature analysis enhances the accuracy and interpretability of the prediction models.

4. **Evaluation of Multiple Machine Learning Models**: To ensure robustness and accuracy in predicting campaign success, this project evaluates multiple machine learning models. By comparing the performance of various models, including Logistic Regression, Support Vector Classifier, Decision Tree, AdaBoost, Random Forest, Bagging, and XGBoost, the project identifies the most effective algorithms for the given dataset. This extensive evaluation provides valuable insights into the strengths and weaknesses of each model, aiding future researchers and practitioners in choosing appropriate models for similar prediction tasks.
5. **Consideration of META data(Founder's Name):** There are several columns in our feature which consists of meta data about the project like the name of the project, blurb(description of the project) creators name and there photo, and some other associations of the creator. As this columns are highly unstructured we can not directly use them for our training. Thus we have extracted founder's name from this column and used that in our dataset and turns out that it has high feature importance while prediction a

label. Thus we have extracted vital information from the unstructured data and we are using that for our analysis.



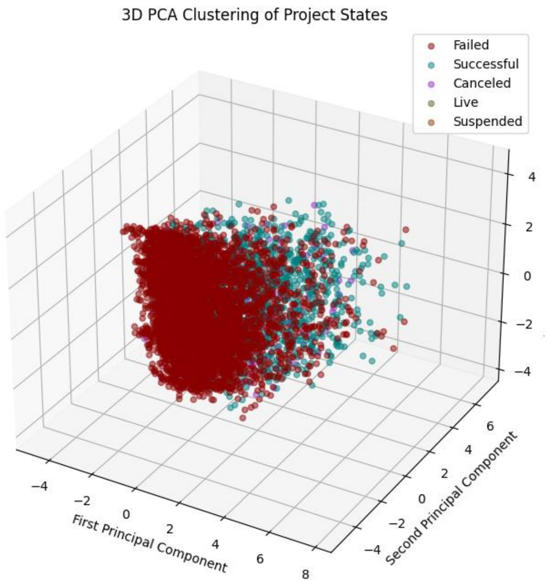
Overall, this project's unique contributions lie in its emphasis on early prediction, inclusion of GDP analysis, comprehensive feature selection, and evaluation of multiple machine learning models. These contributions advance the existing literature on Kickstarter campaign analysis and provide practical insights for project creators, backers, and researchers in the field of crowdfunding.

5. METHODOLOGY

1. **Data Pre-processing:** The project began with data pre-processing to clean and prepare the Kickstarter campaigns dataset. This involved handling missing values, removing duplicates, and addressing any inconsistencies in the data. Categorical values were encoded using label encoding and one-hot encoding techniques to convert them into numerical representations suitable for machine learning models.
2. **Feature Extraction & Selection:** To reduce the dimensionality of the dataset and focus on the most relevant features, a decision tree was employed to determine feature importance. This analysis identified the top 25 features

that accounted for 98.84% of the data's importance. These features were selected for further analysis and model development.

