**TVS Credit Case Study**

Thesis submitted in partial fulfillment of the

requirements for the

**Post Graduate Diploma in Data Science**

By

**Vishal Upadhyay**

**18325760022**

Under the guidance of

Raghavendra

Designation

Manipal Prolearn

Bengaluru



**MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPA**

**TVS Credit Case Study**

Thesis submitted in partial fulfillment of the

requirements for

**Post Graduate Diploma in Data Science**

By

**Vishal Upadhyay**

**18325760022**

Under the guidance of

Raghavendra

Designation

Manipal Prolearn

Bengaluru

(If two guides are there indicate them side by side)



**MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPAL**

**TVS Credit Case Study**

Thesis submitted in partial fulfillment of the

requirements for

**Post Graduate Diploma in Data Science**

By

**Vishal Upadhyay**

**18325760022**

**Examiner 1** **Examiner 2**

Signature: Signature:

Name: Name:



**MANIPAL ACADEMY OF HIGHER EDUCATION, MANIPAL**

**CERTIFICATE**

This is to certify that the project work titled

**TVS Credit Case Study**

is a bonafide record of the work done by

**Vishal Upadhyay**

**18325760022**

In partial fulfillment of the requirements for the award of **Post Graduate Diploma in** **Data Science** under Manipal Academy of Higher Education, Manipal, Manipal and thesame has not been submitted elsewhere for any kind of certification/recognition.

Raghavendra

Designation

Manipal Prolearn

Bengaluru

**TVS CREDIT CASE STUDY**

Acknowledgements:

I would like to express my deep gratitude to MR. Raghavendra, my mentor for the project who clarified my understanding towards the data and guided me in the data preparation and modeling thanks are also extended to professors who taught us subjects in Data Science Diploma program specifically Exploratory Data Science, Unstructured Data, Deep Learning, Machine Learning and Python Programming. The teachings in these subjects helped me to complete the project. Finally, I wish to thank Manipal Global Academy team who supported with Infrastructure, logistics and helped me in completing the project and Post Graduate Diploma in Data Science program on time.

**Table of Content**

**1. Acknowledgements**

**2. Table of Contents**

**3. Introduction**

**3.1 Motivation**

**3.2 Project Scope**

**3.3 Project Goal**

**4. Project Description**

**4.1 Business Domain and Understanding**

**4.2 Data Limitation**

**4.3 Data Description**

**4.4 Benefits of Project**

**5. EDA**

**5.1 Data Collection**

**5.2 Complexity of Data**

**5.3 Data Transformation**

**6. Design**

**6.1Analytical Method and Technology used**

**6.2 Descriptive Analysis with Visualization**

**7. Modelling**

**7.1 Prediction Methodology**

**7.2 Selection of model/technique**

**7.3 Variable Selection**

**7.4 Algorithms**

**7.5 Analysis of Result and Performance Model**

**8 Future Scope**

**9. Conclusion**

**10. Reference**

**11. Appendix**

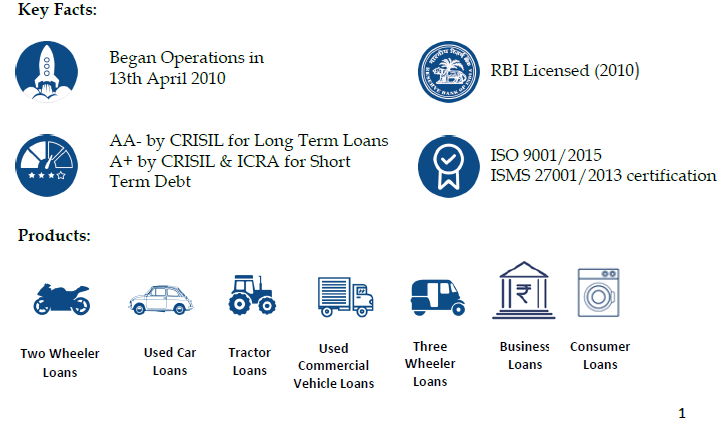
**3.Introduction:**

From the largest cities to the smallest villages, India is filled with ambition and enterprise. As Indians from all walks of life set out to write their growth story, our timely and affordable credit empowers them to bring their dreams alive.

As part of the $8.5 billion TVS Group, we empower Indians from various socioeconomic backgrounds with financial products that serve their needs. In doing so, we further the cause of financial inclusion.

Our two wheeler, used car, three wheeler, and tractor loans are designed for Indians in small towns and the rural heartland, for our nation's growth is powered by their prosperity. Our foray into the used commercial vehicle and consumer durable finance is yet another step in this direction.

With over 5 million customers and a long-term CRISIL rating of AA-, our growth is built on firm foundations. We have won several awards, including the Flame Award for Excellence in Rural Marketing, The Best BFSI Company Award at the ET Now Makers of Developed India Awards 2018, and the Most Effective Employee Engagement Strategy Award at the World HRD Congress.



**3.1 Motivation:**

The loan is one of the most important products of the banking. All the banks are trying to figure out effective business strategies to persuade customers to apply their loans. However, there are some customers behave negatively after their application are approved. To prevent this situation, banks have to find some methods to predict customers’ behaviors. Machine learning algorithms have a pretty good performance on this purpose, which are widely-used by the banking. Here, I will work on loan behaviors prediction using machine learning models.

**3.2 Project Scope:**

Loans default will cause huge loss for the banks, so they pay much attention on this issue and apply various method to detect and predict default behaviors of their customers. In this blog, I am going to talk about the basic process of loan default prediction with machine learning algorithms.

**3.3 Project Goal:**

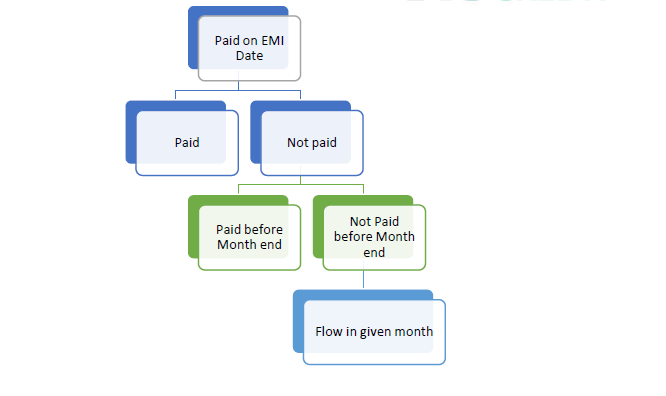
Objective of this exercise is to predict which customer is going to flow to N+1 bucket in coming month based on their past payment history, previous month collection, demographics, profile, external data etc.

**4.Project Description:**

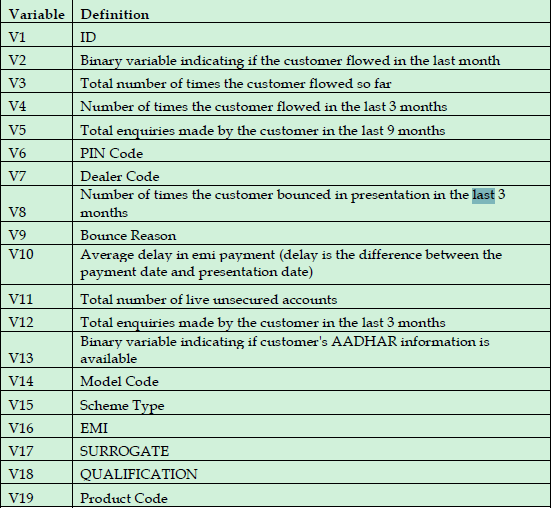
**4.1 Business/Domain Understanding:**

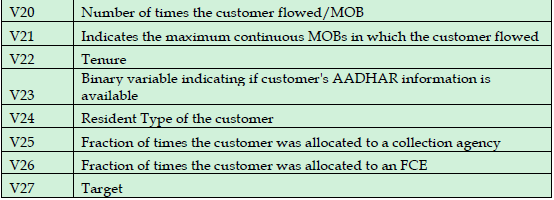
The primary thing of this Organization is to provide their wealth in the safer hands. In recent times, banks approve loan after verifying and validating the documents provided by the customer. Yet there is no guarantee whether the applicant is deserving or not. This paper classifies the customers based on certain criteria. The classification is done using Exploratory Data Analysis. Exploratory Data Analysis (EDA) is an approach to analyse the datasets that summarizes the main characteristics with visual methods. The purpose of using EDA is to uncover the underlying structure of a relatively larger set of variables using visualizing techniques. Whenever the bank makes decision to give loan to any customers then it automatically exposes itself to several financial risks. It is necessary for the bank to be aware of the clients applying for the loan. This problem motivates to do an EDA on the given dataset and thus analyzing the nature of the customer. The dataset that uses EDA undergoes the process of normalization, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualizing the data in the graphical format. Python is used for easy and efficient processing of data. This paper used the pandas library available in Python to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualization of the results and for better understanding. For obtaining the graph Mat plot library is used.

