Credit Risk Classifier

6th Semester Training Project Report

Department of Information Technology, University Institute of Engineering and Technology
Panjab University, Chandigarh

Submitted to :

Dr. Veenu Mangat,

Associate Professor

I.T. , U.I.E.T ,Panjab University

Chandigarh

Submitted by :

Anshul Gupta

BE- IT Section-1

(2017-2021 batch)

Roll number - UE178022

Executive Summary

The Project aims to build a model to predict whether an entity is good or bad risk and uses machine learning supervised classification technique to predict the outcome.

The model is trained on the dataset from UCI Machine Learning repository and datasets consists of about One thousand instances. Data preprocessing is done before the data modeling is done. Data visualizations are used to get various insights from the dataset. Two models are built and evaluation of them is done so as to compare the performance.

The main motive to build the project was to disprove the hypothesis that rich must get credit and poor is a bad risk by providing the necessary evidence and thus building a model which is able to increase the profits and decrease the losses.

Table of Contents:

1.Introduction

1.1 Problem setting:

In current banking world, both in formal and informal sector, with growing amount of people needing loans for various purposes, be it productive or not, the main problem with giving out credit is the Risk involved.

The main problem is due to the fact that many a times even rich are unable to pay the loan back and seemingly poor pay back easily.

This can be due to a variety of factors like the rich person is old and wants loan for non-productive purpose and thus turns out to be bad risk rather being a good risk.

1.2 Potential Solution of problem:

So, it becomes extremely necessary to have the background information of the entity to be given credit.

Also, if the risk involved can be measured or approximated using the background information, it will work wonders for the bank since the loss will be minimized and the profit can be maximized, thus optimizing the decision of giving out credit and this project aims to build the same.

1.3 Potential Audience:

Government and private banks lending out money to various entities (individuals, big & small corporations etc)

Local moneylenders (they usually loan small amount of credit at lower interest than a bank) can also check the risk involved with a particular entity.

1.4 Knowledge Gap:

The assumption that if someone is rich is a good risk and if someone has less savings is bad risk can be quite misleading sometimes since various factors can have varying effect on the risk classification

Knowing these factors is a huge plus point and can optimize many decisions related to giving the credit.

AIM

To make a model using the dataset which classifies new entries as good risk or bad risk and thus helping to identify which people should be given credit or not

2. Methodology

2.1 Data Source:

- The dataset used here is originally "Statlog (German Credit Data) Data Set" from UCI Machine learning repository.
- While there are plenty of credit approval datasets like Poland Credit data, Taiwan Credit data, etc available on UCI ML repository, German credit data was chosen since other datasets contained the information of credit companies as well and thus the information in them was not well aligned to fill the knowledge gap presented.
- Since the original source contains data in a complicated numeric format that had to be decoded using a cost matrix, the dataset is imported here from Kaggle website and is in a format which can be understood (not clean or normalized), only understandable!

About the data- This dataset classifies people described by a set of attributes as good or bad credit risks.

Glimpse of dataset (top 10 rows)

	Unnamed: 0	Age	Sex	Job	Housing	Saving accounts	Checking account	Credit amount	Duration	Purpose	Risk
0	0	67	male	2	own	NaN	little	1169	6	radio/TV	good
1	1	22	female	2	own	little	moderate	5951	48	radio/TV	bad
2	2	49	male	1	own	little	NaN	2096	12	education	good
3	3	45	male	2	free	little	little	7882	42	furniture/equipment	good
4	4	53	male	2	free	little	little	4870	24	car	bad
5	5	35	male	1	free	NaN	NaN	9055	36	education	good
6	6	53	male	2	own	quite rich	NaN	2835	24	furniture/equipment	good
7	7	35	male	3	rent	little	moderate	6948	36	car	good
8	8	61	male	1	own	rich	NaN	3059	12	radio/TV	good
9	9	28	male	3	own	little	moderate	5234	30	car	bad

2.2 Data Description

This is the description of the attributes as provided by the source:

