Credit Card Approval Prediction

Muhammad Ahmed Rao

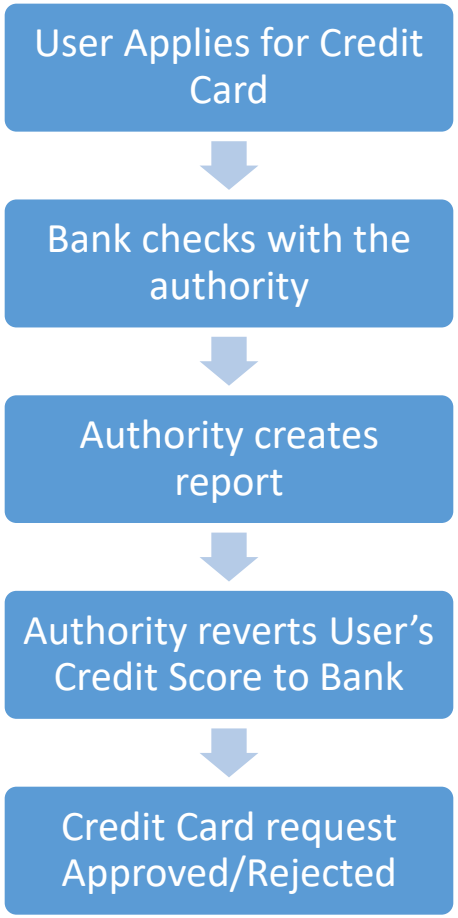
Senior Data Scientist

Introduction

- This exercise is to analyze customer application and credit information to extrapolate the type of applicants which makes a Good or Bad Candidate for a Credit Card.
- The dataset is from Kaggle.
  - application_record.csv contains appliers personal information, which you could use as features for predicting.
  - credit_record.csv records users' behaviors of credit card.

Dataset	# of records
Application	438,557
Credit	1,048,575
After merging both datasets by ID	77,715
Unique ID (Client Number)	36,457

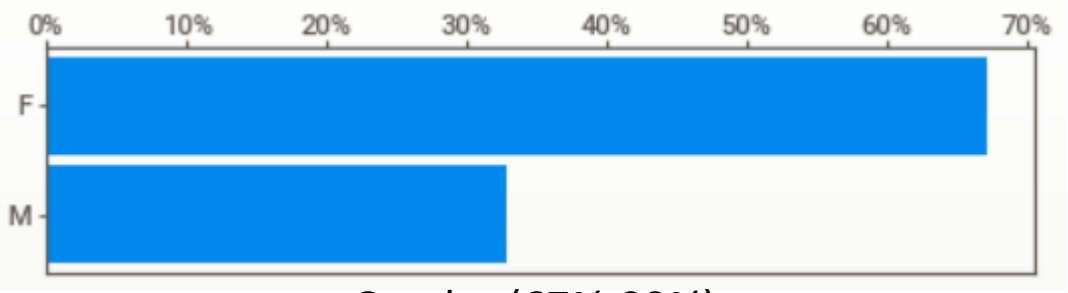
Credit Card Approval Process



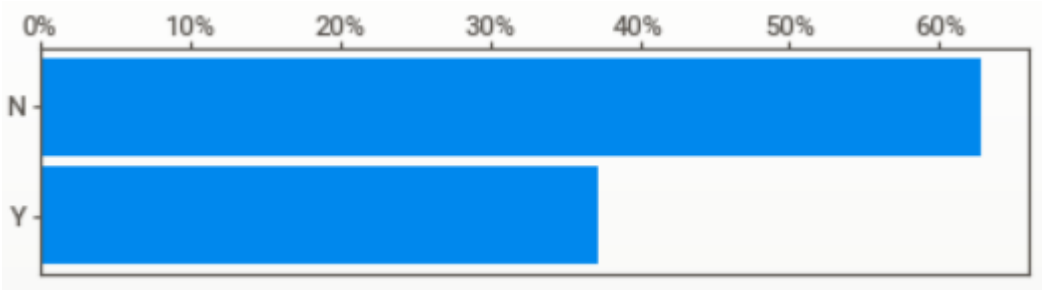
Problem

- A Credit Scorecard is a type of Risk Model used in the classification (scoring) of Credit Risk for individuals, corporations or other legal entities.
- Bad applicants are a major risk for the bank since they are not able to pay and get default at the end hence loss of the bank.
- In order to make this process more robust and lower the magnitude of risk, we need to pass on the historical data and train the model. This will supplement the business make correct decisions.

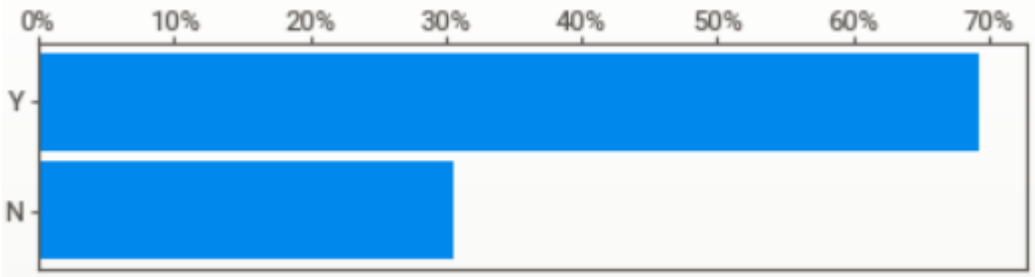
Data Analysis



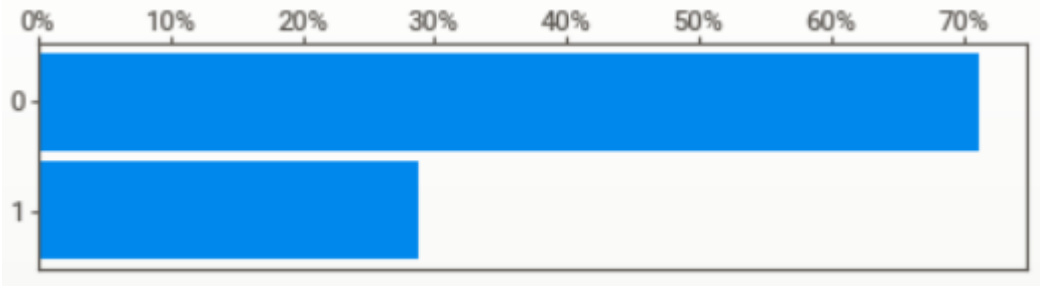
Gender (67%-33%)



