Home Credit Default Risk Prediction

- BHARATESHA N S

PROBLEM STATEMENT:

- Building a model to predict how capable each applicant is of repaying a loan, so that sanctioning loan only for the applicants who are likely to repay the loan.
- Based on the applicant data we have to predict “ **TARGET** “.
- Target variable contains two values.
- Target variable value 0 means loan is repayed, value 1 means loan is not repayed.
- Build a Model and Calculate best classification accuracy measures, for this problem statement.

ABOUT DATASET

DATA SOURCE :

- For this project , an Home Credit Default Risk dataset has been picked, which is available on Kaggle website .
- In this website two files picked application_train.csv and application_test.csv.
- In application_train.csv data contains records of 307511 and 94 fields like SK_ID_CURR, TARGET, NAME_CONTRACT_TYPE, CODE_GENDER

  CNT_CHILDREN, OCCUPATION_TYPE , …… etc .
- In application_test.csv data contains records of 48744 and field of 93

  here TARGET field not their ,need to predict .

TOOLS AND TECHNIQUES

- We have selected **python3** as our analytics tool .
- Platforms used for this project Jupyter Notebook .
- Python include many packages such as pandas , numpy , seaborn , matplotlib , Scikit-Learn , XGBoost ,… etc .
- Algorithm used for this project such as Logistic Regression , SGDClassifier, Random Forest , Extreme Gradient Boosting , Gradient Boosting Classifier, K-Nearest Neighbors , Decision Tree Classifier .

Road To Achieve

STEP-1

Collecting the data and Analyzing the data

STEP-2

Exploratory data analysis

STEP-3

Data Pre-Processing

STEP-4

Model Building and Testing

STEP-5

Model Evaluation

Methodology

1. CORRELATION

- Checking the Correlation by plotting heap map.
- Removing the High **Negatively** and **Positively** Correlated Columns such as AMT_GOODS_PRICE, APARTMENTS_MODE,APARTMENTS_MODE, BASEMENTAREA_MODE, YEARS_BUILD_MODE , ….. etc .

2. HANDLING MISSIG VALUES

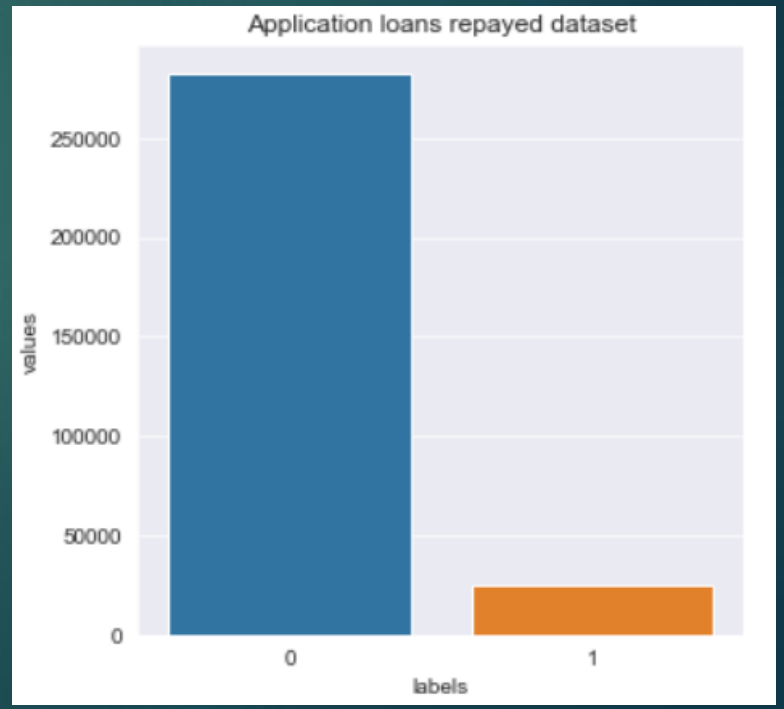
- Checking the missing values by plotting pointplot .
- Here some Categorical and Numerical columns having more than 60% missing values .
- To resolve this problem of missing values treat with “ **MODE** “ for Categorical and “ **Simpleimpute** “ from sklearn for Numerical columns .

