International Journal of Engineering &Technology, 7 (3.12) (2018) 1038 -1045

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Predictions for Startups

Eliganti Ramalakshmi[1*], Sindhuja Reddy Kamidi[2]

[1]Assistant Professor, [2]Student
[1,2]Department of Information Technology, Chaitanya Bharathi Institute of Technology
Hyderabad, India
**Corresponding Author E-mail: [1]eramalakshmi@cbit.ac.in , [2]sindhujareddy1997@gmail.com*

Abstract

In general 90 out of 100 startups fail to gain expected funding. There can be several reasons like bad management, lack of sufficient funds, good working team etc. which leads to failure of startup. This work aims to create a machine learning model for predicting the range of funding for the startups based on many key attributesthat are involved at different stages in the startup functioning. It is very important to predict the range of funding prior to the implementation of project and till today not much work is done in this respect. This paper proposes implementing a model to predict the funding of a startups based on many important factors like idea of the startup, place where the startup established, domain vertical to which the startup belongs, prior investors, type of funding the organization is expecting. A model is developed by working on real time data of startups from 2015 to 2017. Classification and regression algorithms are used to build the model.

Keywords: *Crunchbase, Random Forest, Startup Funding.*

1. Introduction

As today's world is making big transformation by startups, on an average 90 out of 100 startups fail to attain the expected range of profits. . There can be several reasons likeinefficient planning, inefficient way of using the funds, lack of good team to work, insufficient funds, etc. which leads to failure of startup. This work aims to create a machine learning model for predicting the range of funding that can be expected by the organization based on many important factors that play a major role in all stages of a startup. It is very important to increase the success rate of startups. It helps in predicting the future of startups by involving machine learning and increases the productivity and decrease the failure percentage.

The main factor that influence the success of the startup is the idea which further gets implemented. A significant number of startups fail because of lack of awareness about the progress of that particular domain. Every entrepreneur aims at development and success of his/her organization. They want their idea to be helpful and appreciated by their customers so that the organization will receive the enough profits to expand share in market and to make further modifications required.

The main elements of any startup are:

1. Funding
2. Appointing best employees.
3. Reduce the employee attrition.

This paper aims at building models for predicting the amount of funding the startup can receive based on the factors like product idea, domain, sub vertical to which it belongs to etc.

It is desired to know the future of the project and it's plausibility to give success far before the implementation of the idea. Much work has not been addressed to solve this problem. The machine learning algorithms will help to handle such cases more effectively.

This work mainly aims in guiding to build the model which predicts the range of funding a startup can expect far before the implementation process. Many prior works have been provided but which fail in some extreme cases. A summary of previous work related to this issue is presented below.

2. Related Work

The machine learning includes supervised, unsupervised and semi supervised learnings. The supervised learning provides many different regression and classification techniques to implement a machine learning model based on the labeled data. The existing solutions for this problems include all the algorithms briefly explained below[1].

The existing system to solve this problem of predicting the future of the startups used the following algorithms.

Lazy lb1

The lazy algorithms trains the model only when there is demand for the prediction. The algorithm runs only when the unknown tuple is provided to predict the output. The Lazy algorithm demands to fix the number of clusters before starting building the model. The number of data items per cluster will change with the size of the dataset. The number of instances chosen before implementation depends on the count of final clusters. Each instance is treated as the median and the other instances are included in the cluster to which they are close to according to the values of their attributes. The paper titled Lazy Association Classification the author discusses the solution to this problem using lazy ib1 algorithm and focuses more on the feature selection to attain accuracy of 75%.

Random Forest
Random Forest classifier is the collection of multiple independent decision trees. The disconnected decision trees are formed by taking different starting nodes. The initial nodes are selected based on the gini index and many different criteria. the individual trees are built independent of the other trees. when an unknown data item is given to the model, the individual outputs of the decision tree is send to a optimiser which finds the maximum favorable class label and gives it as the output. As the output of multiple trees is considered the accuracy of the model is expected to be high and resist the underfitting and overfittingproblems[3]. This algorithm is very robust and handles the highly imbalanced classes very effectively.

Naive Bayes
Naive Bayes classifier assume that the features of the data items are independent to each other. The hidden correlation between the features are not addressed effectively. Naive bayes classifier is trained on the supervised learning settings depending on the probability model. The accuracy of the output can be highly dependent on the supervised learning settings and fails to find the patterns and dependency of features.

