**Data Science Assignment - REUNION**

**KanikaBhardwaj**

**ABSTRACT**

The objective of our model is to predict the person’s creditworthiness is often associated (conversely) with the likelihood they may default on loans.

We have anonymized data on about 1000 loan applications, along with a certain set of attributes about the applicant itself, and whether they were considered high risk.

0 = Low credit risk i.e high chance of paying back the loan amount

1 = High credit risk i.e low chance of paying back the loan amount

So, in our methodology, we propose a method to make predictions of credit risk, in which we have used several algorithms, tune the corresponding parameters of the algorithm by analyzing each parameter against F1 score and predict the credit risk. To make our prediction we used Random Forest, Decision Trees and Logistic Regression. We improved the accuracy by tuning hyper-parameters and Random Forest gave the best accuracy. We also analyzed several data mining techniques to handle missing data, remove redundancy and resolve data conflicts.

**INTRODUCTION**

There are many possible methods of moving between two given points in a city; however, the taxi trip has found wider applications in urban cities when compared to any other mode of transport. It hence becomes very important

to analyze and predict trip duration between two points in the city when provided with the required set of parameters that affect the trip duration. For a good taxi service and its integration with the existing transportation system the project serves as a right means to comprehend the traffic system in the New York City. For prediction purposes factors such as pick up latitude, pick up longitude, drop off latitude, drop off longitude etc. is considered. These geographical locations clubbed with other important factors such as pick up date, pick up time are used for the overall trip duration prediction. The primary focus of this project is in depth analysis of the factors associated with a taxi trip in NYC. The different algorithms used are: Linear Regression, Random Forest and Decision Trees.

1. LINEAR REGRESSION

It is a linear model that establishes the relationship between a dependent variable y (Target) and one or more independent variables denoted X (Inputs). Linear regression has been studied at great length, and there is a lot of literature on how your data must be structured to make best use of the model.

1. DECISION TREES

Decision Trees (DTs) are a non parametric supervised learning method used for classification and regression. The goal is to create the model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

1. RANDOM FOREST

The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance. However, random forests also use another trick to make the multiple fitted trees a bit less correlated with each other: when growing each tree, instead of only sampling over the observations in the dataset to generate a bootstrap sample, we also sample over features and keep only a random subset of them to build the tree. Sampling over features has indeed the effect that all trees do not look at the exact same information to make their decisions and, so, it reduces the correlation between the different returned outputs. Thus, Random forest algorithm combines the concepts of bagging and random feature subspace selection to create more robust models.

**MATERIALS AND METHODOLOGY**

Materials that we have used include: Python software for coding and credit risk Data. Our methodology involves the use of machine learning techniques such as: Linear Regression, Decision Trees and Random Forest Classifier.

1. DATASET

We selected the following features: Months\_loan\_taken\_for as it gives the information for how long time customer has taken the loan.

The columns that include are:

* applicant\_id
* Primary\_applicant\_age\_in\_year
* Gender
* Marital\_status
* Number\_of\_dependents
* Housing
* Years\_at\_current\_residence
* Employment\_status
* Has\_been\_employed\_for\_at\_least
* Has\_been\_employed\_for\_at\_most
* Telephone
* Foreign\_worker
* Savings\_account\_balance
* Balance\_in\_existing\_bank\_account\_(lower\_limit\_of\_bucket)
* Balance\_in\_existing\_bank\_account\_(upper\_limit\_of\_bucket)
* loan\_application\_id
* Months\_loan\_taken\_for
* Purpose
* Principal\_loan\_amount
* EMI\_rate\_in\_percentage\_of\_disposable\_income
* Property
* Has\_coapplicant
* Has\_guarantor
* Other\_EMI\_plans
* Number\_of\_existing\_loans\_at\_this\_bank
* Loan\_history
* high\_risk\_applicant

METHODOLOGY

We analyzed several data mining techniques to handle missing data, remove redundancy and resolve data conflicts. The Data Mining techniques are used to handle missing data. After analyzing we found that there were so many missing data. So, In order to get rid of redundant data, we perform correlation analysis with the help of plots to check if the attributes are positively or negatively correlated if not redundant. To train the model we used Random Forest Regressor.

**ANALYSIS**

In order to undergo the analysis part we examined several algorithms on regression and concluded that Random Forest is the well-suited Regression technique for our respective proposed model.

To begin with, we first examined the missing data, which was much less compared to the whole data so we decided to exclude the missing data.

By Exploratory Data Analysis we observed that customers who have co-applicant and guarantor are low risky than any other.

We did correlation analysis to check for relation between two attributes which helped us to find the redundant data.

