CS-580L

PROJECT REPORT ON

CREDIT CARD DEFAULT PREDICTION

Term: Spring 2020

Team members:

Aditya Sawwalakhe

Harshal Rasal

Shubham Patwa

Instructor:

Prof. Arti Ramesh

1. Abstract

Credit cards have been issued without background checks to increase their market share and hence, there is a good chance of customers defaulting the payment or overusing the credit card in the sense that debt cannot be repaid. Thus, for the bank, identifying risky and non-risky customers is very crucial. In this project, we used different machine learning algorithms for coming up with solutions whether the customer will default or not. In this paper we have compared five machine learning algorithms, Logistic Regression, Linear SVM, RBF SVM, Random Forest and Neural Network.

2. Introduction

Credit card issuers have become one of the major consumer lending products in the U.S., representing roughly 30% of total consumer lending. Credit cards issued by banks hold the majority of the market share with approximately 70% of the total outstanding balance. There are differences in the credit card charge off levels between different adversaries. Over the period of time credit cards have become one of the formable means by which consumers can use a bank's capital for a short period of span. [4]

If consumers accept a credit card, then he\she agrees to pay their bills by the due date listed on the credit card statement. If not, the credit card will get defaulted. When a customer is not able to pay back the loan before the due date and the bank is totally assured about not being able to collect the payment, then it will usually consider selling the loan. Following, if the bank notices that they are unable to sell it, they will lower it down, and this is called a charge-off in Credit Card terms. This derives in compelling financial losses to the bank which also damages the credit rating of individual customers and hence it is a critical problem to deal with. [4]

Anticipating precisely which consumers are most plausible to default shows compelling business opportunity for all banks. Bank cards are the most accepted credit card type in the U.S., which highlight the jolt of risk prediction to both the consumers and banks. In a well-advanced financial structure, risk prediction is very crucial for predicting business performance or individual consumer’s credit risk and to reduce the damage and concern. [4]

The bottom-line target of the research is to carry out a dedicated default avoidance guideline to avail banks determine and take action on customers with high probability of defaulting to boost their credit scores. The challenge is to help the bank to advance its credit card services for the bilateral gains of consumers as well as the businesses. Even though plenty of solutions to the default prediction using the full data set have been previously done, even in published papers, the scope of our project expands beyond that, the eventual goal is to provide an easy-to-predict default mitigation program for the banks. [4] [5]

In addition to default prevention, the project includes a set of learning goals. And the key considerations in selecting analytics methods and how these analytics methods can be used efficiently to create direct business value. Also the objective of learning how to communicate complex topics to people with different backgrounds. An accurate predictive model can help the Banks and Companies to analyze customers who might default their payment in the future use so that the companies and banks can get

involved prior to handling risk and reducing loss. It is even better if a model can assist the company on credit card application approval to minimize the risk at sincere. However, credit card default prediction is never easy work. It is dynamic and a customer who paid their payment on or before time in the past few months may suddenly default their next payment. It is also unstable considering the fact that default payment is rare compared to non-default payments. Irregular dataset will easily fail using most machine learning techniques if the dataset is not treated properly. [4] [5]

3. Methods:

3.1 Dataset

We are using the dataset on the UCI website, and taking consideration of various attributes of customers including education, payment history, gender, marital status, age, bill statements etc.

We have used binary variables, default payment which denotes Yes = 1 and No = 0 respectively, for the response variance. We used the following 23 variables as information attributes:

A1: Amount for the allotted credit: this includes individual customer credit and their family credit information
A2: Sex (which is indicated as, 1:male; 2: female)
A3: Education (which is indicated as, 1: graduate college; 2: university; 3: school; 4: others).
A4: Marital Status (which is indicated as, 1:married; 2: single; 3 : others).
A5: Age (year).
A6 - A11: Previous payment History. In this dataset we have used tracked information of customer form his/her previous monthly payment records (from April to September 2005) as follows:
A6: the repayment status in September 2005
A7: the repayment status in August 2005
.
.
A11: the repayment status in April 2005.
The indicators for the repayment status measuring is: -1: paid on time; 1: payment delay for a month;
.
7: payment delay for eight months;
.
9: payment delay for nine months and above.
A12-A17: Bill statement amounted.
A12: The bill statement amounted in September, 2005;
A13: The bill statement amounted in August 2005
.
A17: The bill statement amounted in April, 2005.
A18-A23: Previous payment amounted
A18: Paid amount in September 2005;
.
A23: Paid amount in April, 2005. [7]

