Enhancing Crowdfunding Prediction Success Using Combinational Approach of Classification and NLP Techniques

MSc Research Project
Data Analytics

Umesh Patil
Student ID: X22216481

School of Computing
National College of Ireland

Supervisor: Abubakr Siddig

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Mr. Umesh E. Patil

Student ID: X22216481

Programme: Msc. Data Analytics **Year:** 2023-2024

Module: Msc Research Project

Supervisor: Professor Abubakr Siddig

Submission Due Date: 12/08/2024

Project Title: Enhancing Crowdfunding Success Prediction Using Combinational Approach of Classification and NLP Techniques

Word Count: 6567 **Page Count** 20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Umesh E. Patil

Date: 12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

Msc Research Project

Project Report/ Thesis

Your Name/Student Number	Course	Date
Umesh Patil/x22216481	MSc Data Analytics	12/08/2024

AI Acknowledgment

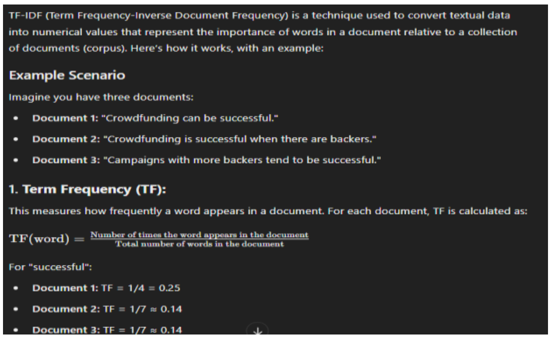
Tool Name	Brief Description	Link to tool
Grammarly	Grammar and Punctuation Checks	https://www.grammarly.com
Chat GPT	To understand How exactly TF-IDF works	https://chat.openai.com/

Description of AI Usage

[Insert Tool Name]	
Chat GPT is used to understand the Working of TF-IDF with example	
Explain with example how tf idf works	TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used to convert textual data into numerical values that represent the importance of words in a document relative to a collection of documents (corpus). Here's how it works, with an example: 1 Example Scenario Imagine you have three documents: • Document 1: "Crowdfunding can be successful."

Grammarly Go	
Grammarly was used throughout the writing process for checking grammar, punctuation, and readability of the text.	
NA	NA

Additional Evidence:



TF-IDF (Term Frequency-Inverse Document Frequency) is a technique used to convert textual data into numerical values that represent the importance of words in a document relative to a collection of documents (corpus). Here's how it works, with an example:

Example Scenario

Imagine you have three documents:

- **Document 1:** "Crowdfunding can be successful."
- **Document 2:** "Crowdfunding is successful when there are backers."
- **Document 3:** "Campaigns with more backers tend to be successful."

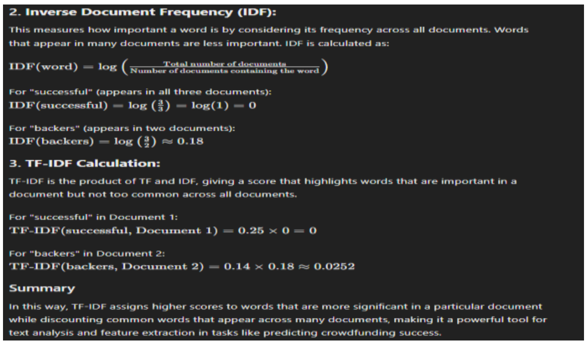
1. Term Frequency (TF):

This measures how frequently a word appears in a document. For each document, TF is calculated as:

$$\mathrm{TF(word)} = \frac{\text{Number of times the word appears in the document}}{\text{Total number of words in the document}}$$

For "successful":

- **Document 1:** TF = 1/4 = 0.25
- **Document 2:** TF = 1/7 ≈ 0.14
- **Document 3:** TF = 1/7 ≈ 0.14



2. Inverse Document Frequency (IDF):

This measures how important a word is by considering its frequency across all documents. Words that appear in many documents are less important. IDF is calculated as:

$$\mathrm{IDF(word)} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents containing the word}}\right)$$

For "successful" (appears in all three documents):
$\mathrm{IDF(successful)} = \log\left(\frac{3}{3}\right) = \log(1) = 0$

For "backers" (appears in two documents):
$\mathrm{IDF(backers)} = \log\left(\frac{3}{2}\right) \approx 0.18$

3. TF-IDF Calculation:

TF-IDF is the product of TF and IDF, giving a score that highlights words that are important in a document but not too common across all documents.

For "successful" in Document 1:
$\text{TF-IDF(successful, Document 1)} = 0.25 \times 0 = 0$

For "backers" in Document 2:
$\text{TF-IDF(backers, Document 2)} = 0.14 \times 0.18 \approx 0.0252$

Summary

In this way, TF-IDF assigns higher scores to words that are more significant in a particular document while discounting common words that appear across many documents, making it a powerful tool for text analysis and feature extraction in tasks like predicting crowdfunding success.

Enhancing Crowdfunding Success Prediction Using Combinational Approach of Classification and NLP Techniques

Umesh Patil
X22216481

Abstract

This study consists of investigation of crowdfunding campaign success by integrating classification models with Natural Language Processing (NLP) techniques, specifically Term Frequency-Inverse Document Frequency (TF-IDF). Machine learning models are evaluated on both balanced and imbalanced datasets, comprehensive assessment is ensured using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results shows that ensemble methods like AdaBoost and Random Forest outperformed which showcased in results with high accuracy and robust classification capabilities. Logistic Regression and SVM also perform effectively on both datasets but particularly in recall, whereas K-Nearest Neighbors (KNN) shows variability in precision and accuracy. This combinational approach significantly enhances predictive accuracy and offers valuable insights for optimizing crowdfunding campaigns. These findings are beneficial to entrepreneurs, investors, and crowdfunding platforms by enhancing the effectiveness and efficient funding process. The research underscores the importance of integrating numerical and textual data for more accurate predictions, contributing to the advancement of predictive modeling in crowdfunding.

2 Introduction

2.1 Background

Crowdfunding is a way to raise funds for fresh or existing ventures through a large number of individuals, friends, and families primarily via social media and crowdfunding platforms e.g. Kickstarter, and Indiegogo. Crowdfunding revolutionized the process for entrepreneurs and innovators to secure the funds for their innovative ideas by authorizing them to appeal directly to a broad audience of potential investors (BaniMustafa, A. et.al). It helps entrepreneurs to get funding at the early stages of their projects, most of them are struggling to get the capital because of a lack of experience and trust (Zhou, M. et.al). Popular websites like Kickstarter and Indiegogo attract big crowds to invest their money into startups and in return get rewards or equity. This not only helps innovators to get funding but also gives them the opportunity to market their products. This process happens online without any interference from middlemen which differentiates it even more from other traditional methods

The financial crisis of 2008 played an important role in the rise of crowdfunding as well, because of the economic challenges and strict regulations in getting loans from the bank and investments from the investor got quite difficult. It led entrepreneurs to seek an alternative way, and crowdfunding rose as a vital solution. As crowdfunding becomes more and more popular many researchers have investigated various methods to understand its underlying mechanisms.

Mollick (2014) investigated that project quality and geography were the most important factors in determining the success of crowdfunding, Whereas Mitra and Gilbert (2014) focused on analyzing the language that is used for crowdfunding, after analyzing 45000 projects they concluded that phrases follow certain principles such as Cooperation, Scarcity, Social Identity increases the chances of success. In summary, predicting crowdfunding success is important to all three entities that are Entrepreneurs, Investors, and crowdfunding platforms. Crowdfunding is a bridge between entrepreneurial goals and financial support, The future for it is still promising and holds great potential for ambitious ventures throughout the world as researchers keep digging into its complexities.

2.2 Research Objectives

The motivation behind this study is the need to amplify the capacity to predict and enhance the effectiveness of the success rate for crowdfunding campaigns. The model till date have been focusing on individual assessment approach i.e. either take numerical data or textual data into consideration while forming predictions. However, this limits the holistic understanding given the one-sided approach. Numerical models lack contextual understanding as they are solely based on data while miss the rich, contextual information imbibed in textual descriptions. For instance, features of numerical like funding goals and backer counts provide quantitative insights but fail to consider the qualitative perspective of project descriptions as this can be vital for one to understand the nature of the project/idea, its appeal and the narrative. Furthermore, restrict the feature space, potentially omitting influential textual elements like sentiment, keywords, or writing style, which are often critical for crowdfunding success. On contrary, models that rely on solely contextual data might overlook the critical quantitative metrics that help signify the success of crowd funding such as the financial goals, amounts raised, and deadlines etc. This definitely proves the worth of project's story and pitch but lacks in exhibiting concrete evidence to meet the performance goals and targets. Both approaches draw conclusion on the chances of bias and imbalanced data. Given the limitation on providing comprehensive insights, the need to come up with a solution that offers a comprehensive understanding of the project's success by considering both numerical and contextual aspects together is important.