ADTree
ADTree is a variation in decision tree which includes AND OR graph. The knowledge is distributed into different multiple paths. When the prediction has to make the different paths must be traversed. each node is splitted into multiple ranges and the prediction from each node is summed to find the final prediction. The sum having positive value falls under one class and the sum having negative value fall under another class. A single tree which have capabilities to predict is obtained as a result.

Bayesian Network
It is a type of Bayesian classification technique having dual nature and consists of two stages: first the network structure is learned, and then the probability tables are formed. The complete Network is used for the structural and conditional probability tables are formed. thus formed tables are used to build the decision trees. The resulting models are collectively called as Full Bayesian network classifiers. By forming such classifiers it captures both independence that is variable and context-specific.

Simple Logistic
The simple logistic regression is built by plotting the graph of the dataset and then forming the boundaries separating the different classes. This is inefficient when the data is linearly inseparable and very sensitive to the underfitting and overfitting problems. It uses a stage-wise fitting process.

Comparison of previous work in terms of parameters

Algorithm used	accuracy	Advantages	Drawbacks	details
Lazy lb1	75%	More the dataset more the accuracy	It is hard to fix the number of cluster prior to the process implementation	need to focus on feature selection of individual data item.
Random Forest	80% with variations	It can handle the input data even if it has large number of data features and it can produce more accurate results.	Very sensitive to the outliers in dataset	Individual desition trees are formed using split criteria and output of every individual tree is considered to find the final result.
Naive Bayes	79% after tuning algorithm	It is the easiest approach to train the model if the data provided has class labels (have supervised data)	It is not suitable when the correlated data is used. As it treats all the features as unrelated	It assumes every feature can be calculated individually and not dependent on other features of the same data.
ADTree	80%	multiple paths are traversed to form predictions. So makes easy to avoid local maximas. And gives global solutions	Inefficient when multiple classes are present in dataset.	A single tree which have capabilities to predict is obtained as a result.
Bayesian Network	77%	By forming such classifiers it captures both independence that is variable and context-specific.	Very sensitive to the biased dataset as it uses probability function	Full Bayesian network classifiers
Simple Logistic	75%	The simple logistic regression is built by plotting the graph of the dataset and then forming the boundaries separating the different classes.	Fails when the dataset is unstructured and linearly inseperable	The graph fits the dataset in an incremental and step wise process

A. CrunchBase Data Preprocessing

The crunchbase dataset us a raw startup data which includes different attributes of different types and it is consolidated data. Working with such datasets demand for rigorous data preprocessing. So, the previous work on this dataset focusses on data cleaning, data reduction, data transformation and normalization[2].
The basic steps involved in data preprocessing are
(a) Converting attributes that hold numeric values to symbols (nominal) for example funding type (seed funding, private equity etc.)
(b) Constructing the attributes that represent time like seed funding date etc. The date attributes are replaced with the number of days from a fixed date so that the data remains integer for better comparison.
Predictive Modeling: The model development starts here. The preprocessed data must be used to train the predictive models for finding the success rate of a particular startup. The steps involved are:
(a) The preprocessed data is splitted into two groups called training data and testing data we can also use cross validation n folds strategy.
(b) The training dataset is used to build q machine learning model by training till the desired accuracy is attained. multiple algorithms like Bayesian network, simple logistic regression, decision trees, etc., are used along with multiple ensembling methods.
Evaluation: once the model is trained using the training dataset, in this phase the accuracy of the trained model is obtained by finding predictions for the testing set.
(a) The unknown success rate of the testing dataset is predicted by comparing the corresponding attribute values to the existing successful startups on which the model is trained.
(b) calculate accuracy, precision, recall/ sensitivity, specificity, area under ROC curve etc.

B. Key Factors Involved

There can be a lot of factors that play a key role in either success or failure of the startups. Here more that 20 key factors has been considered like series A funding, series B funding, seed funding etc. few of them are briefly explained below:

a) Seed funding: Seed funding is required to start the basic services required to build the organization and also helps in managing basic bills, wages etc along with the extending of startup for further development.

b) Time to get seed funding: This represents the number of months that the company spent to collect the required seed funding to start the company. The time is a main factor that plays major role because no startup should spend a lot of time to collect funds for unproductive project.

c) Rounds of Funding: there can be different types of investment and the number of rounds can vary. So this represents the number of rounds including type of funding.

d) Severity factors:This contributes to the authenticity of predictive models. It is divided into two segments. One is positive factors like low burn rate, better management system, better plan and use of funds and time.