3.2 Data Preprocessing

3.2.1 Scaling / Normalizing data

In data preprocessing, we normalized the data using the standard normalization function in SKLearn called StandardScaler() which normalized the each attribute in the scalar. The reason behind normalizing or scaling the data is not each of the attributes we have in the dataset having the same range. For example, Age typically varies between 0-100 while the income can vary from 0-100000. And hence this may result in one attribute contributing more than another towards prediction, which most of the time isn't the case.

3.2.2. Dimension reduction

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. [7]

We used Principal Component Analysis (PCA) to overcome the curse of dimensionality, reducing the dimension of our dataset from (30000,25) to (30000, 15) with retaining 95% of the information.

3.3 Machine Learning Algorithms

We used the following machine learning to create a model which predicts payment defaulters.

1. Logistic Regression
2. Linear SVM
3. RBF SVM
4. Random Forest
5. Neural Network

1. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic loss function to model a binary dependent variable. It is used for binary classification. Problems that are confined to two classes. In regression analysis, logistic regression is estimating the parameters of a logistic model. [7][8]

Here, first we run gridsearch to find the best hyperparameter by splitting the dataset into 70/30 . We get the optimal C as 3.16 by running all combinations and using the scoring metric as AUC. After that, we test the model against the test data and we get the accuracy as 61% ROC-AUC-SCORE. We plot the confusion matrices for the train data as well as the test data from which we can see the true positive and false positive rates.