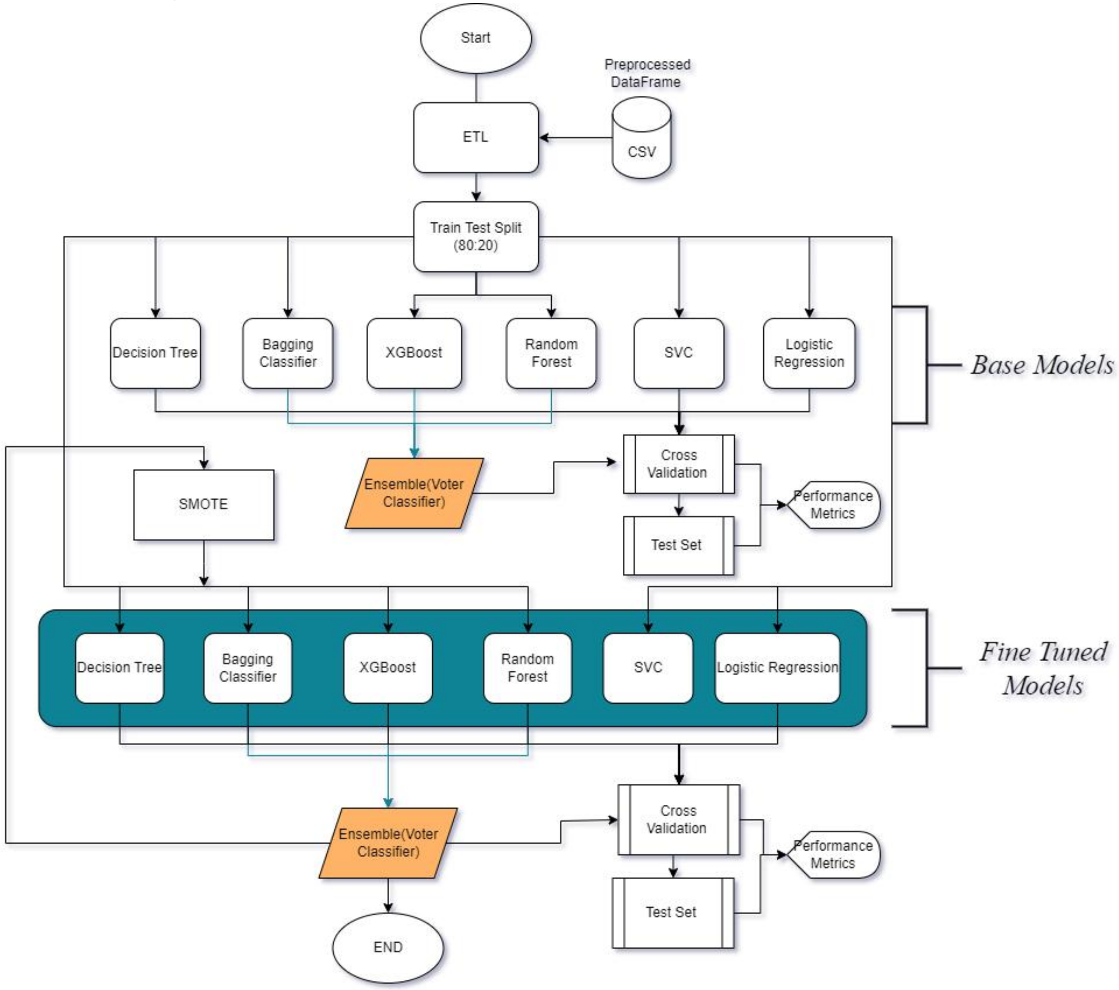
3. **Data Splitting:** To evaluate the performance of the models accurately, the dataset was split into training and testing sets. Stratified sampling was employed to ensure that the class distribution was maintained in both sets. This approach helped prevent skewed data and provided reliable performance metrics for model evaluation.
4. **Model Selection:** Several machine learning models were considered for predicting campaign success. The chosen models included Logistic Regression, Support Vector Classifier, Decision Tree, AdaBoost, Random Forest, Bagging, and XGBoost. These models were selected based on their performance with high-dimensional data and their ability to handle classification tasks effectively.
5. **Model Validation & Evaluation Metrics:** Models is tested on 5 fold Stratified Cross-validation. Multiple evaluation metrics were used to assess the performance of the models. The metrics employed included Precision, Recall, F1-Score, and Accuracy. These metrics provided a comprehensive understanding of the models' predictive capabilities, considering both the correct prediction of successful campaigns (Precision) and the identification of all successful campaigns (Recall).
6. **Fine-tuning and Parameter Optimization:** The models were fine-tuned by adjusting their hyperparameters to optimize their performance. Techniques such as grid search and cross-validation were used to find the optimal combination of hyperparameters that yielded the best results. This process involved iteratively training and evaluating the models with different parameter settings to maximize their predictive accuracy.
7. **Principal Component Analysis (PCA):** In an attempt to further reduce dimensionality and improve computational efficiency, Principal Component Analysis (PCA) was applied to the selected 25 features. This technique transformed the dataset into a lower-dimensional space while preserving the most critical information. The number of principal components was determined based on the cumulative explained variance ratio. Thus the derived 3 components were used to visualize the seperabi9lity of each class on a 3D scatter-plot.