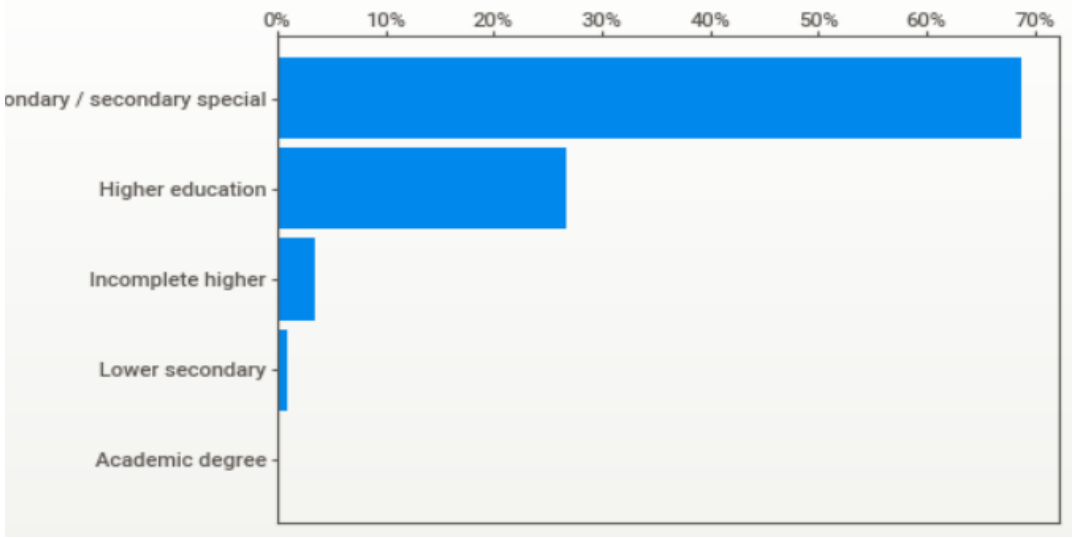
Own Car (63%-37%)



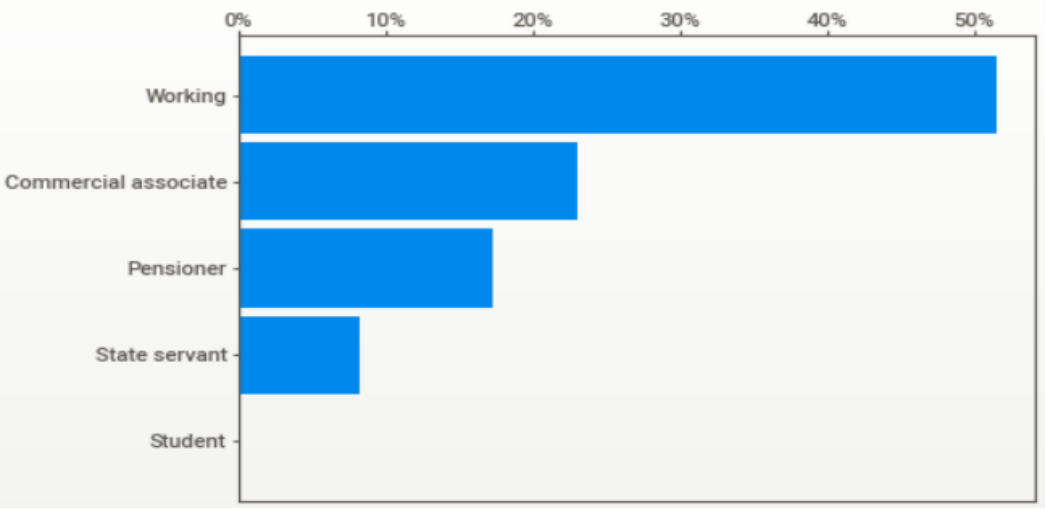
Own Property (69%-31%)



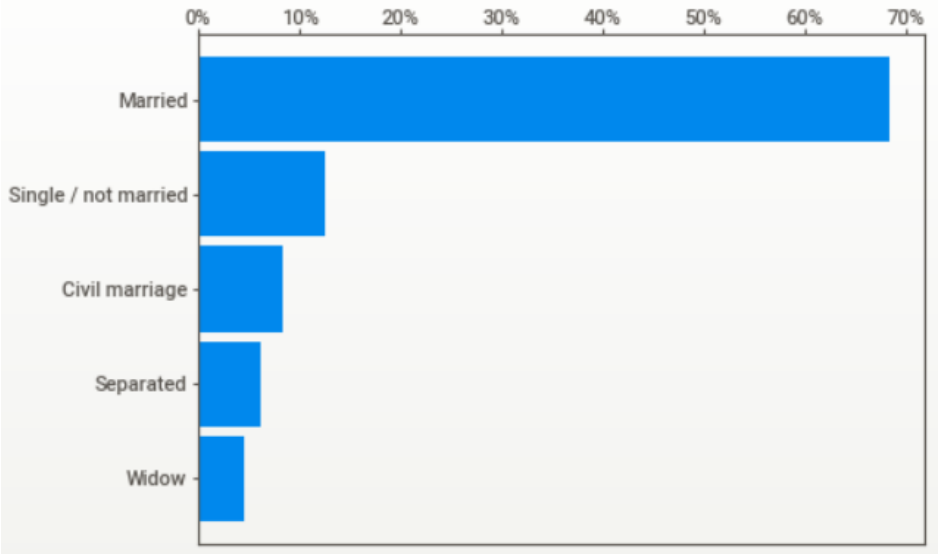
Phone Flag (71%-29%)