**4.2 Data Limitations:**



**4.3 Data Description:**





**4.4 Benefits of project** :

We use Predictive Modeling techniques to figure out which customers are going to flow into the Next Bucket in the current Month i.e. customers moving from Bucket N to N+1.

Once a customer fails to pay an EMI on particular EMI date (Due date) during presentation, TVS Credit uses different channels to recover money from the customer. After all the efforts if there are still some customers who don’t pay in that month, such customers are said to have moved to Next Bucket (N+1) i.e. they have defaulted on (N+1) Payments, N being the Opening Bucket of customer (Bucket at start of the month).

5. **Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an approach to analyze the datasets that summarizes the main characteristics with visual methods. The purpose of using EDA is to uncover the underlying structure of a relatively larger set of variable using visualization techniques.

**5.1. Data collection:**

Data is taken from Crypto. CRYPTO is the Annual Data hackathon sponsored by TVS Credit.

**5.2 Complexity of Data:**

The next step is to look at the data we’re working with. Realistically, most of the data we will get can have errors, and it’s important to identify these errors before spending time analyzing the data. I have identified several aspects in the data set before I continue with my analysis. Let’s review it:

* There are no missing values in any of the variables.
* Looking at the distributions of the data, we noticed no outliers.
* No Duplicates.

**5.3 Data Transformation**

* Converting Categorical Column in Numerical Columns. As Data not in orders so I decided to go for One Hot Encoding.



**6. Design:**

**6.1 Analytical Method and Technology used:**

As the project is developed in python, we have used Anaconda for Python 3.6.5 and jupyter notebook.

• Anaconda

Itis a free and open source distribution of the Python and R programming languages

for data science and machine learning related applications (large-scale data processing,

predictive analytics, scientific computing), that aims to simplify package management

and deployment. Package versions are managed by the package management system

condo. The Anaconda distribution is used by over 6 million users, and it includes more

than 250 popular data science packages suitable for Windows, Linux, and MacOS.

• Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows you to create

and share documents that contain live code, equations, visualizations and narrative

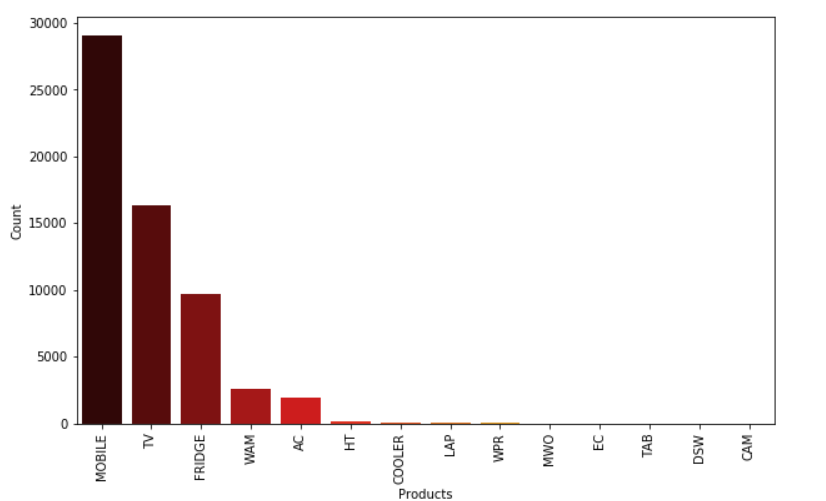
text. Uses include: data cleaning and transformation, numerical simulation, statistical

modeling, data visualization, machine learning, and much more.

**6.2** **Descriptive Analysis with Visualization:**

* Top Products on which emi is taken.

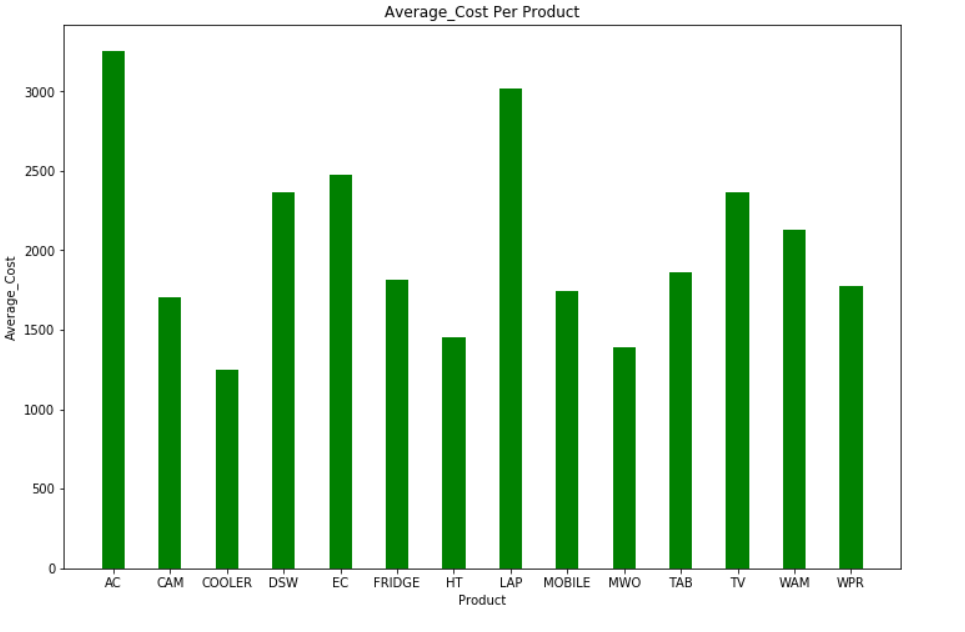




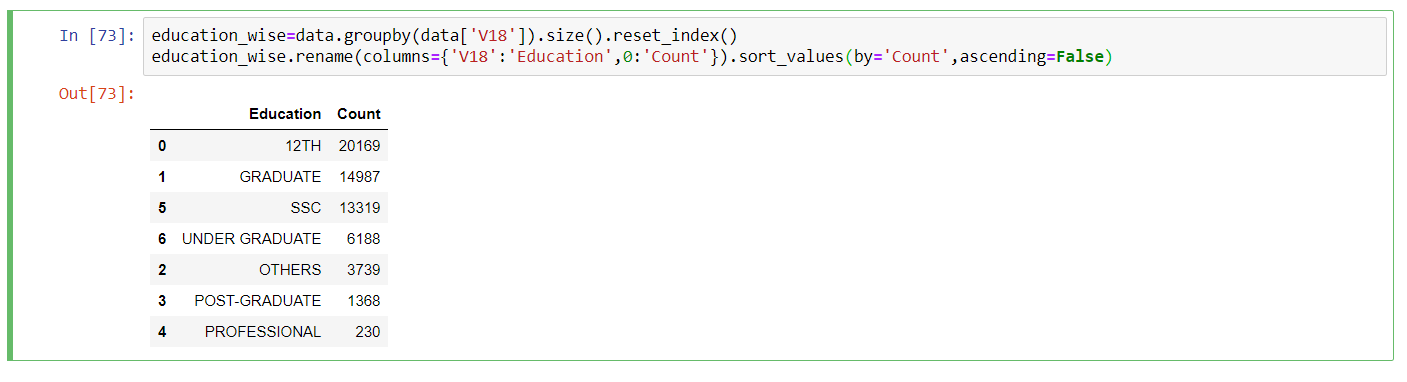
* Average emi cost per Product.