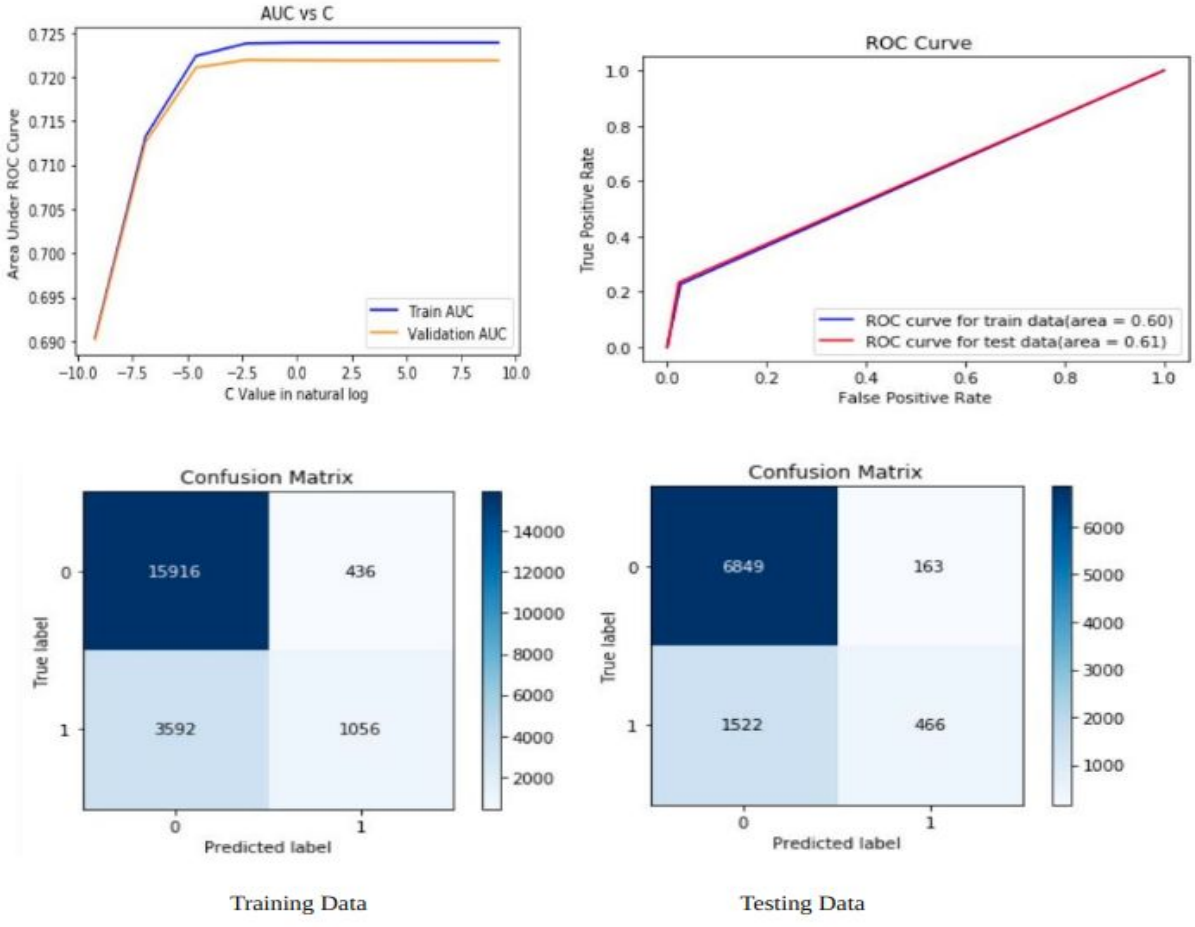
