STARTUP SUCCESS PREDICTION

Hrithikender Reddy Bathula, hbathula@gmu.edu
Praveen Reddy Duligunti, pduligan@gmu.edu
Kushi Nelavelli, knelavel@gmu.edu
Lakshmi Chetana, lgomaram@gmu.edu

Summary

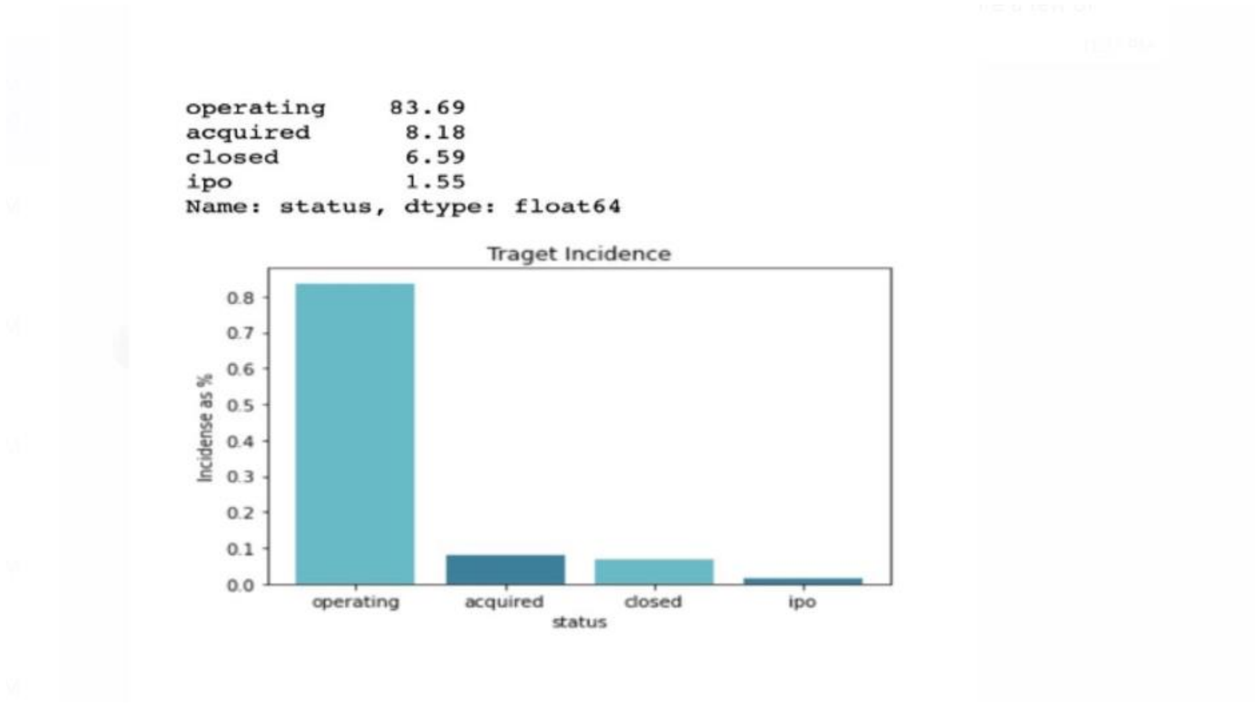
Each year, hundreds of new enterprises spring up all around the world. The emergence of new businesses has significantly increased during the past several decades in the US, India, and China. Making forecasts regarding the success of a company's performance and comprehending the elements that contribute to it are consequently critical jobs. By foreseeing a startup's success, investors are also better able to spot companies with the potential for rapid expansion, providing them an edge over other investors. On average, however, nine out of ten companies fail, according to the industry norm. A startup's failure might occur for a variety of reasons. Lack of funds, poor management, etc. are a few of them. This work aims to develop a 'startup' based predictive model using a variety of important factors that are present at different stages of a startup's life and to predict whether a startup which is currently operating turns into a success or a failure. It is very desirable to boost the success percentage of startups and unfortunately not a lot of work has been done to address the same in general.

We provide a technique to forecast a startup's performance based on several important variables, such as the amount of funding rounds, milestones, relationships, and other elements influencing the success or failure of the company at each milestone.

The dataset is basically imported from Crunchbase. There are multiple datasets like acquistions.csv, degrees.csv, funding_rounds.csv, funds.csv, investments.csv, ipos.csv, milestones.csv, objects.csv, offices.csv, people.csv and relationships.csv. It is important to merge the datasets by their unique values and common columns.

On the preprocessed data, multiple data mining optimizations and validations were applied along with different data mining classification approaches. We used machine learning models like Logistic Regression, KNN, Naive Bayes, Decision Tree Classifier, Random Forest Classifier and Neural Networking to deliver our analyses and make predictions. Based on evaluation metrics like ROC curve, precision, and recall metrics, we evaluated the accuracy of our models. We demonstrate how a startup may utilize our models to determine which elements they should pay closer attention to achieve success. Additionally, we have included a more complex model to find the predictions which is a new stage involving neural networking. Keywords - Startups, Prediction, Classification, Random Forest, Neural Networks.