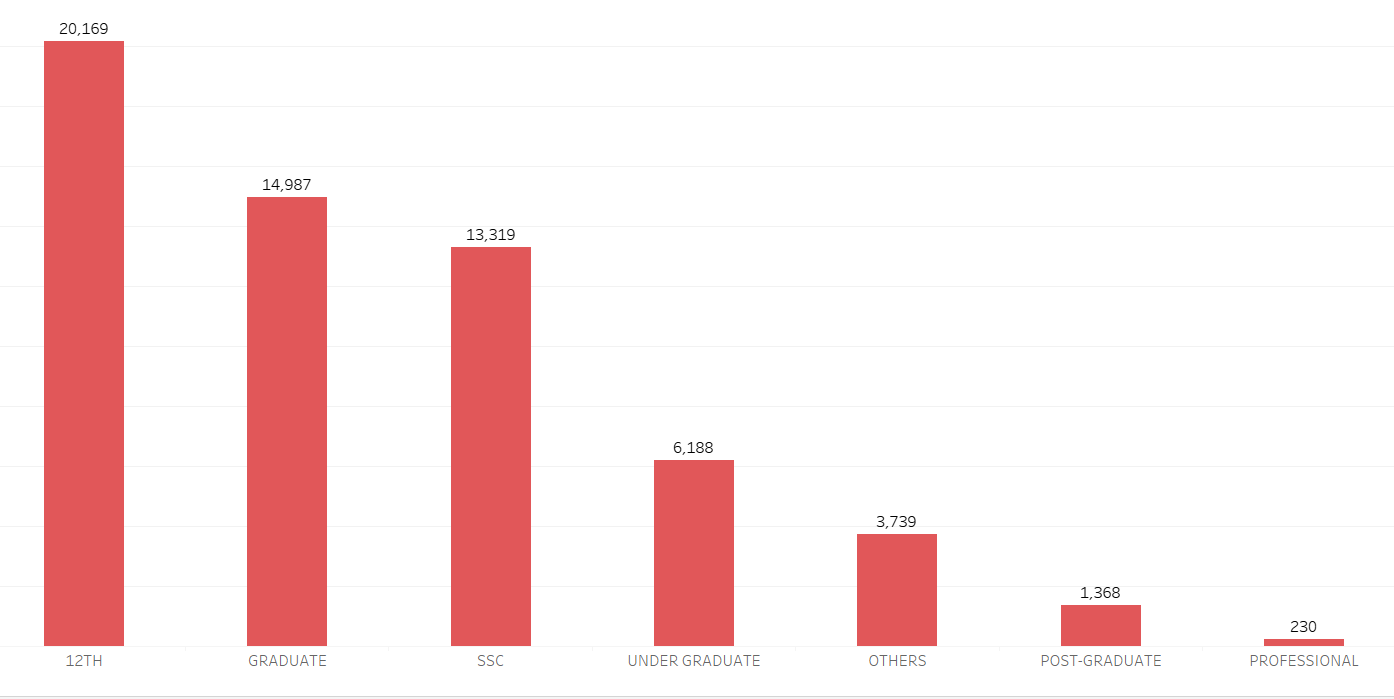
Visualization:



* Education Wise Loan Type.



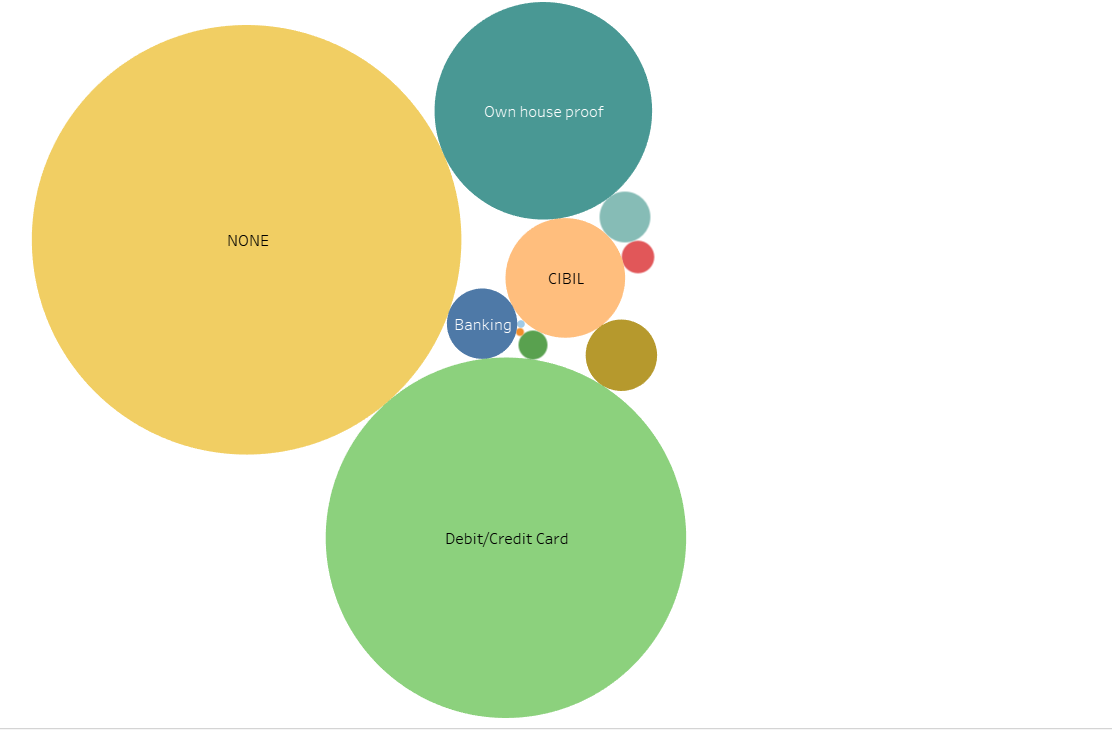
Visualization:



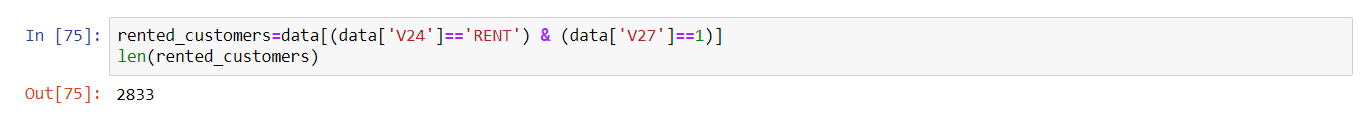
* Most Paytment type:

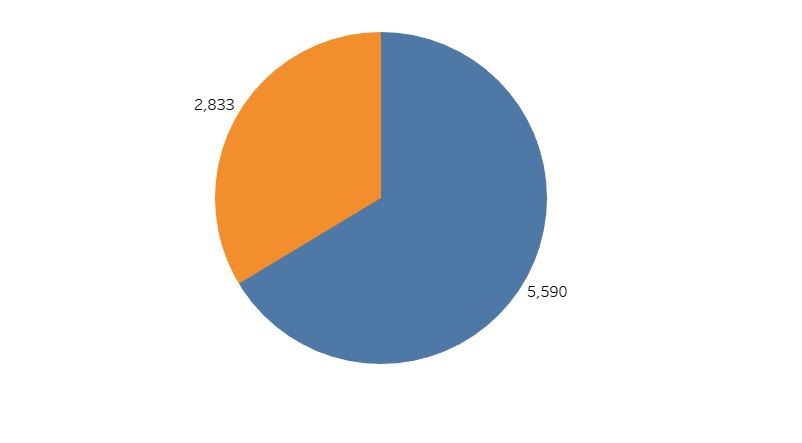


Visualization:

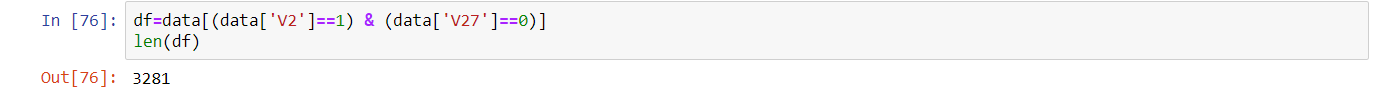


* Customers who are rented the Products are defaulters.





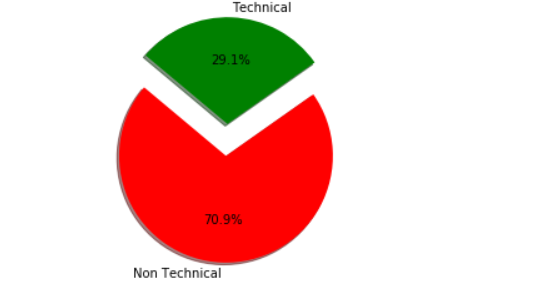
* Customers who flowed in last month will they able to pay loan at next month.



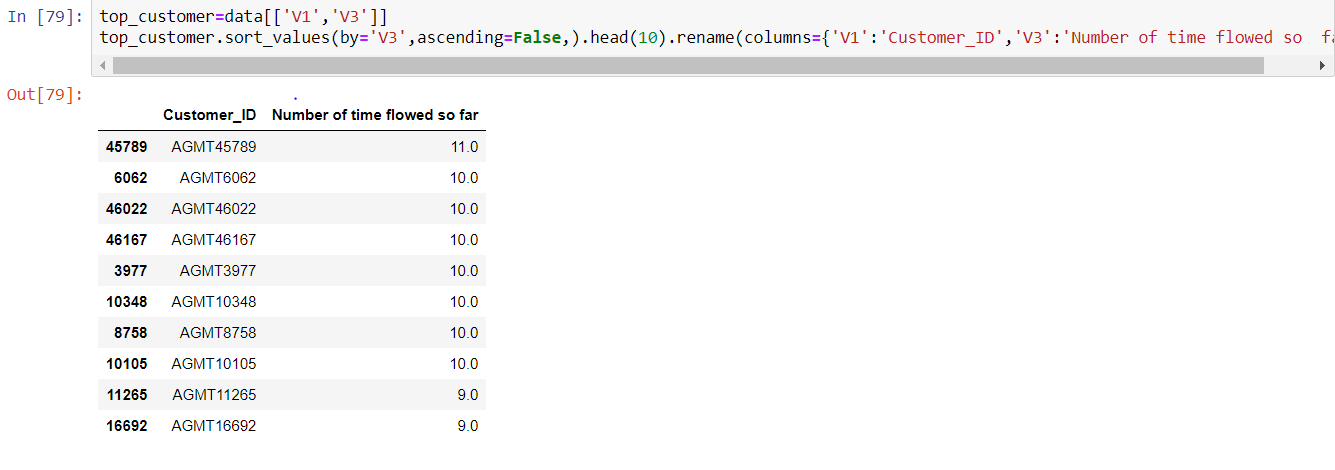
* Bounce Reason.



