
# MICRO_CREDIT_DEFAULTER -MODEL -1843

Submitted by:

RAJESH KAKUMANU

# ACKNOWLEDGMENT

I would like to acknowledge the sources and references to process my assignment during the process of completing the work given by FlipRobo Technologies.

I would like to mention the web and internet sources like-

- Sci-kit libraries
- Research papers in GitHub and Kaggle
- GitHub links of some professionals to whom I followed.
- Google – to get good interaction n understanding.

I would like to thank, DATA TRAINED for making me this activity, to go in a better way in future also.

I would like to thank, FlipROBO Technologies for assigning me the assignment works every week and to submit on time and helped me a lot to solve some issues through their timecards and updates in the websites.

# ***INTRODUCTION***

- ## Business Problem Framing

*A Classic Business problem which helps Micro Financing Institutions and other Lending companies reduce Credit risks by recognizing potential Defaulters. Machine Learning can help lenders predict potential defaulters before approving their candidature using their past data. The candidates' income, past debt and repayment behaviour can be important metrics for the same.*

- ## Conceptual Background of the Domain Problem

*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 insurance of loan. In this case, Label '1' indicates that the loan has been payed i.e., non-defaulter, while, Label '0' indicates that the loan has not been payed i.e., defaulter.*

*Find Enclosed the Data Description File and The Sample Data for the Modelling. Exercise.*

- ## Review of Literature

*For the following project, I took the information from various sources and literature for the completion.*

*-https://seaborn.pydata.org/generated/seaborn.catplot.html*

*-https://scikit-learn.org/stable/search.html?q=metrics*

*-https://github.com/topics/loan-default-prediction*

- ## Motivation for the Problem Undertaken

*Before advancement of Data Science, loan lending companies used to risk a high rate of defaulting. Many a times a perfect candidate would display erratic financial and repayment behavior after being approved for loan. Machine Learning can help lenders predict potential defaulters before approving their candidature using their past data. The candidates' income, past debt and repayment behavior can be important metrics for the same.*

# Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

- ## Data Sources and their formats

  *The various sources and libraries used for finding out the data retrieval ae as follows-*

  *1. NumPy:*
  *It is the fundamental package for scientific computing in Python. A Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.*
  *2. Pandas: Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. A fundamental high-level building block for doing practical, real-world data analysis in Python. The most powerful and flexible open-source data analysis/manipulation tool available in any language.*
  *3. Matplotlib: It is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, web application servers, and various graphical user interface toolkits.*
  *4.Seaborn: It is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.*
  *5. SKlearn -Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.*

- ## Data Pre-processing Done

  *Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.*

  - *The Data Processing is divided into four stages – Data cleaning, Data integration, Data reduction and Data transformation.*
  - *The various types of Data preprocessing are –*
  - *a. Aggregation*
  - *b. Sampling*
  - *c. Dimensionality Reduction*
  - *d. Feature Subset Selection*
  - *e. Feature Creation*

- *f. Discretization and Binarization*
- *g. Variable Transformation.*

## Data Inputs- Logic- Output Relationships

*1. Data Exploration and Cleaning -*
*A) Dataset was imbalanced for the target feature(87.5% for Non-defaulters and 12.5% for Defaulters).*
*B) Data had some very unrealistic values such as 999860 days which is not possible.*
*C) There are negative values for variables which must not have one (example:frequency,amount of recharge etc). All these unrealistic values were dropped which caused a data loss of 8% only.*

*2. Feature Selection -*
*A) Since there were 36 features, many of which I suspected were redundant because of the data duplication.*
*B) The method used was 'Univariate Selection' using chi-square test. I selected top 20 features which were highly significant.*
*3. Data Visualization - a. Imbalance of data. b. Distribution was not normal.*
*4. Data Normalization- Since the data was not normal, except the target variable which was dichotomous (Values '1' and '0').*
*5. Oversampling of Minority class - The data was expensive, oversample the minority class using SMOTE.*
*6. Build Models - It is a supervised classification problem,5 models to evaluate performance of each*
*of them: a. Logistic Regression b. Linear SVM c. Decision Tree d. Random Forest e. Gradient Boost Classifier Since the data was imbalanced, accuracy was not the correct performance metric. Instead, I focused on other metrics like precision, recall and ROC-AUC curve.*

## Hardware and Software Requirements and Tools Used

*The various sources and libraries used for finding out the data retrieval ae as follows-*

*1. NumPy:*
*It is the fundamental package for scientific computing in Python. A Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.*
*2. Pandas: Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. A fundamental high-level building block for doing practical, real-world data analysis in Python. The most powerful and flexible open-source data analysis/manipulation tool available in any language.*

*3. Matplotlib: It is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, web application servers, and various graphical user interface toolkits.*

*4.Seaborn: It is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas' data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots.*

*5. SKlearn -Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.*

- Data Pre-processing Done

*Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for a building and training Machine Learning models.*

- *The Data Processing is divided into four stages – Data cleaning, Data integration, Data reduction and Data transformation.*
- *The various types of Data preprocessing are –*
- *a. Aggregation*
- *b. Sampling*
- *c. Dimensionality Reduction*
- *d. Feature Subset Selection*
- *e. Feature Creation*
- *f. Discretization and Binarization*
- *g. Variable Transformation.*

# Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  *The steps to be followed –*
  *1. Splitting Dataset into Training set and Testing set.*
  *I have split our dataset into training set (80%) and testing set (20%) to train and test rainfall prediction models.*
  *2. Training and Testing*

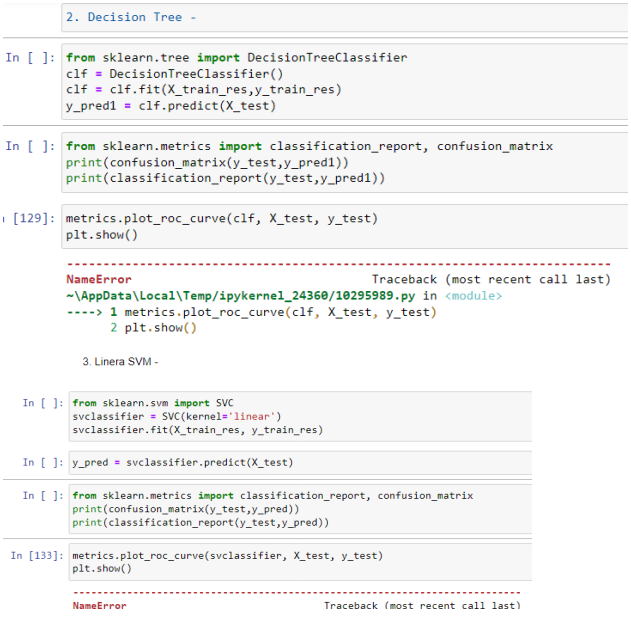
- Testing of Identified Approaches (Algorithms)

  *Analysis of the output of each model are as –*

  1. *Logistic Regression*
  2. *Decision Tree*
  3. *Linear SVM*
  4. *Random Forest Regression*
  5. *Gradient Boosting Classifier*

- Run and Evaluate selected models

  Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.



```
2. Decision Tree -

In [ ]: from sklearn.tree import DecisionTreeClassifier
        clf = DecisionTreeClassifier()
        clf = clf.fit(X_train_res,y_train_res)
        y_pred1 = clf.predict(X_test)

In [ ]: from sklearn.metrics import classification_report, confusion_matrix
        print(confusion_matrix(y_test,y_pred1))
        print(classification_report(y_test,y_pred1))

[129]: metrics.plot_roc_curve(clf, X_test, y_test)
       plt.show()

       ---------------------------------------------------------------------------
       NameError                                 Traceback (most recent call last)
       ~\AppData\Local\Temp/ipykernel_24360/10295989.py in <module>
       ----> 1 metrics.plot_roc_curve(clf, X_test, y_test)
             2 plt.show()

3. Linera SVM -

In [ ]: from sklearn.svm import SVC
        svclassifier = SVC(kernel='linear')
        svclassifier.fit(X_train_res, y_train_res)

In [ ]: y_pred = svclassifier.predict(X_test)

In [ ]: from sklearn.metrics import classification_report, confusion_matrix
        print(confusion_matrix(y_test,y_pred))
        print(classification_report(y_test,y_pred))

In [133]: metrics.plot_roc_curve(svclassifier, X_test, y_test)
          plt.show()

          ---------------------------------------------------------------------------
          NameError                                 Traceback (most recent call last)
```

- ## Key Metrics for success in solving problem under consideration

  *The various key metrices are –*
  *sklearn.metrices- classification report -confusion matrix*
  *svclassifier, randomforest classifier, gradient boost classifier*

- ## Visualizations

  *Data visualization is defined as a graphical representation that contains the information and the data. By using visual elements like charts, graphs, maps. etc. Data visualization techniques provide an accessible way to see and understand trends, outliers, and patterns in data. Some of the Data Visualization methods are: Line plot, Bar plot, Distribution plot, Joint plot, Scatter plot, Violin Plot, Count plot, Pair Plot etc.*
  *Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data. The steps in Data Exploration are–*

  1. *Identification of variables and data types.*
  2. *Analyzing the basic metrics.*
  3. *Non-Graphical Univariate Analysis.*
  4. *Graphical Univariate Analysis.*
  5. *Bivariate Analysis.*
  6. *Variable transformations.*
  7. *Missing value treatment.*
  8. *Outlier treatment.*

- ## Interpretation of the Results.

  *As per the references , what I have observed from the professionals projects are -According to the performance metrics, Random Forrest scores highest in accuracy. Also, the curve is tending towards the ideal shape. Hence, Random Forrest looks like the best fit for this data.*

  *Remarks - I tried lot to resolve the issues , but i am helpless, i sort out all sci-kit library, seaborn , but not solved the issue. At the end after trying with various ML models, We will get a optimized one to suggest.*

# CONCLUSION

- Key Findings and Conclusions of the Study

  Describe the key findings, inferences, observations from the whole problem.

- Learning Outcomes of the Study in respect of Data Science

  *Python implementations do some of the fundamental Machine Learning models and algorithms from scratch. The purpose of this project is not to produce as optimized and computationally efficient algorithms as possible but rather to present the inner workings of them in a transparent and accessible way. What are models of machine learning. A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.*

- Limitations of this work and Scope for Future Work