|  |
| --- |
|  |
| **Loan Success Rate Prediction** |
| **Name: Komal Ashraf**  **Section:A**  **ID: F2019313026**  **Date: 5/2/2020**  **Final Report**  **Submitted to : Miss Satwat Bashir** |

|  |
| --- |
|  |

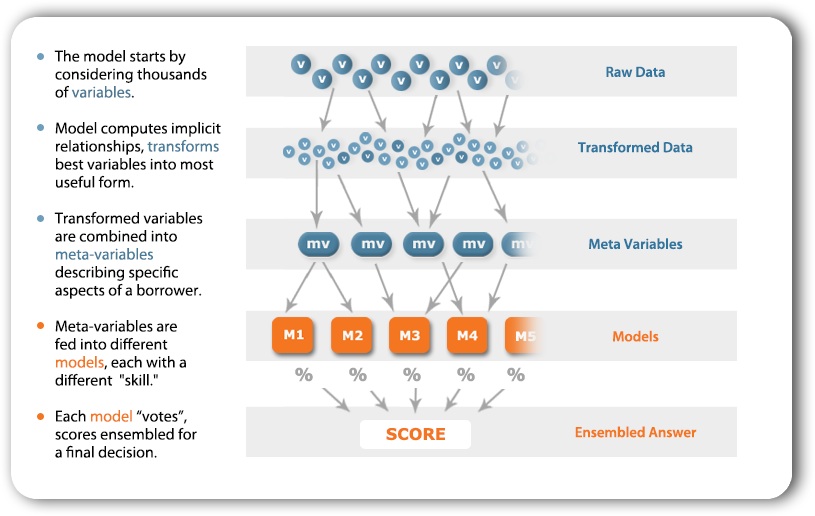
**1. Introduction**

The problem is related to predicting the loan success rate and is a very critical problem as investors have to make decisions that should they invest their heavy amounts as loan to the borrowers or not .Basically a lending party enables the borrowers to create their personal loan between 10000 - 40000 and on the other hand investors can search and browse listed loans on lending party website and select the loan they want to invest in based on the information provided about the borrower, amount of loan grade and loan . Investors makes money from interest and lending party makes money by charging the borrowers an origination fee and investors a service fee.

So basically, this problem occurred in making the decision at the time of investment that either borrower will fully pay the loan or he will be charged off as investors experience high credit loss rates while borrowers get poor experience. By the implementing different suitable machine learning algorithms, investors will be able to predict that:

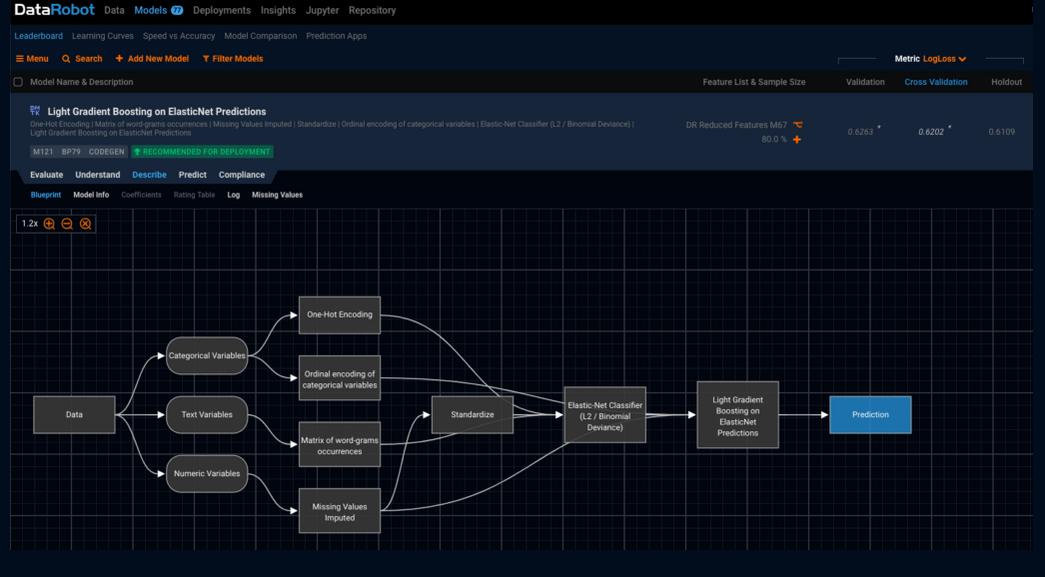
1. About the return on investment before they invest.
2. Quality of the loan.
3. Either borrower will pay his money on time or late.
4. Will the borrower pay fully or partially or charged off.
5. Risk will be identifying at the start.