Keywords - Startups, Prediction, Classification, Random Forest, Neural Networks



```
operating    83.69
acquired      8.18
closed        6.59
ipo           1.55
Name: status, dtype: float64
```

1. Introduction

As Richard Branson correctly stated, "A huge business starts small," and a startup's initial success starts with a wonderful notion that grows into an outstanding hypothesis. A substantial portion of inventors whoattempt to launch a business fail. Nine out of 10 companies fail, according to statistics. Entrepreneurs have always needed to comprehend the essential components involved in establishing a successful firm. Every company owner hopes that their hypothesis will prove to be accurate and will ultimately lead to a successfulenterprise. They want to create a product that consumers like while also enabling the company to get enoughtraction. A firm's performance is influenced by several factors, including marketability, management, skilled labor, and money. The actual causes of startup failure haven't been the subject of a lot of research. Numerous companies, both technical and non-technical, have been concentrating on the same issue for a long time. One of the attempts is to choose the "success failure" criterion during the pre-startup phase. Ourstudy attempts to create an accurate prediction model that forecasts whether an already-running firm will succeed or fail.

The startup industry is booming. Every day, dozens of new businesses are launched, and venture capital has grown to represent a sizable asset class with annual investments topping $100 billion in the US alone. The 2013 dataset from Crunchbase's 2013 Snapshot offers a window into this fascinating industry. This diverse dataset contains information about the startup ecosystem: organizations, individuals, company news, funding rounds, acquisitions, and IPOs. In our work first we downloaded each table, and we selectedonly subset of variabilities, and we merge all the datasets by their unique id which will be useful in our analysis. The dependent variable has multi class values, made up of 4 non-orderable levels, indicating the STATUS of each startup.

The levels of status are closed, acquired, ipo and operating. The Exploratory Data Analysis is performed for getting deep understanding of the data and the relationship between the attributes. The entire dataset issplit into the training and test data sets and the models are trained and tested accordingly on this dataset. The models created are based on the facts like funding rounds, milestones, state, etc. The dates the companyreceived its initial and most recent funding as well as other variables that had an adversely, have been usedto examine this data. For making predictions we used KNN, Logistic regression, Naïve Bayes, Decision Tree Classifier, Random Forest Classifier and Neural Networks . AUC (Area Under Curve), recall values, and precision values are some of the parameters we used to represent for evaluation purposes.

The remaining sections are organized as follows: Following a description of relevant work that explains the shortcomings of the existing approaches in Section 2 and a discussion of the proposed approaches thatdescribes our approach and its potential to address those shortcomings in Section 3, The current project's plan, which covers the development of the code, data sets, and evaluation procedures, is covered in the lastpart.

1.1 Related Work

We have referred to the “https://ieeexplore.ieee.org/document/7836749” as a reference for our current project work. In the IEEE reference paper, the project’s data was collected of 7000 successful companies and 4000 failed companies from Crunchbase (a wiki like database for all companies).

Given that it is compiled from several sources, including TechCrunch and Forbes, this data is made up of avariety of elements. It was concluded that there are numerous important factors that greatly alter the predictive models during the preprocessing stage. The funding rounds it has through is one of them. From 1999 (the height of the dot com bubble) to 2014, data was gathered on the companies. Two recessionarytimes and two dot-com bubbles are the main justifications for looking at this time because they can help usunderstand how some businesses manage to survive economic downturns while others fail. For each organization, they have considered over 70 different factors. Companies that couldn't raise any funds or that might have legal problems (rare examples of failure) weren't taken into consideration. They were ableto create nine different models in total. They created predictive models for classification utilizing more than 30 distinct classification methods. Based on the information received from CrunchBase, the labels for theseclasses indicated whether the company had succeeded or failed. The top 6 schemes that they chose are as follows:

1. Naive Bayes
2. ADTrees

3. Bayes Net
4. Lazylb1
5. RandomForest
6. Simple Logistics

Example of the successful company predictions by the model: Predictive models were initially tested on a startup called Spotify (founded in 2006). Almost all the model's parameters are met by the business. On a scale of 1 to 5, the company's typical severity scores at each stage are close to the company was ableto scale thanks to factors like its sizable customer base, high level of traction, outstanding services, well- built product, etc. Their model indicated an 88.9% chance of success, which qualifies the business as successful.

Example of the failed company predictions by the model: Everpix is one example of a model-predictedunsuccessful startup. The company failed as a result of issues like poor management and low traction. This company fell into the failing company category since the model's projection for it was close to 44.2%.

Deficiency in the existing approaches:

- The primary flaw in the current strategy is that the target variable status has only 2 non orderable levels that is only acquired and closed status. But the operating and ipo status are also important for analysis and prediction because there are large number of startups which are in operating status.

- The investors who are willing to invest in startups had to take wise decisions as it is the matter of money. So, to predict whether the startup will be successful or not for the complex decisions to take it's better to use the Neural Network Modeling.

- The random forest is that it can be too slow and inefficient for real-timeforecasts when there are a lot of trees. While logistic models are susceptible to overconfidence, logistic regression makes predictions based on collection of independent variables. That is, due to sampling basis, the models may appear to have greater predictive power than they do. Therefore, a logistic regression would be "overfit", overstating how accurate its predictions are.

2. Details of the Approach

Enhancing the existing approach by following the below procedures:

We can improve the prediction accuracy by performing better preprocessing and using more complex models like Smote and Neural Networks. Using NEURAL NETWORKS enables us to process parallel andgive a more accurate prediction such that the investor is sure about the start-up's success. In our approach we have four non orderable levels indicating the levels of status. The levels of status are acquired, closed, operating and ipo. It is a clear-cut case of multi label classification. By using SMOTE helps to oversample the data, it is a Synthetic Minority Oversampling technique to deal with the imbalanced data. The advantage of Smote is that you do not need to delete data points, so you do not needto delete valuable information. On the other hand, it is introducing false information into the model to avoidthe bias of the model. The dataset we have is an imbalanced data as most of startup companies are in operating status, so it is very much essential to use SMOTE technique for balancing the data and oversample it. Also, by using smote it helps to increase the records in the dataset by more balanced data it the models can avoid the problem of overfitting and bias.