e) Series A funding: After the seed funding is utilized the series A funding is required to run and manage the organization, to expand the organization, expand the market share.

f) Burn Rate: This value represents the rate at which the company utilizes the funds and investments. The burn rate is inversely proportion to the success rate. The company with more burn rate will demand more funds and has less chances of sustaining in the market.

g) The crunch base dataset includes 7000 successful startups and 4000 failed startups with several factors. After the preprocessing and data visualization step it is concluded that many key factors are involved in success of startups. some include seed funding, date of establishment, raise in funding, the idea, management plans, budget planning etc. the models are developed using multiple algorithms and accuracy and many evaluation factors like precision, recall/ sensitivity, specificity, area under ROC curve etc. are calculated. The accuracy obtained for the model is 73.3%.

3. Methodology

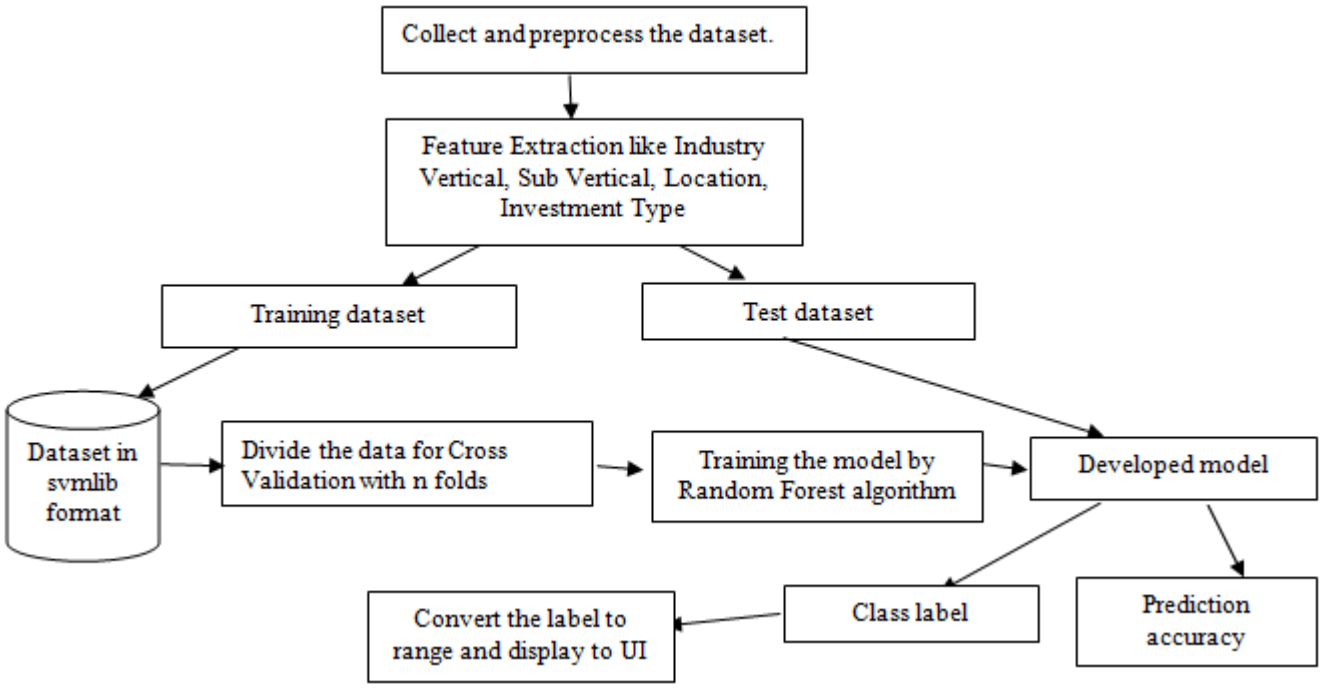

Fig. 1: Model for Startup Funding

A. Data Consolidation

Collect the real time data set related to startups from 2014 to 2017 which includes features like Serial Number, Date of funding, Startup Name, Industry to which the startup belongs, Sub category of the industry type, City Location, Investors Name, Funding Amount etc.

B. Data Preprocessing

Data Preprocessing is one of the essential steps in building machine learning models. It involves transforming the collected data into consistent and understandable format.