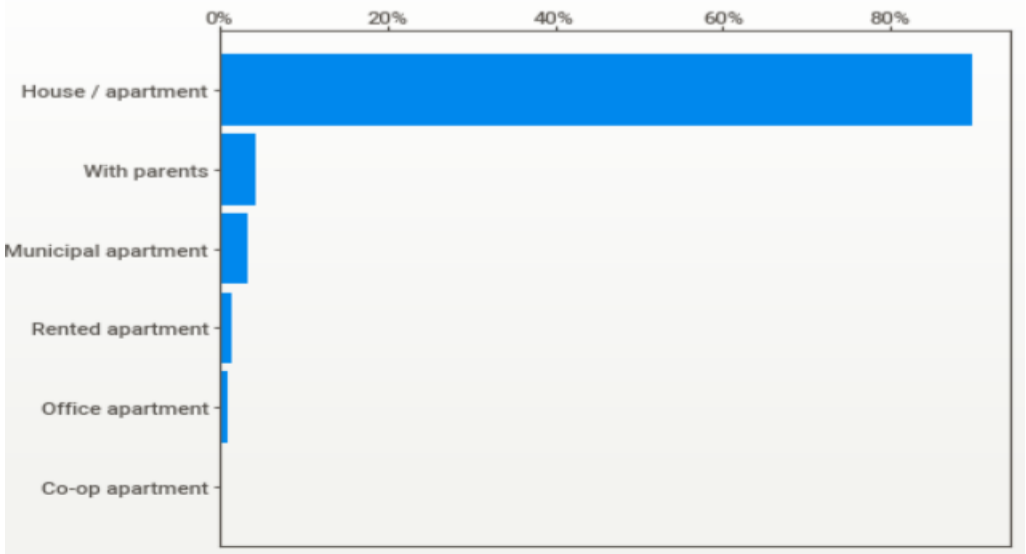
Education Type (69%-27%-3%)



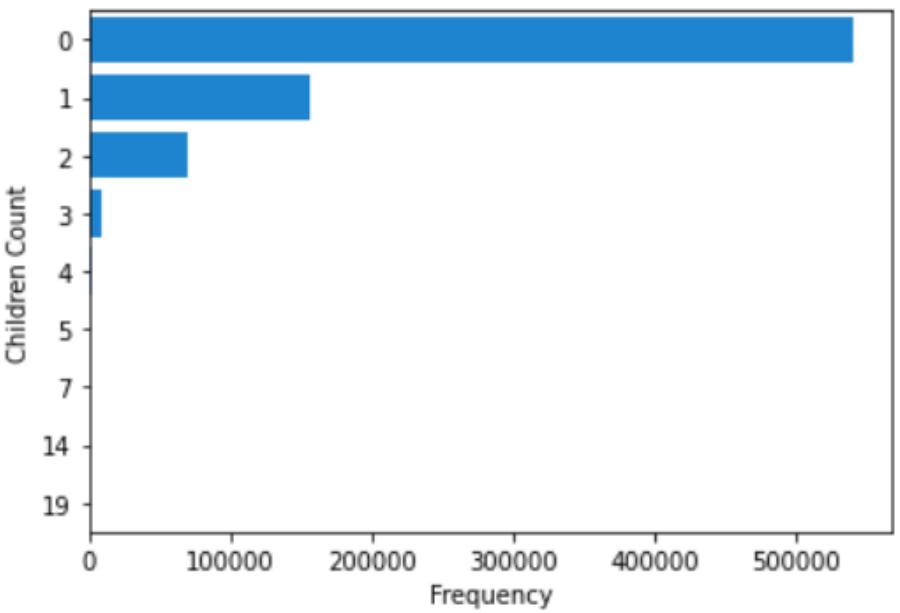
Income Type (52%-23%-17%-18%)



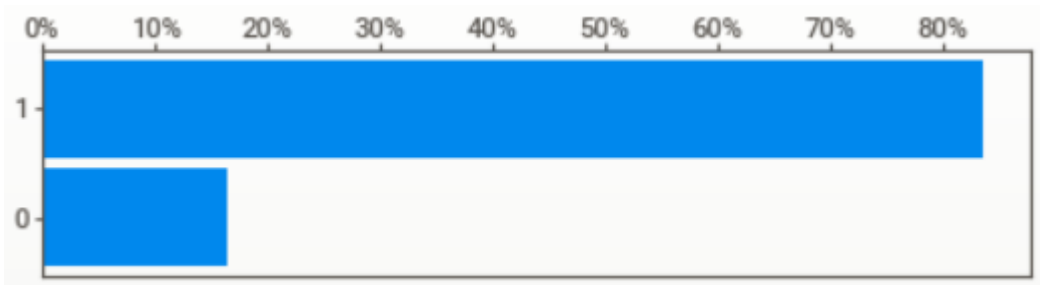
Family Status (68%-13%-8%-6%-4%)