Data Preprocessing:
By preprocessing our data transforms into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. To ensure reliable findings, the techniques are typically applied at the very beginning of the machine learning and AI development pipeline. By providingmore emphasis in data cleansing, data reduction, data transformation and data validation. By carefully preprocessing our data we can remove outliers and transform the data which can be modeled effectively. The objects.csv is main dataset which is around 250MB. The objects.csv dataset consists of all features where it can be used for analyzing and modeling. All the relevant datasets are merged, and all the useful attributes are used for the modeling purpose.

By observing the data, we can extract certain features like investment_rounds, invested_companies, funding_rounds, milestones, relationships etc. Feature selection helps to reduce the amount of redundant data from the data set. In the end, the reduction of the data helps to build the model with less machine effortand increases the speed of learning and generalization steps in the machine learning process. We can extractmore insightful features by using some of the methods like correlation.

With the help of the feature extraction technique, we can create new features by linearly combining the preexisting features. When compared to the values of the original characteristics, the new set of features will have different values. The fundamental objective is to use fewer features to collect the sameinformation. Though it may seem that selecting fewer features would result in underfitting, the extra data in the case of the Feature Extraction technique is typically noise.

KNN: One of the most straightforward machine learning algorithms is K-Nearest Neighbor, which uses thesupervised learning method. The new case is placed in the category that is most comparable to the availablecategories by the KNN method, which assumes that the new instance and the data are like the cases that are already accessible. New data point is classified using the KNN algorithm depending on how similarthe existing data is and is stored. This indicates that the KNN algorithm can quickly classify fresh data as it comes into a suitable category. Because the KNN technique is non-parametric, it makes no assumptionsabout the underlying data.

Random Forest: The decision trees used in the Random Forest classifier are numerous. In a Random Forest, which may offer robust and accurate classification and has the capacity to handles very large number of input variables, the final class of an instance is determined by outputting the class that is the mode of the outputs of individual trees. It can handle datasets with severely unbalanced class distributionsand is reasonably resistant to overfitting.

Logistic Regression: When a dependent variable is dichotomous(binary), the proper regression analysis touse is logistic regression. To describe data and explain the relationship between one dependent binary variable (Success or Failure) and one or more independent nominal, ordinal, interval, or ratio-level variables
i.e., we employ logistic regression.

Neural Networks: Neural Networks are multi-layer perceptron’s. By stacking many linear units, we get neural network. They are remarkably good at figuring out functions from X to Y. In general, all input features are connected to hidden units and NN's are capable of drawing hidden features out of them.

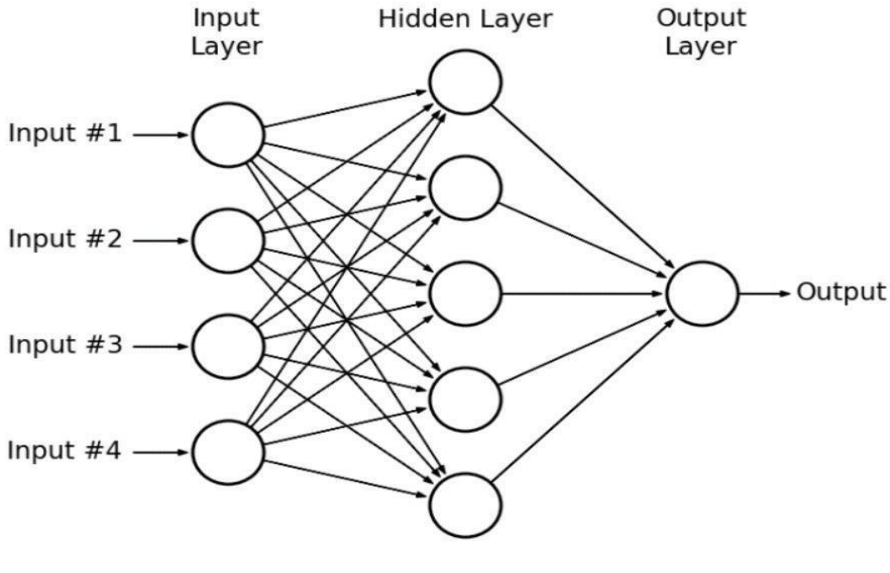


Figure 1: **Neural Network Mechanism**

We are using neural network to make complex decisions in analysis. A single Sigmoid/Logistic neuron at the output layer, which is the foundation of a binary classification neural network, can be used to performbinary classification in neural networks. This is due to the ease with which the estimated probability (p, pronounced p-hat), that the input falls into the "positive" class, may be simply interpreted as the output of a Sigmoid/Logistic function. Any input is condensed by the Sigmoid function into the output range of 0 to

1. So, for instance, if we were developing a "Success (1) vs. not-Success (0)" detector using a neural network, our output layer would still consist of a single Sigmoid neuron that would translate all calculationsfrom earlier layers into p, astraightforward 0–1 output range.