The study mainly focuses on delivering the below objectives:

- **Enhancing Crowdfunding Campaign Success:** Help predict the crowdfunding campaign success backed by improved accuracy and reliable data insights that will benefit all the stakeholders
- **Back decisions with data-driven insights rather than just assumptions:** This model will enable the stakeholders to take informed decisions on whether the idea holds potential or not on the basis of actual data inputs
- **Save time and help take prompt action:** It is important for the model to hold potential that helps save valuable time and mitigate the hassle that goes while deciding on the success of any idea presented. Further, a prompt real-time recommendation backed by strong insights within the category improves its reliability

2.3 Research Question

To what extent can the combination of classification modelling and NLP techniques optimize the prediction of crowdfunding campaign performance?

2.4 Structure of the Report

The research document is divided into seven sections, each section gives a distinct aspect of the investigation done in this research. In the first section detailed information about the Introduction of the study, a background of the research, Objectives, and Scope is given. The second section summarizes the highlights of what has been done in the prior research, followed by the third section consists of the methodology utilized in this study. The fourth section is Design specification which gives information about procedures and methods used. Section 5 gives the implementation information followed by section 6 is Evaluation of the results and findings done in implementation. Finally, in section 7 the research paper wraps up with the conclusion of the study and the potential future scope

3 Related Work

3.1 Regression/Classification Techniques

BaniMustafa *et al.* (2022) did an analysis on Kickstarter crowdfunding success using machine learning which consists of data scraping, wrangling, exploration, engineering, model construction, evaluation, and variable importance analysis. In this research three three different regression techniques has been used such as Random Forest, K-Nearest Neighbour (KNN), and Support Vector Machine (SVM). Finding shows that KNN achieved higher accuracy and AUC metrics, while SVM performed poorly. This research highlights importance of selecting suitable algorithm for numerical features and goals. Oduro, Yu and Huang (2022) used 2009-2017 Kickstarter dataset and based on advanced machine learning algorithm predicted the crowdfunding success. The study applied Logistic Regression, SVM with Linear and Quadratic Discriminant Analysis, and Random Forest with bagging and boosting techniques. For evaluation cross-validation and metrics like accuracy, F1 score, precision, and recall have been used. Where Kaur *et al.* (2022) used a detailed data-driven approach for predicting crowdfunding success. The data has been collected from two big crowdfunding platforms, Kickstarter and Indiegogo which focuses on attributes such as project category, duration, goal amount, number of investors, and social media interactions. For predictive analysis, the study employs several models such as Logistic regression, Naïve Bayes, Support Vector Machines, and Random Forest. Performance evaluations have been done based on the accuracy metrics, in which Logistic and SVM show high accuracy for Kickstarter campaigns while Random Forest performed well for Indiegogo campaigns. Grover *et al.* (2023). used decision tree algorithm on a Kickstarter dataset to predict the success of ongoing projects and achieved 99% of accuracy. This study identified key factors like goal amount, pledged amount, number of backers, FAQs, and count of words in the description. By using WEKA (Waikato Environment for Knowledge Analysis) this study helps to make informed decisions by predicting which projects will succeed.

3.2 Based on Text Analysis / Natural Language Processing

Zhang *et al.* (2023), analyzed the impact of project features and text characteristics on cancer related fund-raising success which hosted by GoFundMe. The study utilized penalized logistic regression and NLP techniques to analyze 92,753 campaigns. The finding show that lower fundraising goals, frequent updates, and active engagement increase success to getting fund. Also, text length had an inverse U-shaped relationship with performance, whereas spelling mistakes shows the negative impact on results. Similarly, Yuan, Lau and Xu (2016) combined text analytics with machine learning Domain-Constraint Latent Dirichlet Allocation (DC-LDA) model. The study shows that prediction is improved accuracy by 11% over traditional LDA. Zhou *et al.* (2015) studied Kickstarter projects using the Elaboration Likelihood Model (ELM) to check the impact of text quality creator trustworthiness on project success. The study explains that well written, positively toned description and project owner's history significantly affect crowdfunding outcomes.

3.3 Based on Deep Learning Techniques

Yu *et al.* (2018), used a deep learning technique to predict Kickstarter project success with dataset consisting of 378,611 projects from 2009 to 2018. MLP (Multi-Layer Perceptron) model was employed it outperformed than other algorithms like Random Forest, AdaBoost, SVM, and Logistic Regression by achieving 93% accuracy and an AUC of 0.9323. Cheng *et al.* (2019) developed a multimodal deep learning model to predict crowdfunding success. The research consists of a combinational approach of textual, visual, and metadata features from Kickstarter profiles. The study used pre-launch profile data, their results showed that by integrating images with text and metadata predictive performance of model is increased. This approach highlights that by leveraging multiple data types crowdfunding predictions success can be enhanced. On other side Saric M. and Simic Saric M. (2023) used Kickstarter project title images, and proposed deep learning methods for predicting crowdfunding success in the pre-posting phase. Study utilizes various CNN architectures such as VGGNet, ResNet, and DenseNet, and finds that deeper networks performed better. This approach helps project creators to optimize their campaigns before launching by leveraging visual content for early success prediction. Yeh and Chen (2022) utilized ensemble neural network method to prevent overfitting and improve accuracy in prediction of crowdfunding success. The study uses social and human capital theories and the Level of Processing (LOP) theory to select features for the model. The ensemble neural network achieved the highest accuracy, which shows that the effectiveness of combining multiple neural networks can enhance the outcomes. Where the authors Shi *et al.* (2021) utilized deep learning framework using audio analytics to predict crowdfunding success. The study uses transfer learning and multi-task learning to extract deep audio features from crowdfunding project descriptions. model showed an 8.28% improvement in F1 score and a 7.35% increase in AUC compared to baseline models, which indicates significant enhancement in prediction accuracy. This approach highlights the value of incorporating audio features alongside traditional data types.

3.4 Based on Clustering Techniques

He, Murray, and Tröbinger (2024) used k-means clustering to predict crowdfunding success. The study identifies key success factors such as campaign type, funding goals, geographic location, and community engagement which have higher impact. The evaluation is done using Silhouette Score and Davies-Bouldin Index and validated significant feature differences across clusters with an ANOVA test. Limitation to this study is found that variability in text quality was affecting the results. Similarly, Fernandez-Blanco et al. (2020) used k-means clustering to categorize crowdfunding projects, Also, it is able to find the patterns and common traits of successful and unsuccessful campaigns. The study is evaluated by the Davies-Bouldin Index. Carbonara (2020) examines the role of geographical clusters in the success of reward-based crowdfunding campaigns. This study investigates whether being a part of geographical clusters can enhance the probability of a campaign's success or not. Outcomes shows that regional economic conditions and social dynamics can influence the prediction of crowdfunding success. These studies show the effectiveness of clustering techniques in improving crowdfunding success predictions.

In summary, existing literature on crowdfunding prediction models highlights notable advancements in applying machine learning, text analysis, deep learning, and clustering techniques. The common limitation that has been found in most studies is that their focus is either on numerical data, or textual data, such as project description and sentiment analysis. This monotonous approach often results in an incomplete understanding of the broader range of factors that contribute to crowdfunding success. By separating these types of data previous research misses the chance to combine strength of both numerical and textual information which results in limiting overall predictive power and accuracy of models.

3.5 Research Gap, Analysis, and Proposed Approach

Despite significant advancement in prediction of crowdfunding success, there are certain gaps remaining particularly in the integration of advanced NLP techniques and diverse data types. Many studies do not uses sophisticated methods like TF-IDF for text feature extraction or integrate numerical, textual, and other data types. This study processes a combinational approach which integrates classification models with TF-IDF-based text analysis. The main aim of this approach is to enhance predictivity and provide reliable data insights by leveraging advanced text analysis and holistic data integration.