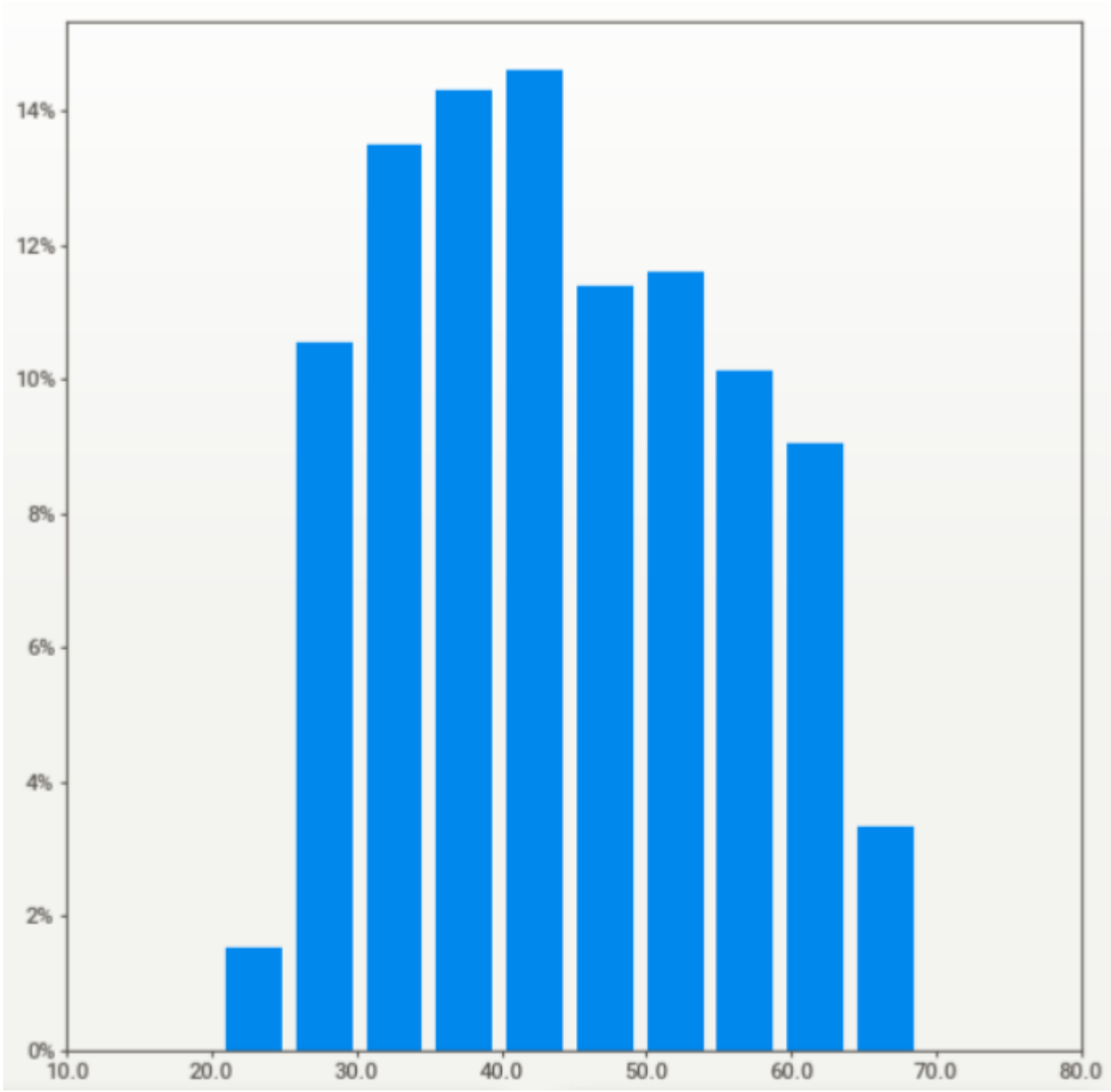
House Type (90%-4%-3%)



