Prediction of the Success of a Startupusing Ensemblemethods and Margin Loss

Kadambri Agarwal

Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, India

kadambri_agarwal@rediffmail.com

Corresponding Author

Aditi Mittal

Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, India

aditim262002@gmail.com

Sachin Kumar

Department of Computer Science and Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, India

imsachingupta@rediffmail.com

Vinita

Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, India

vinita89.cse@gmail.com

All the Authors Have No Conflict of Interest

Abstract:

A startup is a new company started by a person or a group of people to solve a real-world problem uniquely. Startups have dynamic economics and initially require tremendous effort, time, and money. Along with this venture, capitalists have a high amount of risk while investing their money. So, there is necessary to find essential factors that help the startup's success.We propose a machine-learning-based solution for this challenge by keeping a data set of 923 rows and 45 columns, and ensemble methods usingLGM Classifier, XGBoosts, and AdaBoost Classifier are used for the training, while SVM is used for classification instead of the traditional softmax function.

Furthermore, we use margin loss instead of a standard entropy-based algorithm for the loss function, while SMOTE and Tomek's links were used for balancing the data as our dataset is imbalanced. Finally, we compute accuracy, precision, and F1 score,significantly improving various existing models. Our approach produces 90.2 accuracies.

Keywords: Startups, LGM Classifier, XGBoosts,AdaBoost Classifier, SMOTE and Tomek Links, and Margin loss

Statements and Declarations

All the Authors Have No Conflict of Interest

1. Introduction:

With the growing trend of startup culture [1], many youngsters nowadays either leave their job to start new businesses or try to convert their innovative ideas into businesses during their college years.So now, why the youth of today are getting motivated toward entrepreneurship? There are many reasons, some of which are, firstly, many colleges have started helping students implement their ideas by giving them a place to set up, or the startup can provide internships or job opportunities to students within the college. The second reason could be the government has launched various policies to help startups in their initial phase. Also, many youngsters nowadays do not like working under a boss like in the conventional system, so the ones who have some new idea or want to change the traditional way customers are being served toend up starting their startups.

But sadly, only 2 of 5 [2] startups canprofit. Other startups either break even or continue to lose money. Some of the reasons for the failure of a startup are there is "no market need" for that product, the team build was not right or was not motivated enough, or sometimes the marketing strategy applied was not good. There can be many other reasons. In such conditions, it becomes difficult for venture capitalists (VCs) to invest. Also, initially, startups could primarily raise funds from local VCs, butdue to Covid-19 now, more VCs are willing to raise funds outside their geographical area leading to a lot of competition amongst VCs to invest in a better startup than their competitor.

Business failure predictions [3] are made by three primary methods, which are as follows:

1) Theoretical Modelling – Researchers use this type of modeling to find the structure of the studying process and how to approach the problem.
2) Finding the set of explanatory variables using accounting and financial ratios.
3) Using machine learning models.

But recently, the third method has become very popular with the improvement of computer technology, and we no longer have to worry about whether the particular statistic model is feasible and can be applied to the model.

But from our research, we have also found that using a machine learning algorithm alone can sometimes lead to misleading information because of either overfitting or underfitting the data. Sometimes it would be better to collectively judgethe human mind, statistics, and machine learning models. Another challenge is balancing data, as most data sets are imbalanced. We observe that balancing the data set is crucial to get a good performance of the model. Therefore, we need a model that removes such problems.

In this paper, our contributions are as follows:

- Processing of the model and balancing of the data set using the SMOTE, and Tomek Link
- We take an ensemble model to avoid overfitting by ensemblingthree models, LGM Classifier, XG Boosts, and AdaBoost Classifier and using lasso regression
- We usemargin loss as a loss function instead of a traditional function like an entropy

- Finally, we applied SVM for the classification instead of softmax.

2. Related Work:

Ünal et al.[4] have triedsix models to determine which model performs best. The six models are Logistic Regression, Recursive Partitioning Tree (Rpart), Conditional Reference Tree, Random Forest, and Extreme Gradient Boosting. And they found out that the best-performing models are XGB, Random Forest, and Recursive partition tree. Also, the three main variables which they are helpful are the last funding date, the first funding date, and company age. But apart from this, they also claim that according to some people, human instinct, gut feeling, and previous experience of the decision-makers can bring out the best decision. But, at the same time, other sets of people believe that when all valuable and irrelevant data is sent to the decision-makers, it is difficult to process so much information and make the best decisions.Ghassemi M. M et al. [5] have applied six models: Decision Tree, Discriminant Analysis, Logistic Regression, Support Vector Machines, k-Nearest Neighbours (k-NN), Ensemble Learning, and Neural Network. The first best-performing model was Logistic regression. The second best-performing model was Linear SVM.

