VOL. 18, NO. 15, AUGUST 2023 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

# PREDICTIVE ANALYSIS ON STARTUP ACQUISITION STATUS

V. S. Padmavathy[1] and Gowtham Elangovan[2]
[1]Department of Computer Applications and Technology, SRM Arts and Science College, Chennai, India
[2]Data Science and Analytics, University of Hertfordshire, Hatfield, United Kingdom
E-Mail: vspadmavathy@gmail.com

**ABSTRACT**

Start-up companies are newly established companies in a survival struggle. These organisations generally start with brilliant ideas and succeed. Every year, several start-ups start with ideas, but only a very small percentage of these ideas end up lasting. Multiple factors and values affect their survival. Henceforth a concept to forecast if the start-up will be successful or not in the long run is developed. Based on an analysis and forecast, the acquisition status of a startup is determined. This would make it easier for investors to put money into a startup business. By treating the skewed data without under-sampling or oversampling, this is accomplished. To attain more accuracy and performance, a novel model is suggested.

**Keywords:** start-up, machine learning, acquisition status, oversampling.

Manuscript Received 10 June 2023; Revised 28 September 2023; Published 10 October 2023

## INTRODUCTION

Start-ups are new ventures everywhere. Due to the drastic boom, several private companies, government organisations, and even colleges also invest in these start-ups. There are a variety of possible destinations for start-ups, but the two that are most desired are either a successful acquisition from a larger company or an IPO. Statistical information like sector category and funding date is initially given as input. The results of the work let us know whether the company is acquired or it is closed and IPO. The major goal of this work is to solve the imbalanced dataset. The target class of datasets has an uneven sampling of data. Ensemble method XGBoost and QDA are applied to perform the prediction process.

Startups in particular rely on social media platforms like Twitter to build a strong brand and sustain rapid growth [1]. In recent years, Social Media Analytics (SMA) has become a crucial method for gathering and evaluating data from social media networks. It gathers, processes, and analyses Social Media (SM) data using cutting-edge analytics tools and methodologies to find meaningful trends and information [2]. Numerous studies have been conducted to forecast the success of startups at different stages of development and operation. The applicable approaches were found to be based on structured data and include tools for social network analysis, data mining, and machine learning [3].

## RELATED WORK

The objective of this research is to forecast a former start-up's acquisition status using financial data from a corporation. It's remarkable how little research there has been on the subject of utilising machine learning to predict IPO under-pricing. It's remarkable how little research there has been on the subject of utilising machine learning to predict IPO under-pricing. Since it may be enhanced to take into account the likelihood that the business will be acquired, shut down, or become public, the project results to pre-IPO businesses may be of special interest to investors and job seekers.

The findings of this work may also shed light on which characteristics have the greatest bearing on forecasts. It is challenging to examine and understand data to determine a start ups success rate because of how very uncertain and dynamic the start-up ecosystem is [4]. Using data gathered from surveys of US companies, logistic regression, a method of machine learning that has long been used to forecast economic performance, was used to forecast the success of a young firm [5]. A time-aware method in which warm-up and simulation times were assigned to the dataset was developed. Only model training was done during the warm-up session. It contained data that had been published to Crunch base between the companies' establishment and the start of the simulation window [6].

[7] Researchers used Crunchbase data to estimate acquisition or an initial public offering (IPO) for US-based companies using logistic regression, SVM, and random forest algorithms. Startup Initiatives Response Analysis (Sira) [8] was proposed to evaluate the performance of startups based on An Analytics-Based Framework. Capital Venture Exchange [9], a model based on Machine-Learning was generated to formulate the startups outcomes. The assessment was done to check whether the stars exited successfully, failed, or were in private. [10] Proposed a model for foreseeing the success of a company. Several algorithms were compared and it was seen that the Gradient Boosting Classifier showed good results.

[11] Analyses the performance of the machine learning models and assesses its progress over a 3-year time window.

## METHOD

The research involved uses financial indicators to forecast a startup's acquisition status. Avoiding under- or oversampling is necessary to overcome the primary

problem of biased data. Knowledge of the dataset. Crunchbase's "Crunchbase 2013 - Companies, Investors, etc." dataset was used for this. Figure-1 gives the Prediction Process of the startup's acquisition status.

