Credit Worthiness Prediction

Team: IPBA 11 - Group E
Domain: BFSI

Mentored by:
Ms. Ashna Grover
Business Consultant, Accenture

Team Members

Akansha Sinha

Brajesh Kumar Verma

Hrishikesh Bapat

Maneet Singh Saluja

Preeti Asrani

Pritesh Agarwal

Saibal Ghosh

Tanu Khanna Sood

Agenda

- Business Problem & Objectives
- Data Summary & Data Merging
- EDA & ML Methodology
- Challenges & Limitations
- Learnings & Future scope
- Conclusions

Business Problem

A financial institute wants to analyze their customer's eligibility before issuing them a credit card in order to reduce the Credit Risk.

We are building a *"Predictive Model to help financial institutions to derive whether an Applicant is eligible or non-eligible for their product – credit card".*

Credit cards or loans are the core business of banks. The main profit comes directly from the loan's interest. The loan companies grant a loan after an intensive verification and validation process. However, they still don't have an assurance that the applicant can repay the loan with no difficulties.

सिद्धिमूलं प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

Objectives

1. The project team uses personal information submitted by applicants and their past credit history.

2. The team recommends the best model after exhaustive EDA, data pre-processing and showcasing a comparison of various models they have built.

3. The intent of this project is to develop a credit card eligibility model with high prediction accuracy. Hence the project's success will be evaluated on High Accuracy and a High Capture Rate of the final model. By this, the Financial Institutions can decide whether to issue a credit card to the applicant or not.

Model Journey

Tool used :

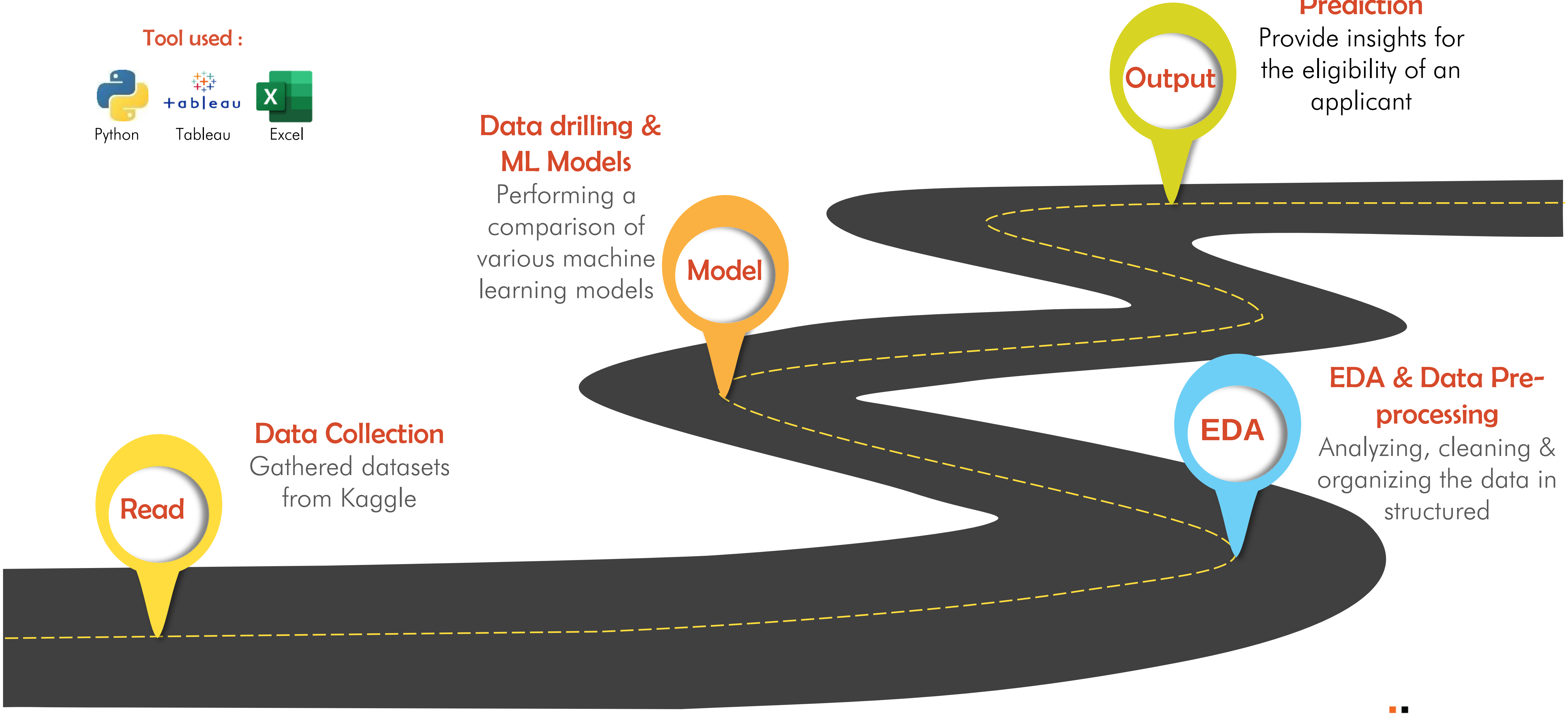
Python Tableau Excel

Data Collection
Gathered datasets from Kaggle

Data drilling & ML Models
Performing a comparison of various machine learning models

EDA & Data Pre-processing
Analyzing, cleaning & organizing the data in structured

Prediction
Provide insights for the eligibility of an applicant



Data Summary

Application Records

Applicants personal & demographics details

Total records – 438557 rows X 18 columns

ID – Unique Id of the row

Categorical Variables - CODE_GENDER ; FLAG_OWN_CAR ; FLAG_OWN_REALTY ; NAME_HOUSING_TYPE ; CNT_CHILDREN ; CNT_FAM_MENBERS; NAME_INCOME_TYPE ; NAME_EDUCATION_TYPE ; NAME_FAMILY_STATUS ; FLAG_MOBILE ; FLAG_WORK_PHONE ; FLAG_PHONE ; FLAG_EMAIL ; OCCUPATION_TYPE

Continuous Variables - AMT_INCOME_TOTAL ; DAYS_BIRTH ; DAYS_EMPLOYED

Credit Records

Credit history of the existing clients

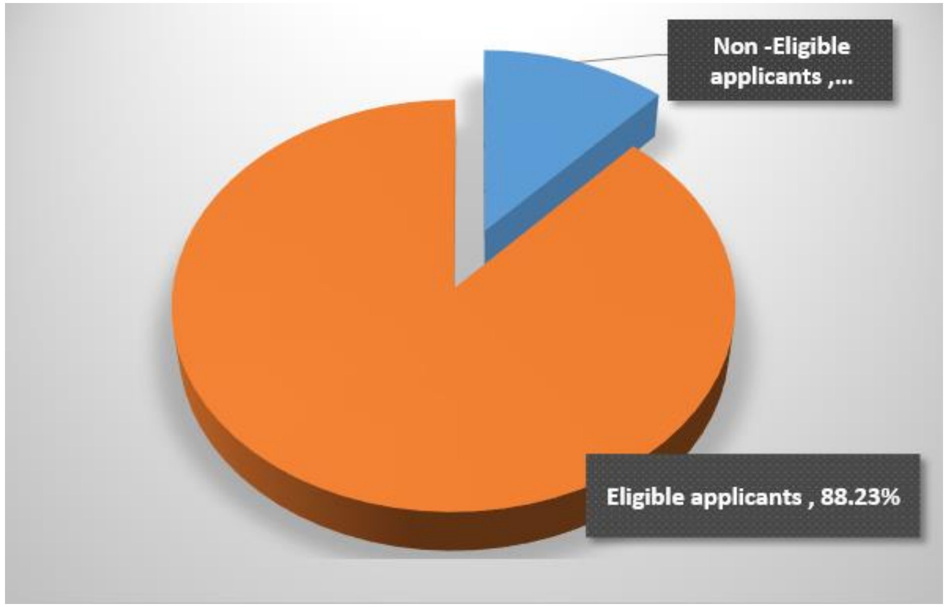
Total records – 1048575 rows X 3 columns