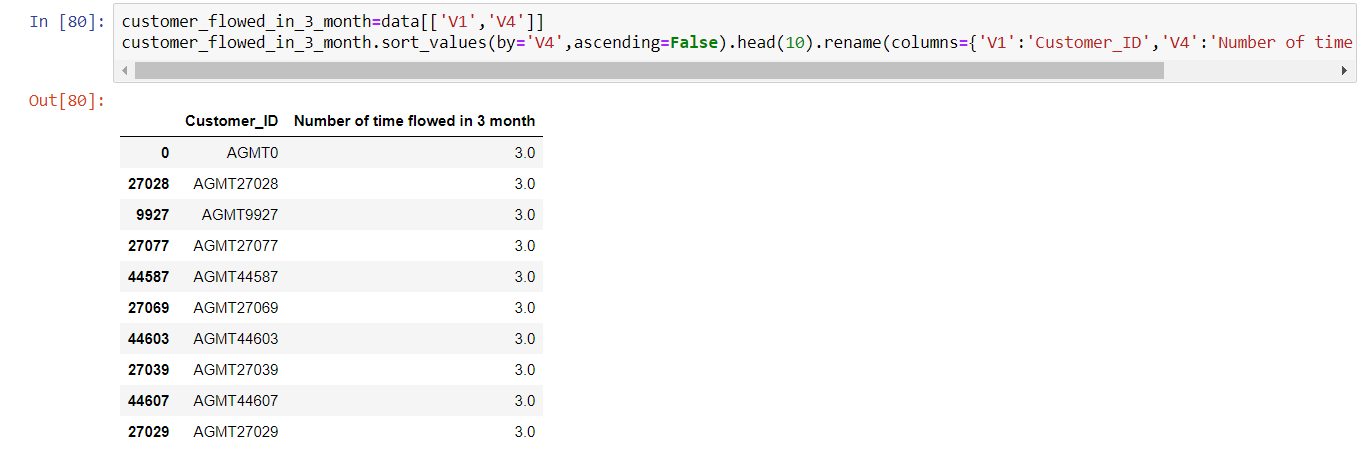
Visualization:



* Top 10 customers who flowed so far most of the times.



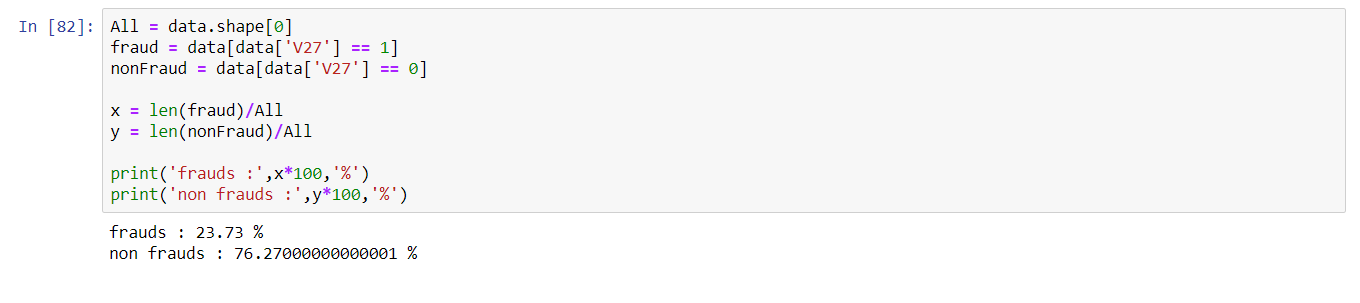
* Number of times the customer flowed in the last 3 months.



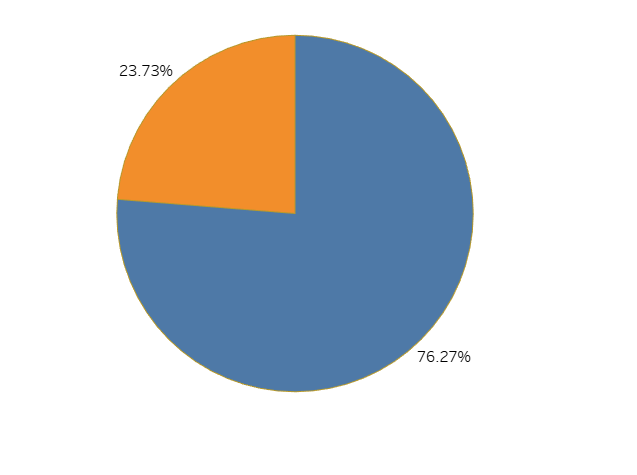
* Total enquiries made by the customer in the last 9 months.



* Total Percentage of Frauds and Non-Frauds:

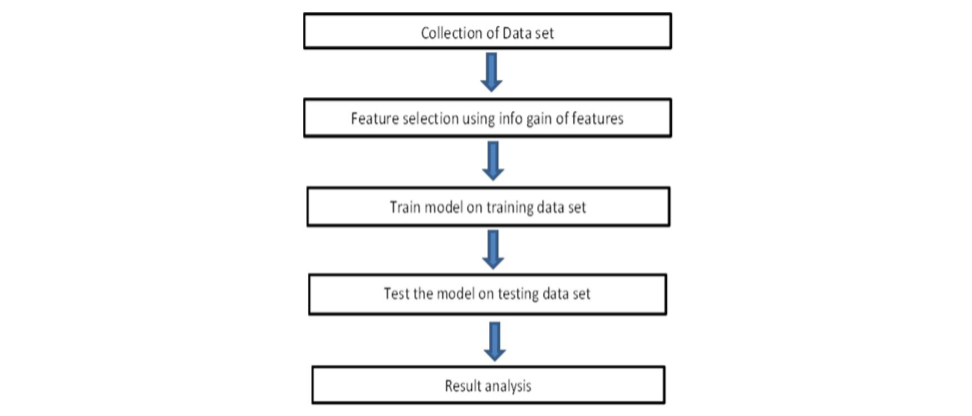


Visualization:



**7. Modelling:**

**7.1 Prediction Methodology:**

****

**7.2 Selection of model/technique:**

The ability of machine learning models to predict whether a customer is going to pay his EMI for his current month or not makes them particularly interesting to Risk team to take some action on it. This expanded post provides how the machine learning process to our dataset in order to predict delinquency. The process includes variable selection, model selection, model evaluation, and model tuning.

For this exercise, the problem is to build a model capable of predicting which EMI will become severely delinquent, defined as falling behind six or more months on payments. This delinquency variable was calculated from the performance dataset for all EMI and merged with the acquisition data based on the EMI’s unique identifier.

Three machine learning classification models have been used for prediction. The models are available in Python Open Source Software called Scikit learn. Scikit-learn is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities. Scikit-learn provides dozens of built-in machine learning algorithms and models, called [estimators](https://scikit-learn.org/stable/glossary.html#term-estimators). Each estimator can be fitted to some data using its [fit](https://scikit-learn.org/stable/glossary.html#term-fit) method. Before Machine Learning Model we have select important feature to predict the target variable. This process is called Feature Selection method.