3.ENCODING FOR CATEGORICAL FEATURES

- The features that had nominal data is converted into binary features by doing “ **One-Hot Encoding** ” using sklearn .

4.CHECKING WHETHER TARGET DATA IS BALANCED OR NOT

- Target variable value 0 means loan is repayed and value 1 means loan is not repayed by the customers.
- Here the target data is imbalanced, if in case the data is imbalanced we need to do oversampling using “ **SMOTE** “ technique .



5. FEATURE SCALING

- **MinMax** Scaler shrinks the data within the given range, usually of **0 to 1**.
- It transforms data by scaling features to a given range.
- It scales the values to a specific value range without changing the shape of the original data .

6. MODEL BUILDING

- For this **Classification** problem I used Logistic Regression , SGDClassifier, Random Forest , Extreme Gradient Boosting , Gradient Boosting Classifier, K-Nearest Neighbors , Decision Tree Classifier .
- Perform the Model Evaluation technique like Confusion Matrix , ROC curve, Accuracy, Precision, Recall, kappa, AUC and F-1 Score

TABULATED THE ALL MODEL

SL.NO	Model	AUC Score	Precision Score	Recall Score	Accuracy Score	Kappa Score	F1-score
01	Logistic Regression	0.761308	0.692819	0.697544	0.695370	0.390747	0.695174
02	SGDClassifier	0.760184	0.671312	0.765615	0.696608	0.393550	0.715369
03	Random Forest	0.979379	1.000000	0.912430	0.956392	0.912751	0.954210
04	Extreme Gradient Boosting	0.966399	0.978107	0.879113	0.930002	0.859945	0.925972
05	Gradient Boosting Classifier	0.974714	0.999804	0.908043	0.954119	0.908202	0.951717
06	K-Nearest Neighbours	0.940334	0.794834	0.988528	0.867221	0.734698	0.881162
07	Decision Tree Classifier	0.821317	0.745534	0.772345	0.755357	0.510777	0.758703

RANDOM FOREST

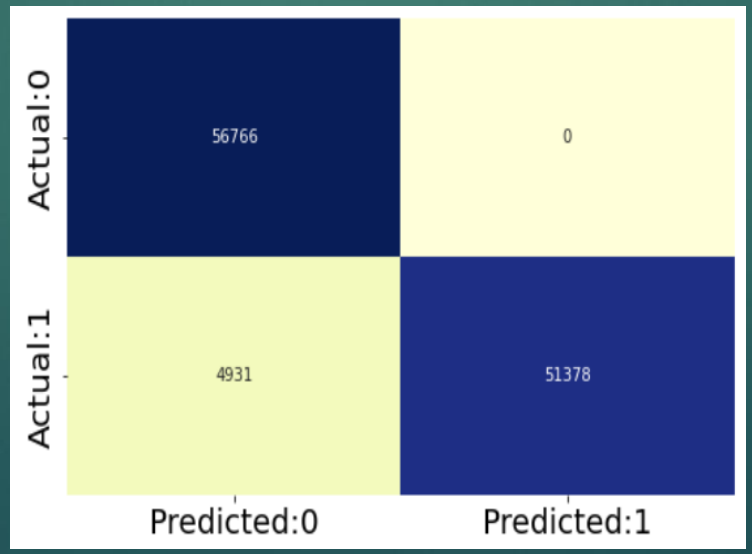
- Random Forest is a supervised learning algorithm
- It creates forest and make it random based on bagging technique.

  It aggregate Classification tree .
- After Build a model performing the Model Evaluation technique.