ID – Unique Id

MONTHS_BALANCE - The month of the extracted data is the starting point, backwards, 0 is the current month, -1 is the previous month, and so on

STATUS - 0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days overdue 3: 90-119 days overdue 4: 120-149 days overdue 5: Overdue or bad debts, write-offs for more than 150 days C: paid off that month X: No loan for the month

Data Merging



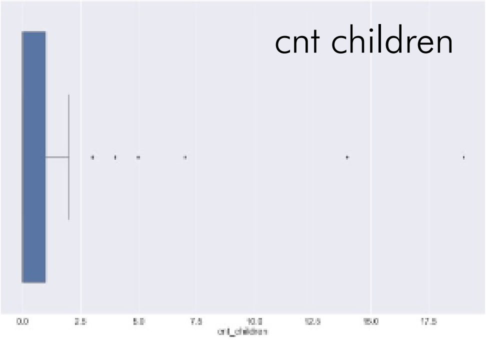
Application Record

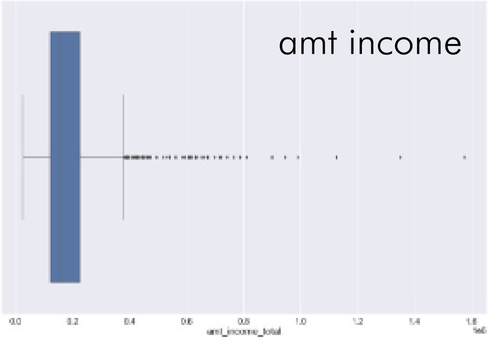
- Two datasets are merged using an inner join.
- Imbalance ratio is 7.5
- Application record dataset has multiple duplicate rows, values for all the IV's are the same across rows except ID. We have kept duplicate records in this approach.
- Occupation_type has 134203 missing values which is 30.60%. As it is an important variable, we have imputed it by creating two categories – retired & others.

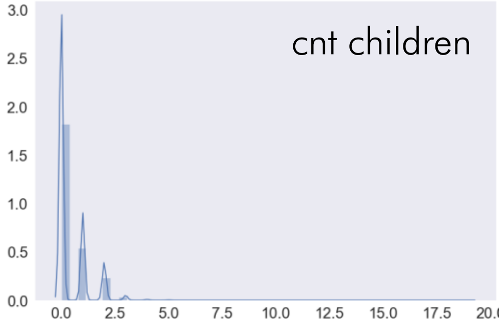
Credit Record

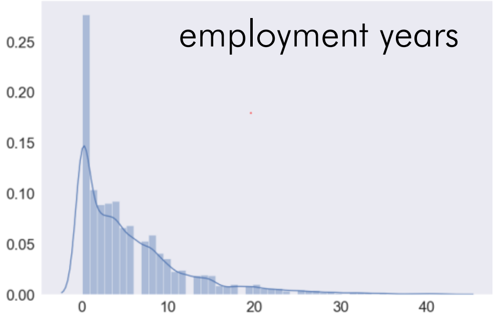
- Credit record dataset is grouped with the 'id' variable by taking the maximum value of the status variable against the applicant's id.
- Status is dependent variable, has changed to the binary output by substituting any value that is equal to 2 or above by '1's and below 2 by '0's.
- '0's means Eligible (including customers that are 0-29 days past due date) and '1's means Non-eligible.