Te, Y. F. et al. [6] used two datasets: one from "Crunchbase" and the second from LinkedIn. They have used two methods of training data first, by using LinkedIn data as a "standalone," and second, by using it as complementary data to "Crunchbase data." In the Baseline-1 model, they used the full Crunchbase dataset; in Baseline-2, they used a small part of the Crunchbase dataset. After their study, they found that only using the LinkedIn dataset was its worst performance among all cases. However, using both Crunchbase and LinkedIn datasets together had a better performance than the baseline-2 model, and the baseline-1 model had the best performance. So, they suggested using the LinkedIn dataset in addition to the Crunchbase dataset to improve the performance of the resulting model, and the AUC value went up to 84%. Baskoro H. et al. [7] have applied Logistic Regression, Random Forest, and Support vector machine for the analysis. These algorithms consider different dependent variables. For example, when "The startup already has an IPO or M&A (merger and acquisition)" was taken as a dependent variable, the accuracy of logistic regression and support vector machine came out to be 92 %, and for the random forest, it was 93%. Whereas when "Total funding above 1 Million Euro" was taken as the dependent variable, the accuracy came out to be 71% in the case of logistic regression.

Sevilla-Bernardo[8] et al. observed the four main findings. First, they conducted a literature survey of 60 articles and found the intersection between business practice and scientific research on entrepreneurship. Secondly, they have discovered seven core factors from their literature survey as significant factors for the success of a startup. They are "Idea,""CEO Leadership,""Business Model,""Marketing approach," and "Entrepreneurial Team." Also, cultural factors affect the weight given to various factors. Their third point says that, through statistical analysis, two additional factorsjeopardize a startup's success: CEO Decisions and Marketing. The fourth contribution of their paper is based on geographical location. They say that "Idea" is the most critical factor irrespective of geographical location. And rest of the elements depend on the geography of the place.
David D. et al. [9] claimed that India had become the third-largest startup ecosystem in the world. With an estimate of 26000 startups, out of which 26 are unicorns. This rise is mainly because of private investments increasing. The government is also trying to stimulate the process by getting flagship startup India initiatives. But from studies, it can be seen that most

of these startups are located in tier-1 cities, and there is less information about these government policies in tier-2 and tier-3 cities. Many Indian startups face challenges regarding a fragmented market, lack of knowledge and exposure, etc. They have concluded that the government should focus on nurturing top-notch talent and impart global business skills by "reverse braindrain." Also, the government should provide some relief to startups trying to implement macroeconomic policies. Many startups have become unicorns because they focus on quantity rather than quality to disrupt the market. But zebra startups require more incubation support to expand beyond those overcrowded categories of startups. Lastly, the government should protect new startups in their initial phase so they do not have to face international competition and can excel without being acquired.

Tripda Rawal et al.[10] explored the reason behind the inorganic growth of startups in India and the significant contribution of startups to the economy. Startups play a very prominent role, and startups in India enjoy a perfect and supportive environment. Inorganic growth is becausehighly skilled youngsters with high proficiency are interested in building their businesses. Along with that, many Indian startups have started getting massive funding nationally and internationally. But the startups can rise even more if they take the help of incubation and acceleration centers. Greg Ross et al. [11] have used data from two sources, first from Crunchbase and second from US Patent Office (USPTO), and produced output for IPO, acquired, failure (which is a three-way model), or remain private (i.e., four-way model). After applying Deep Learning, XGBoost, Random Forests, and KNN for training. The accuracy for training the three-way model was 90%, and the four-way model was 80%. Their best accuracy for the funding model was produced using ensemble, i.e., 0.88, and their best exit model accuracy was predicted using XGBoost, i.e., 0.894

Kamil Żbikowskiet al. [12]used a publicly available Crunchbase data set and applied logical regression, support vector machine, and XGBoost for comparison. The best outcome was produced using XGBoost. Their best training accuracy came out to be 0.86 from logical regression and XGBoost both. And best testing accuracy was 0.86 from logical regression. Ulrich Kaiseret al. [13] have used publicly available data from the university of Danish. By simply running logical regression, they have found that "survival,""employment growth,""patenting activities," and some more factors can be predicted using publicly available data. Their model only requires basic textual information that startups must give during registration. However, they say that initial firm size, initial patents, and word score index can help improve the accuracy of startups.
As we observed various research, none of the studies pointed out the small data set, and their performance is not upto mark as balancing of data was not carried out, and none of the papers used marginloss as a loss function as it has its several advantages.

3. Proposed Model:

Our proposed approach includes data collection, preprocessing steps, balancing of data, and training of the models using ensembling approaches. The margin loss function is used instead of standard methods like entropy for loss computation, while the SVM is used for the classification in place of softmax as SVM has the fast training capability**.** Fig 1 shows the architecture of our proposed approach.