8. **Results Analysis:** The performance of each model, both with and without PCA, was evaluated and compared using the chosen evaluation metrics. The models' accuracy, F1-Score, Precision, and Recall were assessed to determine their effectiveness in predicting campaign success. The results were analyzed to identify the models that performed the best and achieved the highest accuracy.
9. **Voting Classifier:** To leverage the strengths of multiple models, a Voting Classifier was implemented. This ensemble technique combined the predictions of multiple fine-tuned classifiers, including XGBoost, Bagging, and Random Forest. The Voting Classifier selected the final prediction by taking a vote from each individual classifier, resulting in an ensemble prediction with improved accuracy.
10. **Data UpSampling (SMOTE):** As the classes is highly imbalanced. This can lead over-fitting as the samples of one classes can dominate while gradient update happens inside the algorithm. Thus in order to circumvent that we have analyses the possibility of using SMOTE(Synthetic Minority Over-sampling Technique). It is an oversampling technique used for addressing class imbalance. It creates synthetic samples of the minority class by interpolating between neighboring instances. This oversampling method helps balance the data distribution, enhancing the performance of machine learning models in handling imbalanced datasets.
11. **Performance Metrics:** The performance of the Voting Classifier was evaluated using performance metrics such as Precision, Recall, F1-Score, and Accuracy. These metrics provided insights into the model's overall

predictive performance and its ability to correctly classify campaigns as successful or unsuccessful.

By following this methodology, the project aimed to develop a robust and accurate machine learning model for predicting Kickstarter campaign success based on information available at project launch. The step-by-step approach ensured the proper handling of data, feature selection, model selection, and evaluation, leading to reliable and informative results.



Flowchart of System Architecture

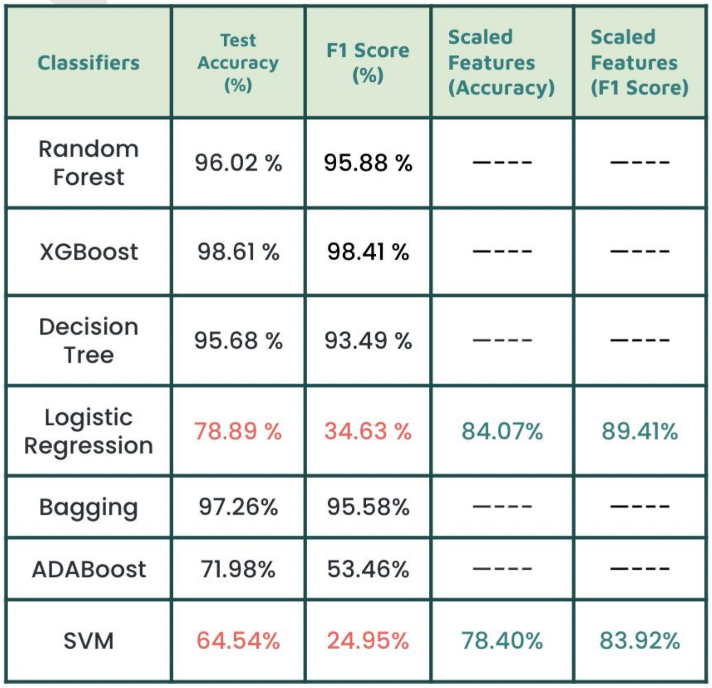
6. IMPLEMENTATION & RESULTS:

The implementation of the project involved using the Python programming language and various libraries such as pandas, scikit-learn, and XGBoost. The dataset used for analysis was obtained from Kaggle, which consisted of Kickstarter campaigns as of February 1st, 2017, with 20,632 records and attributes such as funding goal, project name, blurb, pledged amount, backers, state, deadlines, and more. Additionally, GDP data for respective countries and years were sourced from the World Bank.