Children Count

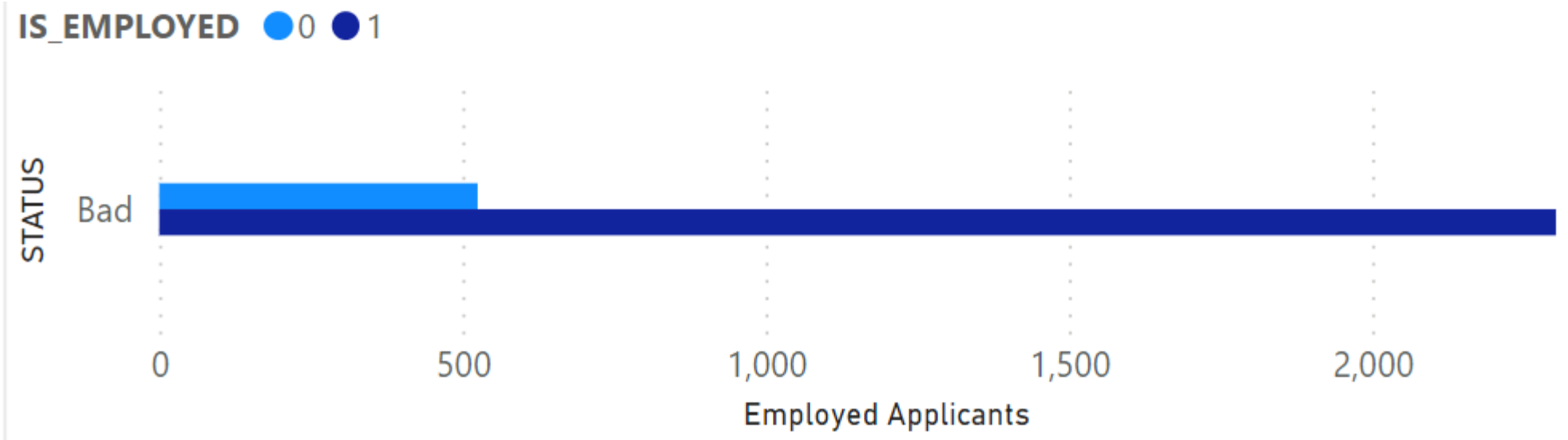


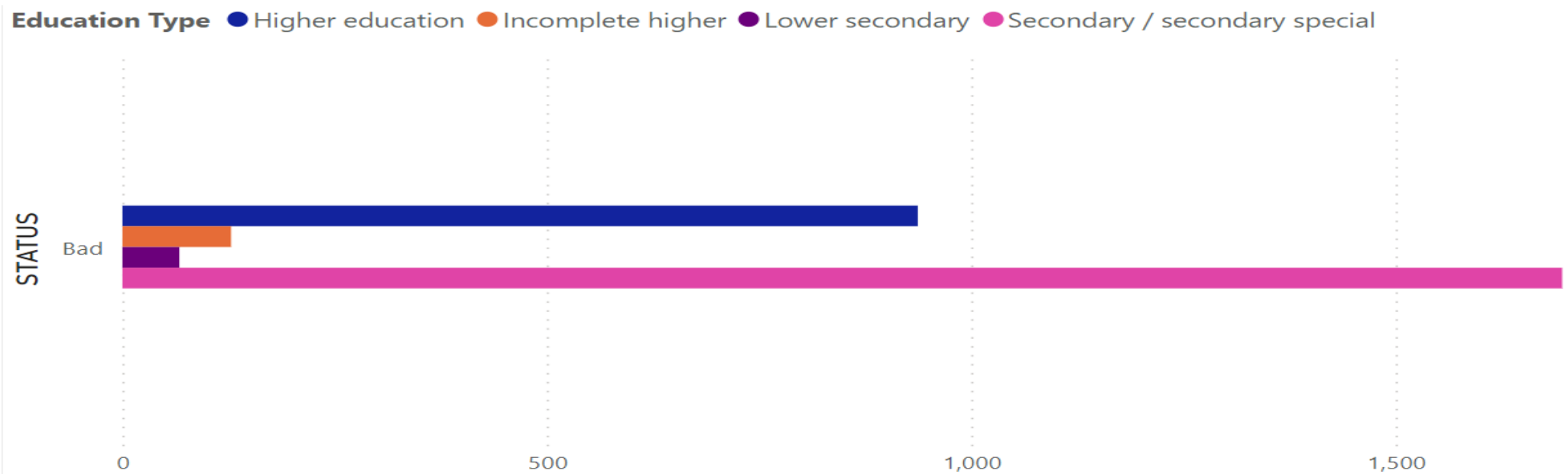
Is Employed (84%-16%')

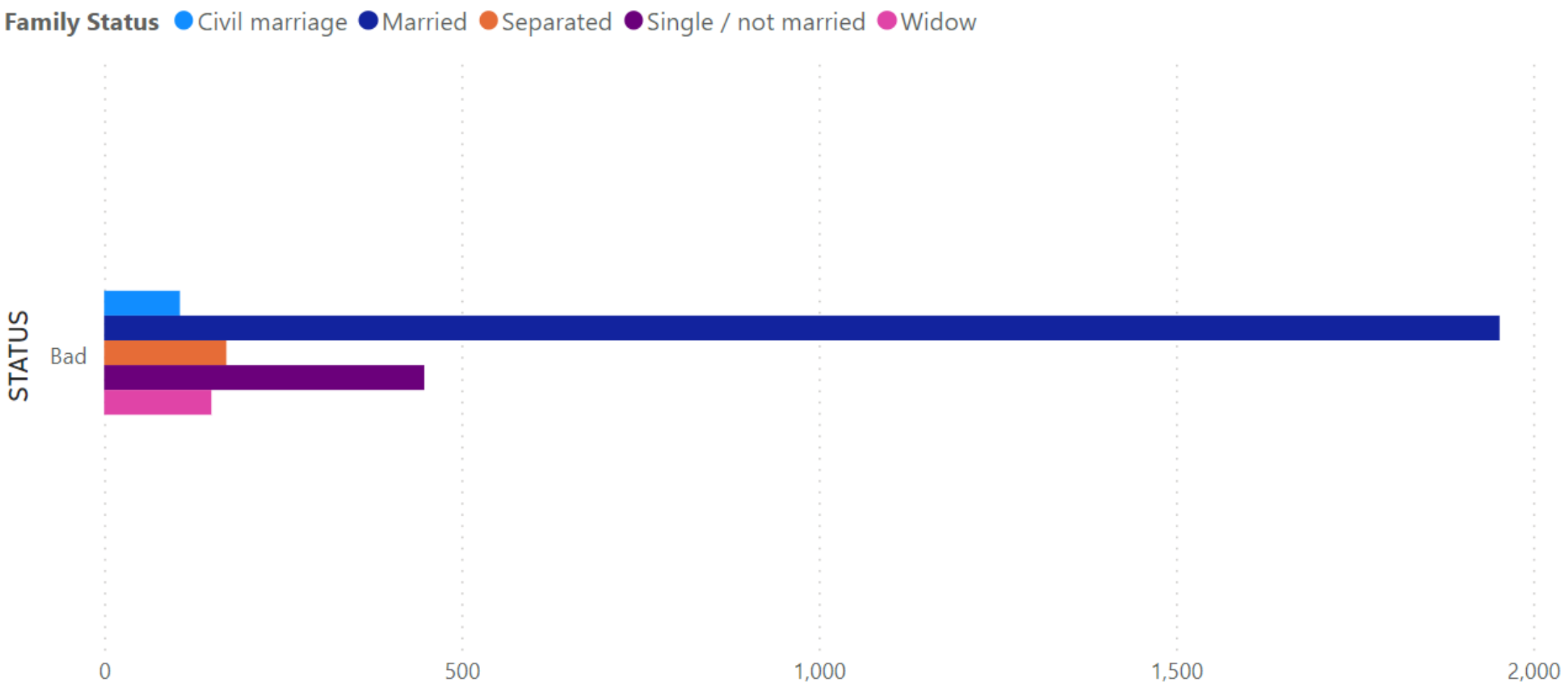


The age ranges from 20.5 to 69 years

- 25% of applicants are under 35 years of age
- 50% are under 45 years of age
- 95% are under 64 years of age