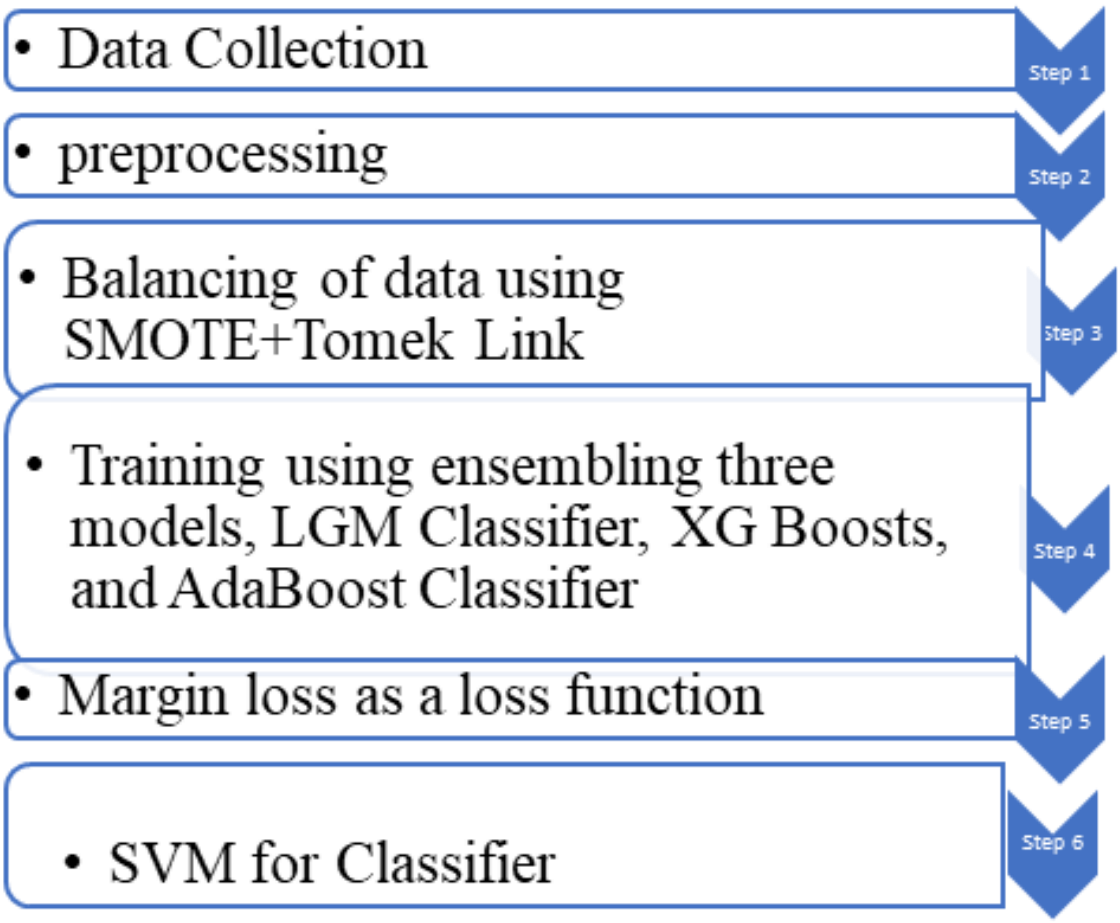

Fig 1: Architecture of our Proposed model

3.1 **About the dataset:**
The dataset used in this research paper is taken from Kaggle, which consists of 923 rows (923 companies) and 45 columns. All the companies are categorized as "acquired" or "closed." So, we have taken companies acquired by others as successful and closed as failed.The columns in the dataset are'Unnamed: 0', 'state_code,' 'city,' 'Unnamed: 6', 'name,' 'labels,' 'founded_at,' 'closed_at,''first_funding_at''last_funding_at,"age_first_funding_year,"age_last_funding_year ,"age_first_milestone_year,' 'age_last_milestone_year,' 'relationships,' 'funding_rounds,' 'funding_total_usd,' 'milestones,' 'state_code.1','is_CA,' 'is_NY,' 'is_MA,' 'is_TX,' 'is_otherstate,' 'category_code,' 'is_software,' 'is_web,' 'is_mobile,' 'is_enterprise,"is_advertising,' 'is_gamesvideo,"is_ecommerce,"is_biotech,"is_consulting,' 'is_othercategory', 'object_id', 'has_VC', 'has_angel', 'has_roundA', 'has_roundB', 'has_roundC', 'has_roundD', 'avg_participants', 'is_top500', 'status'.