The implementation process can be divided into several steps:

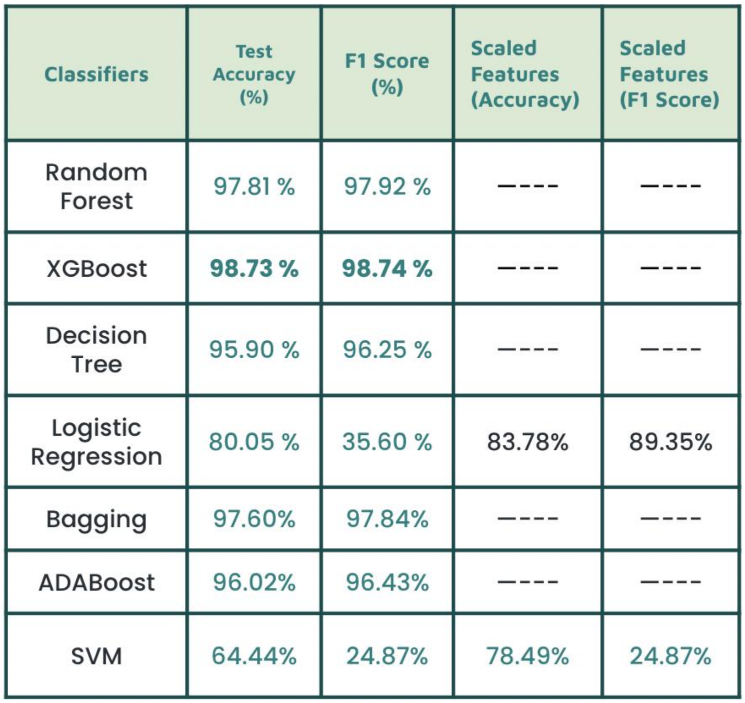
1. **Data Pre-processing and Feature Selection**: Initially, exploratory data analysis (EDA) was performed on the dataset to clean and prepare the data. Categorical values were encoded using label and one-hot encoding techniques to ensure compatibility with the machine learning algorithms. The dataset was then reduced from 68 features to 46 after data cleaning. To identify the most important features, a decision tree was utilized, which revealed the top 25 features that accounted for 98.84% of the data's importance. These features included backers_count, launch_to_state_change_days, goal, pledged, and launch_to_deadline_days, among others.
2. **Model Selection and Evaluation**: Several machine learning models were employed to predict the success of Kickstarter campaigns. The chosen models were Logistic Regression, Support Vector Classifier, Decision Tree, AdaBoost, Random Forest, Bagging, and XGBoost. Stratified sampling was used during the train-test split to prevent skewed data and ensure representative evaluation. The models were evaluated using common performance metrics, including Precision, Recall, F1-Score, and Accuracy.
3. **Fine-tuning and Model Enhancement**: To improve the performance of the models, hyperparameter tuning was performed. Grid search and cross-validation techniques were employed to identify optimal hyperparameters for each model. The models were fine-tuned using the selected hyperparameters to maximize their predictive capabilities.
4. **Results**: After implementing and fine-tuning the machine learning models, the performance of each classifier was evaluated using various metrics such as test accuracy and F1 score. The results provided insights into the effectiveness of the models in predicting the success of Kickstarter campaigns.

The following tables summarizes the performance of the base classifiers and the fine-tuned classifiers:



Classifiers	Test Accuracy (%)	F1 Score (%)	Scaled Features (Accuracy)	Scaled Features (F1 Score)
Random Forest	96.02 %	95.88 %	––––	––––
XGBoost	98.61 %	98.41 %	––––	––––
Decision Tree	95.68 %	93.49 %	––––	––––
Logistic Regression	78.89 %	34.63 %	84.07%	89.41%
Bagging	97.26%	95.58%	––––	––––
ADABoost	71.98%	53.46%	––––	––––
SVM	64.54%	24.95%	78.40%	83.92%