The consolidated raw data may include some outliers, values that are out of range, few missing values, error values and soon. without handling such error will lead to wrong results. The quality of data directly proportional to the accuracy of the model. While training the model the ambiguity raises due to redundant and unimportant data. This makes the necessity of data preprocessing before training the model. This can take considerable amount of time to preprocess. This data preprocessing step includes activities like cleaning the data, transforming the data, selecting the instance, extracting the required features, normalizing the data if it is not within the acceptable range. Thus preprocessed data can be used for training the model.

- Data cleaning: consists of filling the missing values using binning methods, the noisy data must be smoothed, outliers must be identified and resolve if any inconsistency of data present.
- Data integration: building the data cubes which help in analysing the data easily.
- Data transformation: normalizing the attribute values and performing aggregation.
- Data reduction: reduce the volume of the data is it is huge and redundant and aim for the same analytical values as original data.

- Data discretization: If any numerical data exist replace it with the nominal values or some unique symbols.

C. Cleaning the Data

Filling the missing values
- Ignored the tuple with missing values for multiple columns.
- Used the attribute mean to fill in the missing value.

Identify outliers and smooth out noisy data:
- Binning: Sorted the attribute values and partitioned them into bins .Then smoothed by bin means, bin median, or bin boundaries.
- Correcting and removing inconsistent data: Domain knowledge will help in identifying such data and taking expert suggestion to remove inconsistent data.

D. Transforming the Data

- Normalization: If the attribute values are not within the acceptable range or due to some programming restrictions scale the values of such attributes to new specific range.
- Aggregation: combining logically dependent and similar attributes.
- Generalization: using concept hierarchy the attribute values are replaced with more generalized values.
- Here the Amount is exceeding the limit of integer. So, converted the values into USD denomination.
- Generalized few features like location in startup funding dataset. Removed town names and replaced with state and city names.

E. Data Reduction

- Deleted the features which are duplicate
- Removed the features which have same value for all tuples.

4. Building the Model

- Converting the problem into classification problem by selecting the range of values and labeling them with some class.
- Building spark Session: Adding name for the application running on spark, Calling builder on current spark Session and adding configuration settings
- Loading the dataset:The dataset is in the format of svmLib. The Loading to svmLib dataset is done using spark variable setting format and calling read function by passing path of dataset.
- Using parallelize function convert the dataset loaded as RDD on cluster nodes. Converting all the categorical features data into integers using One Hot Handler in spark.
- Splitting the dataset for training and testing

A. Random Forest

Random forests are used to solve classification problems, regression problems that operate by building multiple decision trees at training time and the output of new data item is obtained by calculating median or mode of the results of all individual trees. Decision trees are more likely sensitive to the overfitting problem but this approach will handle such cases intelligently.
Basically the aim of random forest is to build multiple independent decision trees from which the output is collected and final output is given based on some strategy to increase the rate of accuracy of prediction.
Random Forest pseudocode:
1. Initially it selects few features out of all ‘m’ number of features. Suppose the selected feature set has ‘k’ number of features
2. Using best split point strategy a node is obtained among the selected feature set.
3. This node is splitted into its sub nodes based on split point.
4. The above three steps are repeated until all the nodes are reached.
5. The forest is built by repeating the above 4 steps for number of times to obtain multiple decision trees.

Initially Ensembling methods like bagging boosting are used to remove the bias in dataset. then multiple trees are formed using the above process. the voting or averaging methods are used to obtain final prediction value from all the trees that are formed.

Increasing the accuracy of the model:

- **Add more data**

The accuracy is increased when new features which are removed in the feature engineering are added and included to train the model. The accuracy have increased by 5%.

- **Feature Engineering**

The unnecessary features like which holds same value to all the tuples and have no significant relevance to the problem. For example, feature established date where the dataset holds information about the companies established in the same year 2016 and 2017.

- **Multiple algorithms**

It is not possible to select an algorithm and expect it to give maximum accuracy. We need to use multiple algorithms and make average of all the models outputs. We have built multiple random forests with different rules and considered the average of all the models outputs. The accuracy increased by 10%.

- **Algorithm Tuning**

As the algorithm used here is Random forest, this step includes changing of the number of iterations per each data item. Freezing the variables values when the threshold of the accuracy is reached.

- **Ensemble methods (Bagging and Boasting)**

As the data is huge the bagging is applied to sample the data and check if algorithm works for the problem. The dataset collected is biased. So boosting of the class items which are less in number increased the accuracy by 5%.