1. Age (numeric)

2. Sex (text: male, female)

3. Job (numeric: 0 - unskilled and non-resident, 1 - unskilled and resident, 2 - skilled, 3 - highly skilled)

4. Housing (text: own, rent, or free)

5. Saving accounts (text - little, moderate, quite rich, rich)

6. Checking account (text - little, moderate, rich, quite rich)

7. Credit amount (numeric, in DM)

8. Duration (numeric, in month)

9. Purpose (text: car, furniture/equipment, radio/TV, domestic appliances, repairs, education, business, vacation/others)

Above 9 attributes are independent attributes (Predictor Variables)

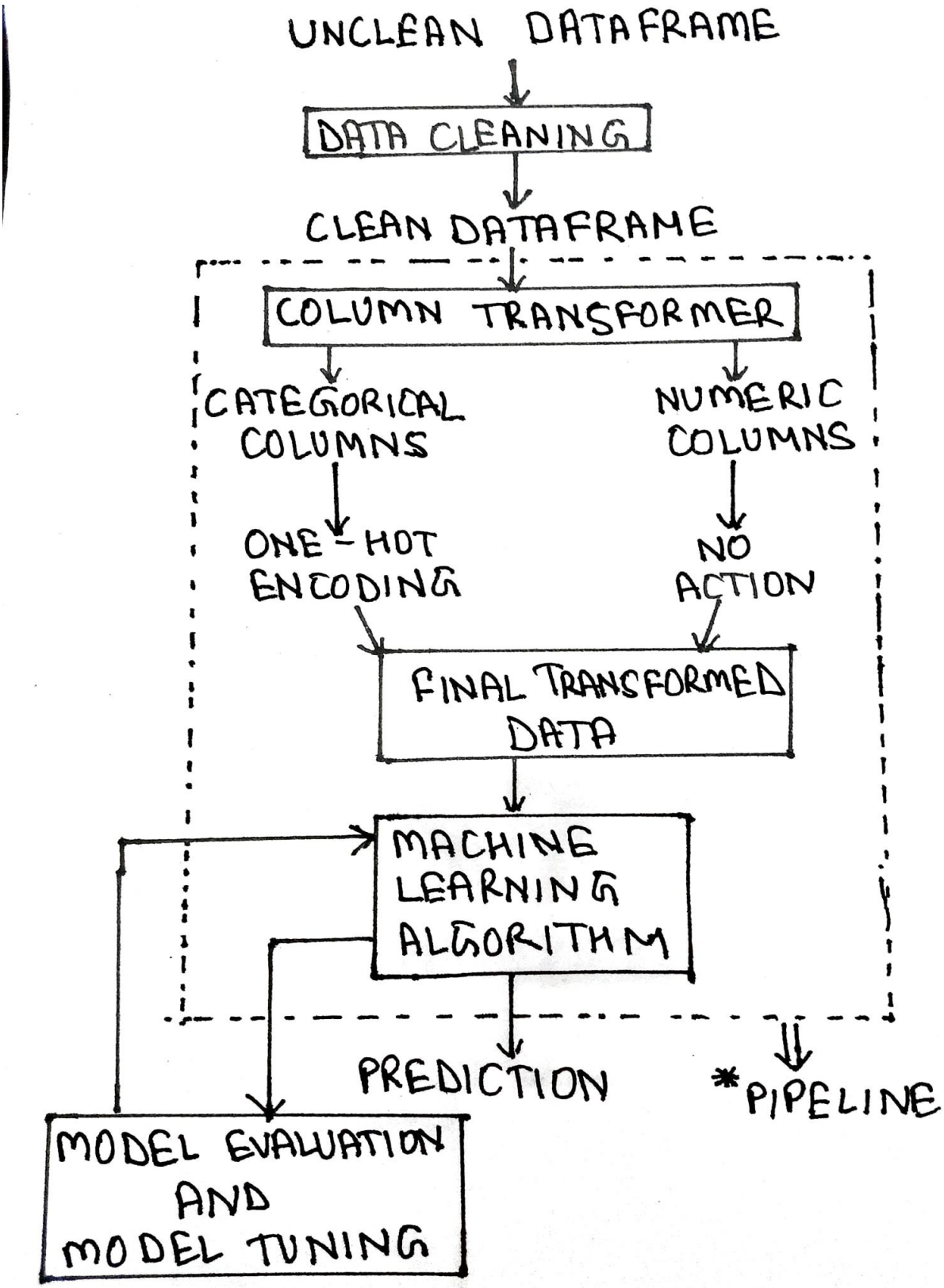
10. **Risk** (text: 'good' or 'bad')