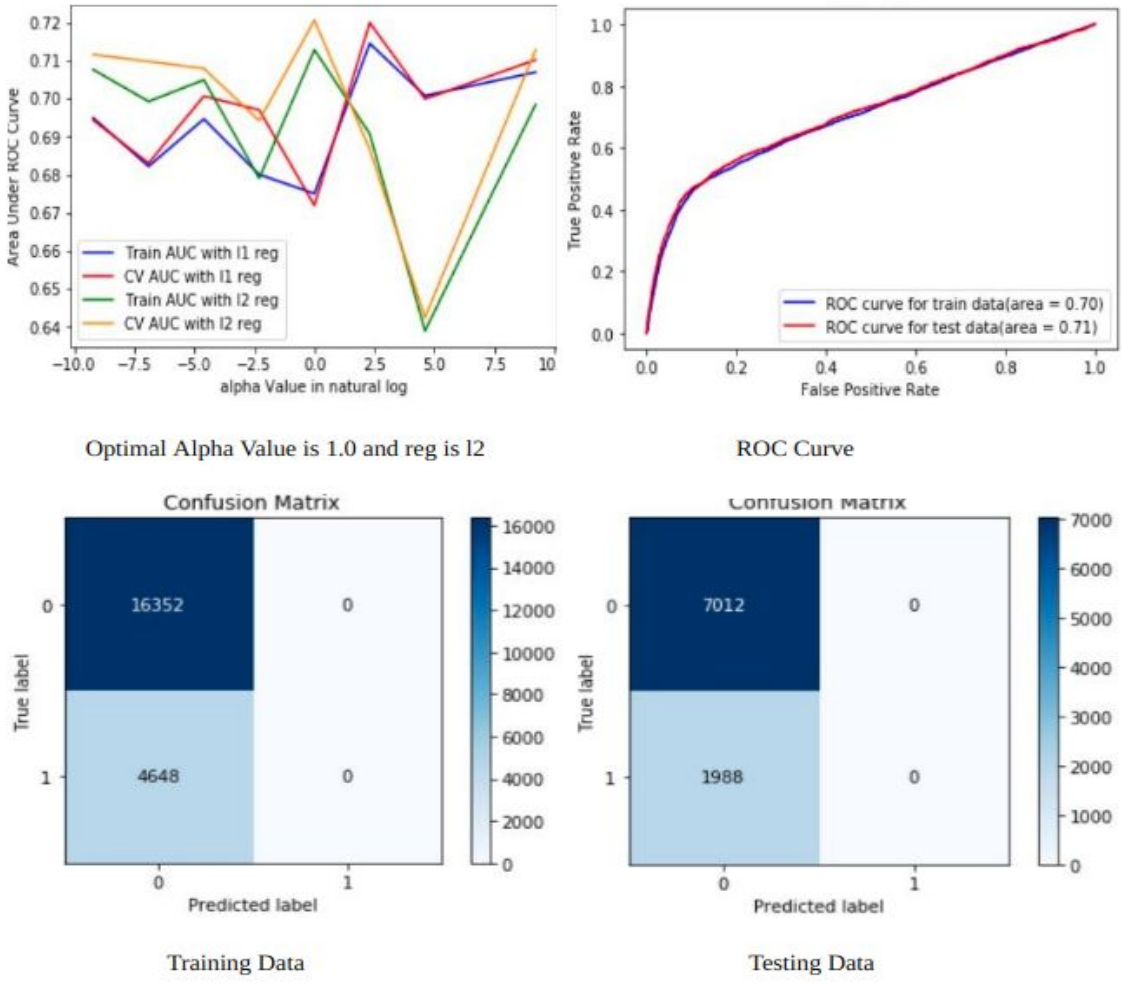