IS_EMPLOYED
0
1
STATUS
Bad
0
500
1,000
1,500
2,000
Employed Applicants


Education Type
Higher education
Incomplete higher
Lower secondary
Secondary / secondary special
STATUS
Bad
0
500
1,000
1,500


Family Status
Civil marriage
Married
Separated
Single / not married
Widow
STATUS
Bad
0
500
1,000
1,500
2,000

Data Cleaning

- I dropped column FLAG_MOBIL since it only contained 1 constant value which does not give any information to separate Bad and Good application added field.
- Occupation type contained 240,048 missing values. I put them with ‘Other’ occupation types

Feature Engineering

- Computed following columns:
  - AGE from DAYS_BIRTH
  - IS_EMPLOYED from DAYS_EMPLOYED
  - EMPLOYMENT_YEARS from DAYS_EMPLOYED
  - EMPLOYMENT_START_DT from DAYS_EMPLOYED (later I removed it since I did not used it in training the model)

Data Labelling and distribution

- The labelling below is based on the description provided and also had feedback with my past colleagues who are in this domain.

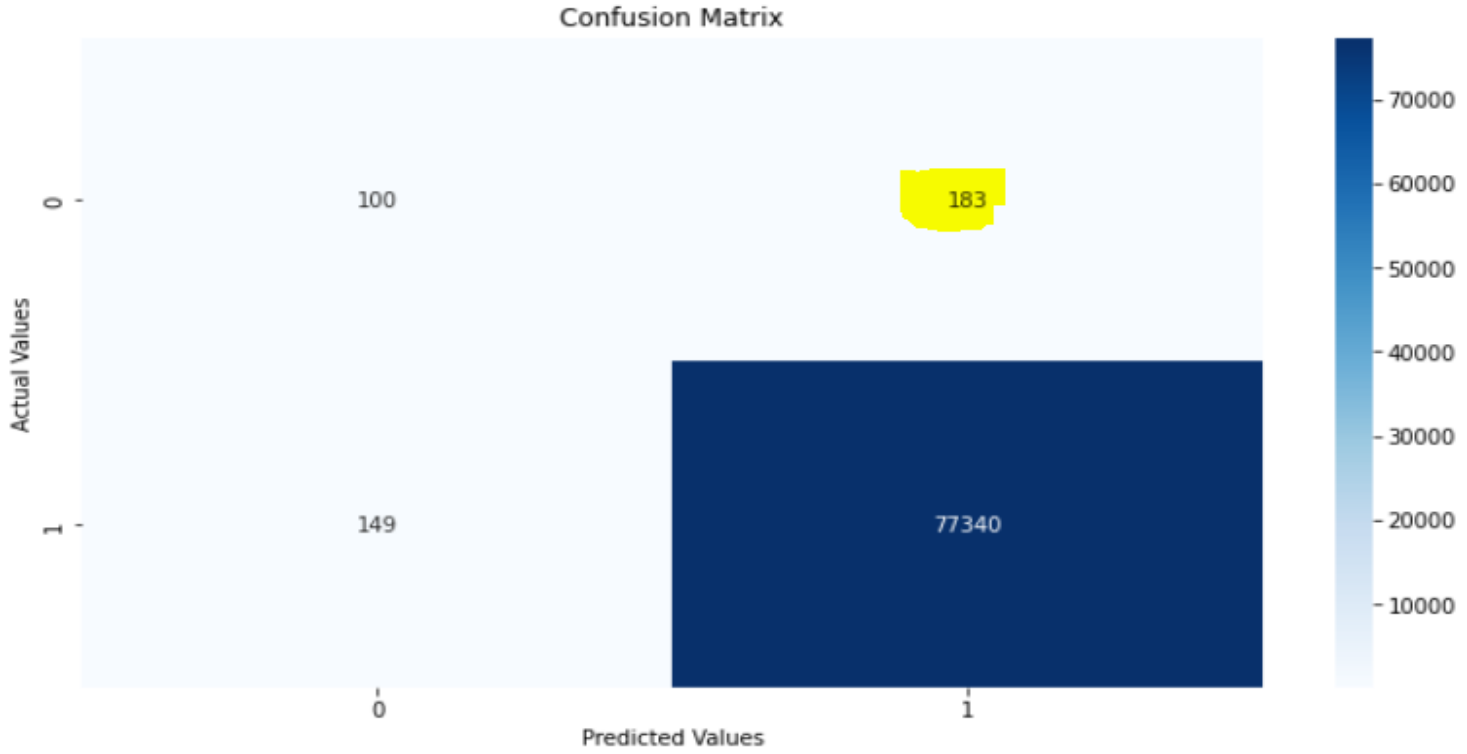
Status	Description	Label	# Applicants	% Applicants
0	1-29 days due	Good	290,654	37.37%
1	30-59 days past due	Good	8,747	1.12%
2	60-89 days overdue	Bad	801	0.10%
3	90-119 days overdue	Bad	286	0.04%
4	120-149 days overdue	Bad	214	0.03%
5	Overdue or bad debts, write-offs for more than 150 days	Bad	1,527	0.20%
C	Paid off that month	Good	329,536	42.37%
X	No loan for the month	Good	145,950	18.77%
Total			77,715	100%