**7.3 Variable Selection**:

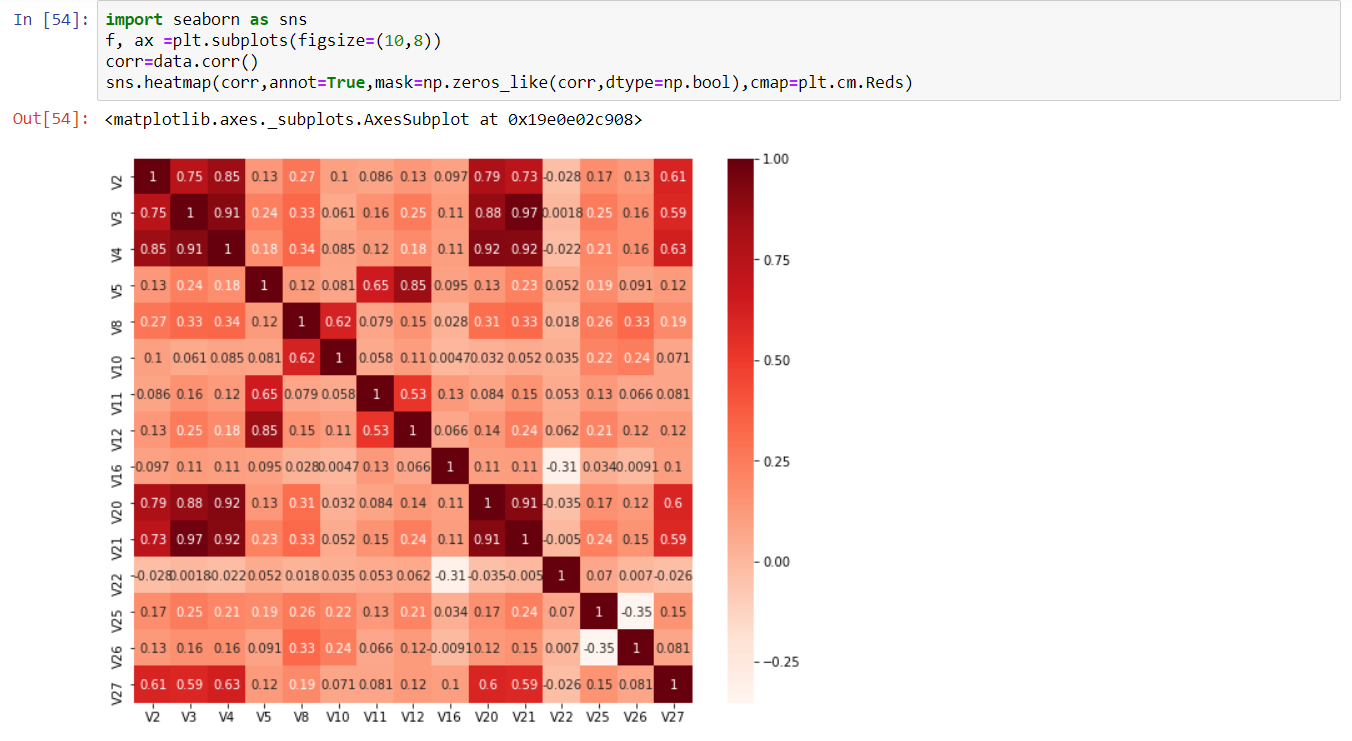
1.Filter Method

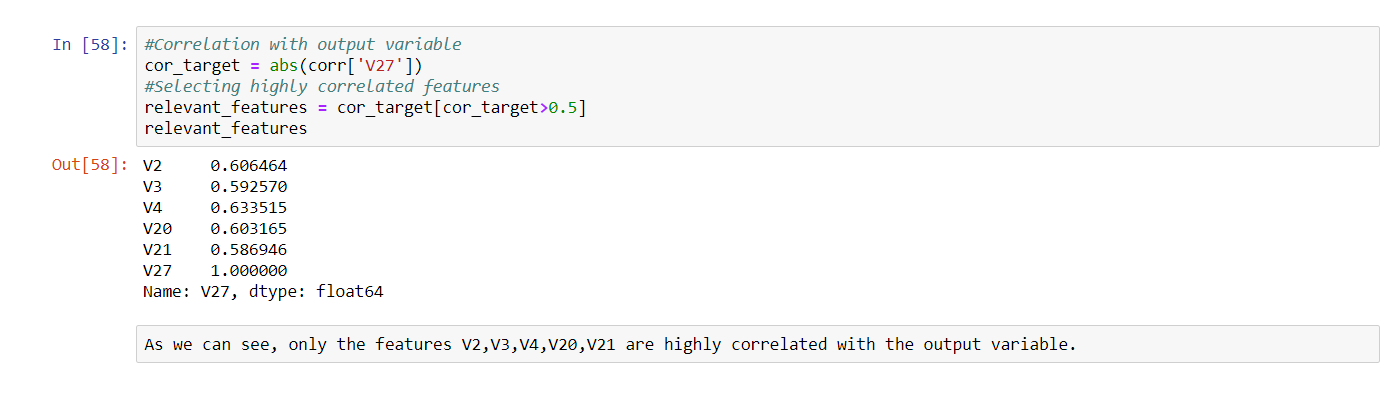
As the name suggest, in this method, you filter and take only the subset of the relevant features. The model is built after selecting the features. The filtering here is done using correlation matrix and it is most commonly done using [Pearson correlation](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient).

Here we will first plot the Pearson correlation heatmap and see the correlation of independent variables with the output variable MEDV. We will only select features which has correlation of above 0.5 (taking absolute value) with the output variable.

The correlation coefficient has values between -1 to 1  
— A value closer to 0 implies weaker correlation (exact 0 implying no correlation)  
— A value closer to 1 implies stronger positive correlation  
— A value closer to -1 implies stronger negative correlation

As we can see, only the features V2,V3,V4,V20,V21 are highly correlated with the output variable V27. Hence, we will drop all other features apart from these. However this is not the end of the process. One of the assumptions of linear regression is that the independent variables need to be uncorrelated with each other. If these variables are correlated with each other, then we need to keep only one of them and drop the rest. So let us check the correlation of selected features with each other. This can be done either by visually checking it from the above correlation matrix or from the code snippet below.



