FLIP ROBO

MICRO_CREDIT_PROJECT

Submitted by:

Amit Narhare

ACKNOWLEDGEMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

INTRODUCTION

- Business Problem Framing

  - ❿ Business problems are current or long term challenges and issues faced by a business. These may prevent a business from executing strategy and achieving goals. In some cases, business problems also threaten the long term survival of a firm.
  - ❿ We need to do analysis as per data collected for loan applied people.
  - ❿ Now a days taking loan from private banking as well as government bank is trend as well as need with increase in standard life.

- Conceptual Background of the Domain Problem

  - ❿ Defining characteristics of a problem domain continues to challenge developers of visualization software although it is essential for designing both tools and resulting visualizations.
  - ❿ Additionally, effectiveness of a visualization software tool often depends on the context of systems and actors within the domain problem.
  - ❿ The nested blocks and guidelines model is a useful template for informing design and evaluation criteria for visualization software development because it aligns design to need.
  - ❿ Characterizing the outermost block of the nested model the domain problem is challenging, mainly due to the nature of contemporary domain problems, which are dynamic and by definition difficult to problematize.
  - ❿ We offer here our emerging conceptual model, based on the central question in our research study what visualization works for whom and in which situation to characterize the outermost block, the domain problem, of the nested model.
  - ❿ We apply examples from a 3-year case study of visualization software design and development to demonstrate how the conceptual model might be used to create evaluation criteria affecting design and development of a visualization tool.

Review of Literature

- Google.com for most of the topics whose refereed from
- Document provided with project.

Motivation for the Problem Undertaken

- ➓ Problem is conducted to identify problems or to find answers to ‘uncertainties’.
- ➓ Manage time and work systematically. For example, in time management, a systematic timetable will make life more manageable. To identify who loan can be taken
- ➓ Developer must keep in mind that the main motivation in developing their Problem is their deep interest in the field, not because of money.
- ➓ Every researcher must have a high degree of confidence and must never give up easily even at one time a development will seem to reach a dead end. However, if the developer is sincere about gaining new knowledge, the development will eventually be a success.
- ➓ Developer should never keep quiet about their newly acquired knowledge and must always be willing to share information with their colleagues. Cooperation is an important asset for the success of a team project.
- ➓ There is no shortcut to gain excellent research results, thus time and energy sacrifices are essential.
- ➓ Most important people taking a loan and tracking for that is very difficult.

Analytical Problem Framing

Mathematical/ Analytical Modelling of the Problem

❿ EDA done on whole data set

❿ Correlation coefficient matrix

Data Sources and their formats:

❿ Collected from client source.

Data Preprocessing Done:

- Import all the Required libraries

A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a list for most popular Python libraries for Data Science. Here's a snippet of me importing the pandas library and assigning a shortcut "pd".

❿ Import the dataset

We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program) and read it using a method called *read_csv* which can be found in the library called *pandas.*

❿ Taking care of Missing Data in Dataset

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.

- ❿ Encoding the categorical data using label encoder as well as simple imputer

Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

So the way we do it, we will import the scikit library that we previously used. There’s a class in the library called *LabelEncoder* which we will use for the task.

❿ Splitting the Dataset into Training set and Test Set

Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import *test_train_split* from *model_selection* library of scikit.

• Data Inputs- Logic- Output Relationships

- Below Scatter plot shows relation of input column with dependent or output column label.

It will tries to set linear relation with dependent column

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

NA

- Hardware and Software Requirements and Tools Used

- Jupyter Notebook IDE
- import pandas as pd
- import numpy as np
- import matplotlib.pyplot as plt
- import seaborn as sns
- from sklearn.model_selection import train_test_split
- from scipy.stats import zscore
- from sklearn.impute import SimpleImputer
- from sklearn.preprocessing import LabelEncoder
- import warnings

- from sklearn.linear_model import LogisticRegression
- from sklearn.neighbors import KNeighborsClassifier
- from sklearn.tree import DecisionTreeClassifier
- from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
- from sklearn.naive_bayes import MultinomialNB
- from sklearn.ensemble import RandomForestClassifier,GradientBoostingClassifier,AdaBoostClassifier
- from sklearn.model_selection import cross_val_score
- from sklearn.metrics import roc_curve
- from sklearn.metrics import roc_auc_score

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.

  Our problem is best framed as:

➓ Binary classification

❿ Unidimensional regression

❿ Multi-class single-label classification

❿ Multi-class multi-label classification

❿ Multidimensional regression

❿ Clustering (unsupervised)

❿ Other (translation, parsing, bounding box id, etc.)

❿ Then, after framing the problem, explain what the model will predict.

❿ Determine Where Data Comes From

Assess how much work it will be to develop a data pipeline to construct each column for a row. When does the example output become available for training purposes? If the example output is difficult to obtain, you might want to revisit your output, and examine whether you can use a different output for your model.

Make sure all your inputs are available at prediction time in exactly the format you've written down. If it will be difficult to obtain certain feature values at prediction time, omit those features from your model.