**2. Existing Work**   
  
**a) ZestFinance** is a financial services technology startup that uses machine learning and large-scale big data analysis that works on linear models such as those produced by logistic regression as well as ML models including deep neural networks, random forests, and gradient-boosted decision trees. **Model Representation:**



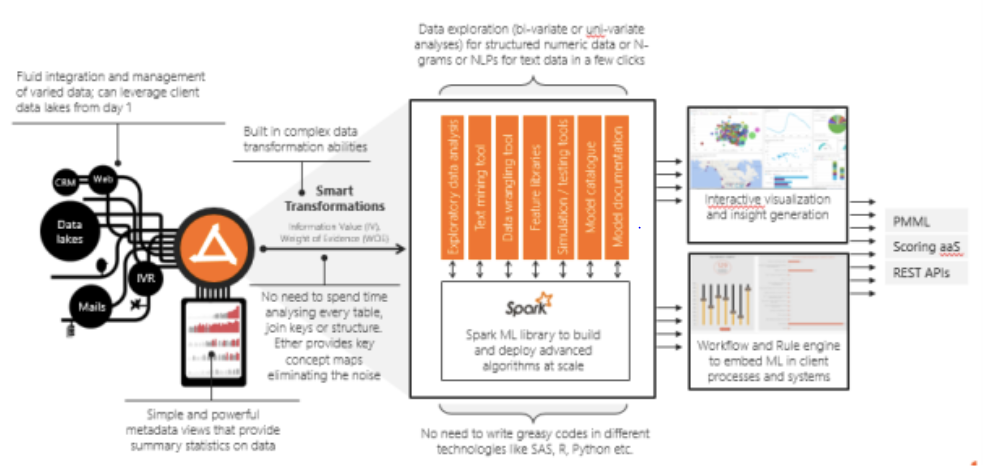
**b)** The **DataRobot** enterprise AI platform democratizes data science and automates the end-to-end process for building, deploying, and maintaining AI at scale.

**Model Representation:**



**c) Scienaptic Systems Inc**. is a Foreign Business Corporation in the State of New York that provide secure credit decisioning instruments (data , credit algos and underwriting technology) using Ether AI to make predictions.

**Model Representation:**

****

**3. Algorithm Definition:**

In order to make the accurate predictions, I have implemented 6 different machine learning techniques and check the accuracy of each model.

**1: XGB Classifier**

The dataset is consists of about 20K rows and 30 columns which is a very huge data to deal with and to make accurate model. Due to huge data set and performance issues I used XGB Classifier as it works by boosting up the data and solve problems in a fast and accurate way. By implementing this algorithm, the data processing time became fast and model accuracy came out as 83 %

**2: Support Vector Classifier**

The second algorithm is found suitable for my problem is the SVC. As the data set i used consist of many columns having categorical data that were encoded before building the model .So for data that is categorized we can use SVC classifier in which data is categorized by drawing the hyper line between them. In subjected case SVC will divide the data by fully paid and partially paid by the regression line.

**3: Logistic Regression**

In the subjected problem, the dependent variable is categorical variable named as “loan\_amnt”. As this variable is consists of 2 category, Fully paid and Charged off so I used Logistic regression on the data to train the model in order to make prediction . Then I identify the actual and predicted results using confusion matrix so that I can calculate the sensitivity and specificity and then check the accuracy of the model.

**4: Decision Tree Classifier**

The data set consists of few columns like emp\_ title, hom\_ ownership , pub\_rec\_bankruptciesand grade etc., which explains the financial condition of the borrower asking for loan and to make decision for loan approval based on the financial history of the borrower i need to use an algorithm thoroughly goes to each predicting variable and then predict the final accurate answer. I used Decision Tree Classifier which works by splitting the variables(node) until it gets the prediction by the help of most frequent prediction.

**5. Random Forest Classifier**

As the data consists of making decision based on different parameters, previously I use Decision Tree Classifier but to make more accurate prediction I used Random Forest Classifier. Basically random forest makes multiple decision trees models that operates as an assemble where each individual tree in the random forest makes a class prediction and the class with the most frequent votes becomes the model’s prediction. For Example, in the subjected loan success rate prediction case, Random forest will randomly select a variable and split it with other randomly selected variables till it gets the final result for loan\_Status as fully paid or charged making a tree. Similarly, the more trees will be made using the same randomly selected variables and then the most frequent answer will be the final result i.e. for loan\_Status as fully paid or charged

**6: Balanced Bagging Classifier**

In order to aggregate the decision from Random Forest and identify the accuracy of random forest, we need to perform bagging. In bagging we identify the missing entry while performing the random forest tree and called out of bag dataset. Now this classifier will build out of bag sample through all of the other trees that was built without it. Hence it rebuilds the random forest tree and identify the missing entry, fit the missing entry and balance the trees to identify the accuracy of random forest result.