'Risk' attribute is the dependent attribute (Target Variable)

2.3 STEPS AND STRATEGY USED TO SOLVE THE PROBLEM:

- **Data Exploration:** exploring columns types, number of unique values in them, null value presence check, plotting distribution of column values.
- **Data Manipulation:** handling null values by substituting with basic business sense & by dropping some rows, dropping redundant columns, categorizing some numeric valued columns
- **Outlier and Anomaly detection:** Removing outliers with basic logic and by box plot visualization to check and remove outliers.
- **Modelling:**
    1. Built two different classifiers, their performance was compared using metrics such as accuracy, precision and recall.
    2. Model which had better and desirable precision and recall values was chosen to be the better performing model. Factors determining the desirable values of Precision and recall are discussed in Discussion Section.
    3. Applied Column Transformer which applies one-hot encoding on the categorical attributes and leaves the numerical attributes as they were and gives a dataframe upon which machine learning model can be made.
    4. Then Created a Pipeline which first applies Column Transform on the dataframe and then applies machine learning algorithm on it
    5. The main motive to create pipeline was that when input is given to get prediction, all the necessary column transformations and the ML algorithm will be applied on it.
    6. Various metrics used so as to compare the model performance.
    7. Hyperparameter tuning done using GridSearchCV.

Flow Chart of Data Modeling process

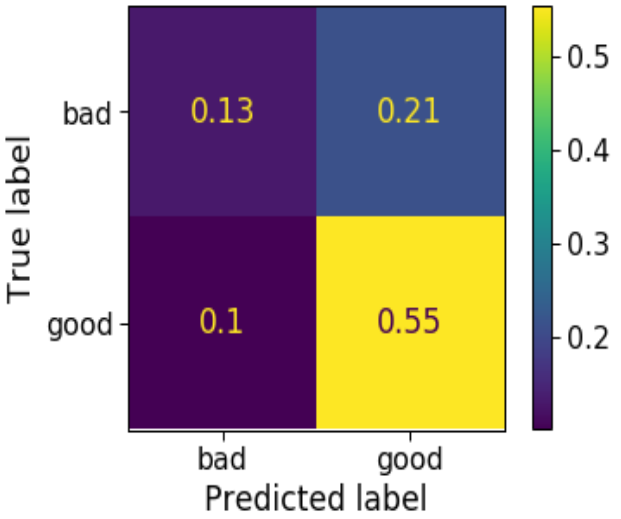


3.Results

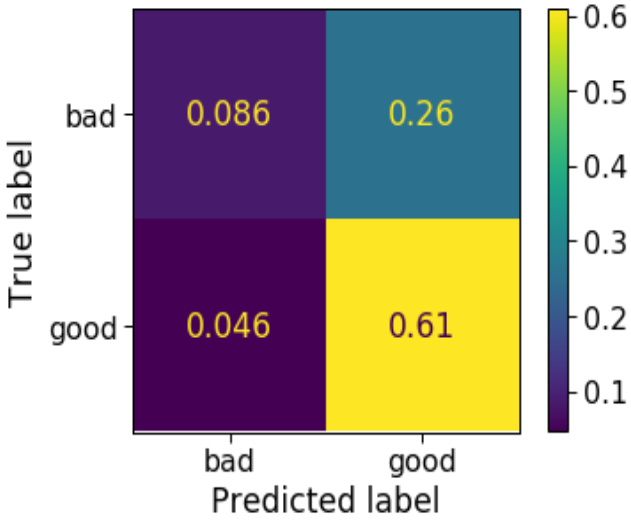
3.1 Random Forest Classifier:

- Accuracy
  a. Before Tuning accuracy is around 66 %
  b. After tuning accuracy went up to 70 %
- Confusion matrix

Before tuning



After tuning



❖Precision and Recall

a. Before tuning

Recall = 0.39,0.83 (bad, good) - in other words, it correctly identifies ~ 83 % of good risks and ~40% of bad risks.