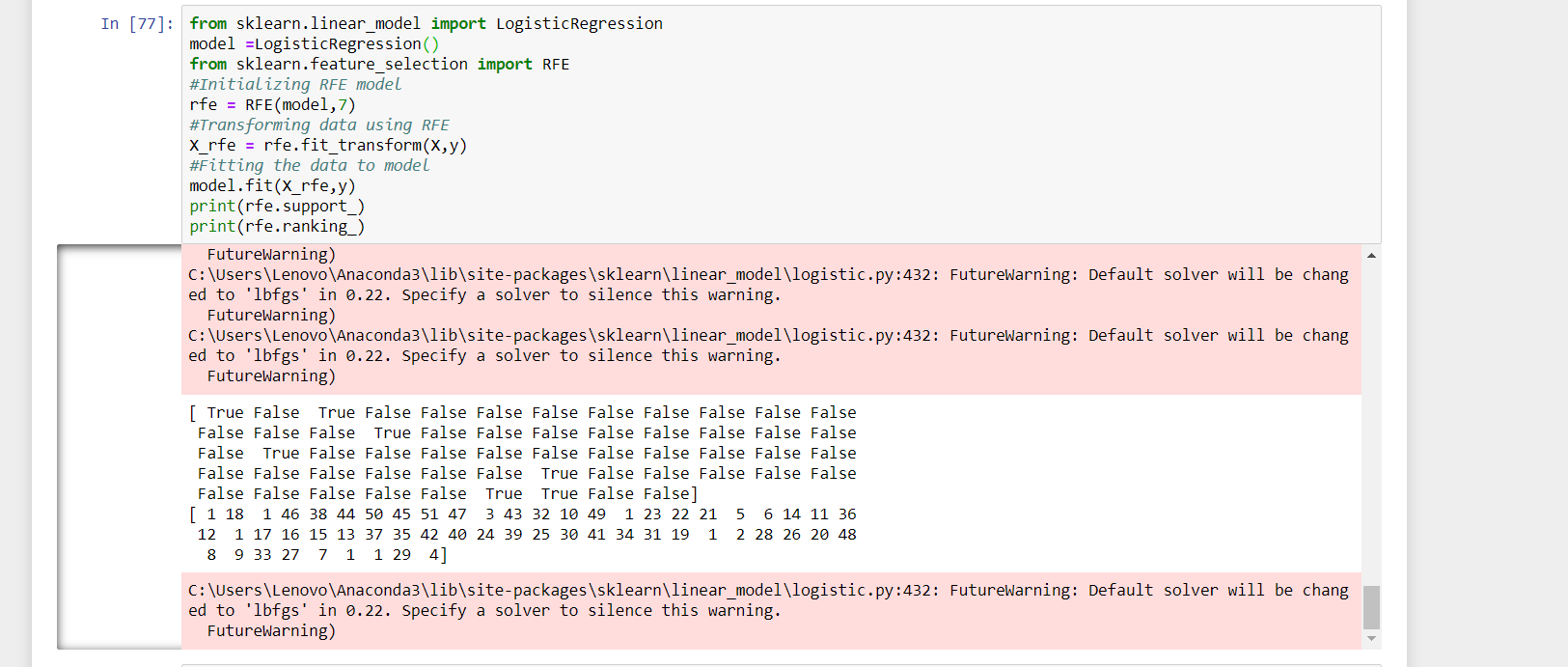


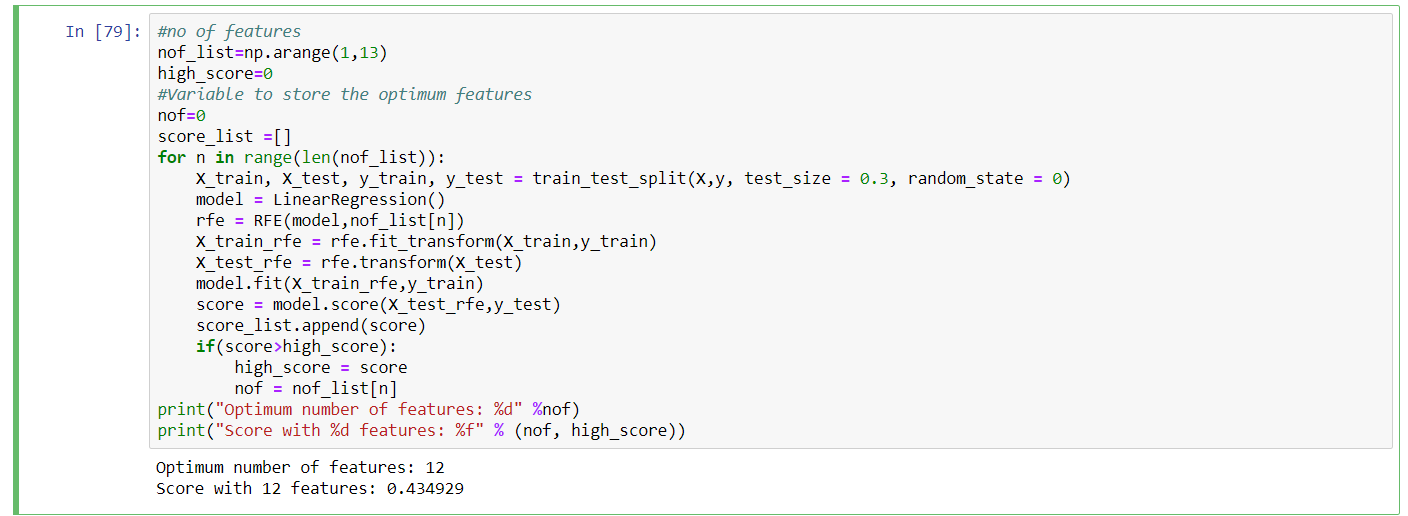
2.RFE (Recursive Feature Elimination):

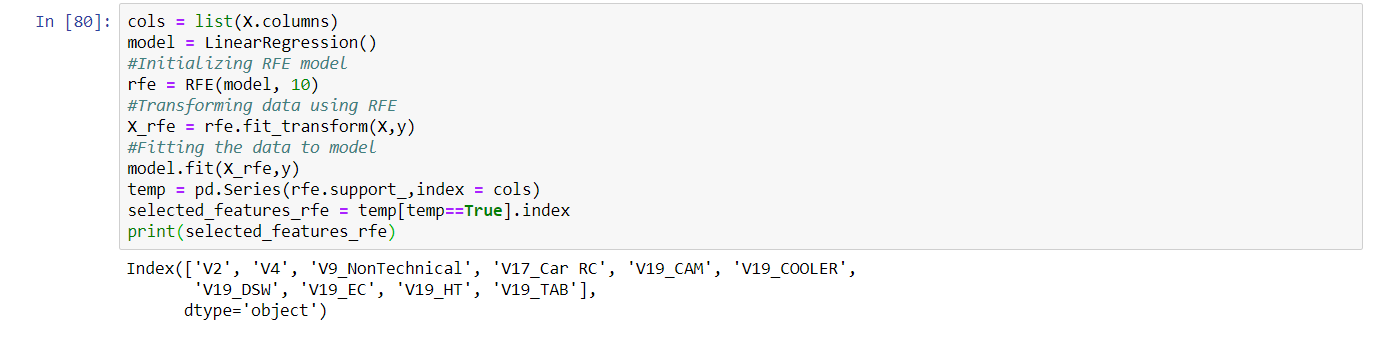
The [Recursive Feature Elimination](https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html) (RFE) method works by recursively removing attributes and building a model on those attributes that remain. It uses accuracy metric to rank the feature according to their importance. The RFE method takes the model to be used and the number of required features as input. It then gives the ranking of all the variables, 1 being most important. It also gives its support, True being relevant feature and False being irrelevant feature.

Here we took Logistic Regression model with 7 features and RFE gave feature ranking as above, but the selection of number ‘7’ was random. Now we need to find the optimum number of features, for which the accuracy is the highest. We do that by using loop starting with 1 feature and going up to 13. We then take the one for which the accuracy is highest.

As seen from below code, the optimum number of features is 12. We now feed 12 as number of features to RFE and get the final set of features given by RFE method, as follows:

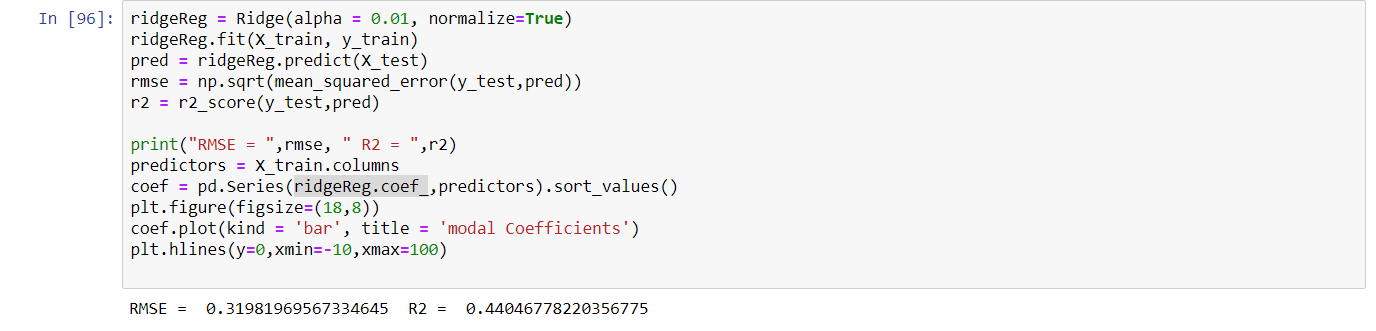


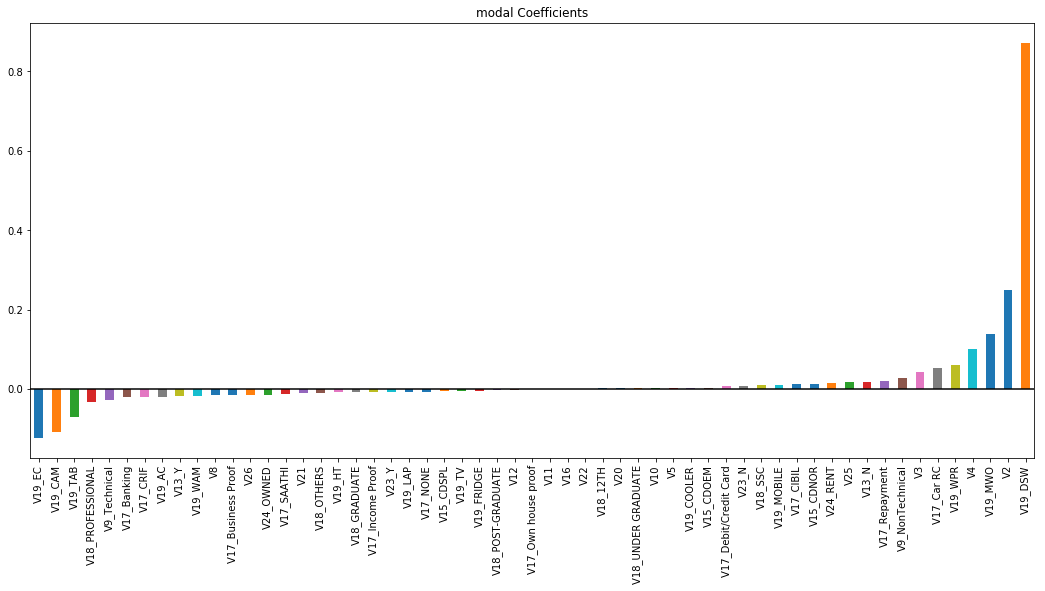




3.Ridge Regression:

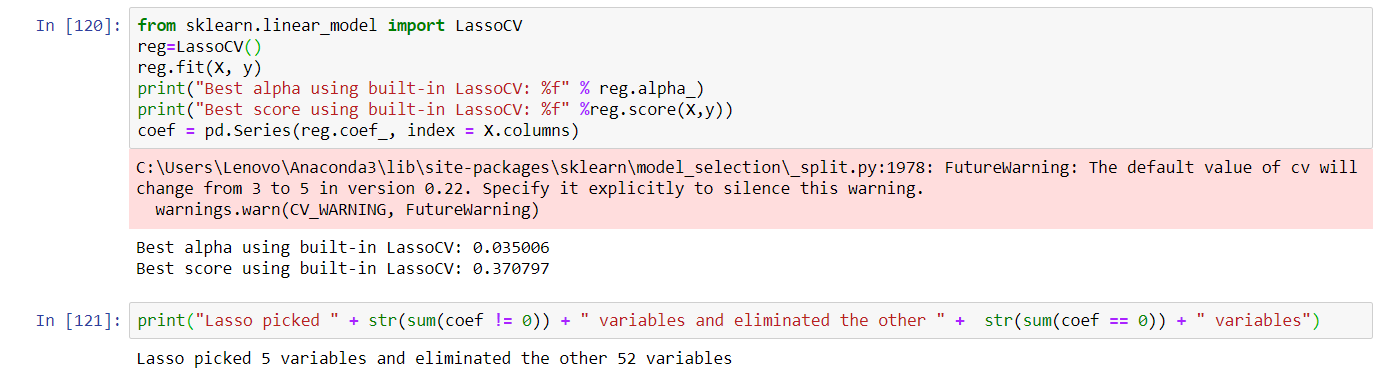
Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value.