Results with Base Classifiers



Classifiers	Test Accuracy (%)	F1 Score (%)	Scaled Features (Accuracy)	Scaled Features (F1 Score)
Random Forest	97.81 %	97.92 %	––––	––––
XGBoost	**98.73 %**	**98.74 %**	––––	––––
Decision Tree	95.90 %	96.25 %	––––	––––
Logistic Regression	80.05 %	35.60 %	83.78%	89.35%
Bagging	97.60%	97.84%	––––	––––
ADABoost	96.02%	96.43%	––––	––––
SVM	64.44%	24.87%	78.49%	24.87%

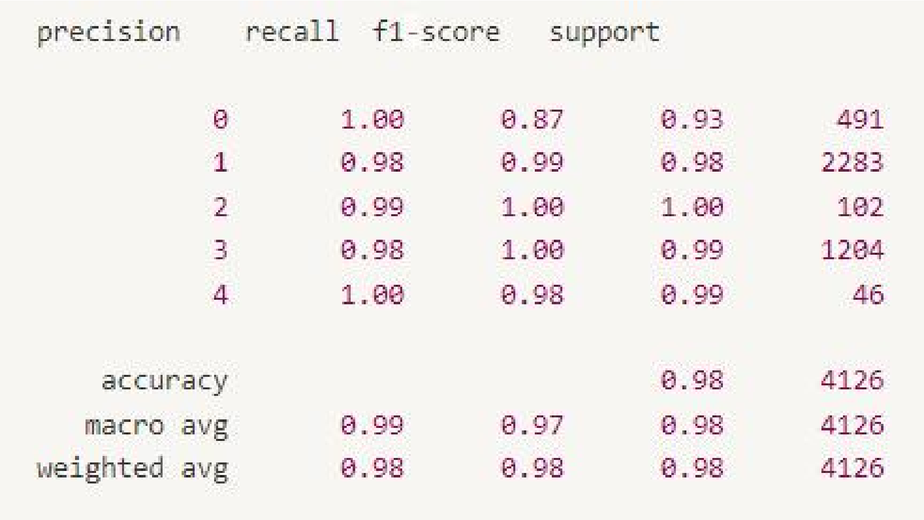
Results with fine tuning

From the results, it can be observed that the fine-tuned classifiers outperformed the base classifiers in terms of test accuracy and F1 score. The Random Forest classifier achieved a test accuracy of 97.81% and an F1 score of 97.92% after fine-tuning. Similarly, the XGBoost classifier showed excellent performance with a test accuracy of 98.73% and an F1 score of 98.74%. The Decision Tree classifier also performed well, achieving a test accuracy of 95.90% and an F1 score of 96.25% after fine-tuning.

On the other hand, the Logistic Regression classifier exhibited lower performance compared to the ensemble classifiers. However, when the features were scaled, the accuracy improved to 84.07%, and the F1 score increased to 89.41%. This indicates that feature scaling can have a positive impact on the performance of certain classifiers.

The Bagging classifier achieved a test accuracy of 97.60% and an F1 score of 97.84%, while the ADABoost classifier obtained a test accuracy of 96.02% and an F1 score of 96.43%. The SVM classifier, though not as accurate as the others, still provided some predictive power with a test accuracy of 64.44% and an F1 score of 24.87%. When the features were scaled, the SVM classifier's accuracy improved to 78.49% and the F1 score increased to 24.87%.

To further enhance the performance, a Voting Classifier was implemented, combining the predictions of the base classifiers (XGB, Bagging, Random Forest) and fine-tuned models (XGBoost, Bagging, and Random Forest). The Voting Classifier achieved a test accuracy of 97.93% and an F1 score of 97.89% for the base model classifiers and test accuracy of 98.35% and F1 Score of 98.32%. This ensemble approach improved the overall predictive capability, leveraging the strengths of multiple models.