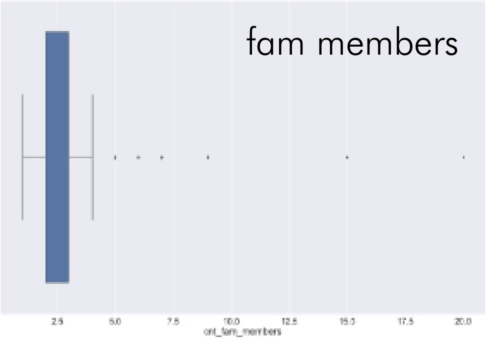
Exploratory Data Analysis

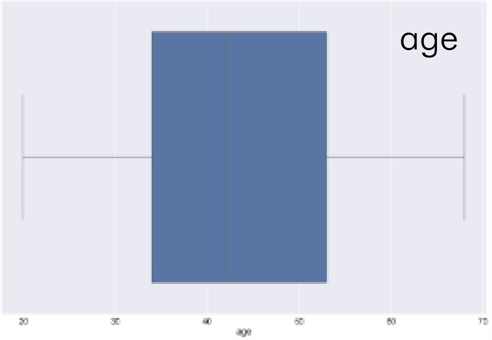


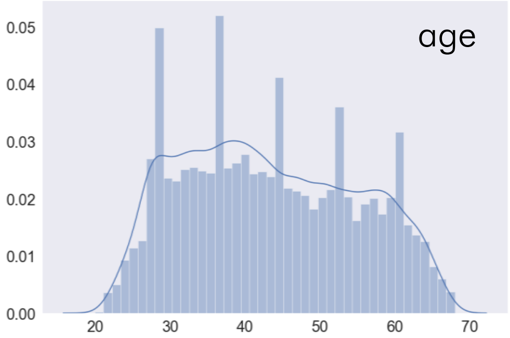


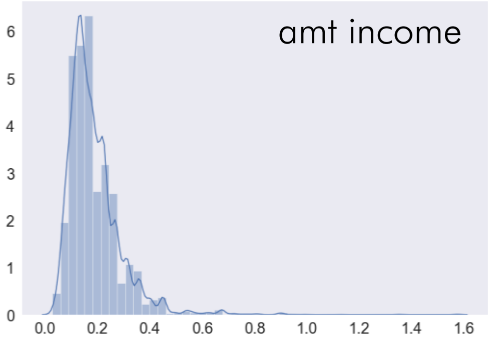


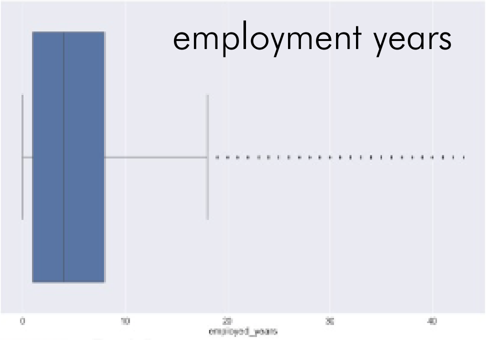




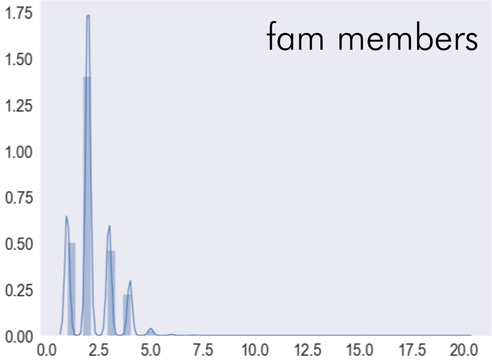






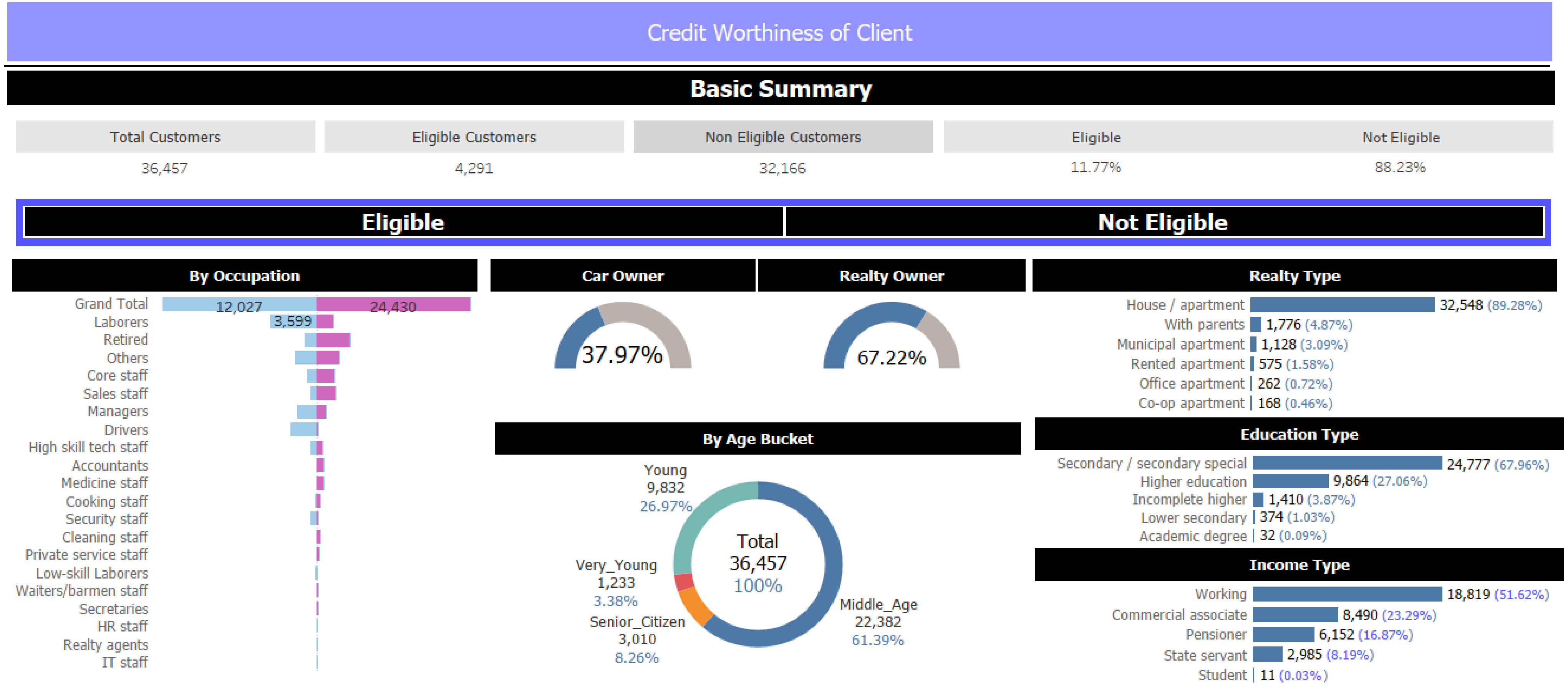


- ✓ Box-plots are for finding outliers in continuous variables.
- ✓ Histograms are to graphical summarize, the distribution of continuous variables.
- ✓ Team had perform the univariate, bivariate & multivariate analysis across different variables. The graphics are available in code file.



Variables Data Visualization

We have developed a Visualization in Tableau using our cleaned and final dataset output. This can help decision makers to understand the numbers of eligible customers by all the dependent parameters that have contributed for their eligibility.

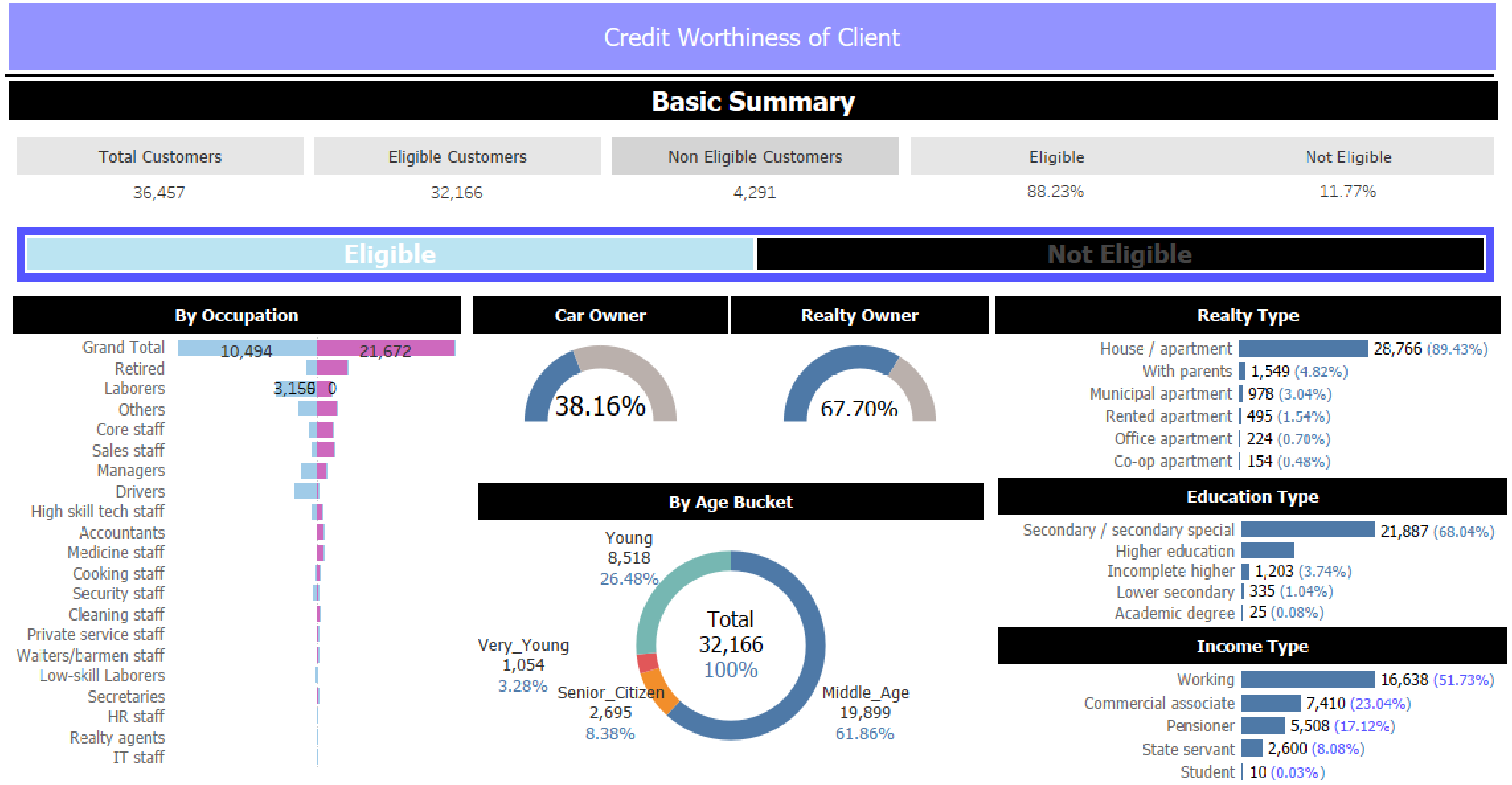


We have published this view in Tableau public as well below is the link for same:
Link to open Tableau Dashboard

Variables Data Visualization

Developed view in an interactive piece showing dynamic values that are filtering each other.