Precision= 0.55,0.72 (bad, good) - in other words, when it predicts the risk is good, it is correct ~ 72 % of time and when it predicts the risk is bad, it is correct ~ 55 % of time.

b. After tuning

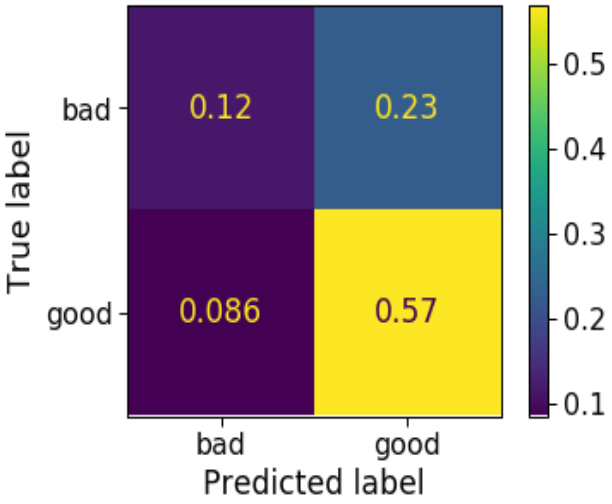
Recall = 0.25,0.93 (bad, good)

Precision = 0.65,0.70 (bad, good)

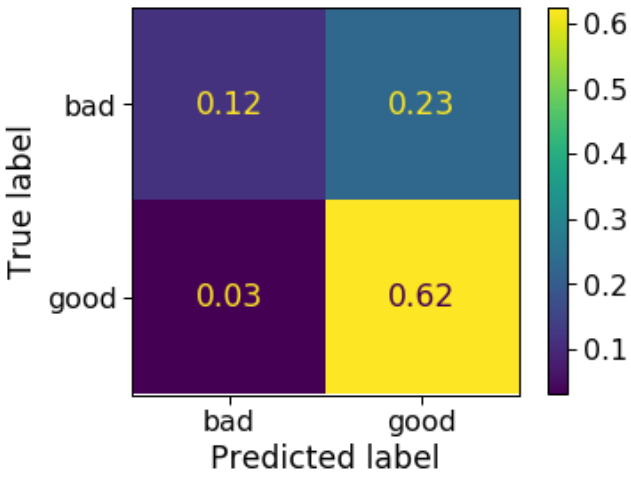
3.2 Logistic Regression classifier

- Accuracy
  a. Before tuning accuracy is ~68 %
  b. After tuning, accuracy went up to 74 %
- Confusion matrix

Before Tuning



After tuning



❖Precision and Recall

c. Before tuning

Recall = 0.34,0.86 (bad, good)

Precision= 0.57,0.71 (bad, good)

d. After tuning

Recall = 0.34,0.95 (bad, good)

Precision = 0.79,0.73 (bad, good)

❖ Hyperparameter Tuning:

Parameters which define the model architecture are referred to as hyperparameters and the process of searching for ideal model architecture is referred to as *hyperparameter tuning.* The searching technique used here is GridSearchCV in which an exhaustive search over specified parameters values is performed and best ones among them are returned.

Random Forest Classifier

- ❖ Accuracy is increased by increasing the number of decision trees, so the averaged predictive value is more closer to the real value. Thus accuracy improves.
- ❖ By this way, overfitting is also prevented since more number of decision tree results are being considered.

Logistic Regression Classifier

- ❖ Accuracy is increased because by changing the hyperparameters, we have used Ridge Regression (modified version of logistic regression)
- ❖ In ridge regression, the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients.
- ❖ This along with low value (0.001 in our case) of C (inverse of strength of regularization), overfitting is prevented here and thus accuracy is improved.