Keras is a minimalist Python library for deep learning that can run on top of TensorFlow. We can implementNeural Networks using Keras.

2.1 PLAN

CODE DEVELOPMENT:

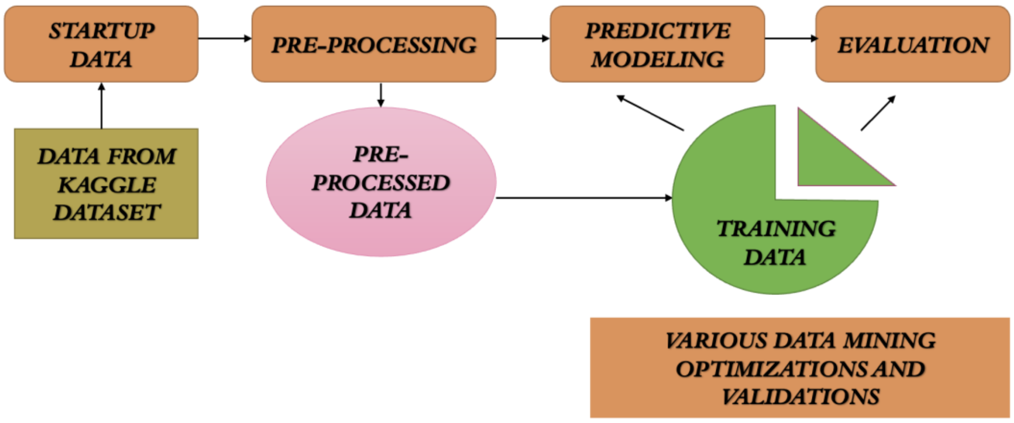


Figure 2: Block Diagram of Success/ Failure Prediction System

Startup Success/Failure Prediction System:
There are the following steps in the suggested predictive systems:

1. Data Gathering and Preprocessing:
The following are the key procedures:
(a) Data gathering: We have selected the relevant Kaggle dataset to perform the data modeling and to generate the required predictions.
(b) Data Preprocessing: Preprocessing the Kaggle dataset data is the next stage where we handle the null, duplicate, categorical values and perform standardization.

2. **Predictive Modeling**:
Using preprocessed data, predictive models are built using data mining classifiers to determine ifcompanies will succeed or fail. The first of these stage's two simple procedures is to divide the preprocessed data into training and testing sets (or use cross validation).
(b) Build a model employing data mining classifiers. In our case, we are building the following models:

- Logistic regression
- Decision trees
- KNN
- Random Forest
- Naïve Bayes
- Neural Network models.

3. Evaluation/Metrics:

Most of this stage is assessing the predictive model using test data.

(a) Contrast the predictive model's success/failure forecasts based on hypothetical data (the testing set) withexamples of known successful and unsuccessful businesses.

(b) Calculate performance metrics such as accuracy (percentage of predictions that are correct), precision (percentage of positive predictions that are correct), recall/sensitivity (percentage of records with positive labels that were predicted as positive), specificity (percentage of records with negative labels that were predicted as negative), area under the ROC curve (a measure of the model's discriminative power), etc.

We will also be calculating the confusion matrix, where we describe the performance of a classification model. Each row of the matrix represents the instances in a predicted class while each column representsthe instances in an actual class (vice versa).

DATASET DETAILS:

We have chosen the database which is comprised of 6 datasets. All of these datasets contain various aspects of the startup world which range from 1901-01-01 to 2014-10-01 time frame.

The main 6 datasets, which we have chosen are:

- **Acquisition**: details about acquisitions (acquired company, acquiring company, price and date of acquisition, payment method)
- **Founding Rounds:** details on the various investment rounds (funded company, date and funding type, total raised amount, number of participants)
- **Investments:** includes the investor's and the funded company's IDs.
- **Objects:** The primary dataset holding the fundamental details of each individual database item Name, entity type, category code, status, founded at, country code, state code, investment rounds, invested companies, funding rounds, and funding total usd are among the 40 variables that make up this dataset. They are also the most crucial.
- **Offices:** location of principal offices (both of the companies and the investment funds).
- **Relationships:** details about connections between individuals and institutions (people, institutions, start and end date of relationship, role held)

Some of the key features of the dataset are:

- id
- status
- normalized_name
- category_code
- founded_at
- closed_at
- tag_list
- country_code
- investment_rounds
- invested_companies
- first_funding_at
- last_funding_at
- funding_rounds
- funding_total_usd
- first_milestone_at
- last_milestone_at
- milestones
- relationships

Experimental Details:

I. Data Gathering:

When we first tried to retrieve the initial data for our project work, we encountered a challengewhere we could only retrieve a 350Mb startup related dataset. In order to increase the quantity of data, we have used the SMOTE Machine learning technique while performing data preprocessing. SMOTE is an algorithm which is used to deal with any problems with the imbalanced data, it can be even considered as an advanced version of oversampling.

II. Data Preprocessing:

By preprocessing our data transforms into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. To ensure reliable findings, the techniques are typically applied at the very beginning of the machine learning and AI development pipeline. In this stage, we provide more emphasis in data cleansing, data reduction, data transformation and data validation.