All the charts on dashboard are working as a filter on each other. For example in below screenshot. We have selected the eligible button in



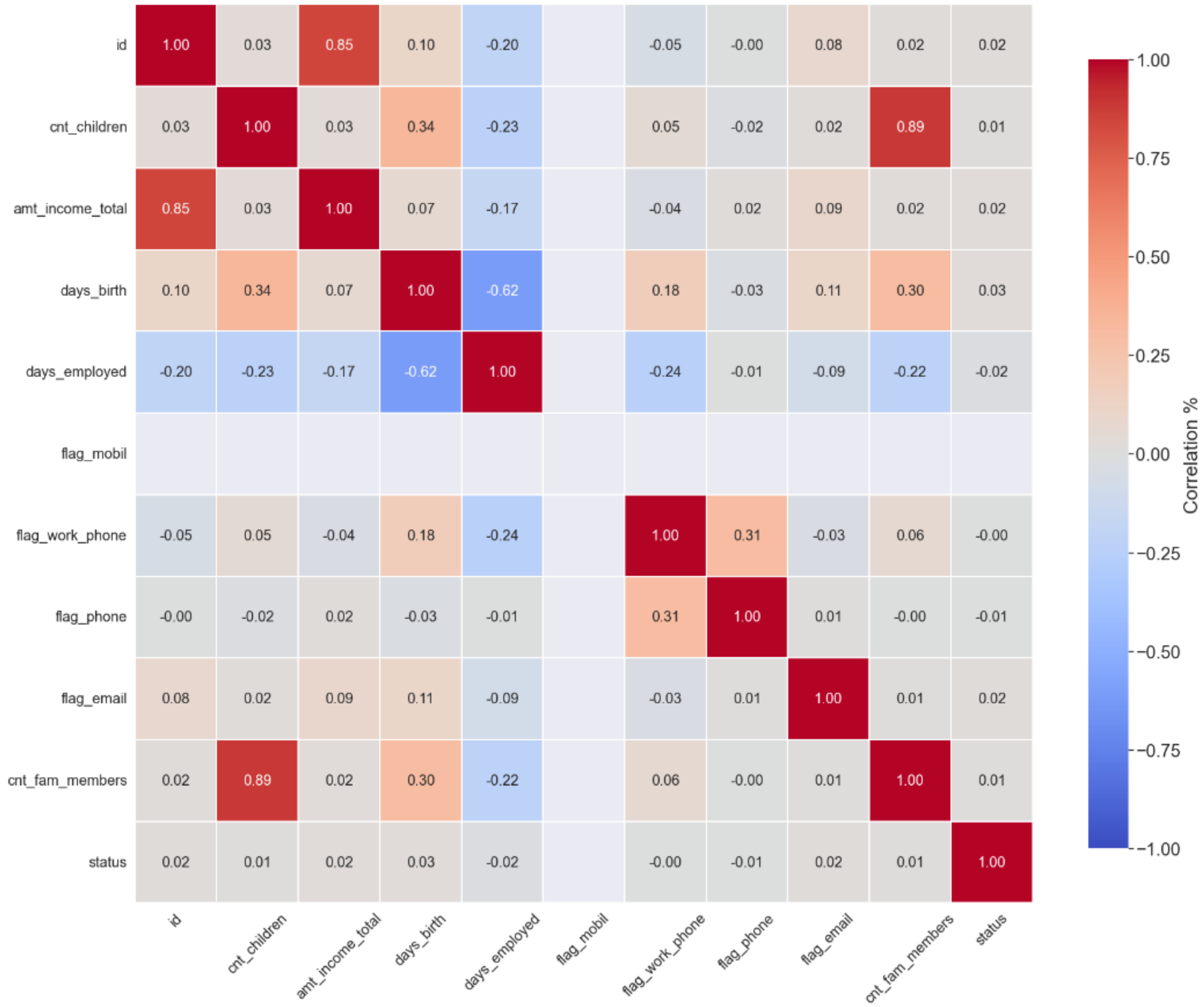
For example:

We can see there are 36457 total customers.
When we select Eligible button as highlighted we can see there are **32166** customers that eligible out of which **10494** are male and **2758** are female out of which **38.16%** Eligible customers own a car and **67.70%** Customers own real estate further we can see the break down by age, realty type, education type, income type and etc.

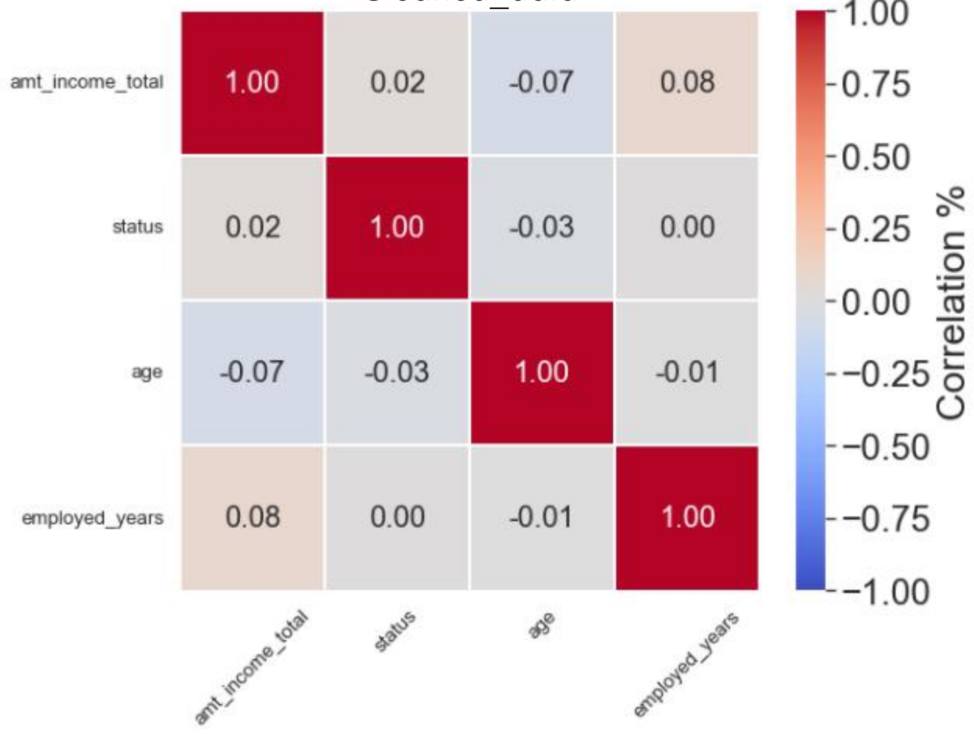
All the charts work as a filter we can select each dimension in chart and it will work as filter.

Correlation – Matrix

Merged_data



Cleaned_data



Variables Selection

Dependent Variable

Status

- ✓ Eligible Applicant (0)
- ✓ Non-eligible Applicant (1)

Dropped Variables

- ✓ ID – unique identifier will lead to overfitting
- ✓ flag_mobil - as all the applicant has the mobile
- ✓ cnt_children - VIF is higher than 10

Independent Variables

- ✓ code_gender
- ✓ flag_own_car
- ✓ flag_own_realty
- ✓ flag_phone
- ✓ flag_email
- ✓ cnt_fam_members_bkt
- ✓ name_education_type
- ✓ name_housing_type
- ✓ flag_work_phone
- ✓ occupation_type
- ✓ name_income_type
- ✓ name_family_status

Key Variables

Age

- Days of birth column has been transformed to understand the age of applicant.
- Mid age people are the most frequent applicants for credit card with most difficulty in pay it back.
- Senior Citizen and Young people below the age of 18 face lesser difficulties.

Amount income

- Applicant's income is a vital component in deciding the eligibility with a mean of 1.8 Lakhs.
- Income has been bucketed into 5 categories with majority of the population falling under Medium bucket

Employed years

- Work experience is another vital aspect to understand person capacity to payback loan.
- Retired people are also considered in this column and replaced to 0.
- With higher years of Employment the rejection rate is decreased.
- Applicants with more than 30 years of service are most likely not to be rejected.

Occupation Type

- Applicant's occupational details are described in this column.
- There are 30% null values and in order to justify this a new category called "Retired" has been formed and rest are filled by NA.
- Low skill labourers had faced the maximum rejection rate of about 19% with Private service staff being the most eligible candidates.