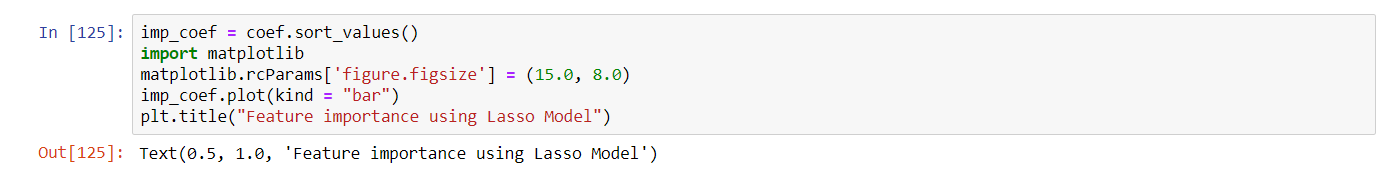




4.Lasso Regression:

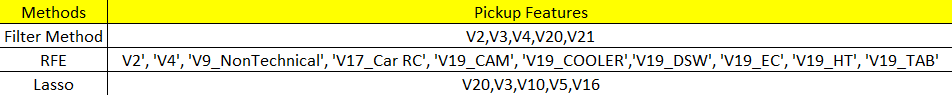
We will do feature selection using Lasso regularization. If the feature is irrelevant, lasso penalizes its coefficient and make it 0. Hence the features with coefficient = 0 are removed and the rest are taken.







Hence from the feature selection methods we got:



**7.4 Algorithms:**

* Logistic Regression:

Logistic regression aims to model the probability of an event occurring depending on the values of independent variables. These independent variables are the various categorical or numerical information available to us regarding the loan, and these variables can help us model the probability of the event (in our case, the probability of default). These variables are also called predictor variables. Some examples of these predictor variables are provided below:

Personal details: Personal details of the borrower such as age, employment status, profession, income, residential status, and number of dependents. Credit history: Length of credit history, number and value of past loans, number and value of past delinquent loans.

All these variables can be used as predictor variables to predict the probability of default. So, using logistic regression, we model the probability of default using other independent variables as described above.

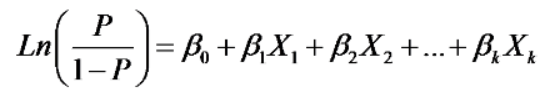
The logistic regression model seeks to estimate that an event (default) will occur for a randomly selected observation versus the probability that the event does not occur. Suppose we have data for 1000 loans along with all the predictor variables and also whether the borrower defaulted on it or not. Here the probability of default is referred to as the response variable or the dependent variable. The default itself is a binary variable, that is, its value will be either 0 or 1 (0 is no default, and 1 is default).

In logistic regression, the dependent variable is binary, i.e. it only contains data marked as 1 (Default) or 0 (No default).

We can say that logistic regression is a classification algorithm used to predict a binary outcome (1 / 0, Default / No Default) given a set of independent variables. It is a special case of linear regression when the outcome variable is categorical. It predicts the probability of occurrence of a default by fitting data to a logit function.

The Link Logit Function a link function is simply a function of the mean of the response variable Y that we use as the response instead of Y itself. In our example, Y represents default.

All that means is when Y is categorical, we use the logit of Y as the response in our regression equation instead of just Y:



The logit function is the natural log of the odds that Y equals one of the categories.  For mathematical simplicity, we’re going to assume Y has only two categories and code them as 0 and 1.

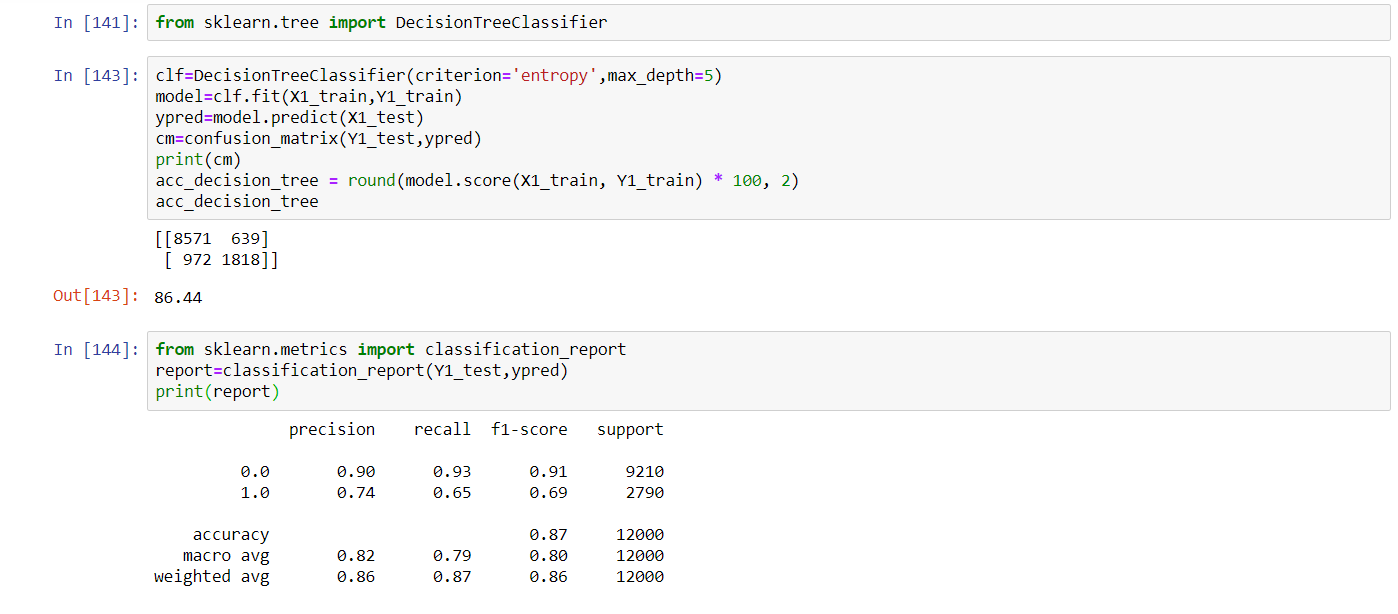
The logit function is the inverse of the logistic transform. When the function’s variable represents a probability p, the logit function gives the log-odds, or the logarithm of the odds p/(1 − p). The log-odds score is typically the basis of the credit score used by banks and credit bureaus to rank people.

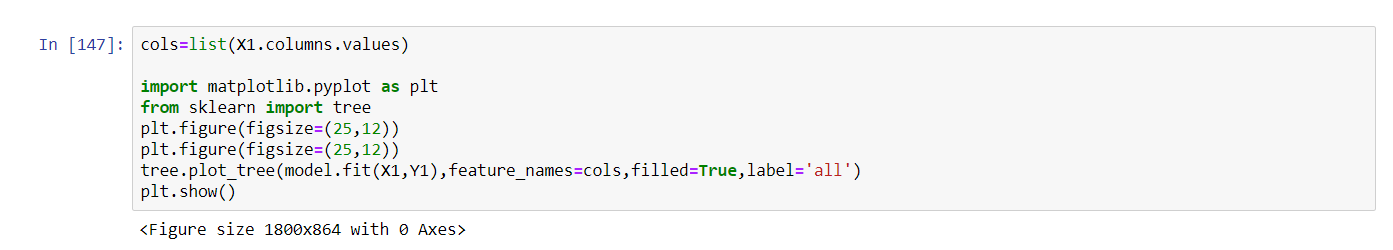
P is defined as the probability that Y=1 (Representing Default).  So for example, those Xs could be specific risk factors, like age, income, employment status, credit history, and P would be the probability that a borrower defaults. B0 is an intercept and ( B1…Bk) is a vector of coefficients, one for each predictor variable.