4 Research Methodology

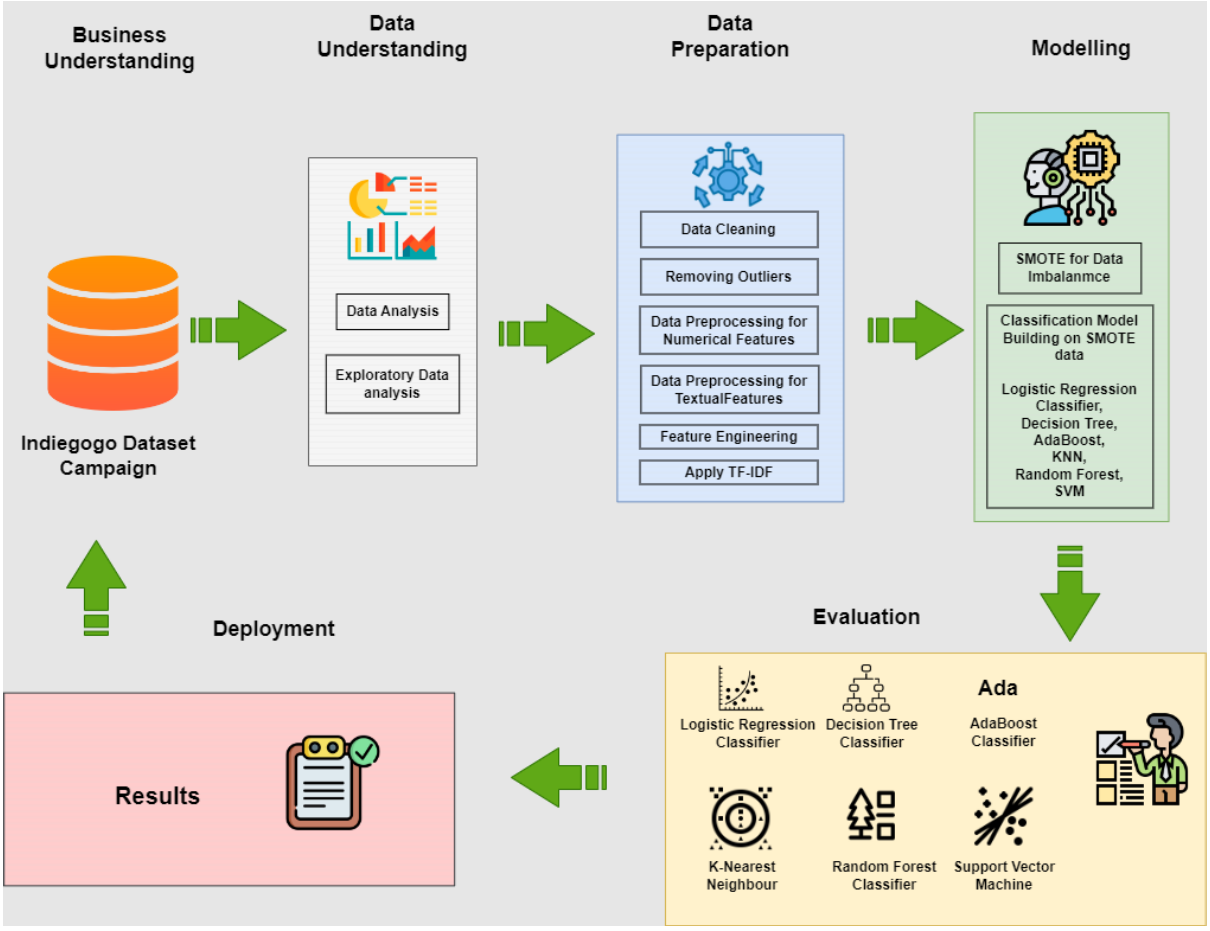

Business Understanding
Data Understanding
Data Preparation
Modelling
Indiegogo Dataset Campaign
Data Analysis
Exploratory Data analysis
Data Cleaning
Removing Outliers
Data Preprocessing for Numerical Features
Data Preprocessing for TextualFeatures
Feature Engineering
Apply TF-IDF
SMOTE for Data Imbalanmce
Classification Model Building on SMOTE data
Logistic Regression Classifier, Decision Tree, AdaBoost, KNN, Random Forest, SVM
Deployment
Evaluation
Logistic Regression Classifier
Decision Tree Classifier
Ada
AdaBoost Classifier
K-Nearest Neighbour
Random Forest Classifier
Support Vector Machine
Results

Figure 1: Flow Diagram of Model

The study uses the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology for development. CRISP-DM consists of six sequential stages as shown in Figure 1.

4.1 Business Understanding

This phase mainly focuses on understanding the objective and requirements of the project. All the stakeholders involved in this process such as Crowdfunding platforms, Entrepreneurs, and investors would benefit from the proposed model since it would enhance the predictability of campaign success. This study uses the combinational approach of classification model and NLP technique (TF-IDF) which provides actionable insights about what factors are affecting successful crowdfunding success. This approach will help three entities in the following way: First, it will assist investors to optimize their campaign strategies, Second, it will help investors to identify the projects that have the potential to be successful and last it will help crowdfunding platforms to improve their overall success rates. These improvements will not only lead to increased user satisfaction but also it will help to build more trust in crowdfunding platforms, which ultimately drive a more effective and efficient funding process.

4.2 Data Understanding

The dataset used in this research is sourced from Kaggle and includes various set of features that are relevant to campaign success[1]. The dataset consists of 20,632 rows with each row representing a unique crowdfunding campaign. The dataset consists of both numerical features like Raised Amount, funded percentage, goal amount, geographical locations, duration, etc., and textual data like tagline and title. Exploratory data analysis (EDA) is performed which involves univariate and bivariate analysis to understand the relationship between variables and create a visualization for insights with respect to Campaign success. Figure 2 shows the count of successful and failed projects which is evident that the data is imbalanced.

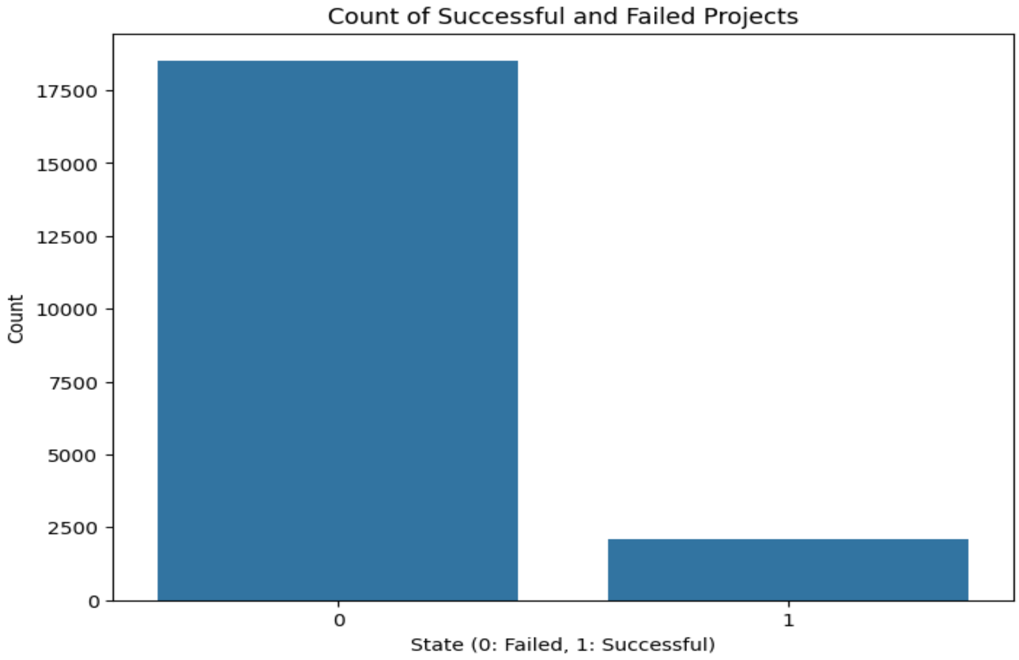


Figure2: Imbalance of class in Dataset

4.3 Data Preparation

In this phase, the data is cleaned, transformed, and prepared for modeling. The data preparation is done for numerical data and textual data separately. In the overall dataset, Tagline and Title columns consist of missing values, other columns don't consist of missing values as shown in Table 1.