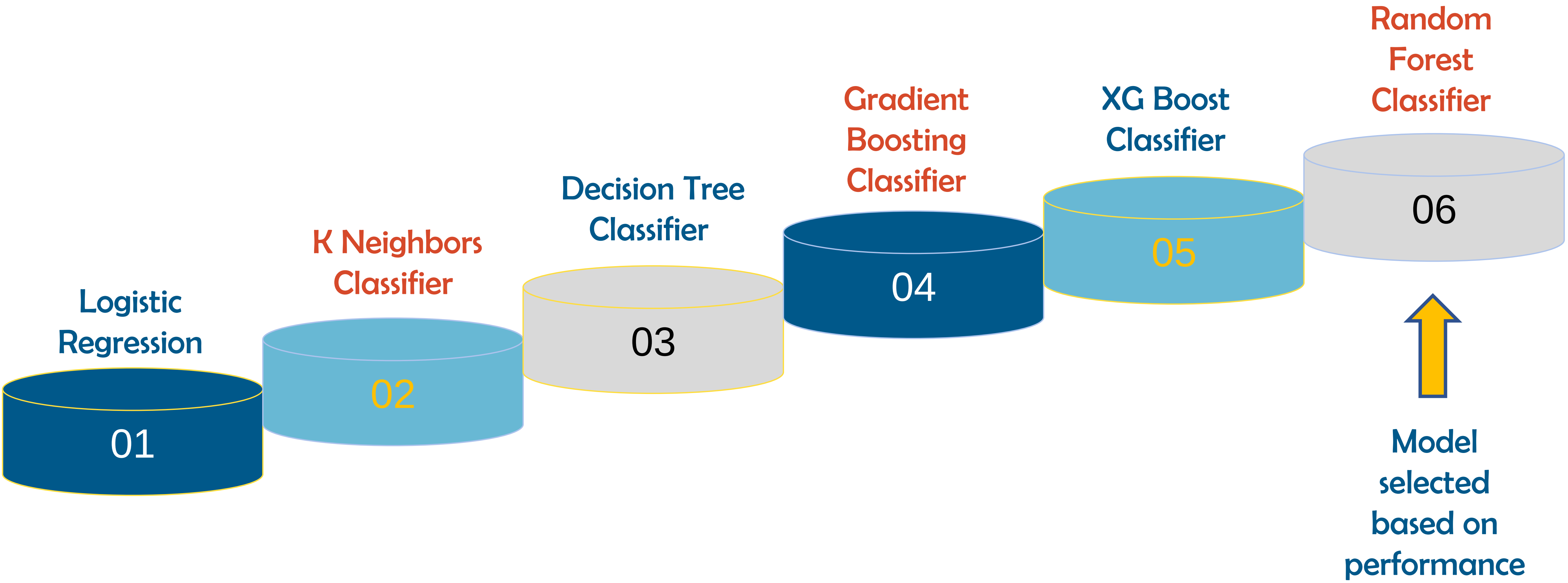
Count fam members

- No of Family members can be vital to understand dependencies of the applicant.
- People having Family members greater than 4 have been treated as outliers.
- The rejection rate of people with family members more than 4 has gone up to 67%.

Family status

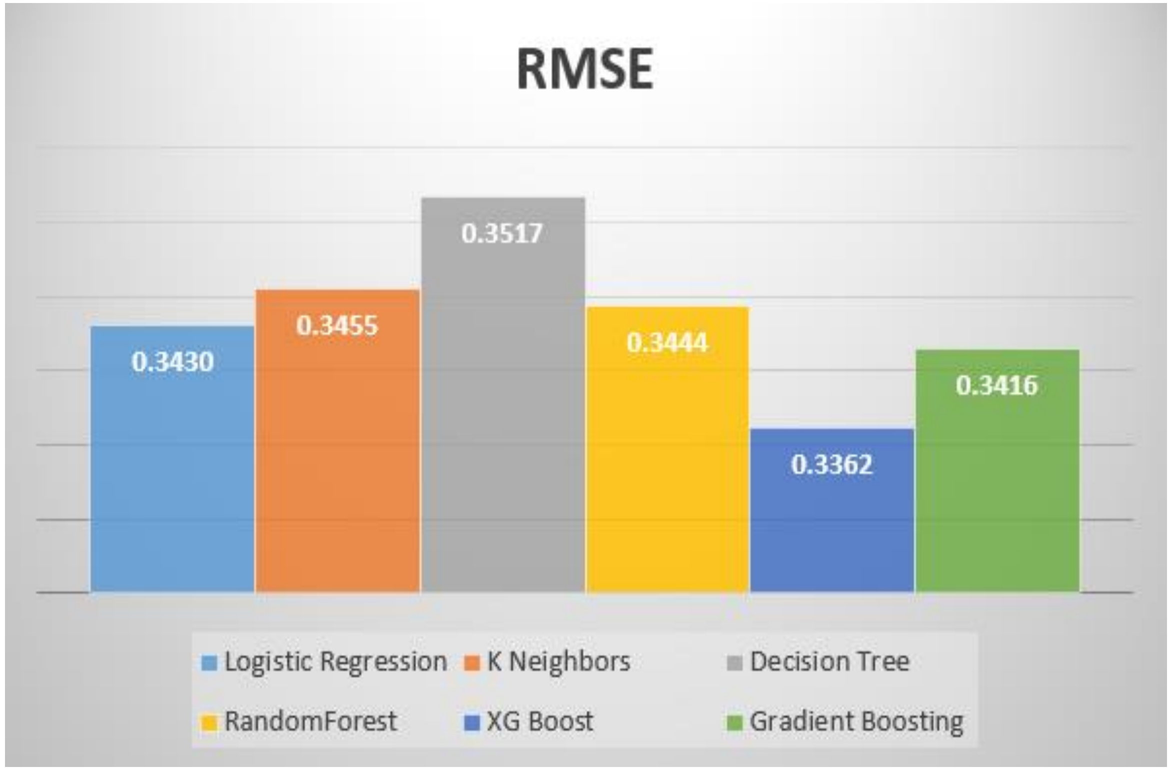
- Marital status is represented by this column.
- 69% of our applicants are married.
- 13% is the maximum rejection rate and is faced by Single applicants followed by Civil marriage.