The dataTypeof all the columns can be seen in fig 2:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 923 entries, 0 to 922
Data columns (total 45 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Unnamed: 0                923 non-null    int64
 1   state_code                923 non-null    object
 2   city                      923 non-null    object
 3   Unnamed: 6                430 non-null    object
 4   name                      923 non-null    object
 5   labels                    923 non-null    int64
 6   founded_at                923 non-null    object
 7   closed_at                 335 non-null    object
 8   first_funding_at          923 non-null    object
 9   last_funding_at           923 non-null    object
 10  age_first_funding_year    923 non-null    float64
 11  age_last_funding_year     923 non-null    float64
 12  age_first_milestone_year  771 non-null    float64
 13  age_last_milestone_year   771 non-null    float64
 14  relationships             923 non-null    int64
 15  funding_rounds            923 non-null    int64
 16  funding_total_usd         923 non-null    int64
 17  milestones                923 non-null    int64
 18  state_code.1              922 non-null    object
 19  is_CA                     923 non-null    int64
 20  is_NY                     923 non-null    int64
 21  is_MA                     923 non-null    int64
 22  is_TX                     923 non-null    int64
 23  is_otherstate             923 non-null    int64
 24  category_code             923 non-null    object
 25  is_software               923 non-null    int64
 26  is_web                    923 non-null    int64
 27  is_mobile                 923 non-null    int64
 28  is_enterprise             923 non-null    int64
 29  is_advertising            923 non-null    int64
 30  is_gamesvideo             923 non-null    int64
 31  is_ecommerce              923 non-null    int64
 32  is_biotech                923 non-null    int64
 33  is_consulting             923 non-null    int64
 34  is_othercategory          923 non-null    int64
 35  object_id                 923 non-null    object
 36  has_VC                    923 non-null    int64
 37  has_angel                 923 non-null    int64
 38  has_roundA                923 non-null    int64
 39  has_roundB                923 non-null    int64
 40  has_roundC                923 non-null    int64
 41  has_roundD                923 non-null    int64
 42  avg_participants          923 non-null    float64
 43  is_top500                 923 non-null    int64
 44  status                    923 non-null    object
dtypes: float64(5), int64(28), object(12)
memory usage: 324.6+ KB
```

Fig 2: Data Type

3.2 Data Preprocessing:

In preprocessing data, we manage null values and balance the imbalanced data.

3.2.1Null Values As this is a large amount of data with a different datatype and some null values, we first need to preprocess the data. So, for preprocessing, we can first see that there are two columns,"label" and "status," which look almost similar, so after checking, we found out that all of their values are the same, so we drop the "labels" column. Similarly,"state_code" and "state_code.1" are also the same, so we drop "state_code.1". Then we study columns' Unnamed: 0', 'Unnamed: 6', and 'object_id' and find out that all of them have all unique values so that no pattern can be found, so we drop them. Then we check all the null values in the data which is shown in fig 3