4.Discussion

Our two major concerns while looking for model with good results should be:

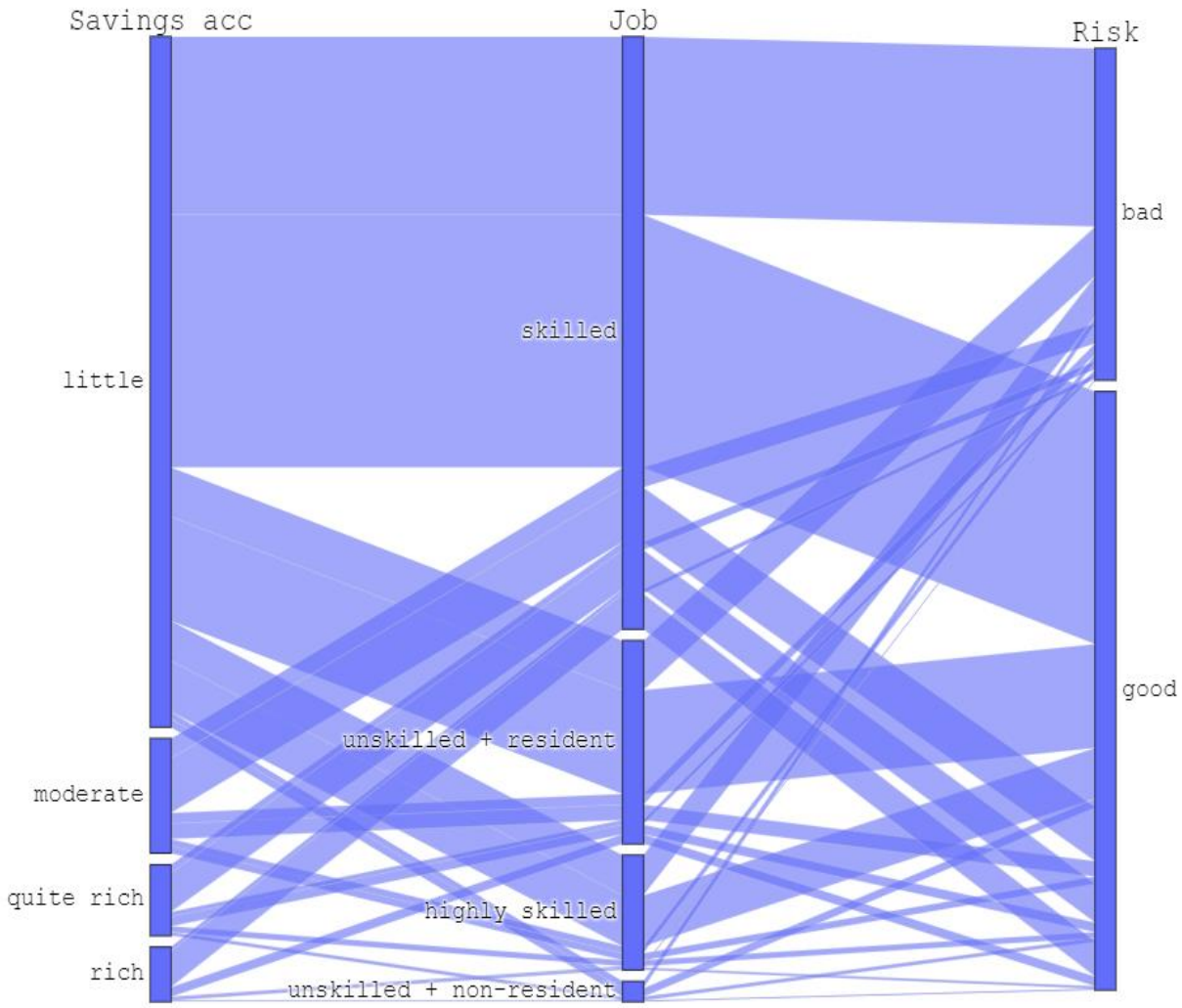
- To recognize and reduce the people who are classified as good risks but are actually bad risk, in order to minimize losses. That is ***to have greater precision in case of recognizing 'bad' risks.***
- And to recognize more people who are actually 'Good' risk, so as to increase profits. Thus, ***to have greater 'Good' risk recall.***

The results we have obtained after tuning logistic regression classifier satisfies both the above mentioned points.