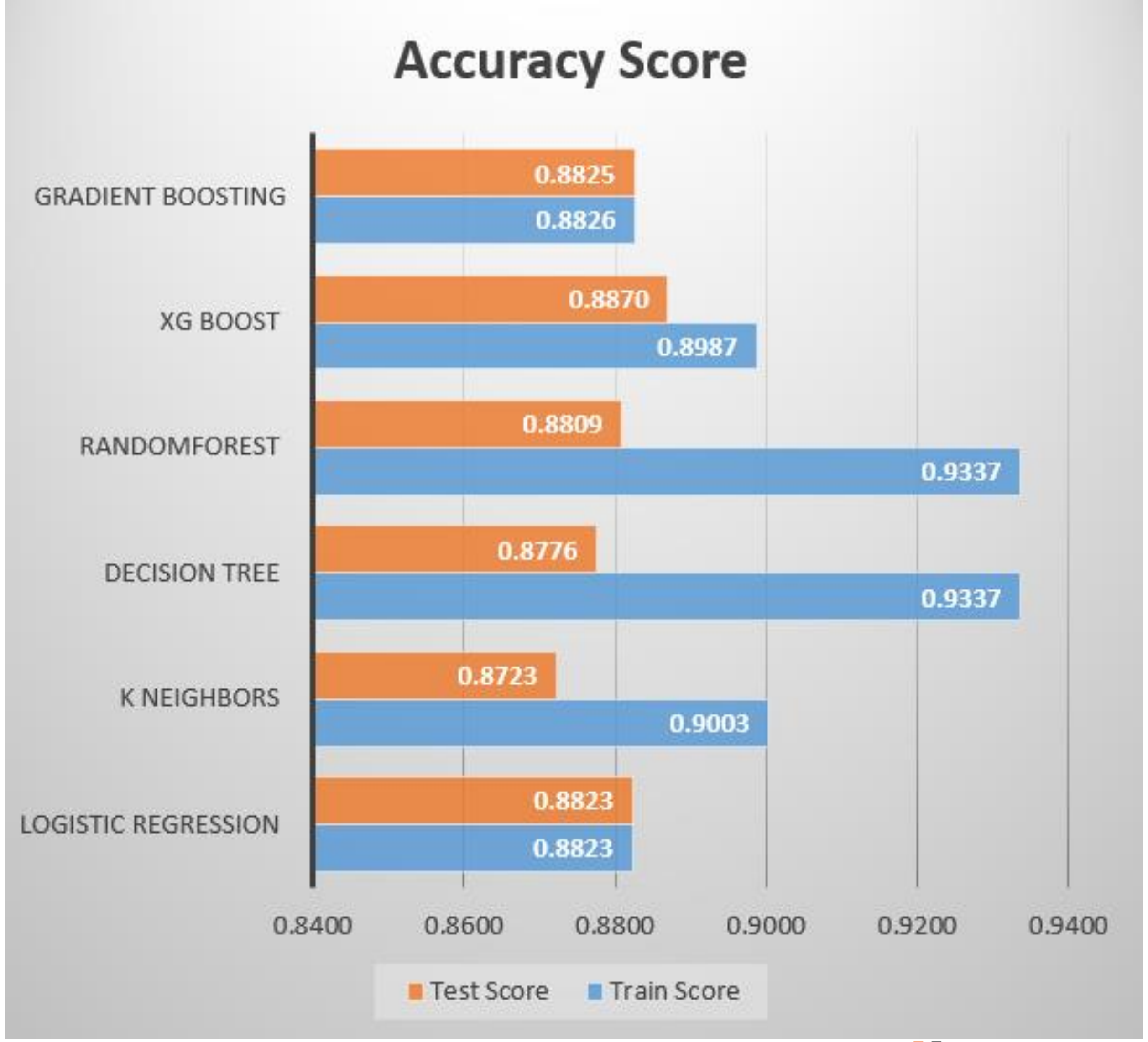
Application of Machine Learning Models



Accuracy & RMSE

- ✓ XGBoost standout among all the models when comparing accuracy score and RMSE
- ✓ accuracy score of train data – 89.87% and test data - 88.70%
- ✓ RMSE - 0.3362

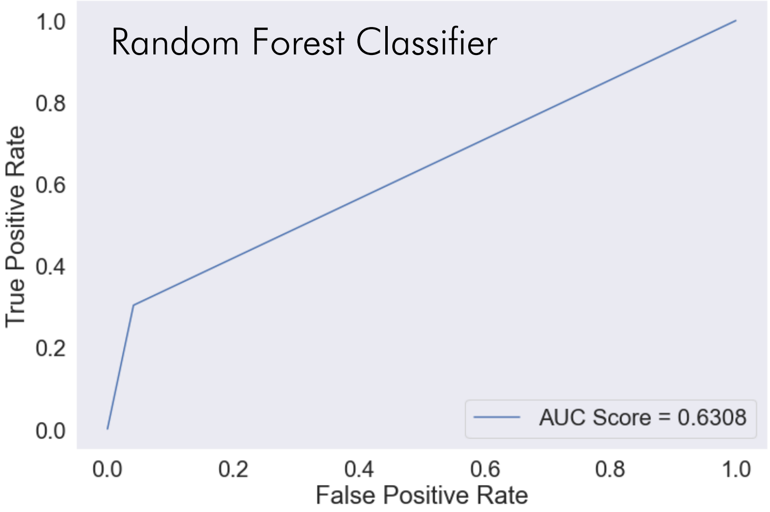


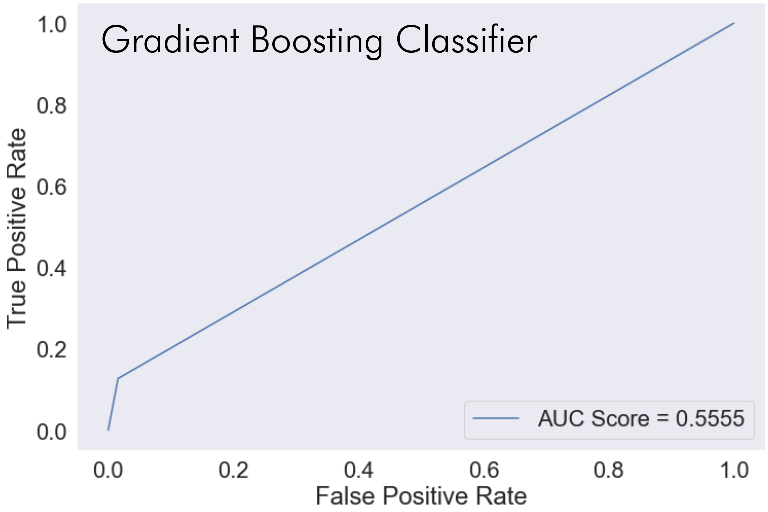


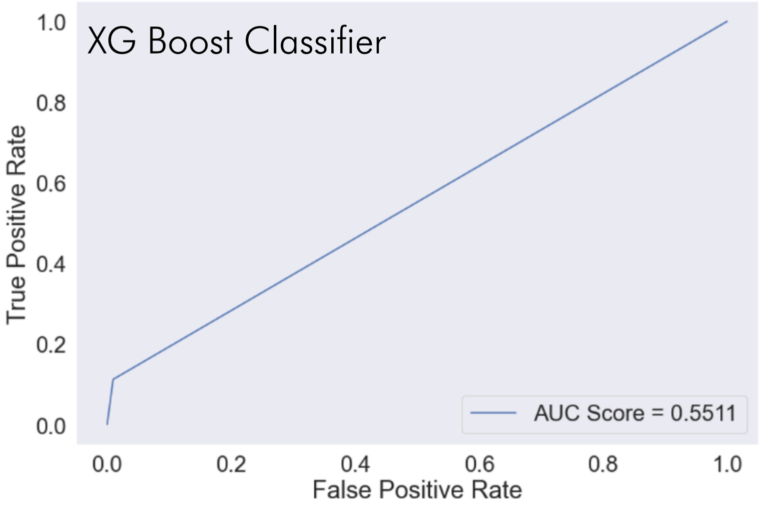
ROC Curve

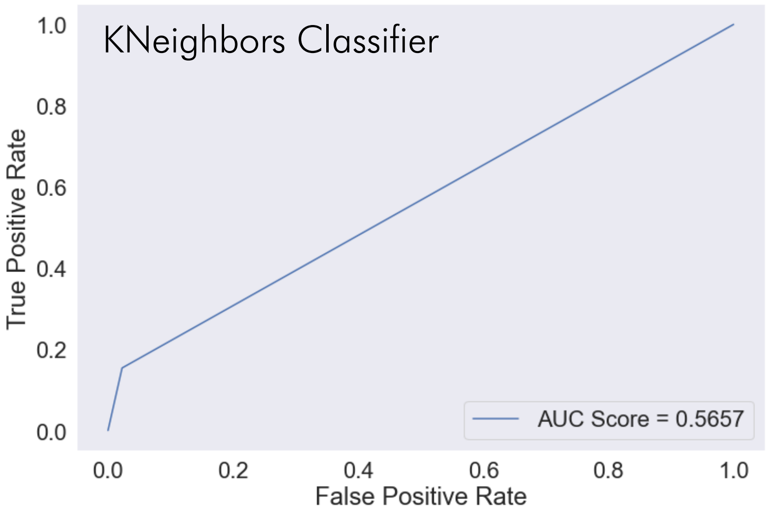

Logistic Regression
AUC Score = 0.5000
True Positive Rate
False Positive Rate


Decision Tree Classifier
AUC Score = 0.6269
True Positive Rate
False Positive Rate


Random Forest Classifier
AUC Score = 0.6308
True Positive Rate
False Positive Rate

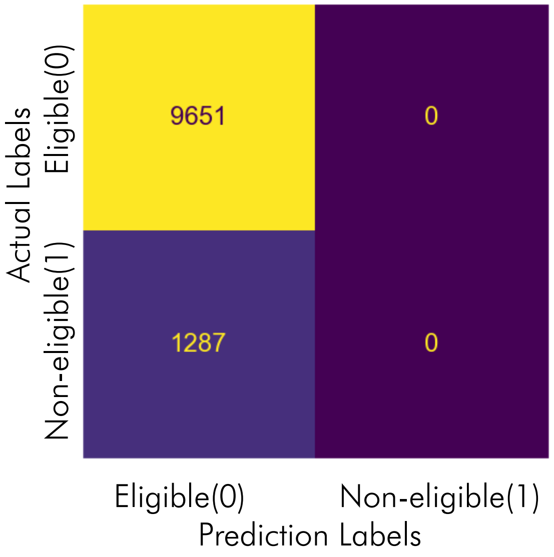

Gradient Boosting Classifier
AUC Score = 0.5555
True Positive Rate
False Positive Rate