```
data.isna().sum()
```

```
state_code                  0
city                        0
name                        0
founded_at                  0
closed_at                 588
first_funding_at            0
last_funding_at             0
age_first_funding_year      0
age_last_funding_year       0
age_first_milestone_year  152
age_last_milestone_year   152
relationships               0
funding_rounds              0
funding_total_usd           0
milestones                  0
is_CA                       0
is_NY                       0
is_MA                       0
is_TX                       0
is_otherstate               0
category_code               0
is_software                 0
is_web                      0
is_mobile                   0
is_enterprise               0
is_advertising              0
is_gamesvideo               0
is_ecommerce                0
is_biotech                  0
is_consulting               0
is_othercategory            0
has_VC                      0
has_angel                   0
has_roundA                  0
has_roundB                  0
has_roundC                  0
has_roundD                  0
avg_participants            0
is_top500                   0
status                      0
dtype: int64
```

Fig 3: Description of Null Values

From the above, we can see many null values in "age_first_milestone_year" and "age_last_milestone_year." But we can see that companies with no milestone year have values blank in "age_first_milestone_year" and "age_last_milestone_year." So, remove the empty values by filling in 0 in place of null values.

Now, null values in the "closed_at" column are replaced by the last date of the "closed_at" column. We can also see that the "found_at,""first_funding_at," and "last-funding_at" columns are in obj format, but we want them in date and time format, so we change their data type.

Then we change the "status" column, which has a string (acquired, closed) to 1 and 0. Also, we make a new column,"age," which is a difference betweenthe "closed_at" year and "founded_at" year. Then we find an error in the data for some companies, and age is coming negative, which is impossible, so we delete those rows.

3.2.2 Balancing of data using SMOTE+Tomek Link

Now When we deal with binary classification problems, the dataset we use is imbalanced. In an imbalanced dataset, the number of values for majority classes is enormous compared to

the minority dataset. So, to improve the model's accuracy, we have to deal with this imbalanced dataset by either undersampling it or oversampling it. So, in this paper, we will try to solve this problem by the Hybridization method, i.e., SMOTE + Tomek Link. The studies related to SMOTE + Tomek Link are mentioned in [13, 15]. In this method, we try to combine both undersampling and oversampling techniques. SMOTE stands for synthetic Minority Oversampling Technique, which solves the problem of overfitting by generating new samples of minority classes. On the other hand, Tomek Link is used to remove the data that falls near the minority classes.

For implementing SMOTE- Tomek Link technique, we first follow the procedure of SMOTE and then the Tomek Link.

For SMOTE analysis, we first make a set of Minority classes. For example, let us say "X." Now suppose we have a value "X1" such that "X1 € X". After that, we find the k nearest neighbor of X1 with every other value in set "X"with the help of the Euclidean distance formula.

$$d(X1, X2) = \sqrt{(X1 - X0)^2 + (X2 - X0)^2} \quad -1$$

Now we set a sampling rate N according to imbalance proportion, and a new set "X_Sub" is made by randomly selecting N values from the k-nearest neighbor.

If there is a value "X_1" such that "X_1 € X_Sub," then we use the following formula to generate synthetic samples:

$$X\_New = X1 + rand(0,1)*|Xi-X\_i| \quad -2$$

Over here, (0, 1) is used for any random number between 0 and 1.

After that, for Tomek Link, we randomly choose data from the majority classes, and if it falls near the boundary of the minority class, then we create a Tomek Link. The rules used to make the Tomek Link are as follows:

If "x" is the sample belonging to the minority class and "y" is the sample belonging to the majority class. Then there should be no "K" such that the following condition holds:

1) $d(x,k)<d(x,y)$ or -3
2) $d(y,k)<d(x,y)$ - 4

4. Training Models:

4.1 Regularization

