

Micro-Credit Defaulter Model Project

Submitted by:

* Prathamesh Vilas Mistry

**ACKNOWLEDGMENT**

I would like to express my special thanks of gratitude to “FlipRobo Technologies” that made this project possible. I would like thank my SME Mr. Nitin Mishra for his guidance in building this project. I would also like to thank Data Trained Institution to making me capable of making proper decisions in the field of damascene and Machine learning. Lastly, I would like to thank my parents to make all of this happen.

References:

* Data Preparation for Machine Learning by Jason Brownlee

## Applied Predictive Modeling by By Max Kuhn and Kjell Johnson

# Feature Engineering and Selection: A Practical Approach for Predictive Models by Max Kuhn and Kjell Johnson

# Medium.com

# AnalyticsVidhya.com

# Research Papers:

# RANDOM FORESTS: Leo Breiman Statistics Department University of California Berkeley, CA 94720 January 2001

**INTRODUCTION**

* Business Problem Framing

We have collaborated with a client from the Telecom Industry that provides micro-credit on mobile-balanced (main as well as data) to its customers. The company provides micro-credit to the customers which is to be paid within a period of 5 Days. We have to build a model that will help the client to Loan-out credit to those population who will be able to pay back the loan back within 5 Days based of the historical data.

* Conceptual Background of the Domain Problem

As we are aware of the targeted customer of our client is the population with low income, we can say almost all of the customers would be mostly taking loan for calling and not for internet services. So, concentrating on the main balance should be the key here.

* Review of Literature

Electronical Communication is the need of the day. For shopping for bread to hollering the Emergency Services there’s no second to communication.

The Companies belonging from the telecom industries are well aware of it, and they want to approach there customers furthermore than just providing telecom services. The Companies wants to lend a small amount of credit to those customers those who have their services expired and need to use the service right away without paying the complete bill at the point of time. This loan is to be paid back within a period of 5 days with a particular interest. If the customer does not pay the amount with interest within the given period of time, the consumer believed to be defaulter.

We have to build a model which can be used to predict in terms of a probability for each loan transaction, whether the customer will be paying back the loaned amount within 5 days of loan.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

The Problem is of Classification.

The Data consists of 32 features in the dataset.

There are more than 200K samples in the dataset.

All the Data is of numerical datatype.

Use various algorithms to build up the model.

Almost all of the features are highly skewed, skewness is to be addresses.

There are some samples with non-negotiable values to the respective features that must be addressed.

The data is unscaled.

There are no missing values in the dataset.

The target classes are highly imbalance, imbalance must be addressed.

The target has 2 classes only, it is a binary classification problem.

As the target contains highly imbalance classes (85%-15%), we may use AUC\_ROC as out primary scoring and evaluation metric.

* Data Sources and their formats

The data was provided by the client to “FlipRobo Technologies”. The data is in the form of a comma separated file (CSV). The data i.e. the features and the target are in the single file.

* Data Pre-processing Done

The Data pre-processing done is as follows:

1. Skew transformation using Cube root Transformation.
2. Reducing Dimensions by removing the features with about 0 percent variance.
3. Reducing Dimensions by removing the features that are independent of the target variable using ANOVA Test (f\_classify)
4. Over Sampling of the minority class using SMOTE. Increased minority class samples by 1.5% only.
5. Last 0.5 Percent of Quantile stripped of some features, 2 to be exact (features had extremely high skew , about 1400 samples removed in total using quantile reduction)
6. Standard Scaling the data

(Note: All of the above processes are carried out strictly on the training dataset only and no test data is exposed to the model.)

* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

Data is inputted in the form of a Pandas data frame to the model. The model is evaluated on a validation set using 3 Time Repeated, 5 fold Stratified cross validation set with AUC\_ROC as the scoring parameter , And the model which preforms the best on the validation set is used for prediction of the classes and their respective probability per record on the test set.

* State the set of assumptions (if any) related to the problem under consideration

There are no such formal assumptions as we are Random-Forest to be the best model, producing better results than any other algorithms. Random Forests Algorithms are non-parametric and can thus handle skewed and multi-modal data as well as categorical data that are ordinal or non-ordinal.