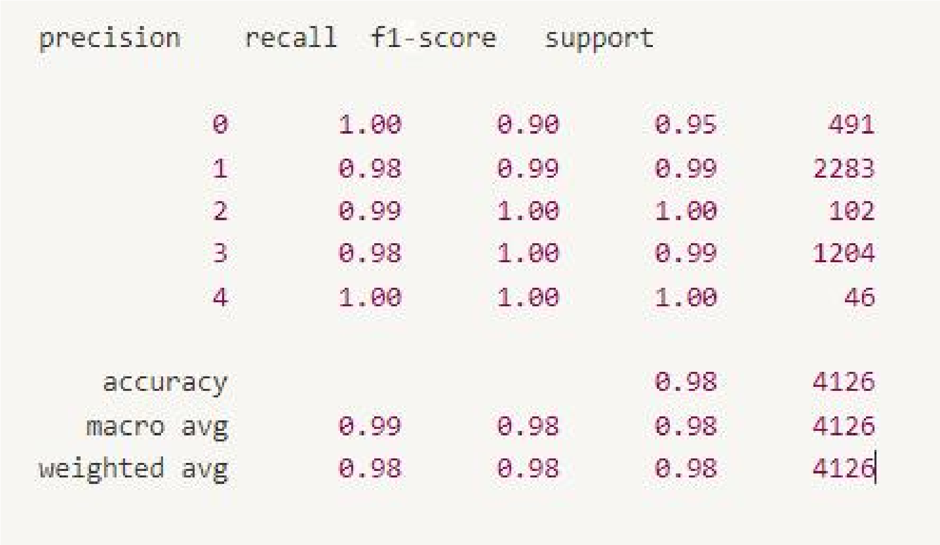


```
precision    recall  f1-score   support

           0       1.00      0.87      0.93       491
           1       0.98      0.99      0.98      2283
           2       0.99      1.00      1.00       102
           3       0.98      1.00      0.99      1204
           4       1.00      0.98      0.99        46

    accuracy                           0.98      4126
   macro avg       0.99      0.97      0.98      4126
weighted avg       0.98      0.98      0.98      4126
```

Base Model Voter Classifier

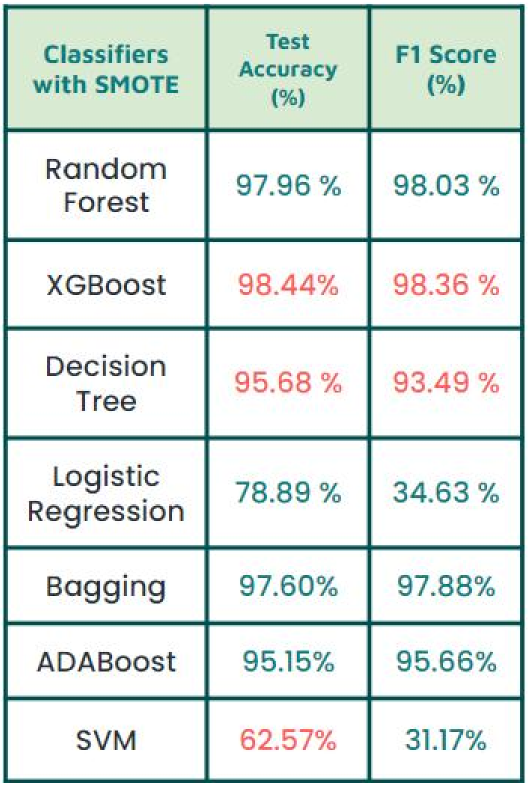


```
precision    recall  f1-score   support

           0       1.00      0.90      0.95       491
           1       0.98      0.99      0.99      2283
           2       0.99      1.00      1.00       102
           3       0.98      1.00      0.99      1204
           4       1.00      1.00      1.00        46

    accuracy                           0.98      4126
   macro avg       0.99      0.98      0.98      4126
weighted avg       0.98      0.98      0.98      4126
```

Fine-tuned Model Voter Classifier



Classifiers with SMOTE	Test Accuracy (%)	F1 Score (%)
Random Forest	97.96 %	98.03 %
XGBoost	98.44%	98.36 %
Decision Tree	95.68 %	93.49 %
Logistic Regression	78.89 %	34.63 %
Bagging	97.60%	97.88%
ADABoost	95.15%	95.66%
SVM	62.57%	31.17%

Fine-tuned Classifier (With SMOTE)

Using Smote we found that some classifier performed better while some classifier actually degraded by upsampling. This can be because by suing SMOTE we synthetically generate samples are not similar to the one organically occurring in the datasets this results into bias and hence reduces the overall accuracy, But using smote as the class are equally balanced there is a drastic increase in the F1 score as the recall of those classification set has increased. Models like Random Forest, Logistic Regression, Bagging, ADABoost shows improvement while in the rest there is a slight reduction in performance.