Scope of increasing the accuracy is still present but here our focus is to fill the knowledge gap presented in the Introduction Section.

We try to fill the knowledge gap by presenting the evidence that even rich and skilled people have been classified to be 'bad' risk and poor and unskilled people be 'good' risk with help of visualizations.

Plot between 3 categorical features (Savings account, Job and Risk):

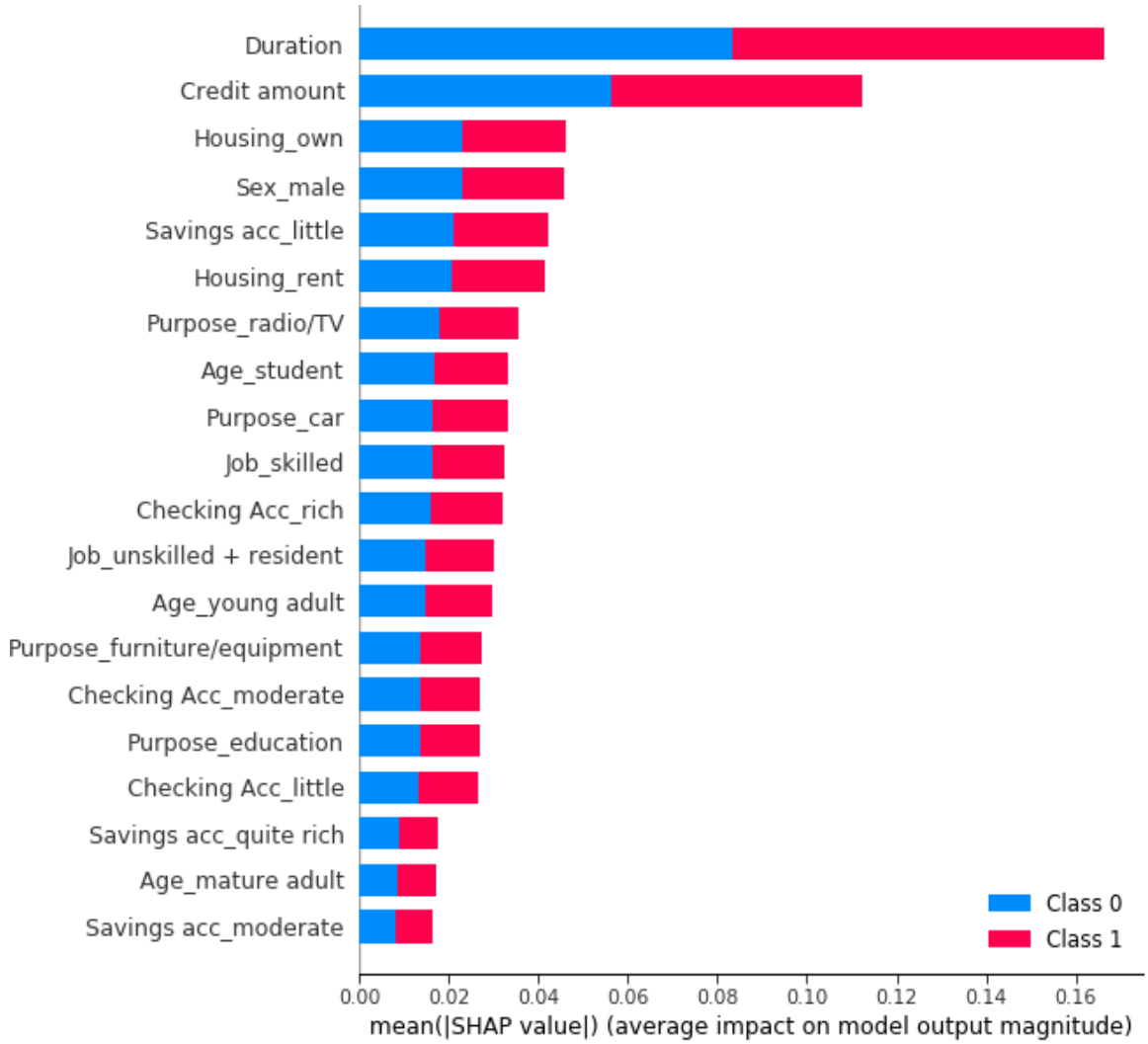


The above plot is a <u>parallel categories</u> diagram

→Each attribute is represented by a column of rectangles where each rectangle corresponds to a discrete value taken on by that variable.

→The relative heights of the rectangles reflect the relative frequency of occurrence of the corresponding value.

SHAP (SHapley Additive exPlanations) analysis



The above plot is Shapley additive explanation analysis diagram:

→Shapely values are obtained by incorporating concepts from *Cooperative Game Theory* and *local explanations*.

→Given a set of palyers, *Cooperative Game Theory* defines how well and fairly to distribute the payoff amongst all the payers that are working in coordination. The analogy here is: players are equivalent to independent features and payoff is the difference between the average prediction of the instance minus the average prediction of all instances.

→SHAP values for each feature represent the change in the expected model prediction when conditioning on that feature.

Insights from previously shown parallel categories diagram and SHAP diagram:

- Some people who are poor (savings acc value='little') and are unemployed as well, are still good risk. This proves that a *person being a good risk and bad risk depends on variety of factors and thus filling the knowledge gap.*
- At the same time the number of people who are rich as well as employed being good risk is much higher. Thus the "job" and "savings account" factor also play a key role in determining the risk involved.
- The idea that one should build a credit risk classifier model using only the features having more impact, stands against the main idea.
- Considering the less impacting features in addition to features having more impact will increase the needed recall and required precision and thus increasing profits as well as reducing losses.
- ***Thus, It is much better to consider greater number of factors which play a role in determining the risk involved and make a better and concise prediction, improving the overall business.***