❿ Determine Easily Obtained Inputs

Pick 1-3 inputs that are easy to obtain and that you believe would produce a reasonable, initial outcome.

Which inputs would be useful for implementing heuristics mentioned previously?

Consider the engineering cost to develop a data pipeline to prepare the inputs, and the expected benefit of having each input in the model. Focus on inputs that can be obtained from a single system with a simple pipeline. Start with the minimum possible infrastructure.

- Testing of Identified Approaches (Algorithms)

  Below are the list of algorithm where we used in project.

  model=[lg,KneighborsClassifier(),DecisionTreeClassifier()]

  model=[rf,GradientBoostingClassifier(),AdaBoostClassifier()]

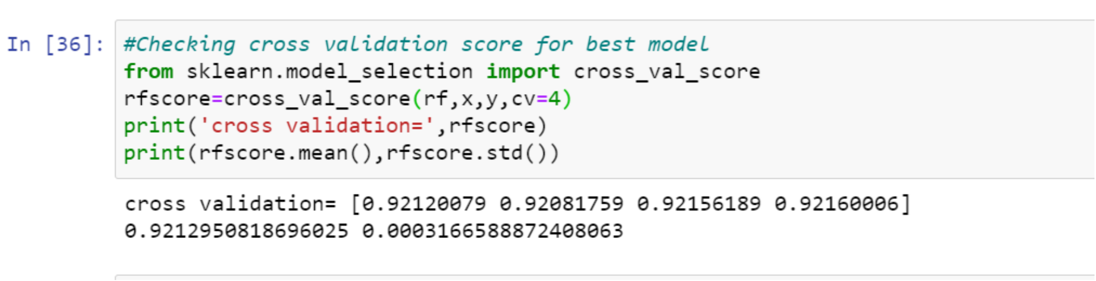
Where:

lg=logistic regression

rf=Fandom Forest Regressor

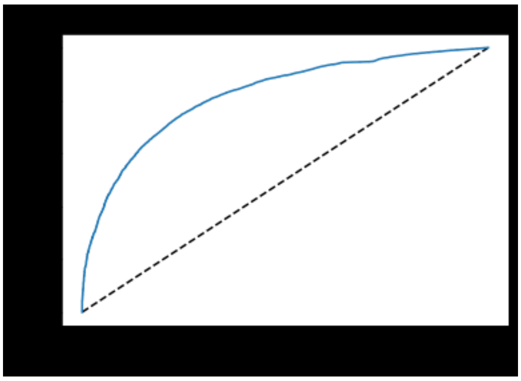
- Run and Evaluate selected models

1. Cross validation Score



```
In [36]: #Checking cross validation score for best model
         from sklearn.model_selection import cross_val_score
         rfscore=cross_val_score(rf,x,y,cv=4)
         print('cross validation=',rfscore)
         print(rfscore.mean(),rfscore.std())

         cross validation= [0.92120079 0.92081759 0.92156189 0.92160006]
         0.9212950818696025 0.0003166588872408063
```