**3.1 Methodology**

The loan data I used consists of 30 columns including dependent variables as follows:

**Dependent variables:** loan\_amnt

**Independent variables:** id , loan\_amnt , term , int\_rate , installment , grade , sub\_grade , emp\_title , emp\_length , home\_ownership , annual\_inc , verification\_status , purpose , title , zip\_code , addr\_state , dti , earliest\_cr\_line , fico\_range\_low, fico\_range\_high, open\_acc , pub\_rec , revol\_bal , revol\_util , total\_acc , initial\_list\_status , application\_type , mort\_acc, pub\_rec\_bankruptcies

* **Data Preprocessing:**

The dataset I use consists of about 20k rows with 30 columns and it contained huge amount of Nan values. Among 30 columns, there were few columns that are not necessary for our prediction. So through data visualization Using PowerBI , I visualize the impact of each independent variable on the dependent variable using graph. After the detailed analysis of unnecessary variables using PowerBI and some python functions like head() , tail () , shape () and plot(), i dropped them from my data frame.

Now the 2nd step was identifying the Nan values .The Nan values can cause problem during model building and evaluation . The dataset consists of huge amount of Nan values in mp\_length and emp\_title columns so I removed Nan values from categorical variable using mode and Nan values from numerical variables using mean.

Now 3rd step is to perform label encoding. The data set consists of many categorical variables where grade and emp\_length follows the order so I apply ordinal encoding in them and apply simple categorical encoding on the rest of categorical columns.

* **Scaling and normalization the data**

Now for building the model and evaluating results from to make prediction data needs to be normalized and scalable so I applied scaling using “from sklearn.preprocessing import StandardScaler”

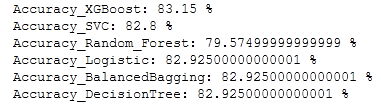
* Splitting the dataset into train and test

Now to build and train in order the model , I split the data set into train and test data set using “from sklearn.model\_selection import train\_test\_split” with split ratio of 20 and 80

* **Building Models and fitting Machine Learning algorithms**

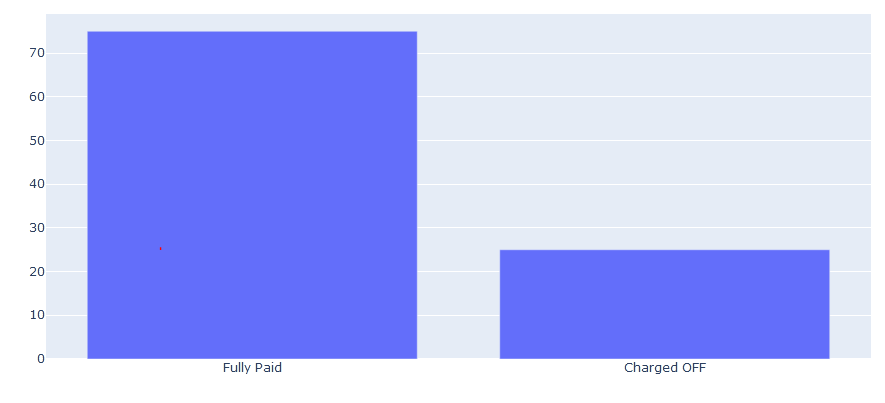
I used six different models and also analyze the sensitivity, specificity and accuracy of each model. 1st I trained the model and then predict on test data. After that I build confusion matrix and at the end analyze the accuracy using sensitivity and specificity.

Following are the accuracy rate of each model.



**3.2 Results**

Results are quite positive as it says for the given dataset loan success rate is 82% of the borrower will pay their loan back on time whereas 18% of the borrower will not be able to pay on time hence they will be charged off



**3.3 Discussion**

The dataset consists of mostly fully paid borrowers and charged of are comparatively very less. Hence it shows that as per the given data set and information provided about the borrowers, investors will get back their money and there is less chance of being credit loss.

The six models i used gave the good accuracy but I found the XGBoster ,Random Forest and Balanced Bagging Classifier are more better in terms of performance , making prediction and accuracy as per the problem . SVM and Logistic regression does not really fit in this dataset as we have better classifier like XGBoster, Random Forest and Balanced Bagging.

**4. Conclusion & Future Work**

Hence I concluded, for given dataset, number of fully paid borrowers is greater than that of charged off. There is need to test the models on the dataset contains more rows but due to tool limitation , python take a lot of time to execute each model when working on large dataset. The results can help finance companies and banks to predict their credit worthiness and loan investments at the time they invest.