- **Treat missing and Outlier values**

The data worked on is containing many outliers. For example, the same feature values for multiple data items are holding different and extreme investment values (one with very low value and other with very high value). This cause the ambiguity. Basically such case raise due to wrong values and outliers. They are handled using clustering method and excluded from the dataset. The accuracy increased drastically and reached 84%.

5. Results

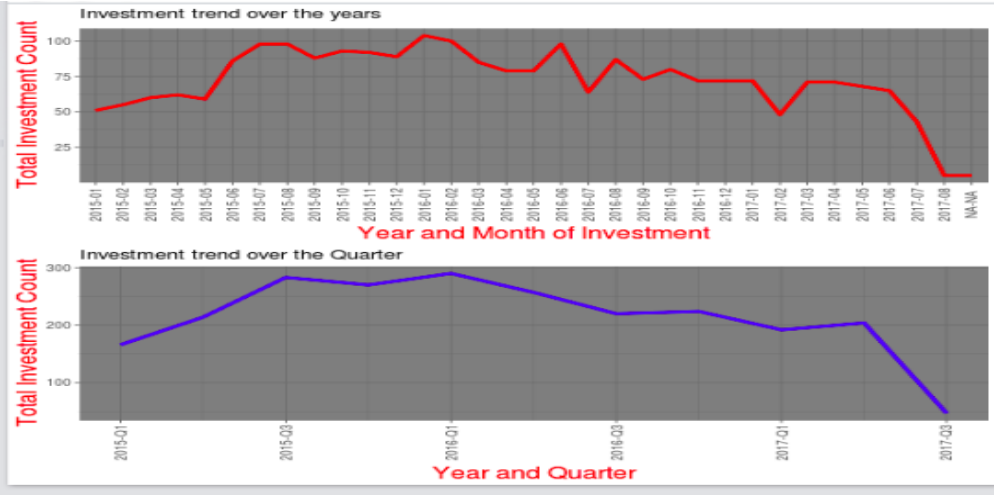

Fig. 2: total investment for every year and quarter

Fig2 shows how total investment for every year is varying according to the dataset. From the Fig2 we can conclude that the year 2015 first quarter has maximum funding and this can be due to multiple external reasons. If the reasons are known that can help to predict if the change occurs again in future.

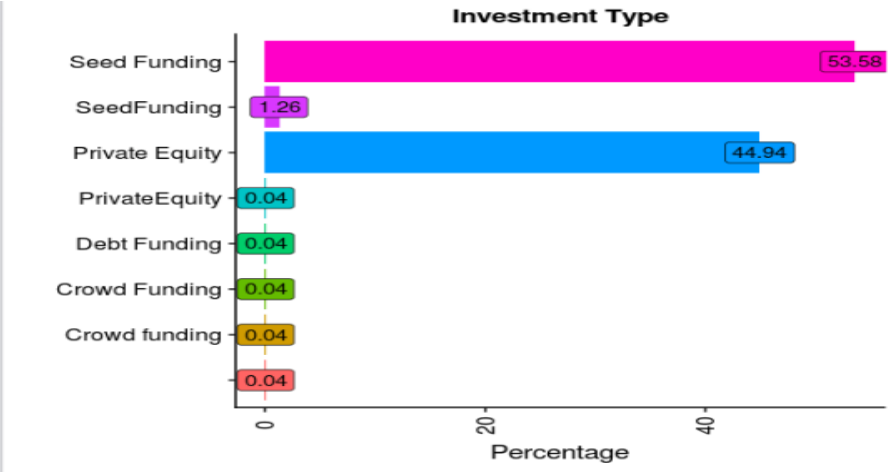

Fig. 3: Investment type and percentage of number of startups

Fig3 shows that 53% of startups are expecting seed funding as investment type. So, the startup can expect the more seed funding compared to private and public equity investment.