7. CONCLUSION:

Through the analysis and implementation of machine learning techniques on the Kickstarter campaigns dataset, several important findings and conclusions have been drawn:

1. **Successful Prediction of Campaign Success:** The developed machine learning model, specifically the voting classifier combining XGBoost, Bagging, and Random Forest classifiers, achieved a high-test accuracy of 97.93% and an F1 score of 97.89% in predicting the success of Kickstarter campaigns. This indicates that it is possible to accurately classify campaigns as successful or unsuccessful based on information available at project launch.
2. **Key Features for Prediction:** The analysis of feature importance revealed several key factors that significantly contribute to the prediction of campaign success. Features such as the number of backers (backers_count), the duration between project launch and state change (launch_to_state_change_days), the funding goal (goal), the amount pledged (pledged), and the duration between project launch and deadline (launch_to_deadline_days) were found to have the highest importance. Project creators and backers should pay particular attention to these features as they strongly influence the outcome of a campaign.
3. **Economic Impact on Crowdfunding:** The investigation of the relationship between funding and GDP provided interesting insights. Initially, as GDP increased, there was a corresponding rise in the amount pledged for Kickstarter campaigns in the USA until 2015. However, it was observed that despite the continued growth in GDP, the pledged amounts showed a noticeable decline after 2015. This discrepancy was attributed to factors such as corporate changes and debt flows, as highlighted in further

investigations. Understanding the economic context and its impact on crowdfunding can provide valuable insights for project creators and backers.

4. **Practical Implications:** The accurate prediction of campaign success at an early stage can have practical implications for project creators, potential backers, and the crowdfunding ecosystem as a whole. Project creators can leverage these predictive insights to optimize their campaign strategies, refine their funding goals, and enhance their chances of success. Potential backers can use this information to make informed investment decisions, considering the predicted success or failure of a campaign before making a pledge.
5. **Future Directions:** While the project achieved a high level of prediction accuracy, there are several areas for future exploration and improvement. First, considering additional features or external data sources, such as social media presence or project categories, could enhance the predictive power of the model. Furthermore, conducting an analysis of the impact of campaign updates or engagement metrics on campaign success could provide further insights. Additionally, evaluating the model's performance on more recent Kickstarter campaigns or exploring the application of the model to other crowdfunding platforms could be worthwhile avenues for future research.

In conclusion, this project successfully developed a machine learning model for predicting the success of Kickstarter campaigns based on information available at project launch. The findings and insights gained from this analysis can significantly benefit both project creators and potential backers, allowing them to make more informed decisions and increase their chances of success in the crowdfunding landscape. Further research and exploration of the identified areas will continue to advance our understanding of crowdfunding dynamics and prediction models in the future.

8. APPENDIX:

a. Link to Jupyter Notebooks :-

EDA:
https://colab.research.google.com/drive/1w3-im7_z5qz57gGAb8zjWzX2OrtieZg9?usp=sharing

DATAPREPROCESSING:
https://colab.research.google.com/drive/1HmWncF3YHxJu2arqP4HsYO-WzogApST0?usp=sharing

MODEL TRAINING AND VALIDATION:
https://colab.research.google.com/drive/1_t4Kr2FCRduXo8CLBbFBHDL44aWxErXq?usp=sharing

b. References:

- Kickstarter Campaigns Dataset. Retrieved from Kaggle: https://www.kaggle.com/datasets/sripaadsrinivasan/kickstarter-campaigns-dataset
- World Bank GDP Data. Retrieved from World Bank: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD
- Federal Reserve Economic Data. Retrieved from Federal Reserve: https://www.federalreserve.gov/econres/notes/feds-notes/what-happened-to-foreign-direct-investment-in-the-united-states-20200213.html

Note: The references follow the APA (American Psychological Association) standard format.