Table 1: Overall Missing Values in Dataset

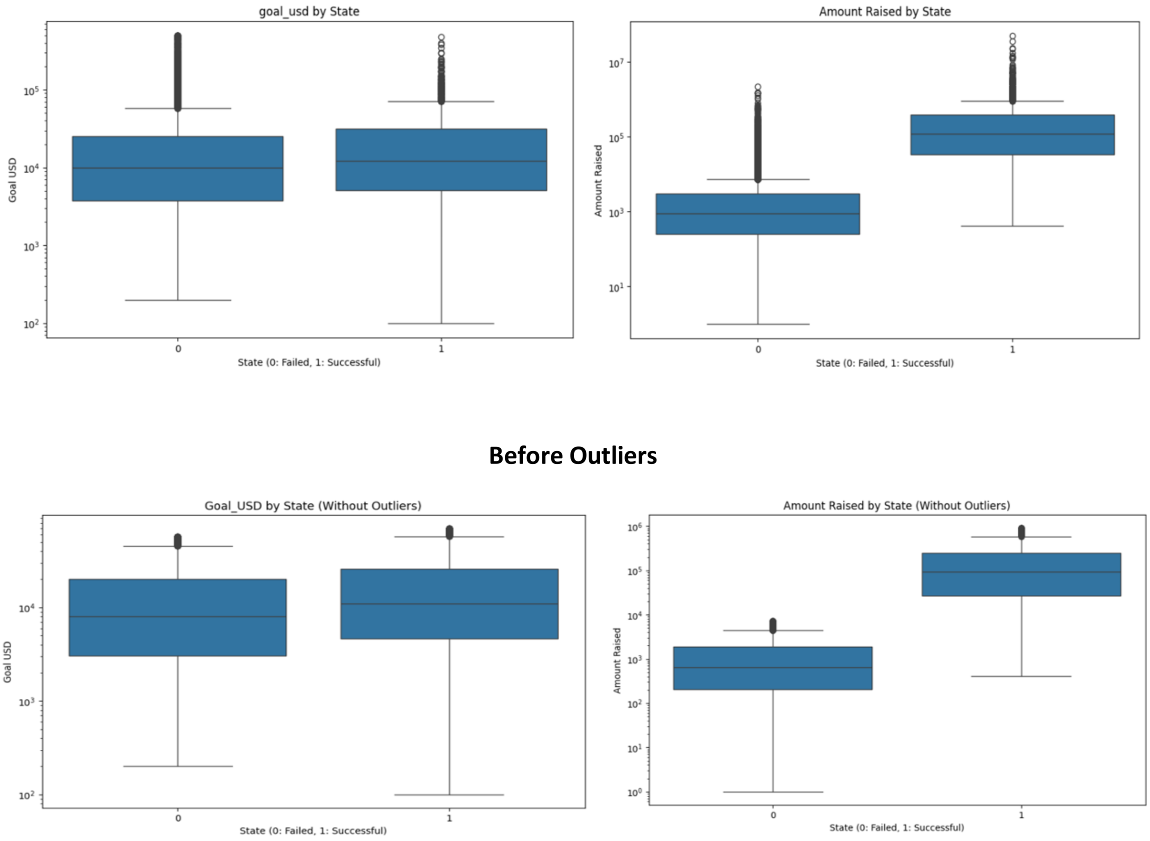
Feature Name	Missing Val
Tagline	14
Title	6

[1] https://www.kaggle.com/datasets/quentinmcteer/indiegogo-crowdfunding-data

Data Preprocessing on Numerical Features

In the data preprocessing stage several steps were undertaken to ensure that the numerical features are ready for analysis and modelling. The first task was to clean the 'funded percentage' column, which consisted of special character that needed to be removed. After removing special characters these values are converted into float format which makes it suitable for model building. Next is to address the categorical columns which are crucial to transform into numerical form. By label encoding categorical columns are converted into numerical values. label encoding has been applied. The ‘Country’ and ‘Categories’ Columns were initially one-hot encoded which were converted back into label-encoded columns. This transformation made the data easier to understand and interpret. Following that column such as URL' and 'Project ID' were not useful for the analysis. These columns don’t provide meaningful insight or contribute to prediction, so they were removed from the dataset. This step helped to streamline the data and focus on the features that matter most.

The outlier detection is another main step of preprocessing, with the help of a box plot the outliers have been identified for numerical features as shown in Figure 3. These outliers could potentially lead to inaccurate predictions. To handle these outliers the Interquartile Range (IQR) method was used which ensured a clean and reliable dataset. Feature engineering played a significant role in enhancing the dataset. Based on the EDA country column is divided into two categories USA and NON-USA. Additionally, the duration of each crowdfunding campaign was calculated by period in months between start and end dates.



Before Outliers

After Normalization

Figure 3: Outlier Detection and Removal

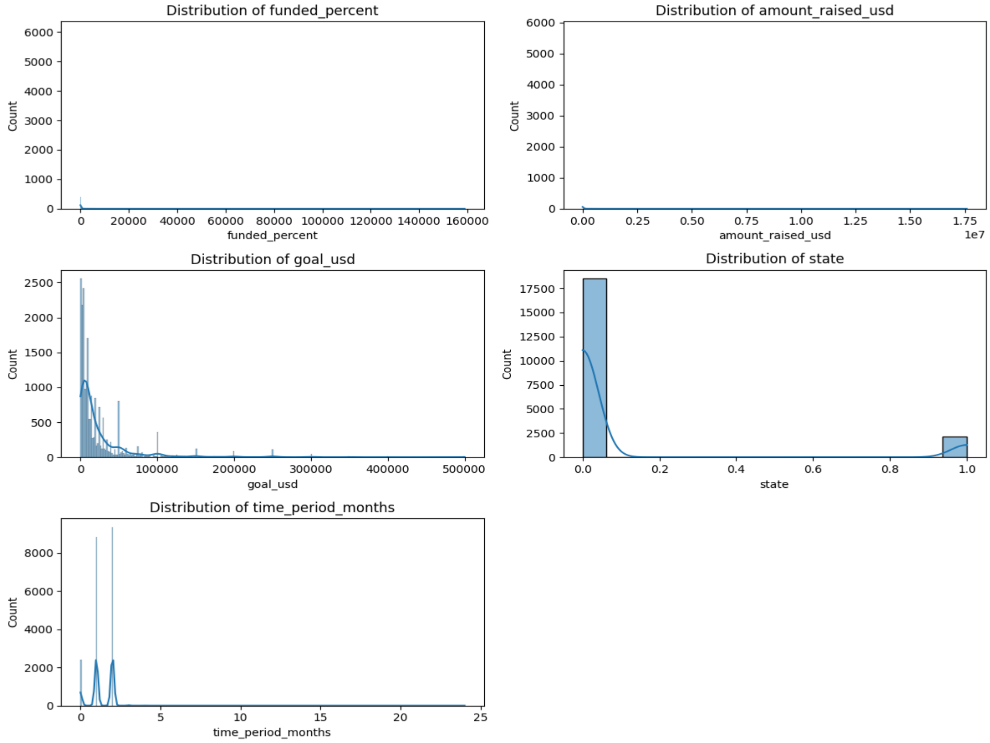
Data Preprocessing on Text Features

The preprocessing of text data consists of several important steps. Initially, all text is converted into lowercase to maintain uniformity. This step is then followed by the removal of punctuation and special characters to clean the data. In the next stage the rows that contains only numeric values have been removed as they don't provide any meaningful information for text analysis. After this check for null value was conducted, to make sure that the removal of numeric rows did not introduce any missing data issues. Then, next step consists of the removal of stop words, which are common words (such as "the", "is") . This word does not give any significant meaning to text and can be excluded to remove the noise from data. After this tagline and title are combined into single column to consolidate the text and help in providing a richer context for analysis.