* Hardware and Software Requirements and Tools Used

Hardware Required:

* A computer with a processor i3 or above.
* More than 4 GiB of Ram.
* GPU preferred.
* Around 100 Mib of Storage Space.

Software Required:

* Python 3.6 or above
* Jupyter Notebook.
* Google Collab.
* Excel

Tools/Libraries Used:

1. Computing Tools:

* Numpy
* Pandas
* Scipy
* Sk-learn

1. Visualizing Tools:

* Matplotlib
* Seaborn

1. Saving Tools:

* Joblib

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

* Testing of Identified Approaches (Algorithms)

Listing down all the algorithms used for the training and testing.

The Algorithms used for testing, training and Validating the models are as follows:

* Logistic Regression
* SVC (with an rbf kernel)
* Decision Tree
* K Nearest Neighbour
* Naïve Bayes
* Random Forest
* Gradient Boosting
* Run and Evaluate selected models

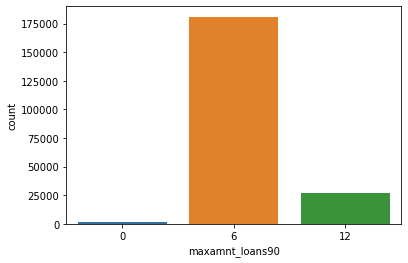
Algorithms used and their Evolutions of the Selected Models:



* Key Metrics for success in solving problem under consideration

The key-metric under considerations is AUC\_ROC although the model was finalized on basis of other metrics as Matthew’s Correlation Coefficient (MCC) as well as F1-score.

* Visualizations
* Popular Loan that are preferred by the customers.



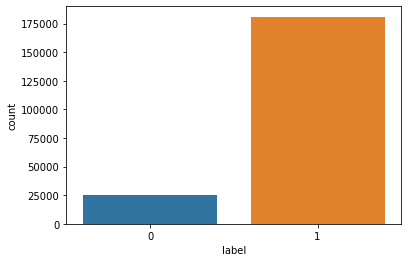
A large percentage of the population prefer loan of 6 Units than 12 Units.

* KDE plot of Defaulters vs Non Defaulters (orange: defaulters)



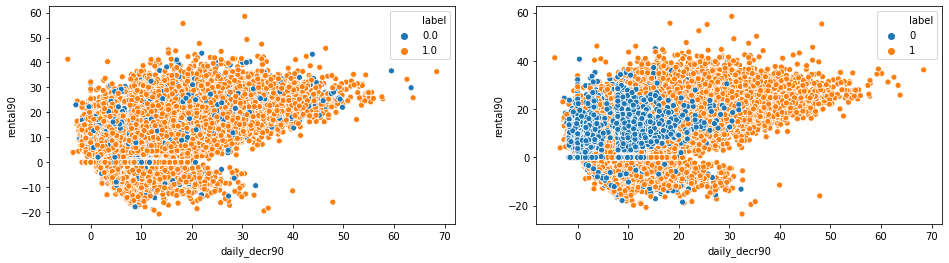
Density of the defaulter increase as the payback time increases. We can visualize that between the values 0 to 25 on X-axis.

* Imbalance of the target classes



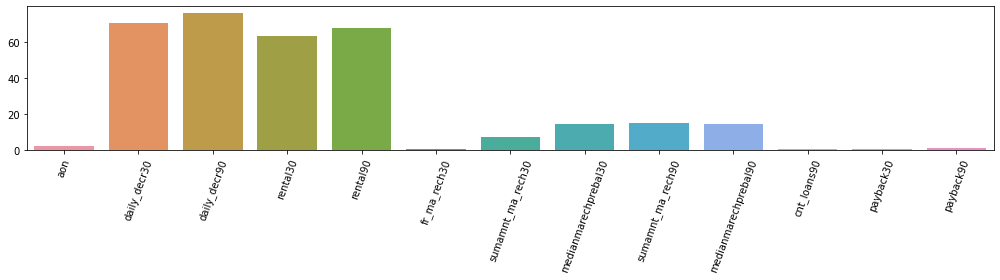
The target classes are highly imbalanced.