When we have a small dataset, overfitting is one of the biggest challengesin training the model. Regularization [16,17] is one of the leading techniques to prevent the model from overfitting by adding some extra information while training the model. Sometimes the model performs well with the training data, but similar results are impossible during the testing. It means that the model performed not upto mark as we have seen in similar studies. Therefore, in our research, we have included Lasso Regression [16,17] and observed that our model

performs well for unseen data after incorporating the Lasso. It mainly regularizes the coefficient of features toward zero.
The working of Lasso Regression is defined as :

$$\sum_{i=1}^{r}(y_i - y_i^{'})^2 = \sum_{i=1}^{r}\left(y_i - \sum_{j=0}^{s} B_j * x_{ij}\right)^2 + l\sum_{j=0}^{s} pos(B_j) \quad -5$$

In the above equation, if the values of l tend to zero, the equation becomes the cost function of the linear regression model, and the pos function returns a positive value. Therefore, we can minimize the coefficients canreduce the overfitting.

4.2 Margin loss as cost function

One of the popular loss functions, cross-entropy, is used to train the model due to its simplicity and excellent performance, but it does not encourage discriminative learning of features [15]. Therefore, as we need more robust features for learning, the ensemble model should have more discriminative information. For this goal, we consider margin lossinspired by the loss function proposed by [19] in his work. Our model outperforms if we take margin loss as a loss function.
Thedescription of the margin loss function is given by:
They set the margin empirically. Eq. (6) shows the margin loss function in which y is a true-label vector, y' is the predicted value, and n is an empirically set parameter.

$$\text{Loss} = [y * \max(0,(0.9 - y'))^2] + [n * (1 - y) * \max(0,(y' - 0.1))^2] \quad -6$$

4.3 Classifiers

In the classification work, we ensemble LGBM Classifier, XG Boosts, and AdaBoost Classifier, and classification work was carried out using SVM in place of the softmax function.

4.3.1 LGBM Classifier:

LGBM Classifier [20,21] (Light Gradient Boosting Classifier) works on the principle of a decision tree to increase efficiency while at the same time reducing memory usage. Primarily it has two variants, Gradient-based and Exclusive Feature Bundling (EFB).
We define verdict trees for memorizinga function, for instance, from the input space (X)tothe gradient space (G). A training set, x_1,$x_{2,}$ and up to x_m is given in the form of a vector of d dimensions in given space X. In this approach, all the negative gradients of a loss function are denoted as g_1, g_2, and up to g_m. The decision tree used in the LGBM classifier divides the node at the most revealing feature. In this model, the data enhancement can be calculated by the variance after isolating and denoted as

$$Y=\text{Inital\_tree}(X)-lr*T_1(X)-lr*T_2(X)-lr*T_3(X) \quad -7$$

4.3.2 Xgboost:

In Xgboost[22] approach, weights are assigned to all the variables, which are then used by the decision tree to predict the result. Now the eight of the variable that was predicted wrong increase, and it then sent to the second decision tree. In this model, each variable has a specific weight, which is supplied to the decision tree to obtain the results. The prediction scores of individual decision treesare given by

$$y = \sum_{i=1}^{m} f_i \in F \quad \text{-8}$$

4.3.3 AdaBoost Classifier:

AdaBoost [23,24], or Adaptive Boosting, is based on the Ensemble method. This also uses a decision tree that only splits to one level. They are also called decision stumps. This will build a model in which all data points are given equal weight. After that,a higher weight is assigned to the wrong classifiers, and then points with higher points are given more importance in the next model. It keeps on doing this until the lowest error is received.

4.3.4 Support Vector Machine Classifier (SVM)

We use SVM for the Classifier. The reason for selecting SVM [25] in place of the traditional Classifier, softmax, is that SVM outperforms softmax for our small data set.SVM has faster learning than softmax. As our data set is small, we need faster learning algorithms, motivating us to choose SVM. The SVM createsa decision boundary, which is used for classification, that can separate n-dimensional space into various classes. This best decision boundary is termed a hyperplane. SVM picks the extreme points, called support vectors, for making the hyperplane.

The hyperplane equation can be defined as