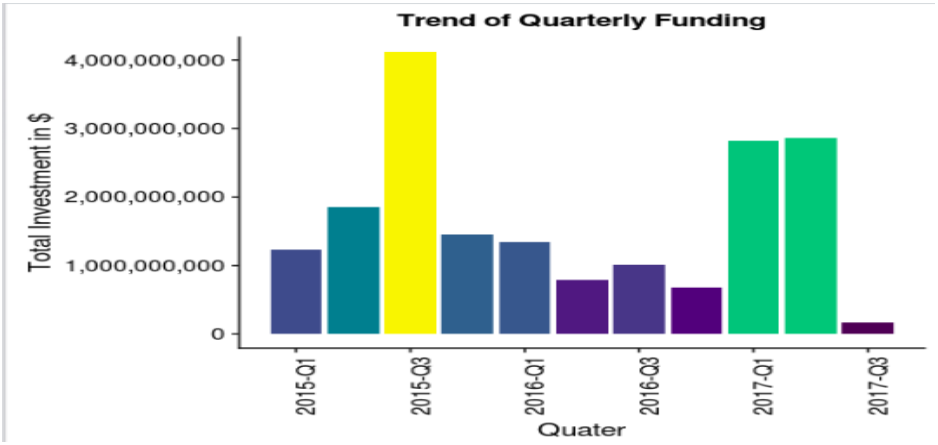

Fig. 4: Trend of quarterly funding

Fig4 shows that in year 2015 3rd quarter has maximumfunding. If any seasons change is responsible for the highest total investment the same pattern can be expected every time the change occurs.

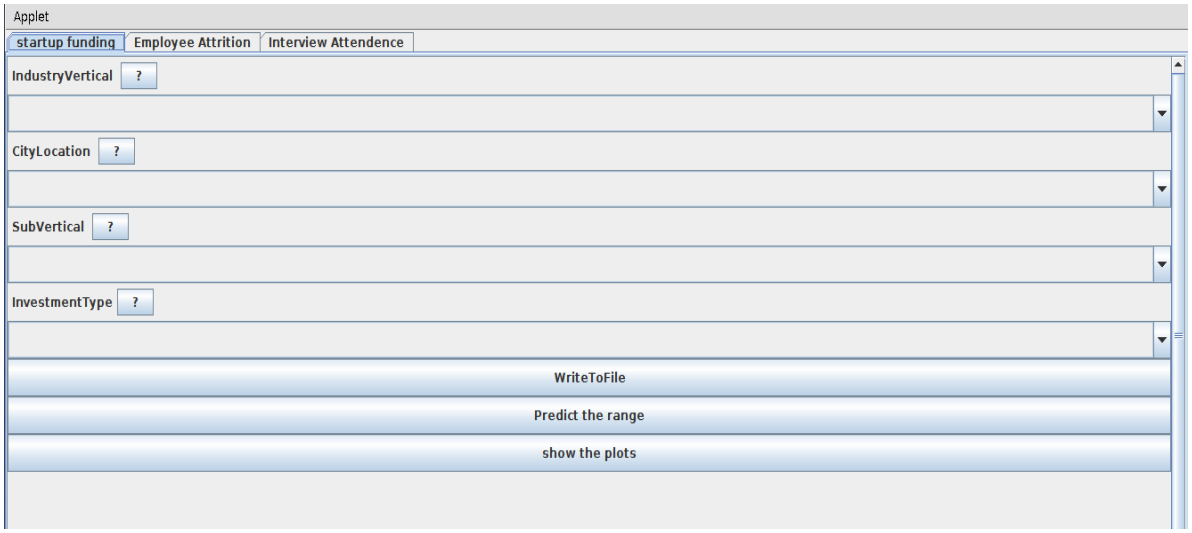

Fig. 5: User interface to give input to the model

Fig5 shows that the attribute values like IndustryVertical, City location, sub vertical etc. are taken from user to predict the expected range of funding.

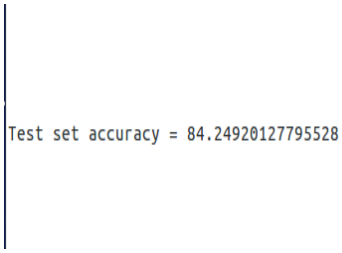

Fig. 6: Test set accuracy of startup funding

Fig6 shows that Test accuracy of startup funding model is 84.24 variance. That means the predicted range has 84% probability to give correct prediction value. The variance in prediction accuracy indicates that the accuracy may change every time the model is executed. This is because the test and train set are chosen randomly by using random separating function into 70% and 30%. So, the tuple which included in training is included in testing the other time.