* Raw Dataset vs Over Sampled Dataset using SMOTE



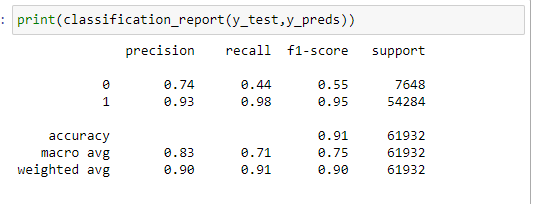
The scatter plot on right is the one which is over sampled.

* Prominent Features chosen for data modelling



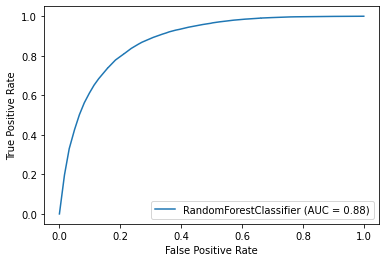
These are the selected features which contribute a weighted amount for prediction of the classes in the target .

* Interpretation of the Results



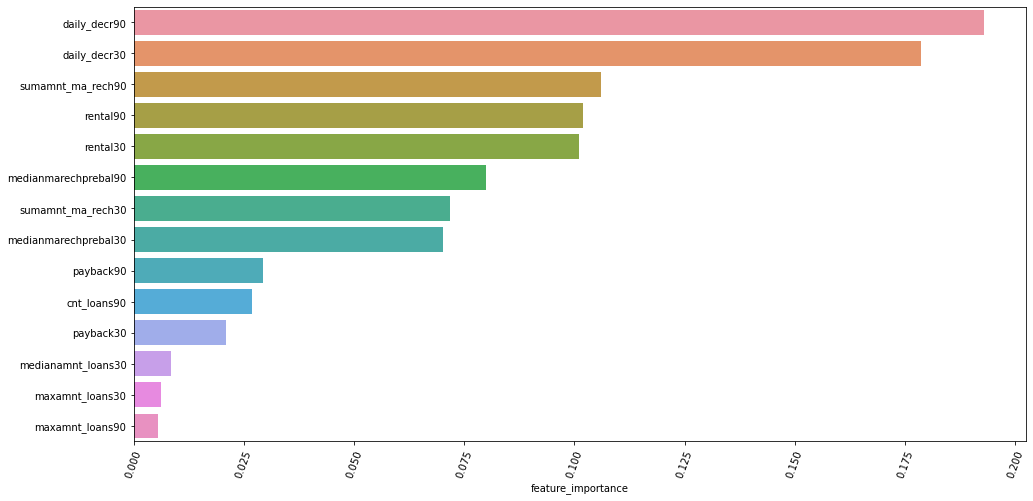
This is the classifications report on the test set. Since we have high imbalance in our target classes we used AUC\_ROC to evaluate the model.

* ROC curve on the test data



We received a AUC score of 88% on the test data using Random Forest.

The below horizontal bar plot gives the feature with the importance in building the model.



**CONCLUSION**

* Key Findings and Conclusions of the Study

A very few of the customer take loan for Internet Services.

Most of the features pay their loan with interest on time.

Most of the population opt for the loan of 06 Units rather 12 Units.

Ensemble Techniques learn large data well without any extra-ordinary requirements.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Outcomes of the Study:

* Almost 90 percent of the time is spent of data cleaning and data modelling.
* Outliers are not to be removed casually as they may contribute to the model and our predictions.
* You do not get a Gaussian distribution in real-word problem.
* Every less than half a quantile of data may make the distributions highly skewed.
* Algorithms like Support Vector Machines and K nearest neighbours may take a long time to converge on a Hugh dataset like this.
* Naïve Bays is very quick as of converging rate.
* Limitations of this work and Scope for Future Work

A data with statistics of beyond 90 days would be even better for data analysis.

Some features such as network-speed received by the user, charges paid for roaming by the customer, may help in better detailed modelling.

The model could be integrated with the analytics app used by the Data Analysts and Statistician of the respective telecom provider for easy decision.

The model could be integrated with the mobile bot that lends loan to the customer to directly fetch the data from the database from the company servers of the particular customer, make predictions and allow or deny the customer loan bases on its predictions.