PROJECT REPORT

CREDIT CARD DEFAULTER'S CLASSIFICATION PROJECT

1. INTRODUCTION

In today's world credit history has become one of the most important factor. Without good credit history financial institutions deny to lend loan. Some financial Institutions also remains in dilemma to whether approve credit card or not for client. Since I am interested in financial sector I decided to implement a Machine learning approach to determine whether a client will be defaulter or not. I found a suitable dataset online, of Taiwan credit card clients.

Financial Institutions need to take best decision regarding loans. In case if financial institute lends loan to client who is not capable of repaying principle along with interest then the institute will suffer a loss. This Machine Learning model will help financial institutions to make better decision whether to lend loan or not.

In this project I implemented Machine learning techniques to classify clients as defaulter or non-defaulter. I've used Logistic Regression, Decision Tree & Random Forest algorithms.

2. RELATED WORK

Abbas Keramati has done a literature survey on similar dataset, It does not extensively analyze any classification algorithm or effects of feature selection but there has been some intense research work on credit card default dataset [5]. Adela Ioana Tudor has done clustering analysis on a similar dataset, the data is clustered related to similarity among different attributes present [6]. Simona Vasilica Oprea evaluated some classification algorithms, but it do not depict the change in efficiency of these algorithms with respect to feature selection [7]. Ajay & Shomona has carried out predictive classification on similar dataset using data mining tool WEKA, they achieved accuracy on Random Forest of 81% [4].

3. APPROACH

This is supervised learning technique of Machine Learning. Goal of this project is to classify defaulter & non defaulter clients, so basic technique I've used is Classification Algorithms.
-Logistic Regression: It is used to predict probability based on given set of independent variables.
-Decision Tree: Used to split population in different groups.
-Random Forest: operate by constructing a multitude of decision tree at training time and outputting the class that is the Mode or Mean.

3.1 Data Preprocessing

- Removing unwanted column, here I deleted column 'ID' from dataset.
- Identifying relevant Categorical, Numerical & Logical Variables and convert them into Proper datatypes.
- Checking for any missing value is very important step in data preprocessing. I checked NA values, but this dataset has no missing values.
- I normalized data by scaling it.
- I carried out Exploratory Data Analysis. EDA is an approach to analyze datasets to summarize their main characteristics.

3.2 Dataset

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan. There are total 24 variables, 1 response (defaulter Yes/No) & 3000 records.

Input: Limit balance, Demographic information (Sex, Education, Age, Marriage), Repayment status of last 5 months (Pay), Bill amount of last 5 months (Bill_Amt) & 5 Previous payment (Pay_Amt).

Output: Defaulter Yes or No.

Source: UCI Machine Learning Repository [2].

3.3 Metrics

I've used Recall & Precision as performance measure (Metrics). Precision refers to the data that is correctly classified by the classification algorithm. Recall refers to the percentage of data that is relevant to the class. Accuracy tells us how accurate our algorithm performs in measure of percentage. I splitted data into Training set & Testing set for Cross Validation. There are 24000 observations in train set & 6000 observations in test set.

3.4 Algorithms

Logistic Regression is a predictive analysis method, it is the appropriate regression analysis to conduct when the dependent variable is binary. Here I used Subset selection as a tuning parameter. I experimented with Forward & Backward method in which by using Forward selection method I got Accuracy of 76.89%. Using Backward selection method I got accuracy of 76.82%. Then I tried Ridge regression & Lasso technique to tune model in which by using Ridge regression technique I got accuracy of 77.52% & using Lasso techniques I got accuracy of 77.29%.

Decision Tree is a decision support method that uses a tree and their possible outcomes. Used to split population in different groups based on conditions. Tuning Parameter is Tree Size. Here I found Decision Tree performance better with accuracy of 81.22%.

Random Forest is an ensemble learning method for classification operate by constructing a multitude of decision tree at training time and outputting the class that is the Mode or Mean. Random forests overcomes drawback of decision tree of overfitting. Tuning parameter is Number of Trees (ntree). Random Forest performs outstandingly on this data it gives accuracy of 81.56%.

4. RESULT

Following table shows results of three classification algorithms:

Algorithm	Accuracy	Precision	Recall
Logistic Regression (Ridge Regularization)	0.7752	0.7746	0.01
Decision Tree	0.8122	0.9599	0.8284
Random Forest	0.8156	0.9535	0.8537

Table 1

By observing above result we can easily conclude that Random forest algorithm performs well as compare to Logistic Regression & decision Tree algorithm.

5. IMPROVEMENT

We can improve this model (accuracy) by using Dimensionality Reduction methods. It is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. Principal Component Analysis (PCA) is one of the techniques used to carry dimensionality reduction.

6. CONCLUSION

Logistic Regression (RIDGE) gives the best model performance at threshold approximately 0.27. The accuracy is seen as approximately 77%.

Decision tree model performs better than logistic regression with accuracy of 81.52%, but worse than random forest and it makes sense because it is prone to both overfitting and under fitting.

Random Forest model on the other hand produces outstanding result with an accuracy of 83.56%, precision of 0.9624 and a recall of 0.8227. This makes sense because random forests average the output from many decision tress which makes it robust to overfitting. Therefore, despite the training error were high the random forest model does an outstanding job in classifying the test data.

7. REFERENCES

1. https://www.wikipedia.com
2. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
3. https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset
4. http://www.ijcaonline.org/archives/volume145/number7/ajay-2016-ijca-910702.pdf
5. http://ieomsociety.org/ieom2011/pdfs/IEOM061.pdf
6. http://www.dbjournal.ro/archive/8/8_3.pdf
7. http://www.wseas.us/e-library/conferences/2012/Vienna/COMPUTERS/COMPUTERS-18.pdf
8. http://www-bcf.usc.edu/~gareth/ISL/