Fig.1

2. Linear SVM

Linear SVM is a machine learning algorithm for solving multiclass classification problems from large data sets that implements an original proprietary version of a cutting plane algorithm for designing a linear support vector machine. LinearSVM is a linearly scalable routine meaning that it creates an SVM model in a CPU time which scales linearly with the size of the training data set. [10]

Using the same procedure as we did on logistic regression, we use the splitted data, run grid search and find the optimal hyperparameter . We use this hyperparameter to run against the test data to get our model's accuracy(ROC-AUC-Score). For linear SVM, we get it as 71%.



Optimal Alpha Value is 1.0 and reg is l2

ROC Curve

Training Data

Testing Data

Fig.2

3. RBF SVM:

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.[10]

Following similar procedure as we did in logistic regression and linear SVM, for finding the optimal hyperparameter, in RBF SVM, we get the accuracy (ROC-AUC-Score) as 71%.

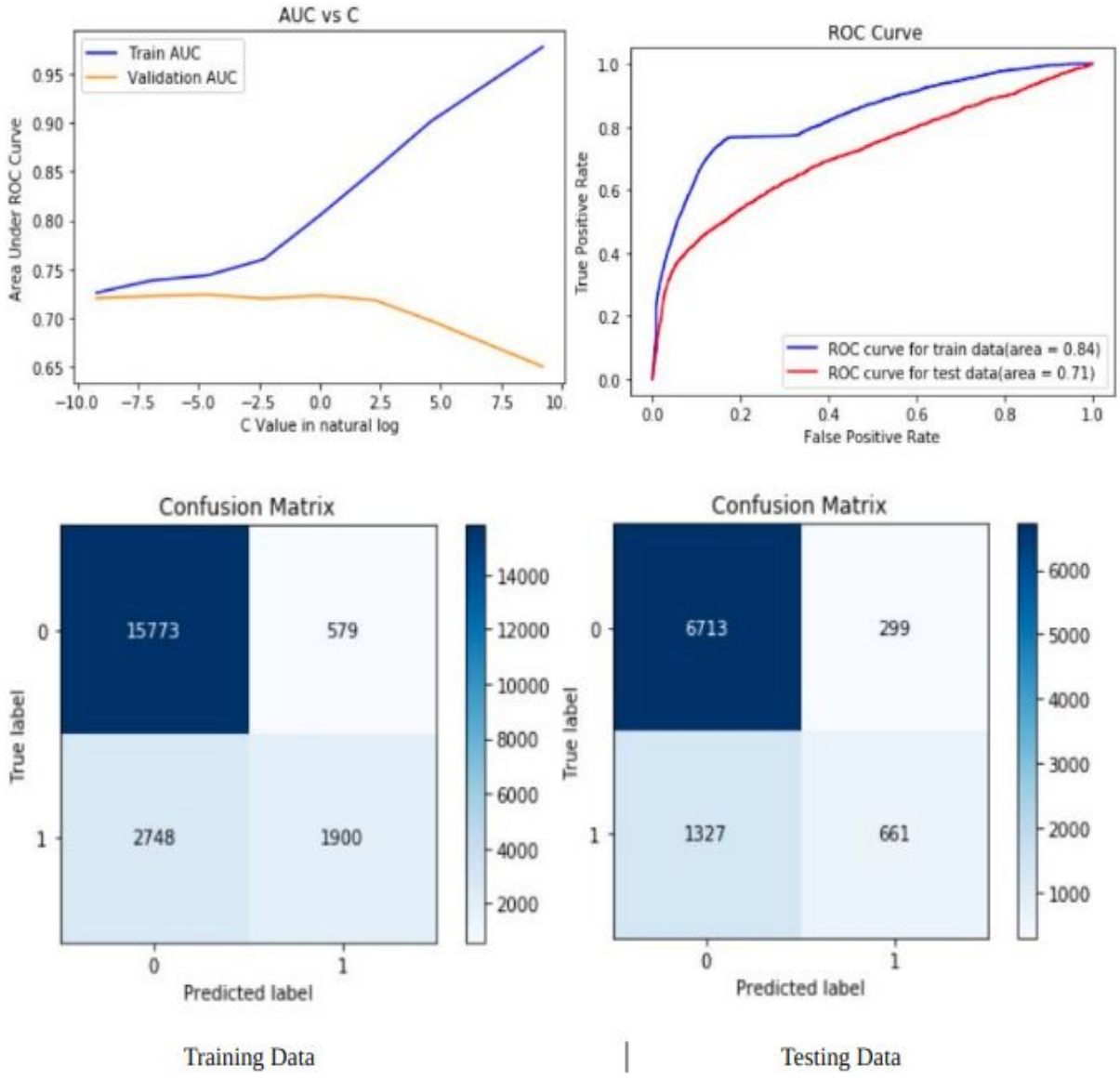


Fig. 3

4. Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. [11]

In random Forest, we have found the best accuracy (for ROC-AUC metric) as 76%.This was found by following similar procedures as before, such as finding optimal hyperparameter by performing grid search on train data and cv data, and then testing it on test data for testing the model. We have used seaborn libraries here to plot heatmaps instead of plotting normal curves for experimental purposes and this was found to be better than a normal curve for finding the hyperparameter.