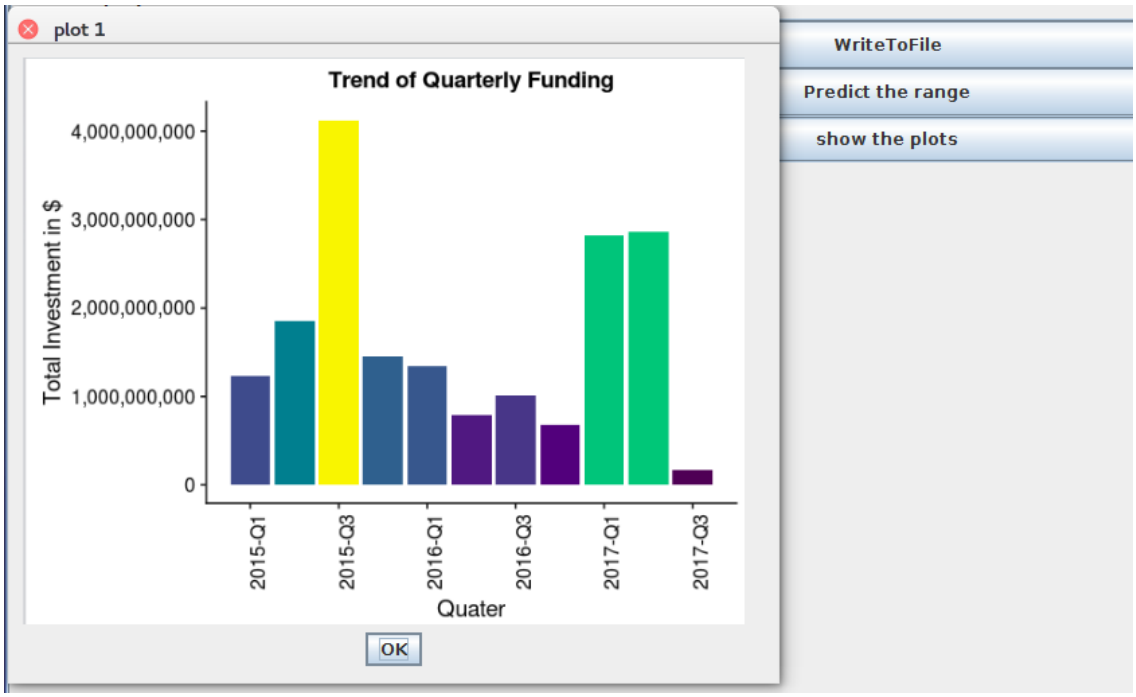

Fig. 7: Trend of Quarterly funding

Fig7 shows that that in year 2015 3rd quarter has maximum funding. This shows that the UI should show the plots so that the user can understand the data much more effectively. When the show plots button is clicked the plots are displayed as shown in Fig7.

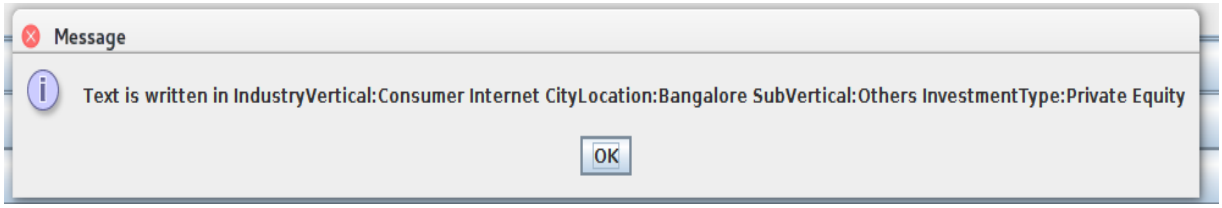

Fig. 8: write to file message

The UI must be connected to the spark such that the input from the UI is sent to machine learning model built on spark and make prediction. For this the data must be in svmlib format. So when the write to file button is clicked the data is converted to required format and sent to the spark session.

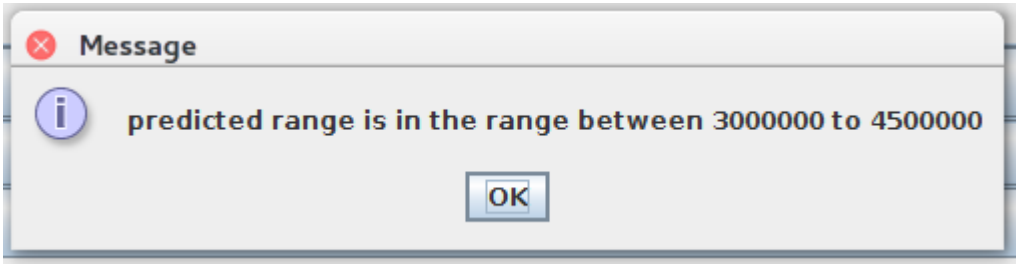

Fig. 9: predicted output dialog box

We can also predict the funding range as shown in Fig9. When the predict button is clicked the spark cluster will start which has 4 nodes. The configured nodes will start and their health will be displayed on the terminal. Once the prediction is done the class label is send to UI and it is converted to user understandable language and the range is displayed.