$$H: w^T(x) + b = 0 \quad \text{-9}$$

Where b is bais and $w^T$is defined as the weight of feature x. We need to compute the distance for the equitation of $ax + by + c = 0$ from a fixed point;for example,(x_0 , y_0) is given by d while the distance parameter of hyperplane equation: $w^T\Phi(x) + b = 0$ from vector $\Phi(x_0)$ can be written as :

$$d_H\big(\emptyset(x_0)\big) = \frac{w^T(\emptyset(x_0))+b}{w} \quad \text{-10}$$

Result:

We have trained the model by taking the following import features, and the selection of the features is based on the p-value used to compute statistically significant features. We have identified the following essential features: the age of a startup, milestones achieved, tier relationship, if it has investors, it is in the top 500, if it has any round of funding, and whether it has seed funding or not into consideration. These features are mentioned in fig 4:

	cols	fea_imp
3	age	984
5	milestones	507
0	tier_relationships	323
2	has_Investor	287
4	is_top500	96
6	has_RoundABCD	77
7	has_Seed	51

Fig 4: Statistically Significant Feature based on the p-value

The training accuracy of our model is given by 90.2, while the same for [11] is 89 and [12] is 86.

The precision, recall, and f1-score of used ML algorithms and our proposed work are mentioned in table 1. Table 1 shows that ensemble methods based on margin loss entropy outperform other classifiers and give 90.2% accuracy, while Precision, Recall, and F1 scores are also provided by 84%, 74%, and 79%, respectively, also outperforming others.

	Precision	Recall	F1 Score	Accuracy
LGBM Classifier	0.76	0.68	0.72	0.85
XGBoost	0.72	0.64	0.68	0.83
AdaBoost Classifier	0.79	0.66	0.72	0.84
Ensemble of LGBM, XGBoost, and AdaBoost using entropy	0.82	0.71	0.76	0.87
Ours (Ensemble of LGBM, XGBoost, and AdaBoost using Margin loss)	0.84	0.74	0.79	0.902

Table 1: Important Results of our computation

Fig 2 gives the accuracy of our proposed model and the accuracy of [11] and [12]. The table also indicates that our proposed approach is better than [11] and [12]. However, [11] and [12] were used on different data set. The [11] gives 89% while [12[gives 86%. In our case, it is 90.2%

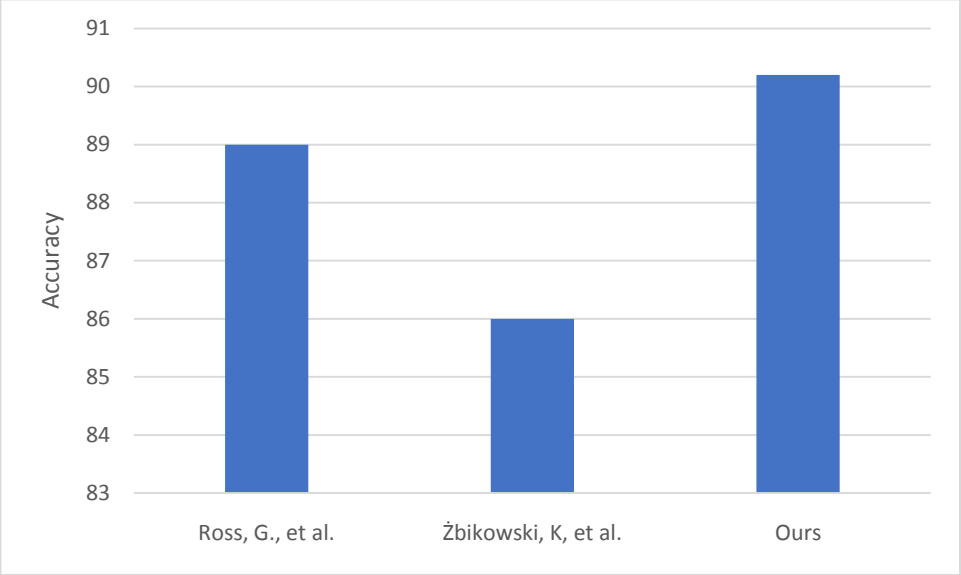

Fig 5: Accuracy of our and other models keeping different data sets

Fig 6 shows our proposed model's accuracy and [11] and [12] on the same data set. The table also indicates that our proposed approach is better than [11] and [12]. In our case, it is 90.2%, [11] gives 86.2%, and [12] offers 85.6%, further enhancement.

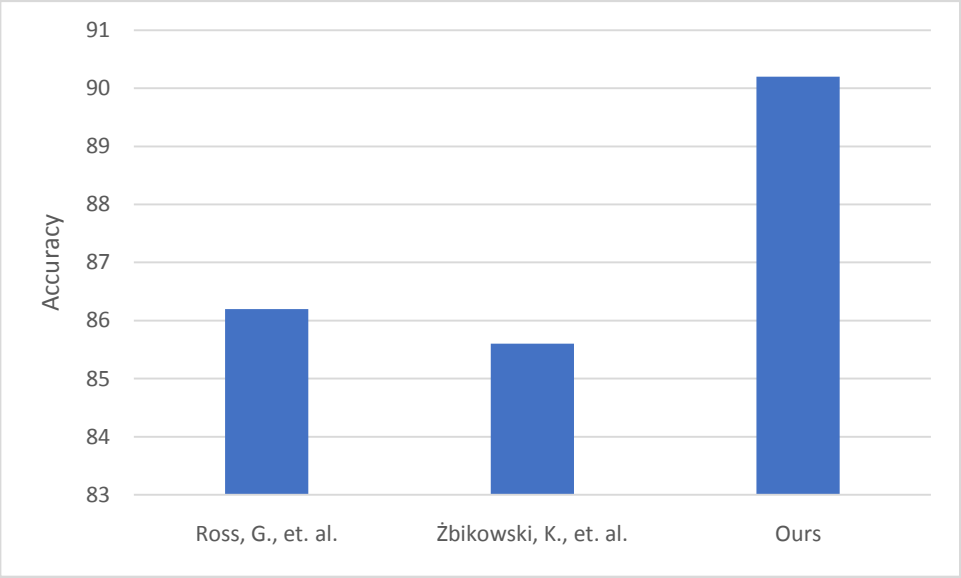

Fig 6: Accuracy of our and other models keeping the same data set

Conclusion:

In this study, we explored different factors that help a startup's success and decided which machine learning model would give the best result for our data set.We also explored various features responsible for the startup's success based on the p-value. First, we balanced the data using SMOTE+ Tomek links as the preprocessing step because the data set was imbalanced. Then, we analyzed the results using LGBM, XGBoost, and AdaBoostand ensembled the

three. For the classification, we have used SVM in place of softmax as our data set is small, and the SVM training rate is faster than softmax. In addition, we have used margin loss instead of entropy, as entropy-based loos functions as certain limitations. Finally, we have computed the efficiency, precision, recall, and f1 scoreand compared our results with the existing state-of-the-art algorithms, showing that our approach outperforms many current algorithms.The accuracy of our model was 92%. Hence, this paper concludesessential features ofthe startup's success and can better predict the success of startups.

Reference:

1. Kalpana Sinha, Rising Trend of Startup Culture And Entrepreneurship, India Education Diary.com, December 18, 2022
2. Matt Mansfield, Startup Statistics – The Number you Need to Know, Small Business Trends, March 28, 2019
3. Sinan Aktan, "Application of machine learning algorithm for business failure prediction", Investment Management and Financial Innovations, June 2011, 8(2).
4. Ünal, Cemre; Ceasu, Ioana, A Machine Learning Approach Towards Startup Success Prediction, IRTG 1792 Discussion Paper, No. 2019-022, HumboldtUniversität zu Berlin, International Research Training Group 1792 "High Dimensional Nonstationary Time Series", 2019, Berlin.
5. Ghassemi, M. M., Song, C., &Alhanai, T. The automated venture capitalist: Data and methods to predict the fate of startup ventures. *Association for the Advancement of Artificial Intelligence*. (2020).