XG Boost Classifier
AUC Score = 0.5511
True Positive Rate
False Positive Rate

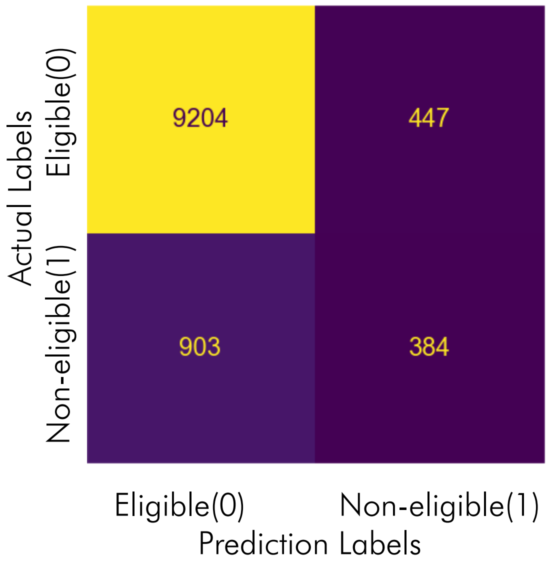

KNeighbors Classifier
AUC Score = 0.5657
True Positive Rate
False Positive Rate

Confusion Matrix

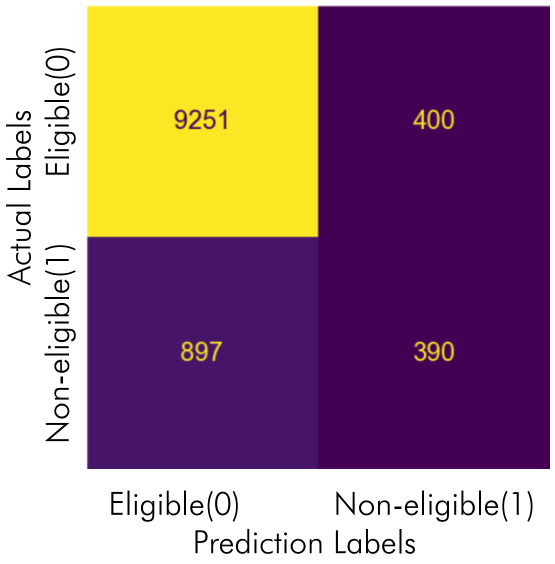
Logistic Regression



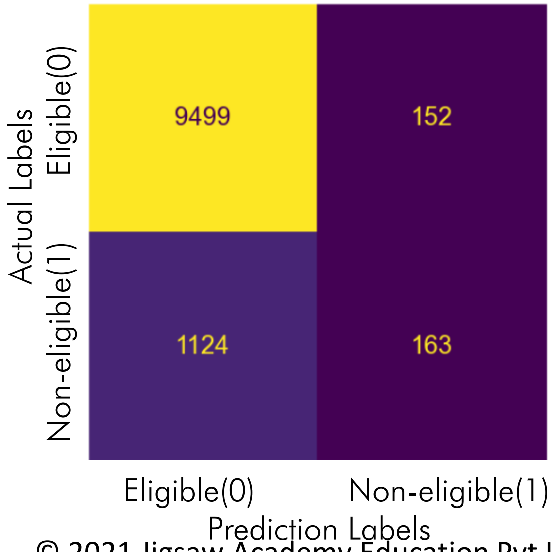
Decision Tree Classifier



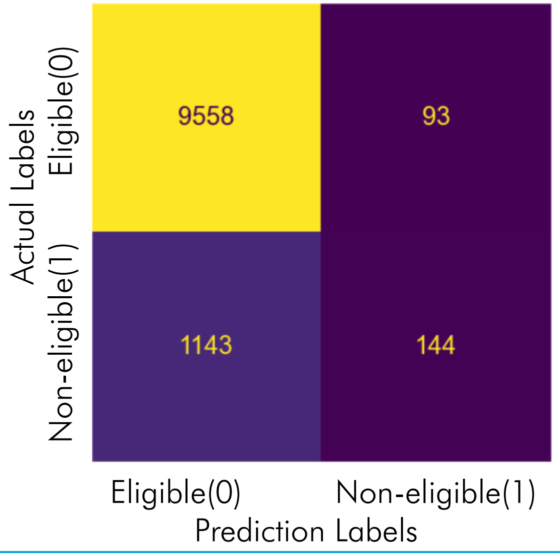
Random Forest Classifier



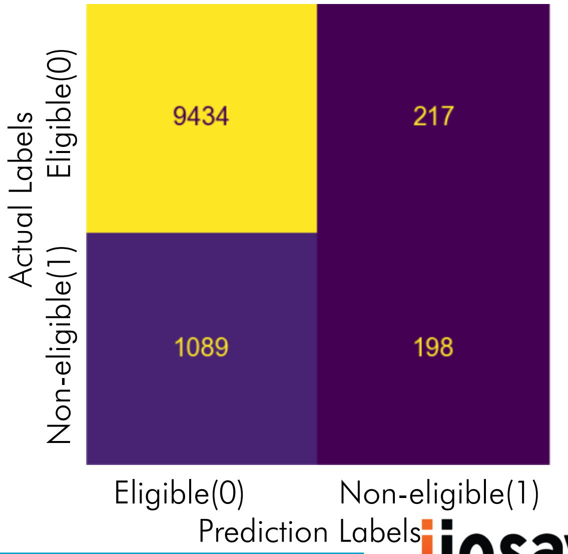
Gradient Boosting Classifier



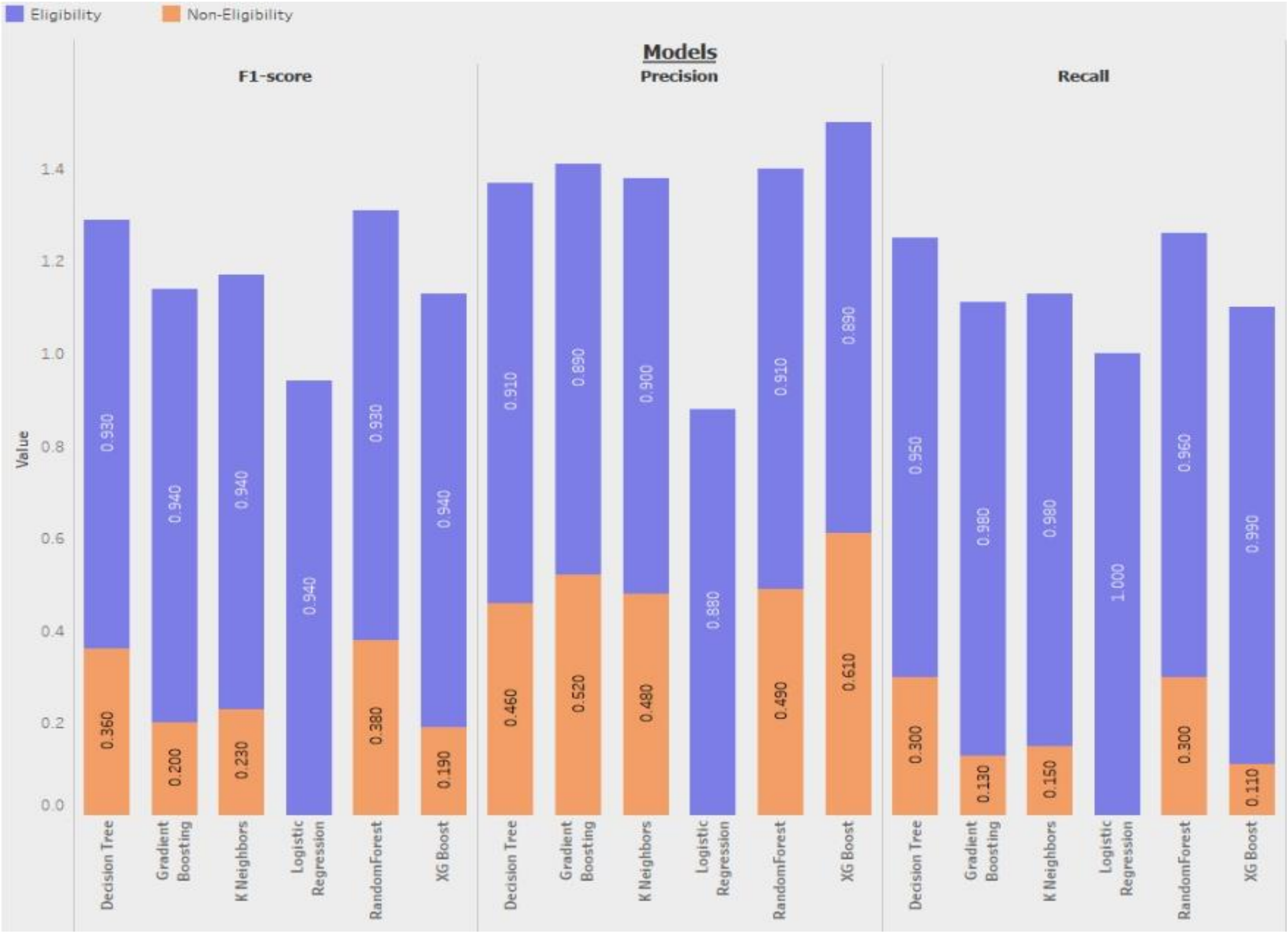
XG Boost Classifier



K Neighbors Classifier



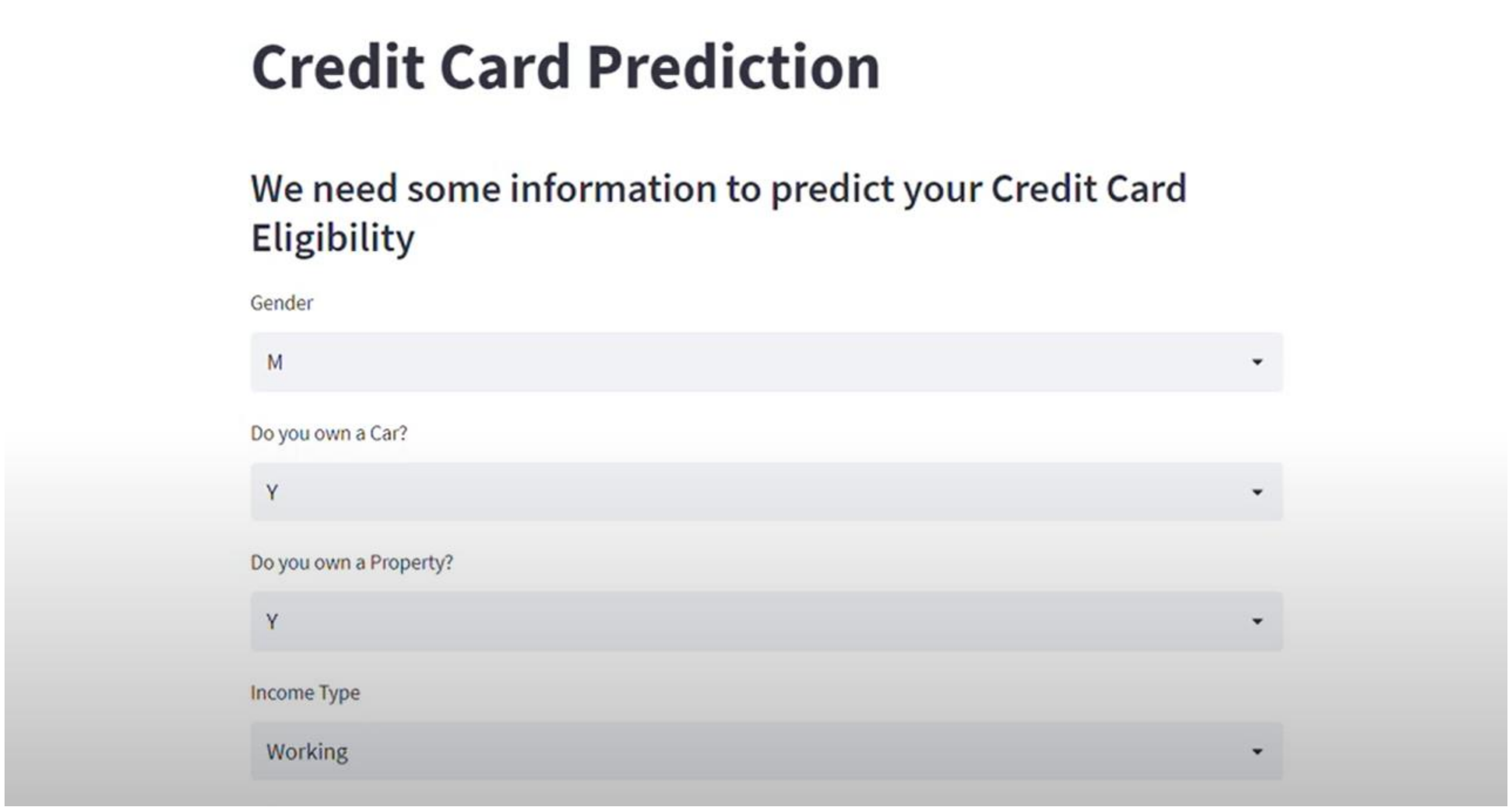
Classification Report



Selected Model

- ✓ As per accuracy, XGBoost was better classification model with 88.7%.
- ✓ We also found out that logistic regression was worst model as per the performance(with recall as 0) although accuracy was 88.23%
- ✓ On comparing other models, as per the performance matrix (like ROC curve & AUC Score, Confusion Matrix and classification report such as recall, precision and F1 score), Random Forest Classifier gave better performance

Machine Learning Model Deployment with Streamlit Web-App



[Link to video of Machine Learning Deployment with Streamlit Web-App]

Challenges, Limitations, Learning & Future scope

CLLF

Challenges & Limitations

- Duplicate data in most of the application records
- Null values in occupation type field is approximately 31%
- Multiple error values in employed years field
- Mean of continuous variable for eligible and not eligible applicants were almost similar

Learning

- Data availability should be to the latest