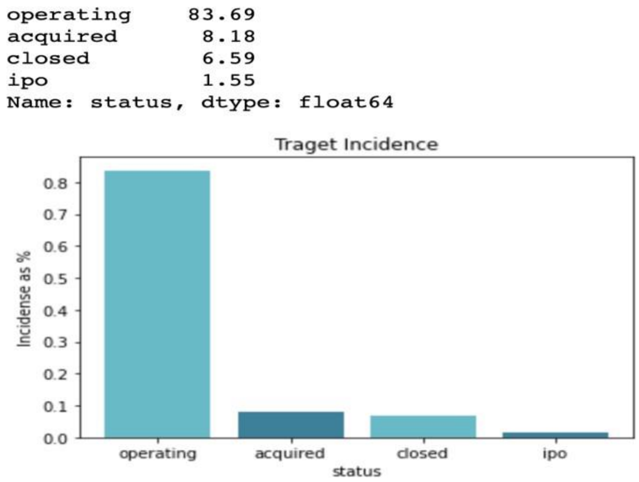
The objects.csv is the main dataset which consists of all the startup-based features and hence it can be used for performing analysis and modeling. The size of objects.csv file is around 250MB. Apart from the objects.csv file, we have other data files (acquisition.csv, degrees.csv, funds.csv etc.) which act as the supporting data files to objects.csv file. All these relevant datasets are merged, and all the useful attributes are used for the modeling purpose. Therefore, feature selection is also done as a part of this stage.

III. Exploratory Data Analysis:

In this step, we have performed the analysis of the datasets using visual techniques by importinglibraries like matplotlib for plotting the graphs and trends. One of the important analysis of thevarious startup's status is as follows:

The dependent variable is a categorial one, made up of 4 non-orderable levels, indicatingthe STATUS of each startup. These levels are:

- **CLOSED**: failed startup
- **ACQUIRED**: acquired startup
- **IPO**: listed startup
- **OPERATING**: startup not acquired or listed



```
operating    83.69
acquired      8.18
closed        6.59
ipo           1.55
Name: status, dtype: float64
```

Figure 3: Plotting graph to demonstrate the startup's status ('operating', 'acquired', 'closed' or 'ipo')

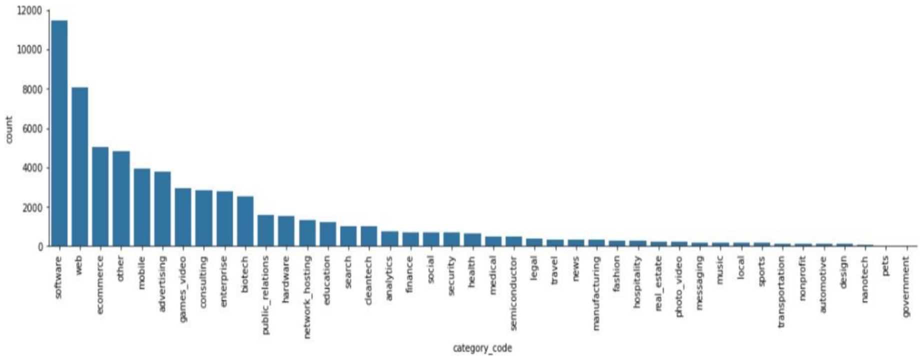


Figure 4: Plotting graph to demonstrate 'category_code' relation with the count

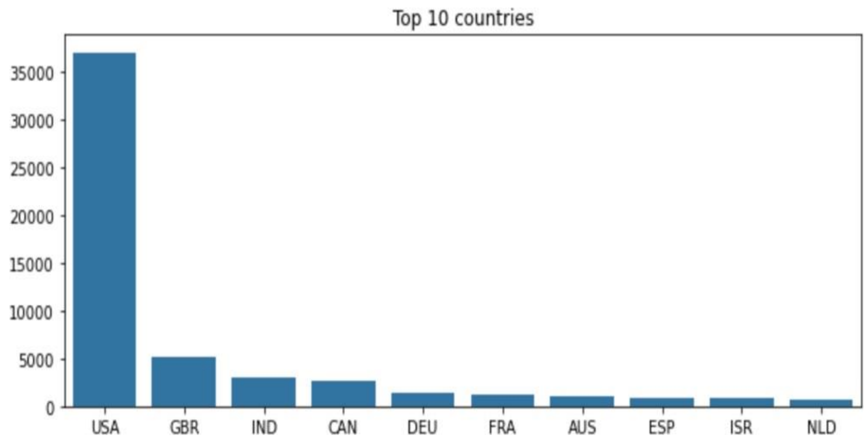
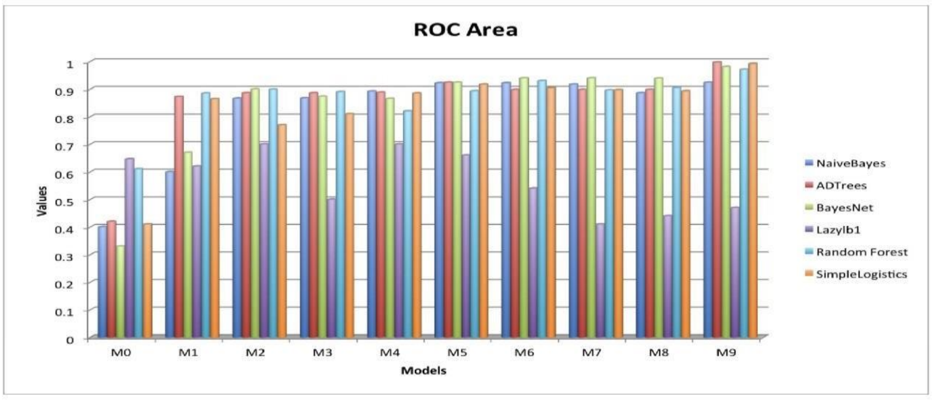