- ***Also increasing the size of data will definitely help in this case by helping in prevention of overfitting of data. The reason is that, as we add more data, the model is unable to overfit all the samples, and is forced to generalize, thereby increasing the overall accuracy of model.***

5.CONCLUSION

- ✓ Main Factors (like if a person is rich or not) do play a role in determining the risk involved in giving out credit, other features should not be ignored.
- ✓ Ignoring less important features will decrease the profits and increase the losses for sure.
- ✓ After tuning of logistic regression classifier, the model we obtained has increased recall in case of good risk and increased precision in case of bad risk.
- ✓ Size of data definitely plays an important role since increasing the data size could provide better understanding of feature impact as well metrics could be improved.
- ✓ Even though the size of data here is not big, the main motive i.e. to fill the knowledge gap is well fulfilled.

Future Scope

- ✓ Increasing the data size can prove to be very beneficial by providing more insights.
- ✓ By considering datasets of various entities (local credit providers and government banks) and gathering the data in such a manner so as a model can be built for every single entity using a single large dataset can be actually huge achievement.
- ✓ Building the model with increased accuracy (close to ~ 95%) will also help further increase the benefits of the same.

6.ACKNOWLEDGEMENT

I would like to express my special thanks of gratitude to my mentor Dr. Veenu Mangat (Associate professor-IT, UIET, Panjab University, Chandigarh) who supervised my project, guided me through the details and also helped in the successful completion of my project.

7.BIBLIOGRAPHY

- UCI ML repository link-

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)/

- Understanding column transformer and pipeline using sklearn -

https://www.youtube.com/watch?v=irHhDMbw3xo&t=1416s

https://scikit-learn.org/stable/modules/compose.html#pipeline

- Poltly python visualization (parallel categories diagram) -

https://plotly.com/python/parallel-categories-diagram/#:~:text=The%20parallel%20categories%20diagram%20(also,taken%20on%20by%20that%20variable.

- Understanding of metrics from –

https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

- SHAP analysis –

https://www.youtube.com/watch?v=ZkIxZ5xlMuI

Thank You