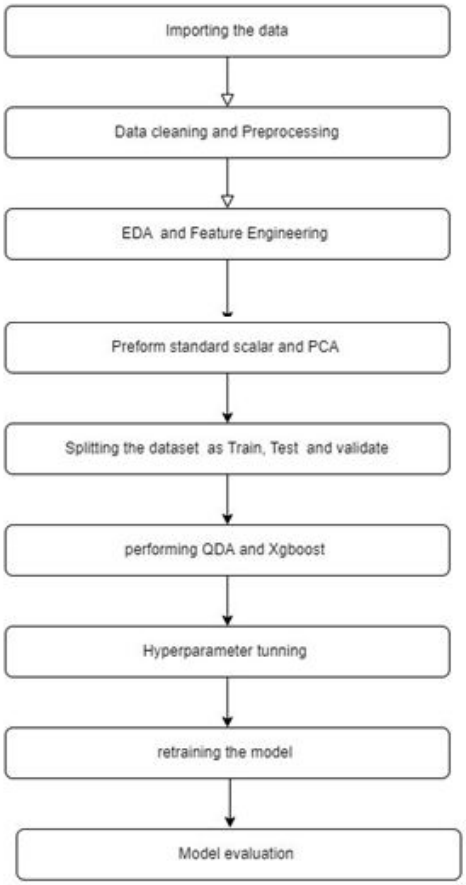


**Figure-1.** Prediction Process of startup's acquisition status.

The dataset has n = 196553 samples and each row of data includes details on a startup. Name of the company, website, sector category, amount of funds received, location of the headquarters (city and state names), financing rounds, date of creation, first and last funding dates, and most recent milestone date are all included. The status of the company ('Acquired', 'Closed', 'IPO', 'Operating') is also indicated in each row. The labels on the dataset demonstrate how seriously skewed the dataset is. The other classes are underrepresented, whereas the 'Operating' class is considerably overrepresented.

## PREPROCESSING

The dataset contains 44 columns. As they contain irrelevant information, remove the columns id, Unnamed: 0.1, entity type, entity id, parent id, created by, created-at, updated-at, domain, homepage URL, Twitter username, logo URL, logo width, logo height, short description, description, overview, tag list, name, normalised name, permalink, and invested companies. Deleted the duplicate values to filter the dataset. Over 96% of the null values in some columns, including initial investment at, last investment at, investment rounds, and ROI, were found and removed.

Delete instances with missing values for 'status', 'country_code', 'category_code', and 'founded_at'. These are the types of data where adding value via imputation will create the wrong pattern only. Checked for outliers and deleted outliers for 'funding_total_usd' and 'funding_rounds'. Converted qualitative data to quantitative data as part of feature extraction. The derived columns are active days and are isclosed.

## EDA AND FEATURE ENGINEERING

- **merged_data** dataset comprises 196553 rows and 44 columns.

- Dataset comprises continuous variable and float data type.

**Information of Dataset:** Using scatterplot found that there is no correlation between funding_total_usd and relationships and also between milestones and relationships.

Using barplot between status and funding_total_usd, it is clear that funding_total_usd is higher for IPO status. Using barplot between status and milestones, it is clear that milestones are higher for IPO status. Using countplot on target variable **Status** we could see that Label 0 has '453' values, Label 1 has '6000', Label 2 has 90, and Label 3 has 936. By this information, we could conclude that there is an imbalance in the data and hence balancing of data is required.

Using counterplot on target variable **is closed** we could see that Label 0 has '1389' values, Label 1 has '6090'. By this information, we could conclude that there is an imbalance in the data and hence balancing of data is required. Generalised the country and state columns and performed one hot encoding for country and state columns.

## UNIVARIATE ANALYSIS

By plotting the distplot it is evident that funding_total_usd, active days are right skewed.