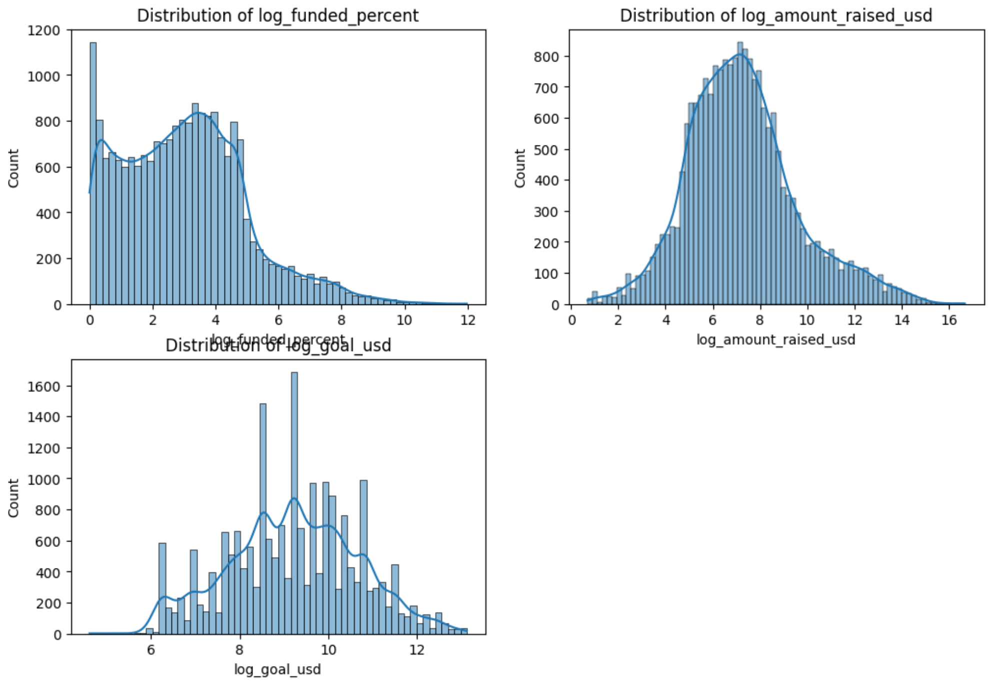
For text feature extraction, TF-IDF (Term Frequency-Inverse Document Frequency) vectorizer is used. TF-IDF is statistical measure, and it helps to transform the textual data into numerical features by considering the frequency of words (term frequency) and how unique or rare they are across all documents (inverse document frequency). This method works on the logic that words that are common give a low weightage and the words that are unique give a higher importance which helps in improving the quality of the text for modeling.

Normalization of Data

Following label encoding and feature engineering, the original columns were dropped to maintain a clean dataset. Then correlation matrix is generated to identify most related features to target variable. Then to asses this most related normalized feature histogram plots are used to visualize their distribution. Based on the visualization it is identified that the data is not normalized which is shown in Figure 4. log transformation has been applied on skewed data to achieve normalization. Once data is normalized it is used for feature selection.



Before Normalization

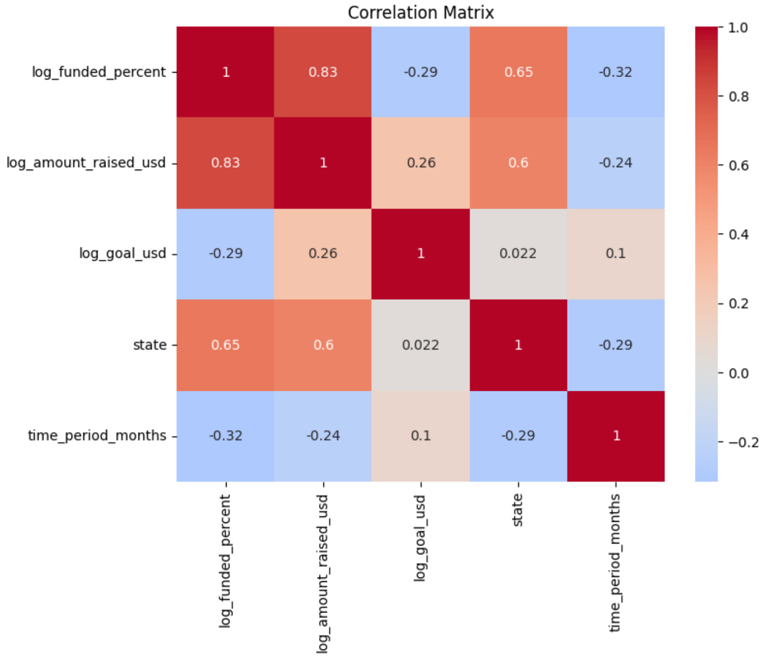


After Normalization

Figure 4: Normalization of funded percent, amount raised, goal usd

Feature Selection

After completion of data preprocessing and normalization, the next stage consists of feature selection for model building. As shown in Figure 5, The correlation matrix is again used after the normalization of numerical features to identify the most relevant features. Additionally, the ANOVA test was used to identify the most relevant categorical features with respect to the target variable. Features generated through TF-IDF vectorization were also included in this selection process. These selected features were then used to build the model.



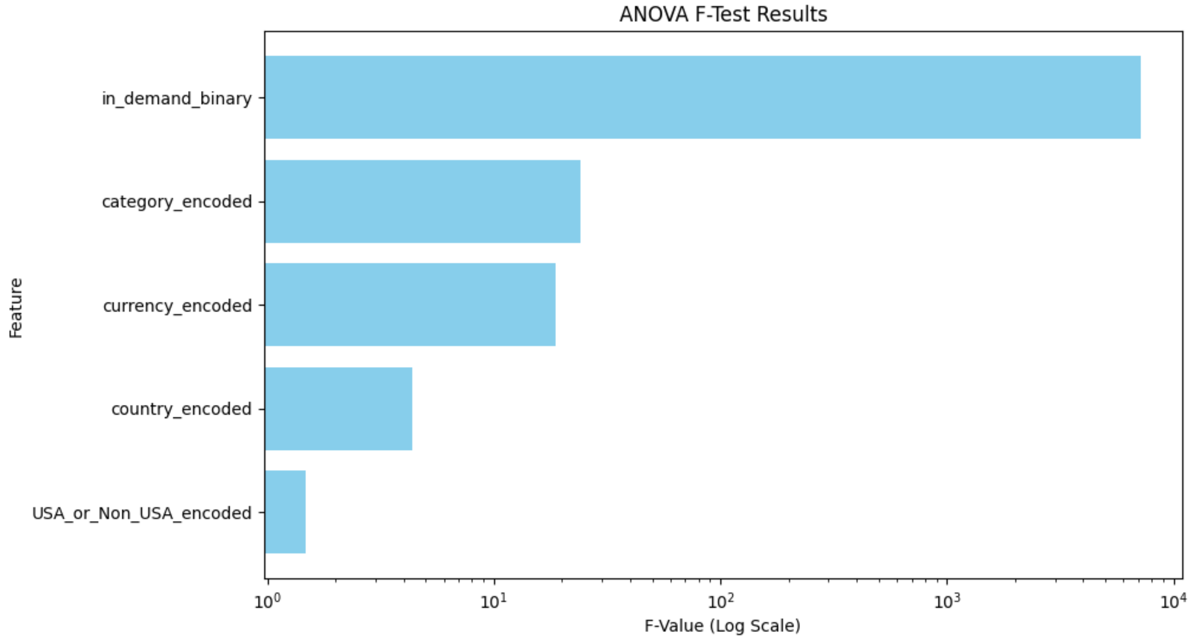


Figure 5: Feature selection based on Correlation Matrix and ANOVA test

4.4 Modelling

Before applying modelling techniques, the class imbalance issues are addressed using the Synthetic Minority Over-sampling Technique (SMOTE) on the cleaned dataset. Figure 6 shows the results of SMOTE before and after applying it. This step makes sure that the minority class which is 'Success / 1' is appropriately represented. This helps in improving the model's capacity to generalize across all classes. After resolving the class imbalance, the data is split into training and testing data with 80% allocated for training and 20% allocated to testing. Several Classification models were applied to the balanced dataset such as Logistic Regression, Decision Tree Classifier, AdaBoost Classifier, K-Nearest Neighbors (KNN) Classifier, and Support Vector Machine (SVM) Classifier.