Good	774,887	99.36%
Bad	2,828	0.36%

Data Splitting & Model Training

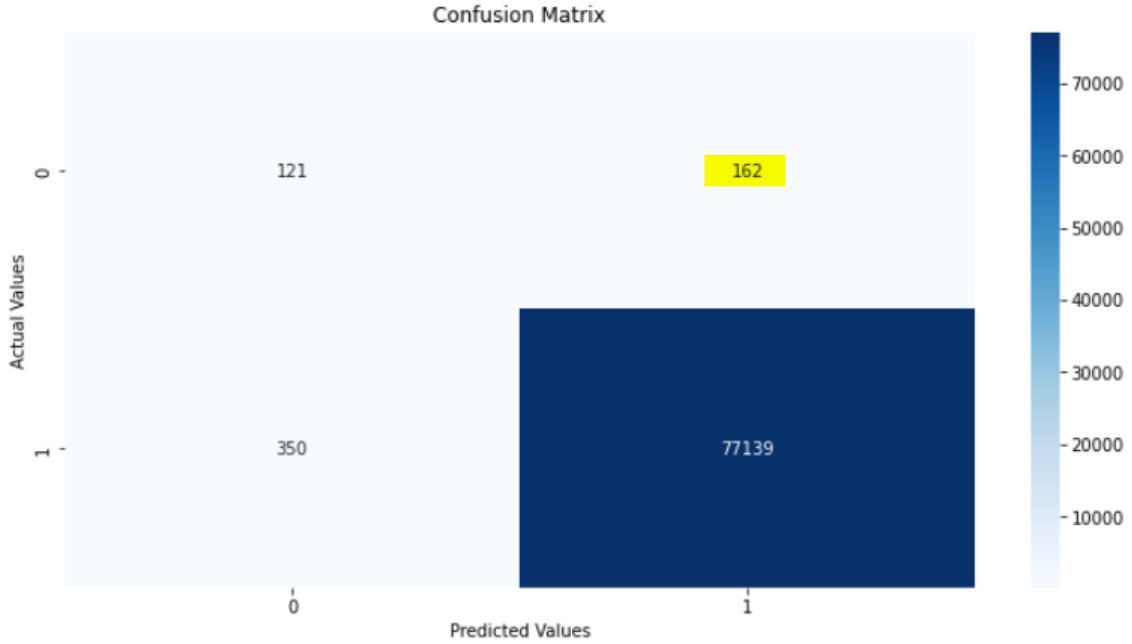
- I split the data in train and test in stratified fashion
- At first, I trained the un-balanced data using Decision Tree and as expected the model did not classify the 64.66% Bad Applicants correctly.

Train	699,943
Test	77,772
Total	**777,715**

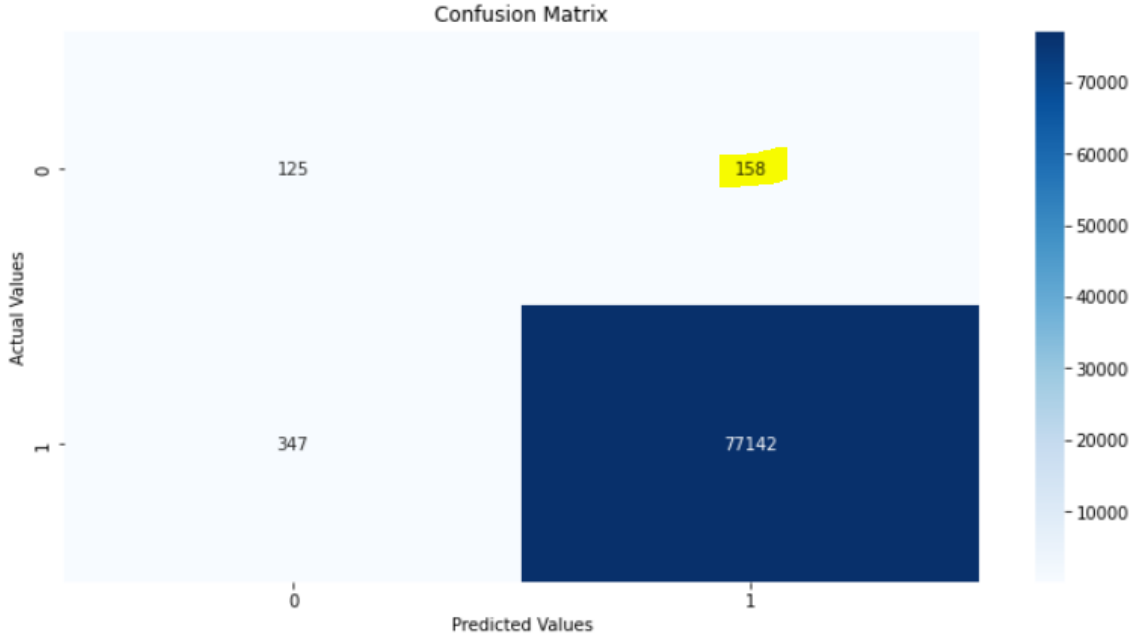


Random Oversampling

- I applied Random Oversampling technique to overcome the un-balancing issue. Then I trained Decision Tree and Random Forest.
- Decision Tree misclassified 57% of the Bad Applicants whereas Random Forest misclassified 55.8% of the Bad Applicants.
- We need further improvement to reduce the misclassification rate