- To predict accurately it's preferred to have a larger dataset and sample size
- Variables with VIF > 5 are usually dropped but we can still consider them based on the business need and its significance

Future scope

- In present scenario we have used worst instance of credit history of the applicant as eligibility criteria. In future we can consider calculating credit score based on payment history to decide his eligibility for the product.
- Hyper-parameter tuning to improve the model performance
- Implementing Neural Network Models to check if it fits best
- Including additional features and test as it may improve the model performance

Conclusion

- Credit cards are a common risk control method in the financial industry. We have used personal information / data submitted by applicants to predict the probability of future defaults and credit card borrowings.

- We have used supervised classification algorithms such as Logistic Regression, Decision Tree, Random Forest, XG Boost, Gradient Boosting, and K Neighbors. After modelling data with each Algo., we have chosen Random Forest Classifier as it performed the best based on evaluation metrics.

- The model can help in decision making by taking input features and by returning the genuineness of the customer based on past delinquency status and personal information.

Hyperlinks

Code zip file:

Final - Credit Card Eligibility - Predictions Model.zip

[Link to open Tableau Dashboard]

[Link to BYOP Capstone Project files on Google Drive]

[Link to video of Machine Learning Deployment with Streamlit Web-App]

Thank You

From:

Team : IPBA 11 - Group E