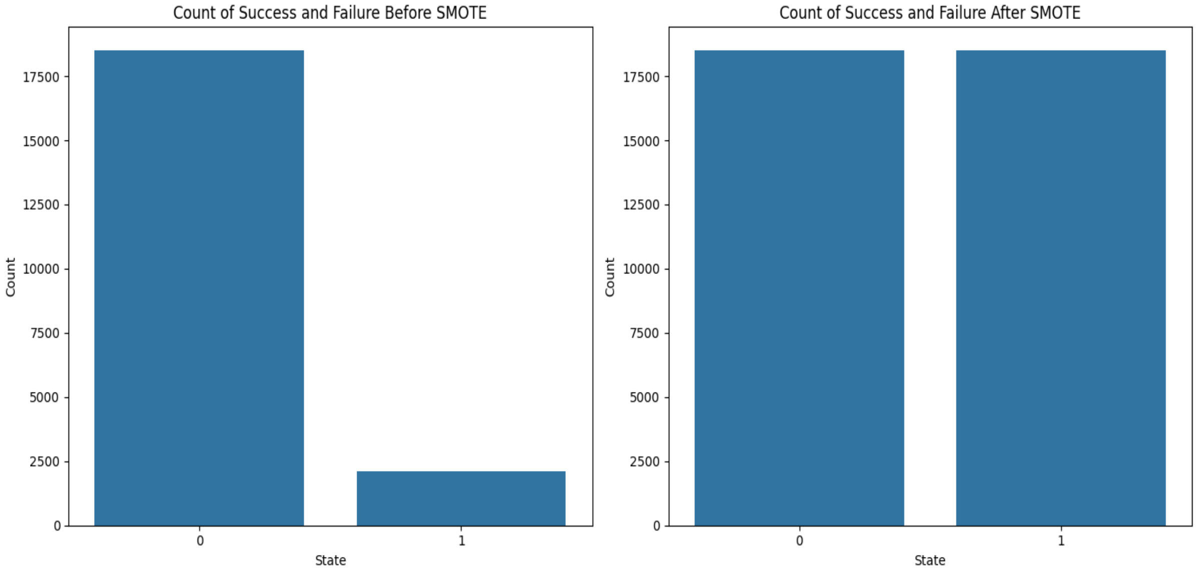


Figure 6: Balancing using SMOTE

4.5 Evaluation

In order to evaluate the performance of model performance matrices has been used which consist of Accuracy, Precision, Recall, and ROC-AUC score and 5 Fold cross-validation. After training the dataset on a balanced dataset the models such as Logistic Regression, Decision Tree, AdaBoost classifiers, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machine (SVM) tested on the original dataset which in imbalanced. This approach ensured that the model's performance metrics reflected their ability to handle real-world data distribution, which provides realistic evaluation.

5 Design Specification and Implementation

5.1 Software and Technologies Used

In order to generate the outputs and results below specifications and software libraries are used:

1. **Programming Language:** Python
2. **IDE:** Google Colab
3. **Python libraries and Modules:**
    A. **Pre-processing:**
    - **Numpy** for numerical operations
    - **Pandas** for data manipulation and analysis
    - **re** for regular expressions
    - **nltk** for natural language processing (e.g., stopwords, lemmatization)

    B. **Feature Engineering:**
    - **matplotlib** for data visualization
    - **seaborn** for statistical data visualization
    - **scipy** for scientific computing
    - **wordcloud** for generating word clouds
    - **sklearn.feature_extraction.text.TfidfVectorizer** for text feature extraction
    - **sklearn.preprocessing.LabelBinarizer, LabelEncoder** for encoding labels

    C. **Modelling and Evaluation:**
    - **sklearn** for machine learning algorithms and model evaluation (e.g., ensemble, tree, linear_model, model_selection, metrics)
    - **imblearn** for handling imbalanced datasets (e.g., SMOTE)

5.2 Modelling Techniques

A total of six modeling techniques were employed to predict the success of crowdfunding campaigns, with the aim of identifying which model performs best. These techniques include:

Logistic Regression Classifier: Logistic regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. This method is mainly used when there are two outcomes typically represented as 0 and 1. It models the probability of the occurrence of an event by fitting data to a logistic curve, thus allowing for the prediction of binary outcomes based on the linear combination of independent variables (Raflesia *et al.*, 2023).

Random Forest Classifier: A Random Forest classifier is a commonly used machine learning algorithm that combines the output from multiple decision trees to reach a single outcome. This method uses the bagging and random feature selection technique to improve classification results. Bagging means it creates multiple versions of dataset using random sample technique, and this versions are then use to build several decision trees. Each decision tree provides its opinion on how to classify data. Then results will be calculated based on combining these trees outcomes which helps to get more accurate and reliable model (Raflesia *et al.*, 2023).

Decision tree Classifier: The decision tree classification is a supervised learning algorithm that uses a particular set of rules to make a particular decision. Decision tree consists of three main things which starts with root node and it follows with branches based on the feature values until it reaches to leaf. The leaf is the final indicator to predict the class for that particular observation. rules (Oduro *et al*., 2022).

AdaBoost Classifier: Ada-boost or Adaptive Boosting is one of ensemble boosting classifier, which basically combines multiple poor performer classifier which helps in to get high accuracy classifier. So concept behind AdaBoost is to set the weights of the classifier and then train the data sample on each iteration which ensures the accurate predictions of unusual observations and accuracy is maximized (Jhaveri *et al*., 2019).

SVM classifier: Support Vector Machine (SVM) is another supervised learning algorithm which is used for both classification and regression problems. SVM works based on mapping the data to high dimensional feature space so it is easy to categorize data points. So once the separator is identified between categories then it transforms the data in such a way that the separator can be drawn as a hyperplane (Banimustafa *et al*., 2022).

KNN Classifier: K-Nearest Neighbors is another supervised learning model which is used for both regression and classification problems. Basically, this algorithm measures the distance between a point to its closest K-Neighbour. Based on the majority of the neighborhood it classifies the sample into that neighborhood class (Banimustafa *et al*., 2022).

5.3 Evaluation Techniques

There are 4 evaluation techniques used to check the model's performance which include:

Accuracy: Accuracy metrics tell that how many predictions are correctly identified by the model. This is calculated as the total number of correct predictions divided by a total number of predictions.

Confusion Matrix: Confusion metrics give the count of the total number of True positives, True Negatives, False Positives, and False Negatives. These values give detailed insights about how the model is performed.

Precision and Recall: Precision is calculated as the percentage of total correctly positive predictions out of all positive predictions. Whereas Recall is calculated as the percentage of actual positive predictions out of correct predictions. These metrics are commonly used in classification problems.

F1 Score: F1- Score is the harmonic mean of precision and recall. This measures the classifier's effectiveness specially when the data is imbalanced.

ROC-AUC Curve: ROC curve is used to check the model's ability to distinguish between the positive and negative classes. This plots the true positive rate against false positive rate at various threshold settings. The higher value indicates the better discrimination between classes. This metric is more useful when evaluating the binary classifiers.

5-Fold Cross Validation: In this technique the dataset is divided into five equal parts then the model is trained on four parts and tested on to the fifth part. This process iterated five times and take the average of this iteration's outcomes. The average provides more reliable and estimates true performance of the model.

6 Results and Evaluation

The evaluation of the model is performed on both balanced as well as imbalanced datasets which ensures comprehensive assessment. By using SMOTE balanced dataset is created which allows for evaluation of the model's performance when the classes are equally distributed. This gives the understanding that how well the model can perform when classes are equally distributed. Conversely, imbalanced data consists of original class distribution, that provides insights into how well the model can perform on real-world scenarios. By evaluation model in both scenarios gives a thorough understanding of the model's robustness and effectiveness.

6.1 Experiment Study 1: Model performance on SMOTE data

Table 2: Model performance on SMOTE data

Classification Models	Accuracy	Precision	Recall	F1-Score	AUC-ROC Curve	Mean of 5-fold Cross Validation
Logistic Regression	0.9924	0.9850	1.0	0.9924	0.9971	0.9930
Decision Tree	0.9971	0.9967	0.9975	0.9971	0.9977	0.9977
AdaBoost	0.9979	0.9962	0.9997	0.9979	0.9997	0.9978
KNN	0.9849	0.9714	0.9991	0.9851	0.9940	0.9844
Random Forest	0.9978	0.9956	1.0	0.9978	0.9999	0.9981
SVM	0.9898	0.9800	1.0	0.9899	0.9970	0.9883

Table 2 shows the evaluation parameters of 6 different models. Logistic regression achieved high accuracy and an excellent AUC-ROC score indicates its effectiveness in differentiating its classes. The perfect recall shows that the model can identify positive instances correctly. The decision tree model showed very high performance across all metrics with near perfect precision with value of (0.9967) and recall (0.9975), these indicates that decision tree model is very strong ability to correctly classify both positive and negative instances. In other side AdaBoost model have produced highest accuracy (0.9979) and F1-score (0.9979) among the models, with near-perfect AUC-ROC (0.9997), precision (0.9962), and recall (0.9997), which shows the robustness and strong classification of ability.

KNN performed well in terms of recall (0.9991), but its precision (0.9714) and accuracy (0.9849) got slightly lower if compared with other models. Random forest achieved almost near to perfect scores in all metrics, which gives the highlights of the model's ability to handle complex scenarios in classification. SVM has shown strong performance with high precision with a value of 0.9800 and perfect recall of 1.0 which indicates effectiveness in differentiating between classes and correctly identifying positive instances. The model's accuracy was slightly lower than other models.