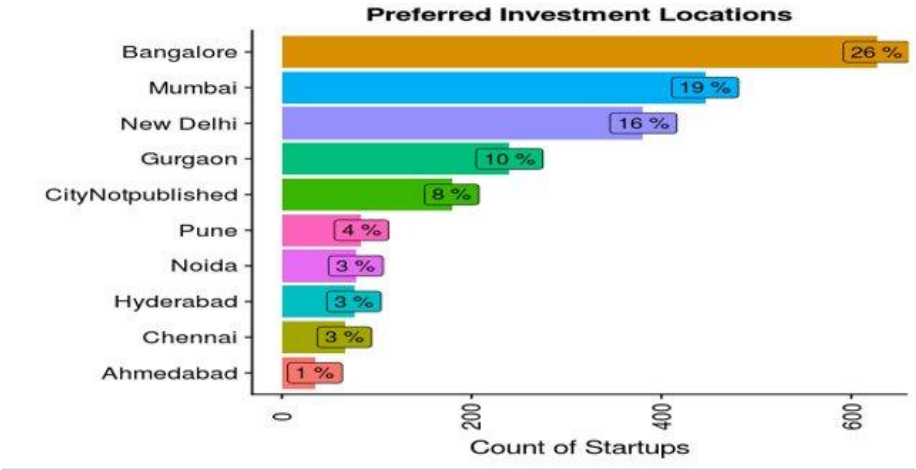

Fig. 10: Investment Locations

The fig 10. Shows the preferred investment locations and count of startups. In this it says that 26% of the startups are established in Bangalore and received expected range of fundings.

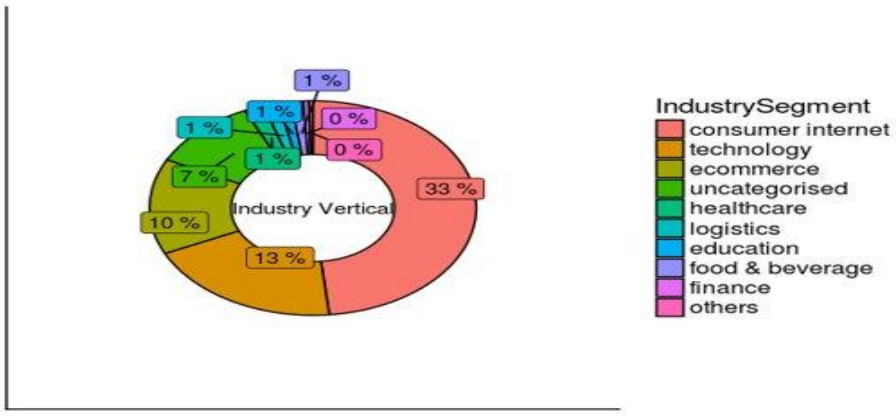

Fig. 11: Pie graph demonstrating industry segment

In fig 11. It says that electronic commerce can expect high range of profits investments and funding compared to other industry segment.

6. Conclusion

In this paper a model which helps startup to predict its future and also help them with suggestions to improve the progress has been proposed and implemented. Models to predict the possible funding range that the startup can expect based on the information like industry vertical, sub vertical to which the startup belong, location where startup started, type of investment the organization is expecting and soon has been implemented. This model gave 87% accuracy using Random forest algorithm. This model may alter the result if any other external factors that affect the funding the external factors can be like psychological reasons and emotional reasons of employee or candidate.

References

[1] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi. Exploration of data science techniques used to predict the strength of steel andIntegrating Materials and Manufacturing Innovation, 3(8):1–19, 2014.

[2] T. M. Begley and W.-L. Tan. The socio-cultural environment for entrepreneurship: it is a comparison between asian and anglo-saxon countries published in Journal of international business studies, pages 537–553, 2001.

[3] Breiman. L. for Random forests. Mach. Learn., 45(1):5– 32, Oct. 2001.

[4] kaggle.com for startup funding data

[5] Amar Krishna, Ankit Agrawal, AlokChoudhary. "Predicting the Outcome of Startups: Less Failure, More Success" , 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016.