2. AUC ROC Curve

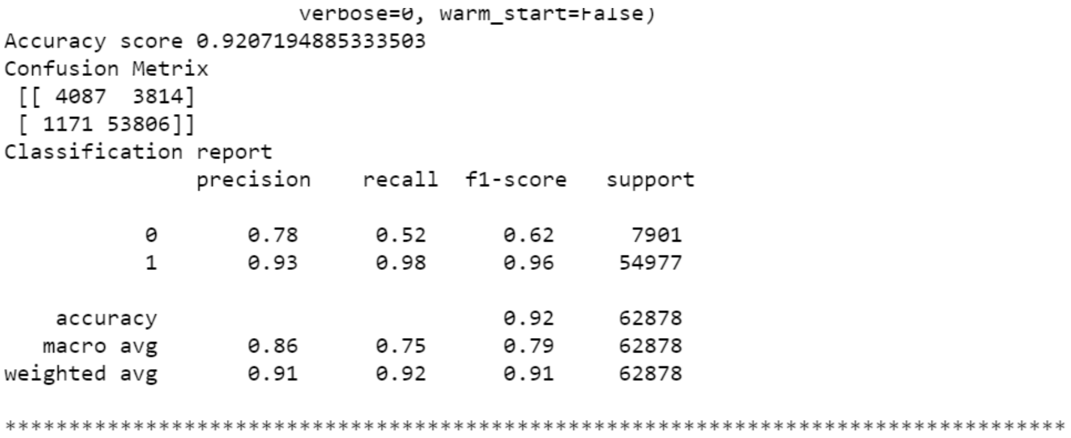


- Key Metrics for success in solving problem under consideration

❿ accuracy_score

❿ classification_report

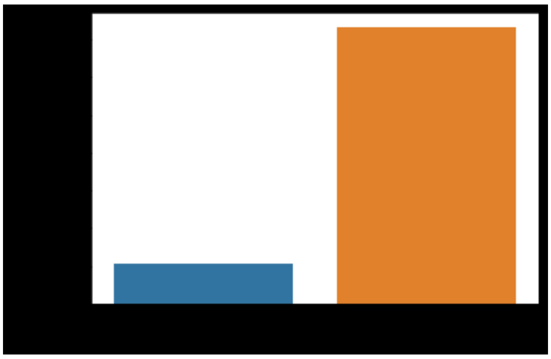
❿ confusion_matrix



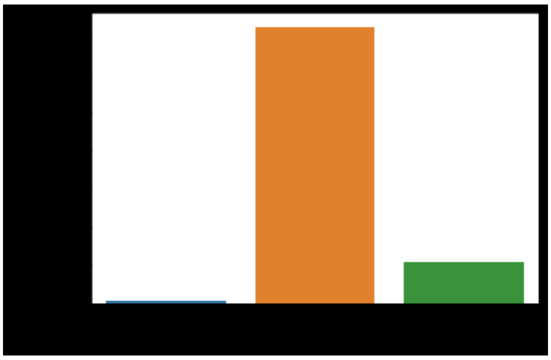
```
                       verbose=0, warm_start=False)
Accuracy score 0.9207194885333503
Confusion Metrix
 [[ 4087  3814]
 [ 1171 53806]]
Classification report
               precision    recall  f1-score   support

           0       0.78      0.52      0.62      7901
           1       0.93      0.98      0.96     54977

    accuracy                           0.92     62878
   macro avg       0.86      0.75      0.79     62878
weighted avg       0.91      0.92      0.91     62878

*********************************************************************************
```

- Visualizations

1) Count plot of label column

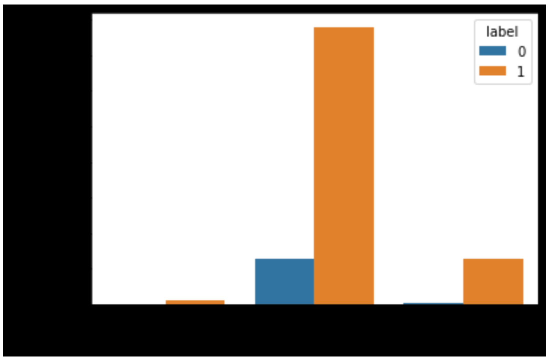


2) maxamnt_loans90 count plot

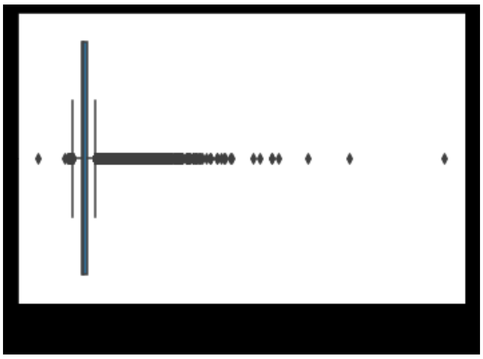


Observations: 6 is around 175000

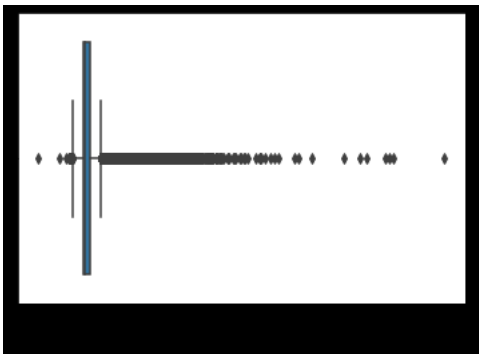
3) sns.countplot(df['maxamnt_loans90'],hue=df['label'])



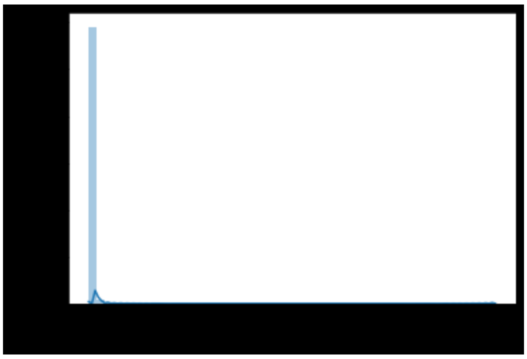
As per above graph more return by 6 as maximum amount of loan



Observation: Outliers is present in data set.



Observation: Outliers is present in data set.



Observation: by above plot we can say data is right skwed.

- **Interpretation of the Results**
- As per visualization we can say dependency of target column with another column.
- Cleaning the data as per requirement, with encoding the data.
- As per model Random forest is the best fit model for the data set.

CONCLUSION

- Key Findings and Conclusions of the Study

Findings:

In recent year multiple people taking the loan and not considering to return the loan interest so creating the model where it will predict if that person who take loan, he will return the loan on time or not.

Observation:

As per observation we can say the model will predict the man who Appling loan, he is fit for take a loan and if he will take a loan he will return amount or not.

- Learning Outcomes of the Study in respect of Data Science

As per model Random forest is the best fit model for the data set.

With provided details.

- Limitations of this work and Scope for Future Work

Accuracy may decrease when lots of difference in data input.

Output is totally depends on input.