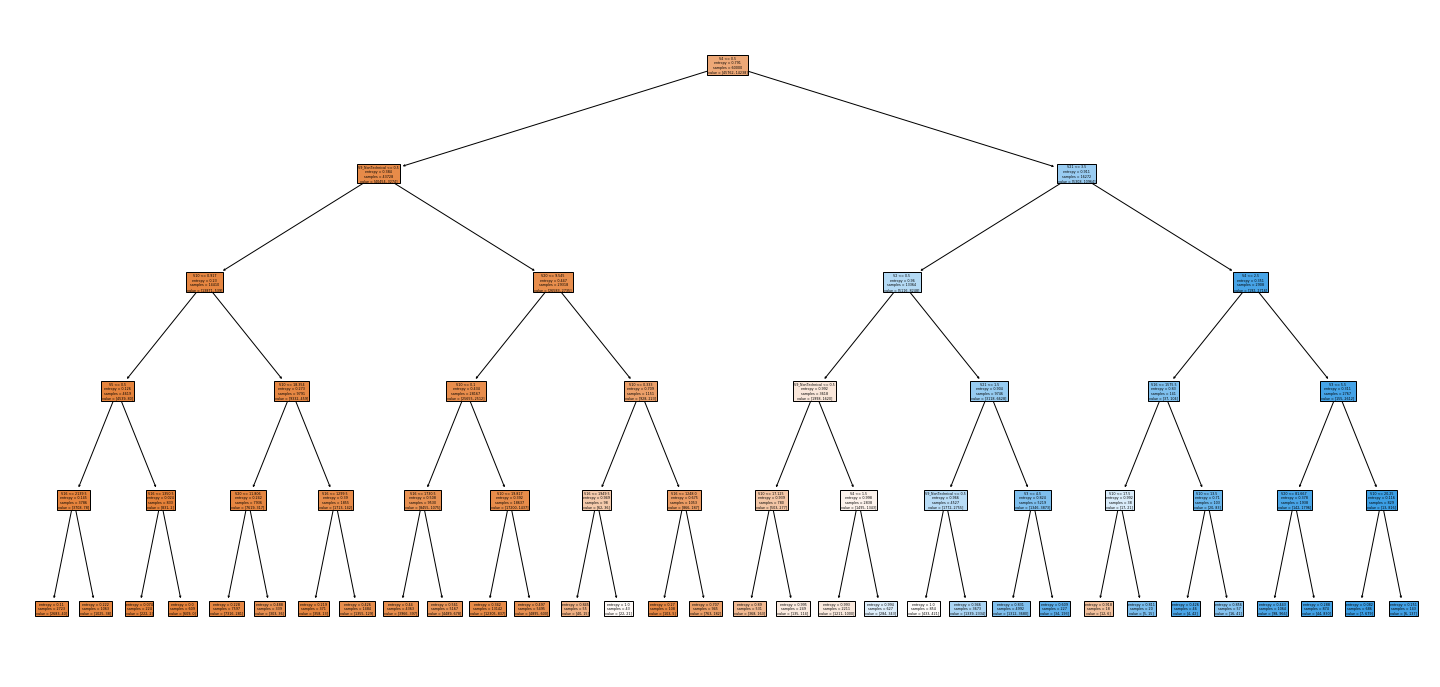


* Decision Tree:

Decision tree is a type of supervised learning algorithm (having a predefined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.





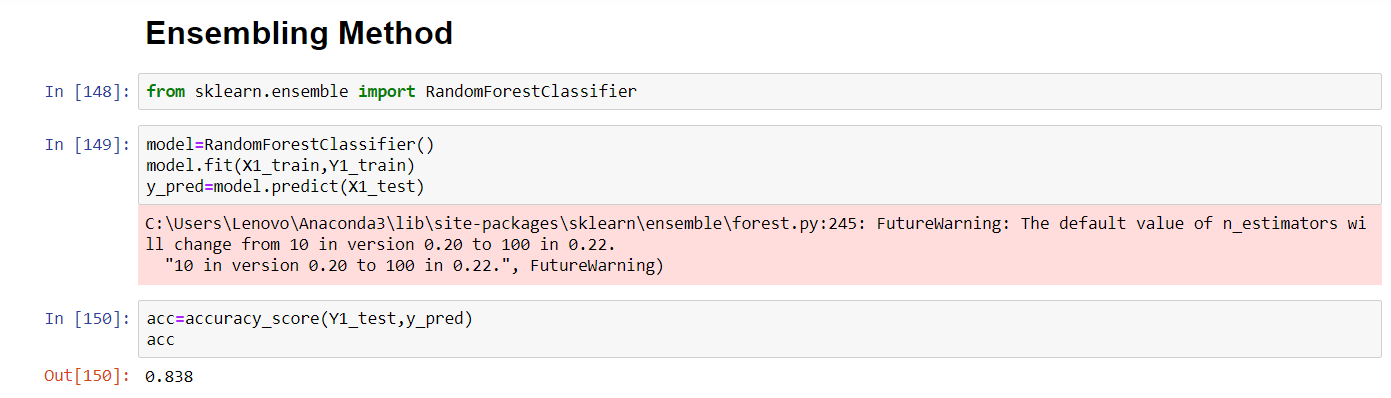


* Random Forest:

Random forests or random decision forests are an [ensemble learning](https://en.m.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.m.wikipedia.org/wiki/Statistical_classification), [regression](https://en.m.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.m.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.m.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.m.wikipedia.org/wiki/Overfitting) to their [training set](https://en.m.wikipedia.org/wiki/Test_set).

Decision trees are a popular method for various machine learning tasks. Tree learning "come closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", say [Hastie](https://en.m.wikipedia.org/wiki/Trevor_Hastie) et al., "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate.

In particular, trees that are grown very deep tend to learn highly irregular patterns: they [overfit](https://en.m.wikipedia.org/wiki/Overfitting) their training sets, i.e. have [low bias, but very high variance](https://en.m.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff). Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.



|  |
| --- |
|  |

**7.5 Analysis of Result and Performance Model:**

**Roc Curve:**

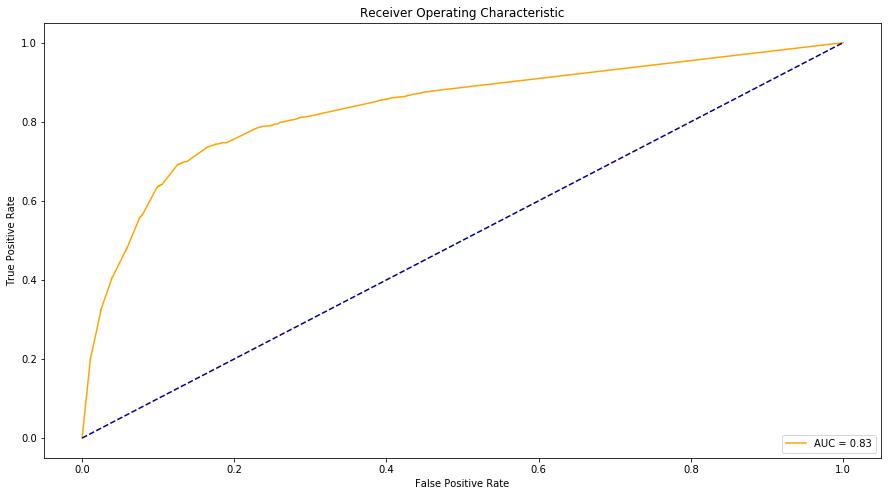
AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. In fact it means it is reciprocating the result. It is predicting 0s as 1s and 1s as 0s. And when AUC is 0.5, it means model has no class separation capacity whatsoever.

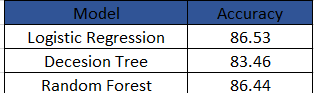
Let’s interpret above statements. As we know, ROC is a curve of probability. So lets plot the distributions of those probabilities:

* Logistic Regression Roc Cu
* rve: Auc is near to 1 which means model is good.



**Model Validation:**

Performance of Each Model is as Follow:



**8 Future Scope:**

To know customer is able to pay EMI or not is very helpful for any finance Organization. Distribution of the loans is the core business part of almost every bank and any organization. The main portion the bank’s and organization assets is directly coming from the profit earned from the loans distributed by the banks. The prime objective in banking environment is to invest their assets in safe hands where it is. Today many banks/financial companies approve loan after a regress process of verification and validation but still there is no surety whether the chosen applicant is the deserving right applicant out of all applicants. Through this system we can predict whether that particular applicant will pay his emi for a recent month or not by validation of past features is autmated by machine learning technique. The disadvantage of this model is that it emphasizes different weights to each factor but in real life sometime customer has different reason for not paying emi, which is not possible through this system. EMI Prediction is very helpful for employee of banks and any organization as if they can predict this segment of customer will not pay emi then they can take precaution according to that.

**9 Conclusion:**

From a proper analysis of positive points and constraints on the component, it can be safely concluded that the model is a highly efficient. This application is working properly and meeting to all Banker and any finance company requirements. It helps organization in making right decision to segment good or bad customer. One can check their profiles carefully later by the credit risk management team. In this way, banks can detect the default behaviors in the earlier stage and conduct the corresponding actions to reduce the possible loss.

**10 References:**

[4]. K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011). [5]. J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086. [6]. Mean Decrease Accuracy https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.hml [7]. J.R. Quinlan. Induction of decision trees. MachinelearningSpringer, 1(1):81–106, 1086. [8]. Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News( http://CRAN.R-project.org/doc/Rnews/ ), 2(3):9–22, 2002. [9]. S.S. Keerthi and E.G. Gilbert. Convergence of a generalizeSMO algorithm for SVM classifier design. M