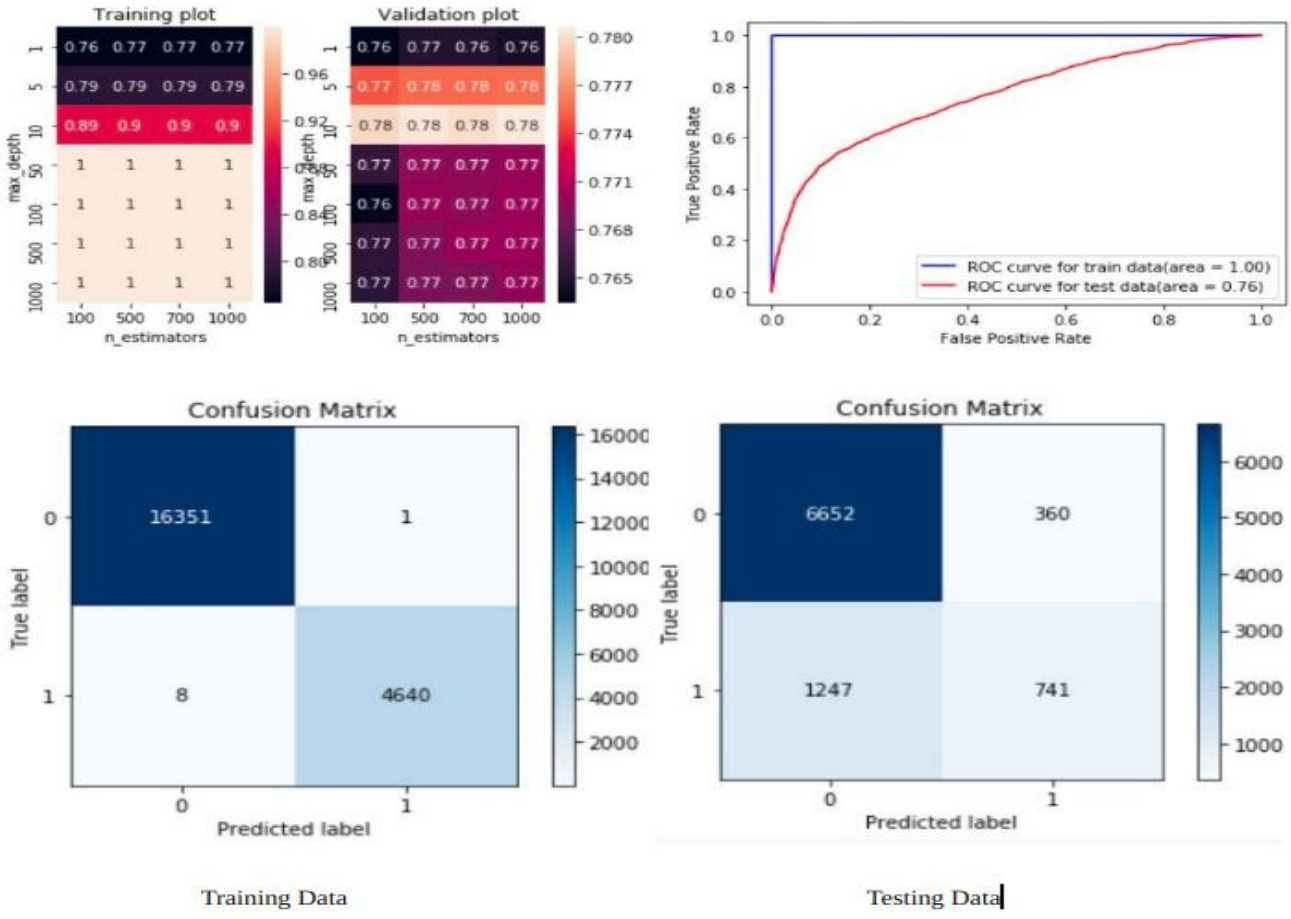


Fig. 4

5. Neural Networks

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labeling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.[12]

Here, we could have used just scikit learn / keras , but we wanted to go ahead and play with the **fastai** library which is state of the art for deep learning today. We have achieved an accuracy of 83%(metric used as accuracy) . However, we could not use the ROC-AUC metric with **fastai** due to some technical difficulties. Though 83% is very high, we cannot truly compare it with random forest's ROC-AUC score of 76%. Hence, we will consider random forest as the better model for now.

epoch	train_loss	valid_loss	accuracy	time
0	0.445409	0.417845	0.837667	00:04
1	0.466801	0.415495	0.837444	00:04
2	0.455183	0.413212	0.835667	00:04
3	0.448780	0.410167	0.838556	00:04
4	0.441445	0.413731	0.830333	00:04
5	0.436116	0.413110	0.837889	00:04
6	0.450096	0.419562	0.834556	00:04
7	0.445459	0.413448	0.829889	00:04
8	0.444447	0.412226	0.832667	00:04
9	0.445253	0.409354	0.836000	00:04

Fig.5

4. Results:

For all the models except Neural Network(fastai), we applied gridsearch to get the best optimal hyperparameter and then used that hyperparameter to test the model for getting the best accuracy in terms of ROC-AUC-score.We found the best model to be Random Forest Classifier with accuracy=76%(ROC-AUC-score). You can see the result in the following table.

Model	**Training Accuracy**	**Testing Accuracy**
Logistic Regression	**61%**	**60%**
Linear SVM	**71%**	**70%**
RBF SVM	**84%**	**71%**
Random Forest	**100%**	**76%**

For the Neural Network, we used a state-of-the-art , deep learning library called fastai. It gave quick results and an accuracy of 83% with metric as 'accuracy'.

Conclusion:

After comparing all the algorithms, we can see that for neural networks, even though it has higher accuracy of 83%, it uses the simple metric 'accuracy'. Hence, we consider the random forest classifier as the better one, since it has an accuracy of 76% with better metric.

References:

[1].https://bradzzz.gitbooks.io/ga-seattle-dsi/content/dsi/dsi_05_classification_databases/2.1-lesson/assets/datasets/DefaultCreditCardClients_yeh_2009.pdf

[2].https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[3].https://en.wikipedia.org/wiki/Accuracy paradox.Accessed on 4−23−2018

[4]. http://salserver.org.aalto.fi/opinnot/mat-2.4177/2018/McKinseyFinal.pdf

[5]. http://vista-analytics.com/wp-content/uploads/2017/05/CreditCard-Branded-V2.pdf

[6]. https://en.wikipedia.org/wiki/Curse_of_dimensionality

[7]. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

[8].https://www.r-bloggers.com/logistic-regression-in-r-a-classification-technique-to-predict-credit-card-default/

[9]. https://medium.com/greyatom/logistic-regression-89e496433063

[10]. https://en.wikipedia.org/wiki/Radial_basis_function_kernel

[11]. https://en.wikipedia.org/wiki/Random_forest

[12]. https://pathmind.com/wiki/neural-network