### Descriptive statistics

Using **describe()** we could get the following result for the numerical features

```
funding_rounds
funding_total_usdmilestones
relationships      lat    lng
count  22889.000000  2.046700e+04  35249.00000
48306.000000  61219.000000  61219.000000  mean
1.805758  1.582132e+07  1.41587  4.442926  37.293151  -
50.708830  std  1.310805  6.990693e+07  0.73856
13.266474  15.812771  70.783600  min  1.000000
2.910000e+02  1.00000  1.000000  -50.942326  -159.485278
25%  1.000000  5.110380e+05  1.00000  1.000000
34.052234  -112.028750  50%  .000000  2.725875e+06
```

1.00000 2.000000 39.739236 -75.898684 75% 2.000000 1.200000e+07 2.00000 4.000000 45.417979 1.801799 max 15.000000 5.700000e+09 9.00000 1189.000000 77.553604 176.165130

Created a cluster with lat and lng columns but there is no significance in the mutual information score hence removed these columns.

**Correlation plot of numerical variables**

All the continuous independent variables are not much correlated with each other hence there is no multicollinearity in the dataset. Before modelling and after splitting we scaled the data using standardization to shift the distribution to have a mean of zero and a standard deviation of one

**Fit transform()** is used on the training data so that we can scale the training data and also learn the scaling parameters of that data. Here, the model built by us will learn the mean and variance of the features of the training set. These learned parameters are then used to scale our test data.

**Transform()** uses the same mean and variance as it is calculated from our training data to transform our test data. Thus, the parameters learned by our model using the training data will help us to transform our test data. We do not want to be biased with our model, but we want our test data to be a completely new and a surprise set for our model.

**PCA transformation**

We reduced the 5 features to only 4.

```
fromsklearn. Decomposition import PCA
pca = PCA(n_components=4) pca.fit(X_train) trained = pca.transform(X_train) transformed = pca.transform(X_train)
```

**Model building**

**Metrics considered for model evaluation**

**Accuracy, precision, recall, and f1 score**

- **Accuracy:** What proportion of actual positives and negatives is correctly classified?
- **Precision:** What proportion of predicted positives are truly positive?
- **Recall:** What proportion of actual positives is correctly classified?
- **F1 Score:** Harmonic mean of Precision and Recall

**MODEL BUILDING**

**Xg boost**

In this analysis, there are two dependent variables ('status' and 'is closed'). QDA is used where isclosed is taken as a dependent variable and Xg boost uses status as a dependent variable. When we apply the Xgboost model the accuracy is 94% (will say the exact numbers) and when we apply Quadratic discriminate analysis the accuracy is 85%.

**RESULTS AND DISCUSSIONS**

The results clearly show that the accuracy score of xg is 1.0 and the accuracy score of Random forest is 0.9024. A comparison of XGBoost and Random Forest depicts that XGBoost is better when compared to Random Forest. The Start-up's acquisition status based on its financial statistics is analysed and predicted more accurately using the XG Boost method. Henceforth this comparison would be very useful in predicting the start-up's acquisition status.

```
In [64]: class_report=classification_report(y_test, pred_xg)
         print(class_report)
         print('Accuracy Score -xg:', metrics.accuracy_score(y_test, pred_xg))

                       precision    recall  f1-score   support

                    0       1.00      1.00      1.00        84
                    1       1.00      1.00      1.00      1239
                    2       1.00      1.00      1.00        16
                    3       1.00      1.00      1.00       157

             accuracy                           1.00      1496
            macro avg       1.00      1.00      1.00      1496
         weighted avg       1.00      1.00      1.00      1496

         Accuracy Score -xg: 1.0
```

**Figure-2.** Accuracy score of XG boost.

```
In [97]: class_report=classification_report(y_test, pred)
         print(class_report)
         print('Accuracy Score -rf:', metrics.accuracy_score(y_test, pred))
```

```
              precision    recall  f1-score   support

           0       0.61      0.33      0.43       168
           1       0.93      0.97      0.95      2402
           2       0.75      0.09      0.17        32
           3       0.80      0.83      0.81       390

    accuracy                           0.90      2992
   macro avg       0.77      0.55      0.59      2992
weighted avg       0.89      0.90      0.89      2992

Accuracy Score -rf: 0.9024064171122995
```