Figure 5: Plotting graph the count analysis for the top 10 countries

3. RESULTS

Existing approach results as per the relative work:

By using the existing approaches, we get the following accuracies:



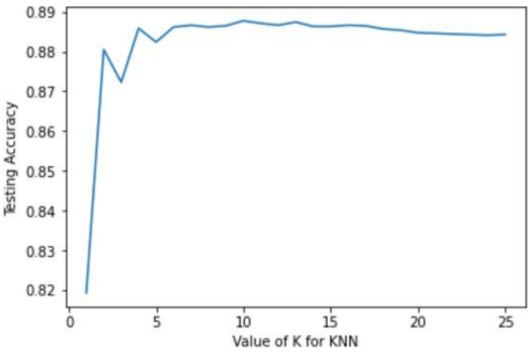
Proposed approach results as per our latest data:

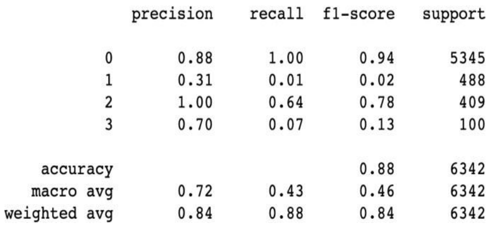
We have currently applied the following models to our startup dataset in order to predict it'ssuccess:

i. **K- Nearest Neighbours Classifier:**

The KNN Classifier is applied on the startup data before and after applying the SMOTEtechnique and the accuracies and classification report is generated as follows:

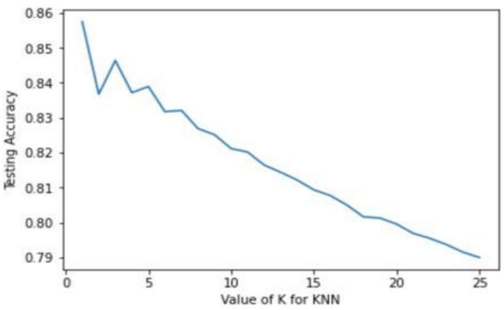
Before applying SMOTE technique

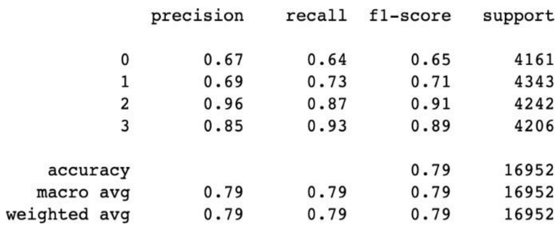




```
              precision    recall  f1-score   support

           0       0.88      1.00      0.94      5345
           1       0.31      0.01      0.02       488
           2       1.00      0.64      0.78       409
           3       0.70      0.07      0.13       100

    accuracy                           0.88      6342
   macro avg       0.72      0.43      0.46      6342
weighted avg       0.84      0.88      0.84      6342
```

After applying SMOTE Technique

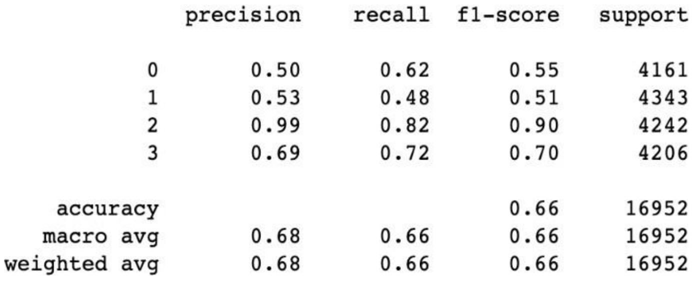




```
              precision    recall  f1-score   support

           0       0.67      0.64      0.65      4161
           1       0.69      0.73      0.71      4343
           2       0.96      0.87      0.91      4242
           3       0.85      0.93      0.89      4206

    accuracy                           0.79     16952
   macro avg       0.79      0.79      0.79     16952
weighted avg       0.79      0.79      0.79     16952
```

ii. **Logistic Regression Model:**

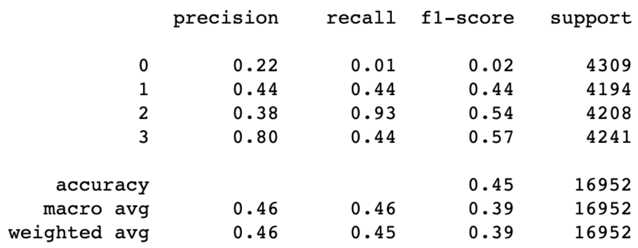
After performing the logistic regression model, we get an accuracy of 66% after applyingSMOTE technique.



```
              precision    recall  f1-score   support

           0       0.50      0.62      0.55      4161
           1       0.53      0.48      0.51      4343
           2       0.99      0.82      0.90      4242
           3       0.69      0.72      0.70      4206

    accuracy                           0.66     16952
   macro avg       0.68      0.66      0.66     16952
weighted avg       0.68      0.66      0.66     16952
```

iii. Gaussian Naïve Bayes:

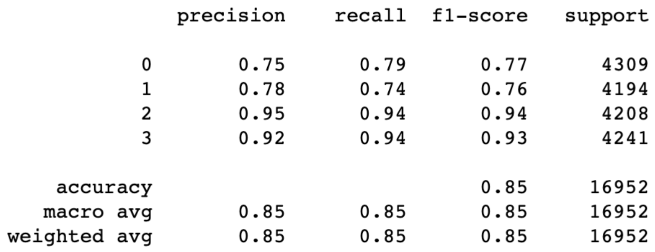
After performing the Gaussian Naïve Bayes model, we get an accuracy of 45% after applyingSMOTE technique.



```
              precision    recall  f1-score   support

           0       0.22      0.01      0.02      4309
           1       0.44      0.44      0.44      4194
           2       0.38      0.93      0.54      4208
           3       0.80      0.44      0.57      4241

    accuracy                           0.45     16952
   macro avg       0.46      0.46      0.39     16952
weighted avg       0.46      0.45      0.39     16952
```

iv. Decision Tree Classifier:

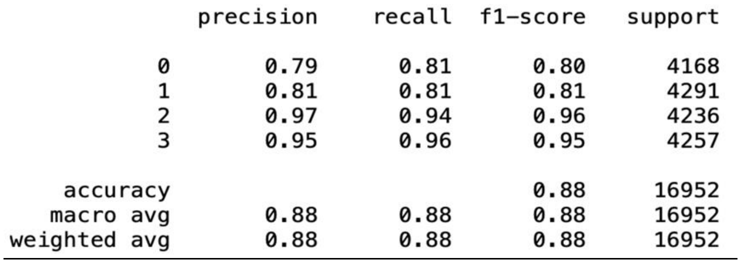
After performing the Decision Tree Classifier model, we get an accuracy of 85% afterapplying SMOTE technique.



```
              precision    recall  f1-score   support

           0       0.75      0.79      0.77      4309
           1       0.78      0.74      0.76      4194
           2       0.95      0.94      0.94      4208
           3       0.92      0.94      0.93      4241

    accuracy                           0.85     16952
   macro avg       0.85      0.85      0.85     16952
weighted avg       0.85      0.85      0.85     16952
```

v. Random Forest:

After applying the Random Forest model, we get an accuracy of 98% before applying the SMOTE technique and 88.07 after applying the SMOTE technique.

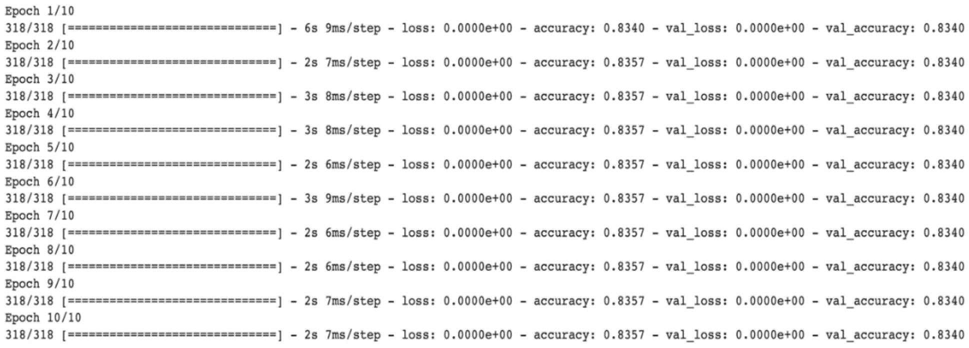


```
              precision    recall  f1-score   support

           0       0.79      0.81      0.80      4168
           1       0.81      0.81      0.81      4291
           2       0.97      0.94      0.96      4236
           3       0.95      0.96      0.95      4257

    accuracy                           0.88     16952
   macro avg       0.88      0.88      0.88     16952
weighted avg       0.88      0.88      0.88     16952
```

vi. Neural Networks:

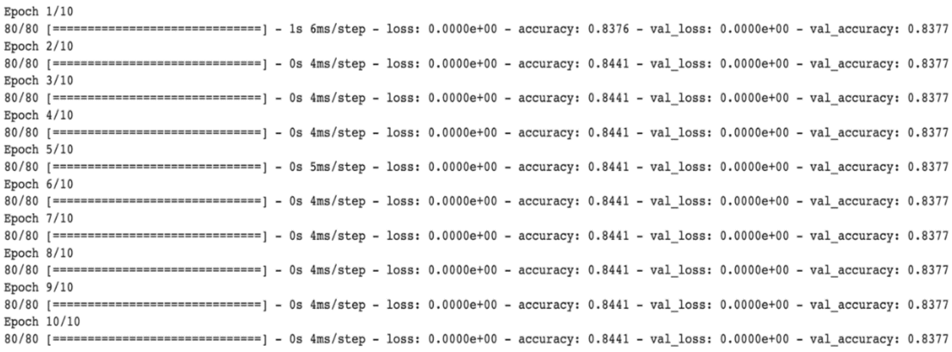
We have applied Neural networks to include complex models which are in turn used to predict the success of a startup. By using the Neural Network models, we get an accuracy of approximately 83.57% before using the SMOTE technique and approximately 84.41% after using the SMOTE technique.