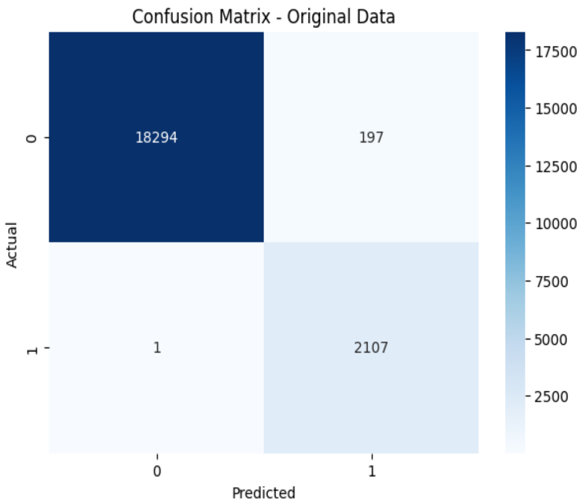
The mean of 5-fold cross validation gives and additional validation for model performance. Amongst all the models AdaBoost achieved a cross-validation mean of 0.9978 which is closely matching to its accuracy score. This indicates consistent performance across different data subsets. Similarly, Random Forest's cross-validation mean of 0.9981 underscores its robustness and reliability. In contrast, KNN's lower cross-validation mean of 0.9844 compared to its original accuracy shows there is some variability in its performance. Variabilit indicates that the model is sensitive to data distribution. This thorough evaluation confirms that AdaBoost and Random Forest are reliable and effective in predicting crowdfunding success across varied data samples.

6.2 Experiment Study 2: Model performance on original data

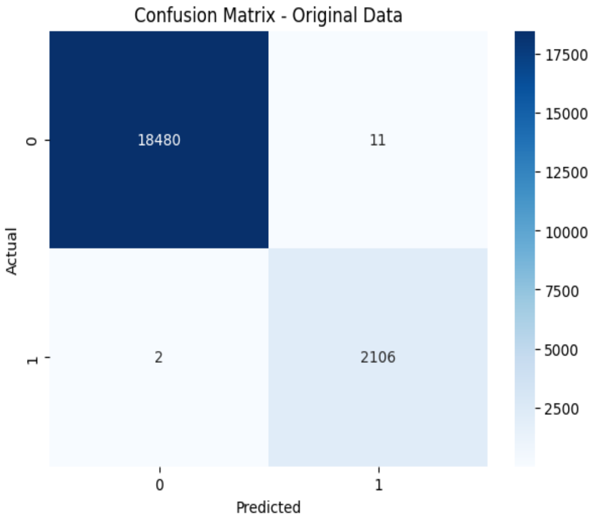
Table 3: Model performance on original dataset data

Classification Models	Accuracy	Precision	Recall	F1-Score	AUC-ROC Curve	Mean of 5-fold Cross Validation
Logistic Regression	0.9906	0.9165	1.0	0.9564	0.9978	0.9868
Decision Tree	0.9991	0.9933	0.9985	0.9959	0.9989	0.9996
AdaBoost	0.9993	0.9934	1.0	0.9966	0.9971	0.9997
KNN	0.9822	0.8526	0.9990	0.9200	0.9984	0.9762
Random Forest	0.9992	0.9924	1.0	0.9962	0.9998	0.9965
SVM	0.9825	0.8544	1.0	0.9215	0.9977	0.9966

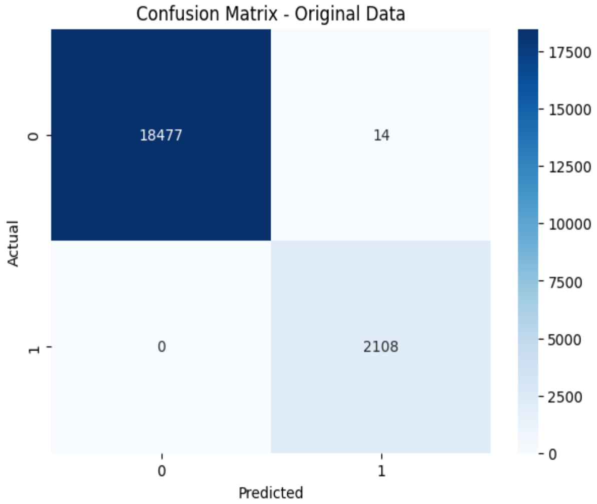
In this experiment, the performance of various classification models has been done on the original dataset without any balancing techniques applied. The Figure 7 shows Confusion matrix of each model.



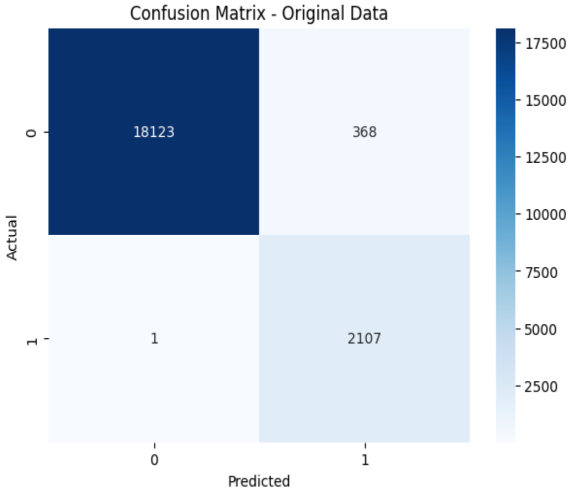
Logistic Regression Classifier



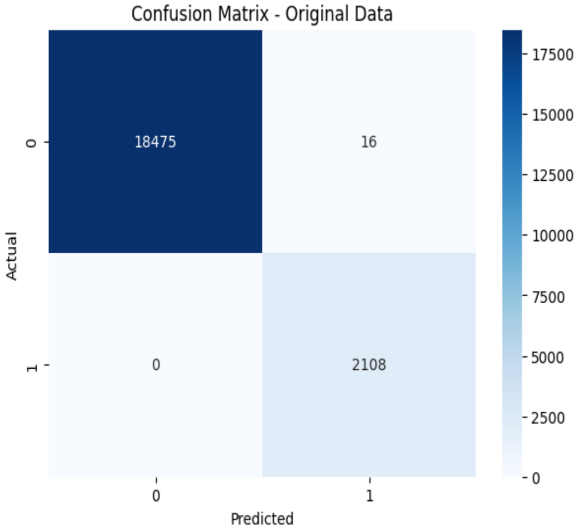
Decision Tree Classifier



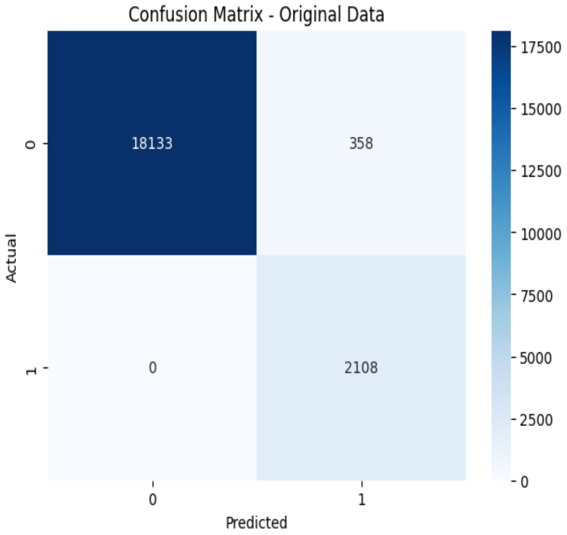
AdaBoost Classifier



KNN Classifier



Random Forest Classifier



SVM Classifier

Figure 7: Confusion Metrics for Crowdfunding Prediction Models

Table 2 shows that Logistic Regression achieved high accuracy with a value of 0.9906 and an excellent AUC-ROC score with a value of 0.9978. The perfect recall score (1.0) suggests that all positive instances were correctly identified, but precision was on the lower side which affected the F1 score. The decision tree model showed high-performance values across all the metrics with an accuracy of 0.9991, precision of 0.9933, recall of 0.9985, and an F1-Score of 0.9959. This shows the strong ability to classify each class properly. AdaBoost performed well with the highest accuracy of 0.9966 among the models with near-perfect precision (0.9934) and recall (1.0). This shows the robustness of the model. The AUR-ROC score was 0.9971.