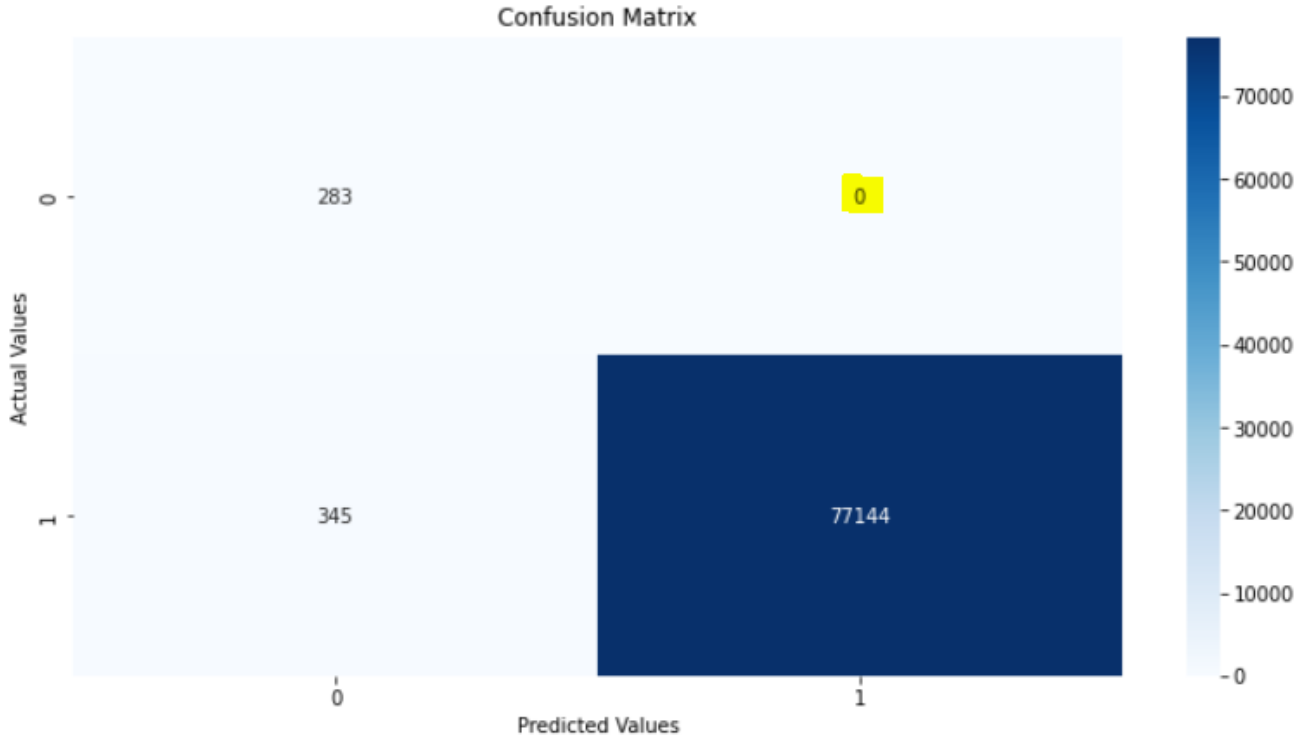
DECISION TREE



RANDOM FOREST

SMOTE

- To further improve I used SMOTE method for oversampling since it maps minority class by generating synthetic examples rather than by oversampling with replacement.
- The results significantly improved using this technique since all bad applicants were classified correctly.



Decision Tree & Random Forest

Feature Importance

- Following top 5 attributes had highest importance after training random forest.
  - AGE
  - EMPLOYMENT_YEARS
  - AMT_INCOME_TOTAL
  - MONTHS_BALANCE
  - CNT_FAM_MEMBERS

Feature importances
Mean decrease in impurity
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
CNT_CHILDREN
AMT_INCOME_TOTAL
FLAG_WORK_PHONE
FLAG_PHONE
FLAG_EMAIL
CNT_FAM_MEMBERS
MONTHS_BALANCE
AGE
IS_EMPLOYED
EMPLOYMENT_YEARS
CODE_GENDER_M
FLAG_OWN_CAR_Y
FLAG_OWN_REALTY_Y
NAME_INCOME_TYPE_Pensioner
NAME_INCOME_TYPE_State servant
NAME_INCOME_TYPE_Student
NAME_INCOME_TYPE_Working
NAME_EDUCATION_TYPE_Higher education
NAME_EDUCATION_TYPE_Incomplete higher
NAME_EDUCATION_TYPE_Lower secondary
UCATION_TYPE_Secondary / secondary special
NAME_FAMILY_STATUS_Married
NAME_FAMILY_STATUS_Separated
NAME_FAMILY_STATUS_Single / not married
NAME_FAMILY_STATUS_Widow
NAME_HOUSING_TYPE_House / apartment
NAME_HOUSING_TYPE_Municipal apartment
NAME_HOUSING_TYPE_Office apartment
NAME_HOUSING_TYPE_Rented apartment
NAME_HOUSING_TYPE_With parents
OCCUPATION_TYPE_Cleaning staff
OCCUPATION_TYPE_Cooking staff
OCCUPATION_TYPE_Core staff
OCCUPATION_TYPE_Drivers
OCCUPATION_TYPE_HR staff
OCCUPATION_TYPE_High skill tech staff
OCCUPATION_TYPE_IT staff
OCCUPATION_TYPE_Laborers
OCCUPATION_TYPE_Low-skill Laborers
OCCUPATION_TYPE_Managers
OCCUPATION_TYPE_Medicine staff
OCCUPATION_TYPE_Others
OCCUPATION_TYPE_Private service staff
OCCUPATION_TYPE_Realty agents
OCCUPATION_TYPE_Sales staff
OCCUPATION_TYPE_Secretaries
OCCUPATION_TYPE_Security staff
OCCUPATION_TYPE_Waiters/barmen staff

Further Work

- If this project had to go on full scale:
  - Filter out the features with less importance.
  - I would work on reducing the Good applicants misclassification rate using Hyper-parameter tuning.
  - Will create automated pipeline of the model and integrate with the system so that this model can supplement business with their decision.