Before applying the SMOTE technique:



```
Epoch 1/10
318/318 [==============================] - 6s 9ms/step - loss: 0.0000e+00 - accuracy: 0.8340 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 2/10
318/318 [==============================] - 2s 7ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 3/10
318/318 [==============================] - 3s 8ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 4/10
318/318 [==============================] - 3s 8ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 5/10
318/318 [==============================] - 2s 6ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 6/10
318/318 [==============================] - 3s 9ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 7/10
318/318 [==============================] - 2s 6ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 8/10
318/318 [==============================] - 2s 6ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 9/10
318/318 [==============================] - 2s 7ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
Epoch 10/10
318/318 [==============================] - 2s 7ms/step - loss: 0.0000e+00 - accuracy: 0.8357 - val_loss: 0.0000e+00 - val_accuracy: 0.8340
```

After applying the SMOTE technique:



```
Epoch 1/10
80/80 [==============================] - 1s 6ms/step - loss: 0.0000e+00 - accuracy: 0.8376 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 2/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 3/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 4/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 5/10
80/80 [==============================] - 0s 5ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 6/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 7/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 8/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 9/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
Epoch 10/10
80/80 [==============================] - 0s 4ms/step - loss: 0.0000e+00 - accuracy: 0.8441 - val_loss: 0.0000e+00 - val_accuracy: 0.8377
```

Table of Accuracies of the proposed approach:

	Model	**Accuracy - Before applyingSMOTE**	**Accuracy - After applyingSMOTE**
1.	**KNN Classifier**	**88.0**	**79.0**
2.	**Logistic Regression Model**	**95.0**	**66.0**
3.	**Gaussian Naïve Bayes**	**95.5**	**45.40**
4.	**Decision Tree Classifier**	**99.5**	**84.99**
5.	**Random Forest**	**98.0**	**88.07**
6.	**Neural Network**	**83.57**	**84.41**

4 Discussions and Conclusions:

The main observation here is that the recall value is more when we use various models on the datasetusing the SMOTE machine learning technique. The accuracy is more in the initial dataset where SMOTE ML technique is not used to add additional data records. But by using SMOTE technique, we can oversample the data and produce variant data-based results and also makes sure that the MLmodel is not biased.

AUC-ROC Curve:
The Area Under Curve is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. As we have multiple labels we are using multi-class approach (ovr approach).

Prediction Model	**SMOTE AUC-ROC Curve Score**
Logistic Regression	**0.87**
Decision Tree Classifier	**0.90**
Random Forest	**0.50**
Gaussian Naïve Bayes	**0.81**
Neural Networks	**0.77**

Future Work:

We have currently applied the Neural Networks and other predictive models. One of the main parts of the future work is to fetch more real time data related to the startups and perform more data cleaning such that the accuracies can be increased, and models can be more precise.

5. Statements of Individual Contribution:

Praveen: As the dataset is large and contains various files on startups. We first had to understand a lot of financial terms like investment rounds (pre-seed, seed, series A), IPO (Initial Public Offering), Unicorn, etc. and then understand the various files in the data set. Later we had to select the features from these files. Cleansing of data is done where all the irrelevant columns and null values are removed. I worked on the part of understanding the datasets, pre-processing the data, and then doing feature-extraction. I have combined multiple features from various files based on Object_id in objects.csv file. Later I have worked on applying Logistic Regression model and improving the accuracy on this model.

Hrithik: In the dataset there are four non orderable levels indicating the target variable 'STATUS' of each startup. The Levels of status are closed, acquired, IPO and operating. The operating level startups are a lot compared to other level startups. So, the model will be more biased towards operating status level. Always model may predict the operating status. So, to overcome this I have worked on SMOTE technique, and I have implemented Neural Networks to better implement this problem of predicting Startup's success and to improve the accuracy.

Chetana: Exploratory data analysis is the crucial process of doing preliminary analyses on data to find patterns, identify anomalies, test hypotheses, and double-check assumptions with the aid of summary statistics and graphical representations. I have worked on exploratory data analysis to find more insights in the data. After that, I have worked on Model Evaluation metrics and have implemented the ML algorithms like K-Nearest Neighbors classifier, Naïve Bayes model to predict and improve the accuracy.

Kushi: Overfitting of the data as most of the companies are in operating status. It is more likely that our machine learning algorithm predicts operating as the status most of the time which leads to a biased model. Less amount of training data will produce inaccurate or too biased predictions. I have worked on SMOTE technique. I have worked on the ML algorithms like Decision Tree and Random Forest model to predict and improve accuracy.

We'll make sure that each of us is responsible for the models we've chosen, as well as for the componentsof data collection and preprocessing, and for ensuring the integrity of the results, reports, and correctness.

6. References:

1. M.KC, "Startup success prediction," *Kaggle*, 16-Sep-2020. [Online]. Available: https://www.kaggle.com/datasets/manishkc06/startup-success-prediction.[Accessed:09-Oct-2022]

2 R. Dickinson. Business failure rate. American Journal of Small Business, 6(2):17–25, 1981.

3 Y. Xie, Z. Chen, K. Zhang, Y. Cheng, D. K. Honbo, A. Agrawal, and A. Choudhary. Muses: a multilingual sentiment elicitation system for social media data. IEEE Intelligent Systems, 99:1541–1672, 2013.

4 Francisco Ramadas da Silva Ribeiro Bento. Predicting start-up success with machine learning. PhD thesis, 2018.

5 Liang Yuxian Eugene and Soe-Tsyr Daphne Yuan. Where's the money? the social behavior of investors in facebook's small world. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pages 158–162. IEEE Computer Society, 2012.

6 Eugene Liang Yuxian and Soe-Tsyr Daphne Yuan. Investors are social animals: Predicting investorbehavior using social network features via supervised learning approach. 2013