KNN performed well, particularly in recall (0.9990), which suggests effectiveness in identifying positive instances. However, its precision (0.8526) and accuracy (0.9822) were lower compared to other models, which resulted in a lower F1-Score (0.9200). Random Forest achieved near-perfect scores in all metrics, this highlights the model's capability to handle complex datasets and also provides robust classification performance through ensemble learning. The model achieved an accuracy of 0.9992, precision of 0.9924, recall of 1.0, F1-Score of 0.9962, and an AUC-ROC score of 0.9998. SVM showed strong accuracy in terms of accuracy (0.9825) and F1-Score (0.9215) but still they were slightly lower than some other models, but the AUC-ROC score remained high at 0.9977. the precision of the model is (0.8544) which is lower than other models. Based on a mean of 5-fold cross-validation it is shown that AdaBoost's and Random Forest align closely with its accuracy, indicating consistent performance on original dataset as well. Similarly, outcome for KNN shows that lower cross-validation mean of 0.9762 compared to its accuracy suggests that there is some performance variability and indicates sensitivity to data distribution.

In summary, the evaluation of models on both SMOTE-balanced and original datasets observed the key insights. Both AdaBoost and Random Forest showed an exceptionally well performance across both the datasets. AdaBoost achieved the highest accuracy and F1-score on the SMOTE data, while Random Forest showed near-perfect metrics on both datasets. The Decision tree model also performed very well on the original dataset, Whereas Logistic regression had high recall but lower precision. In contrast, KNN and SVM were more affected by class distribution which shows lower precision and accuracy and highlighted that they are sensitive to imbalanced data. Overall, AdaBoost and Random Forest consistently demonstrated superior performance, while SVM and KNN were more affected by the class distribution in the data.

6.3 Discussion

The discussion part highlights that the combining the classification and TF-IDF features to build the model is performed exceptionally well in compared to the results of base paper by BaniMustafa *et al.* (2022). The base paper used three machine learning models such Random Forest, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) for predicting crowdfunding success, in which KNN achieved the highest accuracy at 97.9%. But base paper lacked advanced feature extraction and balancing techniques. In contrast, this study employed SMOTE for balancing data and TF-IDF for extracting meaningful text features, which lead to significant improvements in prediction. AdaBoost and Random Forest models showed near-perfect metrics across balanced and original datasets, demonstrating their robustness. Decision Tree and Logistic Regression also performed exceptionally well due to enhanced preprocessing and feature selection. The findings indicate that integrating numerical and textual data through sophisticated techniques significantly improves the prediction of crowdfunding success, and addresses the limitations observed in the base paper.

7 Conclusion and Future Work

This study addressed the question: "To what extent can combining classification models and NLP techniques optimize the prediction of crowdfunding campaign performance?" The research successfully demonstrated that integrating numerical and textual data significantly improves the prediction accuracy of crowdfunding campaign outcomes. Ensemble methods such as AdaBoost and Random Forest are identified as the most effective models that consistently gives high results on balanced data and imbalanced data which represent real-world scenarios. Logistic Regression and SVM models also performed well but their precision and accuracy were more affected by imbalanced class distribution. Decision tree classifier also outperforms too, shows ability to classify positive and negative instances, especially on the original dataset. On the other hand, KNN showed variability in precision and accuracy which highlighted the importance of data balancing.

The integrated approach of using both numerical and textual data addressed the limitation of relying on a single data type. The above research shows that a comprehensive feature set provides a better holistic view of the factors influencing crowdfunding success. The Application of SMOTE plays a vital role in balancing data for model training thus improving its performance. However, the scope is limited as the dataset considered for the study belongs to a single crowdfunding platform thus restricting its universal applicability. Future research can explore advanced NLP techniques such as topic modeling, and deep learning-based text embeddings, to gain deeper insights into the qualitative aspects of crowdfunding campaigns. Additionally, developing hybrid models that combine deep learning with traditional machine learning classifiers can capture more complex patterns and enhance performance. Expanding the study to include data from multiple crowdfunding platforms would help create more generalized models applicable across various platforms and industries.

References

BaniMustafa, A., Almatarneh, S., Bulkrock, O., Samara, G. and Aljaidi, M. (2022) ‘A data science approach for predicting crowdfunding success’, *2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI)*. Zarqa, Jordan, 29 November 2022 - 01 December 2022, pp. 1-6. doi: 10.1109/EICEEAI56378.2022.10050465.

Carbonara, N. (2020) ‘The role of geographical clusters in the success of reward-based crowdfunding campaigns’, *The International Journal of Entrepreneurship and Innovation*, 22(1), 18-32. doi: 10.1177/1465750320918385

Cheng C., Tan F., Hou X. and Wei Z. (2019) ‘Success prediction on crowdfunding with multimodal deep learning’, in *IJCAI'19: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Macao, China, 10-16 August 2019, pp. 2158-2164. doi:10.24963/ijcai.2019/299

Fernandez-Blanco, A., Villanueva-Balsera, J., Rodriguez-Montequin, V. and Moran-Palacios, H. (2020) 'Key factors for project crowdfunding success: An empirical study', *Sustainability*, 12(2), 599. doi: 10.3390/su12020599.

Grover, V., Anbarasi, A., Fuladi, S. and Nallakaruppan, M. K. (2023) ‘Decision tree based crowd funding for Kickstarter projects’, *EAI Endorsed Transactions on Scalable Information Systems*, 11(2). doi: 10.4108/eetsis.4639.

He, V. F., Trobinger, M. and Murray, A. (2024) 'The crowd beyond funders: An integrative review of and research agenda for crowdfunding', *The Academy of Management Annals*, 18(1), pp. 348-384. doi: 10.5465/annals.2022.0064.

Jhaveri, S., Khedkar, I., Kantharia, Y. and Jaswal, S. (2019) ‘Success prediction using Random Forest, CatBoost, XGBoost, and AdaBoost for Kickstarter campaigns', in *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. Erode, India, 27-29 March 2019, pp.1170-1173. doi: 10.1109/ICCMC.2019.8819828

Kaur, P., Deshmukh, S., Apoorva, P. and Batra, S. (2022) ‘Analysis and outcome prediction of crowdfunding campaigns’, *International Journal of Information Retrieval Research*, 12(1), pp. 1-20. doi: 10.4018/IJIRR.289575.

Oduro, M. S., Yu, H. and Huang, H. (2022) ‘Predicting the entrepreneurial success of crowdfunding campaigns using model-based machine learning methods’, *International Journal of Crowd Science*, 6(1), pp. 7-16. doi: 10.26599/IJCS.2022.9100003

Raflesia, S. P., Lestarini, D., Kurnia, R. D. and Hardiyanti, D. Y. (2023) ‘Using machine learning approach towards successful crowdfunding prediction’, *Bulletin of Electrical Engineering and Informatics*, 12(4), pp. 2438~2445. doi: 10.11591/eei.v12i4.5238.

Šarić, M. and Šimić Šarić, M. (2023) ‘Crowdfunding success prediction using project title image and convolutional neural network’, *Interdisciplinary Description of Complex Systems*, 21(6), pp. 631-639. doi: 10.7906/indecs.21.6.8.

Shi, J., Yang, K., Xu, W. and Wang, M. (2021) 'Leveraging deep learning with audio analytics to predict the success of crowdfunding projects', *The Journal of Supercomputing*, 77, pp. 7833–7853. doi: 10.1007/s11227-020-03595-2.

Yeh, J. Y. and Chen, C.-H. (2022) 'A machine learning approach to predict the success of crowdfunding fintech project', *Journal of Enterprise Information Management*, 35(6), pp. 1678-1696. doi: 10.1108/JEIM-01-2019-0017.

Yu P.-F., Huang F.-M., Yang C., Liu Y.-H., Li Z.-Y. and Tsai C.-H. (2018) ‘Prediction of crowdfunding project success with deep learning’, in *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. Xi'an, China, 12-14 October 2019, pp. 1-8. doi: 10.1109/ICEBE.2018.00012

Yuan, H., Lau, R. Y. K. and Xu, W. (2016) 'The determinants of crowdfunding success: A semantic text analytics approach', *Decision Support Systems*, 91, pp. 67-76. doi: 10.1016/j.dss.2016.08.001

Zhang, X., Tao, X., Ji, B., Wang, R. and Sörensen, S. (2023) 'The success of cancer crowdfunding campaigns: Project and text analysis', *Journal of Medical Internet Research*, 25, e44197. doi: 10.2196/44197.

Zhou, M., Du, Q., Zhang, X., Qiao, Z., Wang, A. G. and Fan, W. (2015) ‘Money talks: A predictive model on crowdfunding success using project description’, *Proceedings of the Twenty-first Americas Conference on Information Systems*. Puerto Rico, August 2015, pp. 1-8.