**Figure-3.** Accuracy score of random forest.

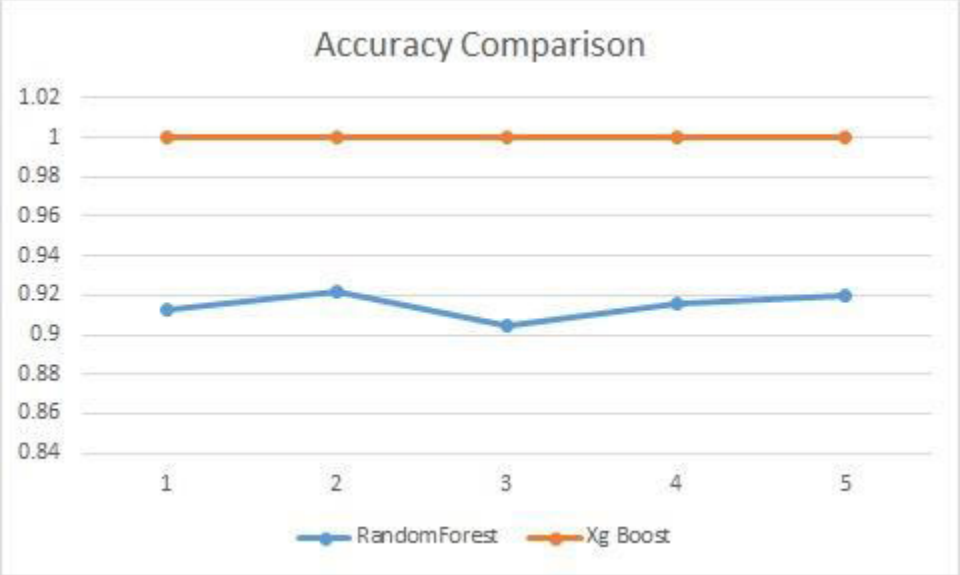


**Figure-4.** Comparison of random forest and XG boost.

## CONCLUSIONS

To retain a company presence, large organisations frequently have fixed budgets and labour forces that make it simple for them to keep pace with industry developments and advancements. However, startups typically have a restricted budget, which makes it difficult for them to monitor the success of their business operations. Start-up's acquisition status based on its financial statistics is analysed and predicted. By addressing the biased data without under sampling or oversampling, a novel model is proposed to achieve a higher accuracy and performance. The accuracy score of the XG Boost is better when compared to the Random forest method.

## REFERENCES

[1] S. Lugovi and W. Ahmed, "An analysis of Twitter usage among startups in Europe", 2015.

[2] W. He, H. Wu, G. Yan, V. Akula and J. Shen. 2015. A novel social media competitive analytics framework with sentiment benchmarks. Inf. Manage. 52: 801-812.

[3] T. Antretter, I. Blohm, and D. Grichnik. 2018. Predicting startup survival from digital traces: Towards a procedure for early stage investors. in Proc. Int. Conf. Inf. Syst. (ICIS), San Francisco, CA, USA, [Online].

[4] Edoc-Server Open Access publication server of the Humboldt University.

[5] 2021. A machine learning, bias-free approach for predicting business success using Crunchbasedata, Kamil Żbikowski, Piotr Antosiuk, Information Processing & amp; Management. 58(4): 102555.

[6] Arroyo J., Corea F., Jimenez-Diaz G., Recio-Garcia J.A. 2019. Assessment of machine learning

performance for decision support in venture capital investments IEEE Access. 7: 124233-124243.

[7] Bento F. R. D. S. R. 2018. Predicting start-up success with machine learning. Universidade Nova de Lisboa.

[8] Bashayer Alotaibi, Rabeeh Ayaz Abbasi, Muhammad Ahtisham Aslam, Kawther Saeedi, and Dimah Alahmadi. Startup Initiative Response Analysis (Sira) Framework for Analyzing Startup Initiatives on Twitter. IEEE Access, Digital Object Identifier 10.1109/Access.2020.2965181.

[9] Greg Ross, Sanjiv Das, Daniel Sciro, Hussain Raza. 2021. Capitalvx: A Machine Learning Model for Startup Selection and Exit Prediction. The Journal of Finance and Data Science. 7: 94-114.

[10] Kamil Żbikowski, Piotr Antosiuk. 2021. A Machine Learning, Bias-Free Approach For Predicting Business Success Using Crunchbase Data. Information Processing & Management. 58(4): 102555.

[11] Javier Arroyo, Francesco Corea, Guillermo Jimenez-Diaz and Juan A. Recio-Garcia. 2019. Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. IEEE Access. 7: 124233-124243.