We trained our model using Linear Regressor which gave us an accuracy of 62.12% and then we compared our model with a Random Forest Regressor and found that Random Forest was giving us more accurate results of 89.14%.

In comparison to older techniques like Linear Regression our model gave a more accurate result by much percentage.

Further, to improve the confidence we tuned the hyper-parameters such as number of trees and maximum depth for the Random Forest Algorithm.

With an increasing number of trees we observed that the F1 score increases.

**RESULTS**

The proposed hierarchy of the workflow model was loading the data, Cleaning the data, Training the model, Making Predictions, Tuning the hyper Parameters to increase Confidence.

1. CLEANING THE DATA

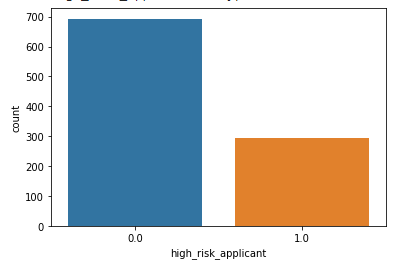
Cleaning the data involves treating the missing values Exploratory Data Analysis (EDA). To fill the missing values, I have used mean and mode for continuous and categorical features. Some features are that have high high number of missing values (nearly 70%-80%) of data, I removed that feature because it is impossible to fill too much large data. But where the data are missed upto 40%, we can treat that.

1. EXPLORATORY DATA ANALYSIS

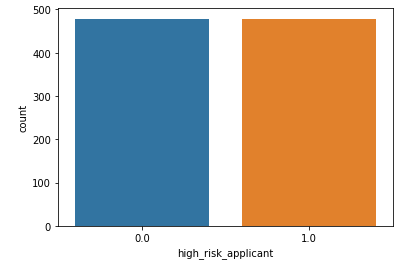
In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images.

**Data imbalance**

Generally, data imbalance often occurs in the credit risk classification due to the huge differences of the number of good borrowers and bad borrowers. SMOTE is one of the most widely used approaches to address this problem. In addition, over-sampling and under-sampling techniques are also employed. Nevertheless, data imbalance has been severely underestimated in many credit risk researches.

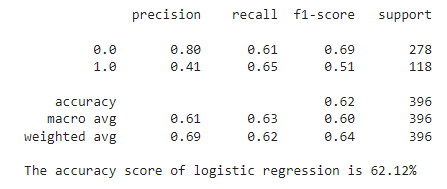


Before using SMOTE

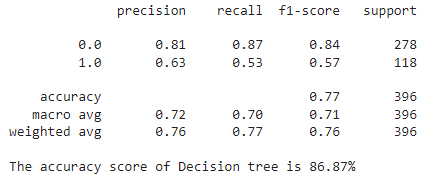


After using SMOTE

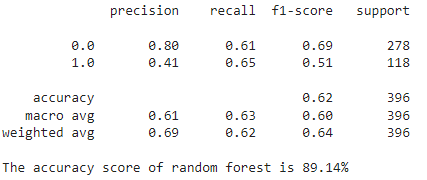
LOGISTIC REGRESSION



DECISION TREES



RANDOM FOREST



Visualizations show us how our model’s

predictions are close to Test Data. It is

evident that decision tree and Random

Forest is performing well.

C. TRAINING MODEL

To train the model we used Logistic Regression and Random Forest Classifier, Decision Tree Classifier algorithm with 80-20 split of dataset for training and testing respectively. It gave an accuracy of 69% and 87-89% percent respectively, to improve the accuracy, tuning of several hyper-parameters such as number of trees and maximum depth

for a random forest algorithm.

**CONCLUSION**

* On the basis of employment Status it be concluded that skilled employees are less risky than any other category.
* Customer who belong to the category existing loan paid back are less risky, they are expected to pay back loan to the company. Their history is also good as well.
* The customer who has high loan amount and having own property and low saving account with high balance account are expected to back (low risky).

**Challenges**

We summarize four major challenges in this project of credit risk.

* First, data imbalance in credit risk is quite severe. Although several approaches such as over-sampling and under-sampling (usually chosen to under-sample the majority) have been proposed to solve this problem, the results are still unsatisfactory in terms of both effectiveness and efficiency.
* Second, the shortage of benchmark datasets is serious. Most existing works use private datasets, thus the results of performance comparison cannot be fair enough.
* Third, most machine learning models are black boxes since they are generally not transparent. Information transparency should be noticed.
* Fourth, the application of deep learning models is still limited in credit risk.

These four challenges are what we are supposed to overcome in future work. We hope more and more deep advanced models will emerge in this area.

**REFERENCES**

* Company link