6. Te, Y. F., Wieland, M., Frey, M., Pyatigorskaya, A., Schiffer, P., &Grabner, H. Predicting the Success of Startups Using Crunchbase and Linkedin Data. *Available at SSRN 4217648*.
7. Baskoro, H. ., Prabowo, H. ., Meyliana, M., &Lumban Gaol, F. . (2023). Predicting Startup Success, a Literature Review. *International Conference on Information Science and Technology Innovation (ICoSTEC)*, *1*(1), 123–129. https://doi.org/10.35842/icostec.v1i1.6
8. Sevilla-Bernardo, J.; Sanchez-Robles, B.; Herrador-Alcaide, T.C. Success Factors of Startups in Research Literature within the Entrepreneurial Ecosystem. *Adm. Sci.* **2022**, *12*, 102. https://doi.org/10.3390/admsci12030102
9. David, D., Gopalan, S., & Ramachandran, S. The startup environment and funding activity in India. In *Investment in Startups and Small Business Financing*, (2021), Asian Development Bank Institute. (pp. 193-232). Accessed on December 1, 2022. https://www.adb.org/sites/default/files/publication/612516/adbi-wp1145.pdf
10. Tripda Rawal, "An inside view in the Indian Startups," International journal of creative research thoughts, February 2018, Vol 6, issue 1, ISSN: 2320-2882
11. Greg Ross, Sanjiv Das, Daniel Sciro, Hussain Raza, CapitalVX: A machine learning model for startup selection and exit prediction, The Journal of Finance and Data Science, Volume 7,2021,PP 94-114,ISSN 2405-9188.
12. Kamil Żbikowski, Piotr Antosiuk, A machine learning, bias-free approach for predicting business success using Crunchbase data, Information Processing & Management,Volume 58, Issue 4,2021,102555,ISSN 0306-4573.

13. Ulrich Kaiser, Johan M. Kuhn,The value of publicly available, textual and non-textual information for startup performance prediction, Journal of Business Venturing Insights,Volume 14,2020,e00179,ISSN 2352-6734.

14. Ahmed Arafa, Nawal El-Fishawy, Mohammed Badawy, Marwa Radad, RN-SMOTE: Reduced Noise SMOTE based on DBSCAN for enhancing imbalanced data classification, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 8, Part A, 2022, Pages 5059-5074, ISSN 1319-1578

15. K. Cheng, C. Zhang, H. Yu, X. Yang, H. Zou and S. Gao, "Grouped SMOTE With Noise Filtering Mechanism for Classifying Imbalanced Data," in *IEEE Access*, vol. 7, pp. 170668-170681, 2019, doi: 10.1109/ACCESS.2019.2955086.

16. Y. Kim, J. Hao, T. Mallavarapu, J. Park and M. Kang, "Hi-LASSO: High-Dimensional LASSO," in IEEE Access, vol. 7, pp. 44562-44573, 2019, doi: 10.1109/ACCESS.2019.2909071.

17. A. Bonnefoy, V. Emiya, L. Ralaivola, and R. Gribonval, "Dynamic Screening: Accelerating First-Order Algorithms for the Lasso and Group-Lasso," in IEEE Transactions on Signal Processing, vol. 63, no. 19, pp. 5121-5132, Oct.1, 2015, doi: 10.1109/TSP.2015.2447503.

18. Liu, Weiyang, et al. "Large-margin softmax loss for convolutional neural networks." *arXiv preprint arXiv:1612.02295* (2016).

19. Sabour, Sara, Nicholas Frosst, and Geoffrey E. Hinton. "Dynamic routing between capsules." *Advances in neural information processing systems* 30 (2017).

20. M. Osman, J. He, F. M. M. Mokbal, N. Zhu and S. Qureshi, "ML-LGBM: A Machine Learning Model Based on Light Gradient Boosting Machine for the Detection of Version Number Attacks in RPL-Based Networks," in IEEE Access, vol. 9, pp. 83654-83665, 2021, doi: 10.1109/ACCESS.2021.3087175.

21. Mohamed Massaoudi, Shady S. Refaat, Ines Chihi, Mohamed Trabelsi, Fakhreddine S. Oueslati, Haitham Abu-Rub, A novel stacked generalization ensemble-based hybrid LGBM-XGB-MLP model for Short-Term Load Forecasting, Energy, Volume 214,2021,118874, ISSN 0360-5442, https://doi.org/10.1016/j.energy.2020.118874.

22. M. Chen, Q. Liu, S. Chen, Y. Liu, C. -H. Zhang and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System," in IEEE Access, vol. 7, pp. 13149-13158, 2019, doi: 10.1109/ACCESS.2019.2893448.

23. Y. Freund and R. E. Schapire, "A short introduction to boosting", *J. Jpn. Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771-780, Sep. 1999.

24. S. Wu and H. Nagahashi, "Parameterized AdaBoost: Introducing a Parameter to Speed Up the Training of Real AdaBoost," in IEEE Signal Processing Letters, vol. 21, no. 6, pp. 687-691, June 2014, doi: 10.1109/LSP.2014.2313570.

25. X. Qi, T. Wang and J. Liu, "Comparison of Support Vector Machine and Softmax Classifiers in Computer Vision," 2017 Second International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2017, pp. 151-155, doi: 10.1109/ICMCCE.2017.49.