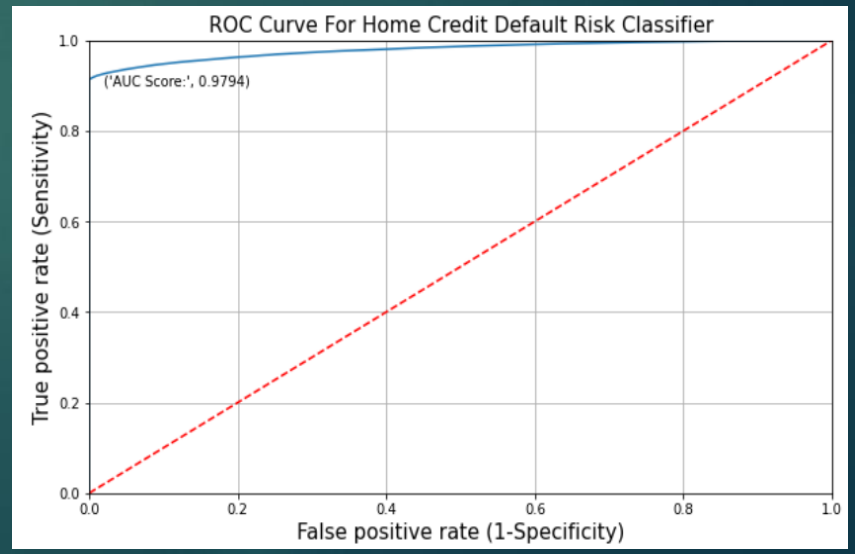
i) Classification Report

	precision	recall	f1-score	support
0	0.92	1.00	0.96	56766
1	1.00	0.91	0.95	56309
accuracy			0.96	113075
macro avg	0.96	0.96	0.96	113075
weighted avg	0.96	0.96	0.96	113075

ii) Confusion Matrix



iii) ROC Curve



Cross validation :

- Cross-validation is a technique for evaluating ML models .
- By training several ML models on subsets of the available input data.
- And evaluating them on the complementary subset of the data.
- Using cross-validation, there are high chances that we can detect over-fitting with ease.

KEY FINDINGS

- In Male and Female highest barrower is Female.
- Here we can see in figure(1) more than 200k Female are the borrower

- And Compare to Male ,Female are more difficulty to repay the loan as shown below figure(2)

Figure 1:

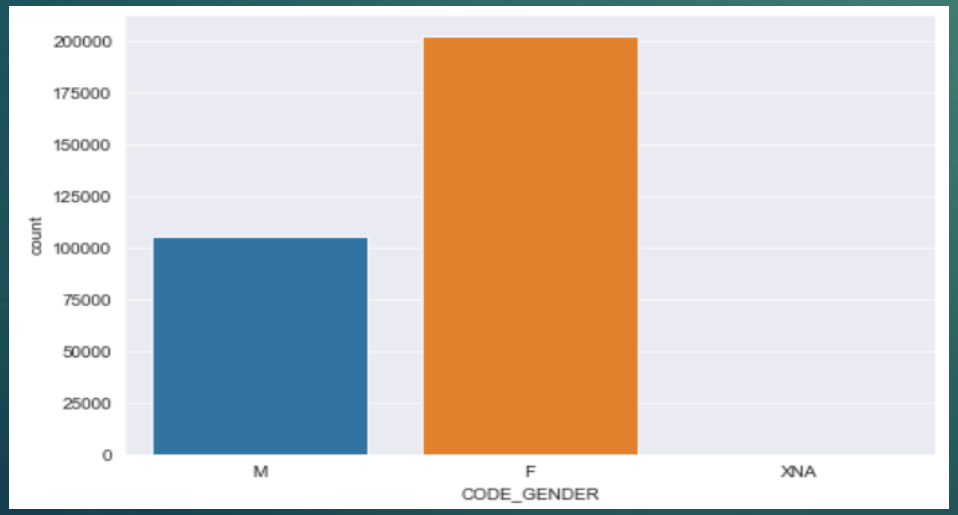
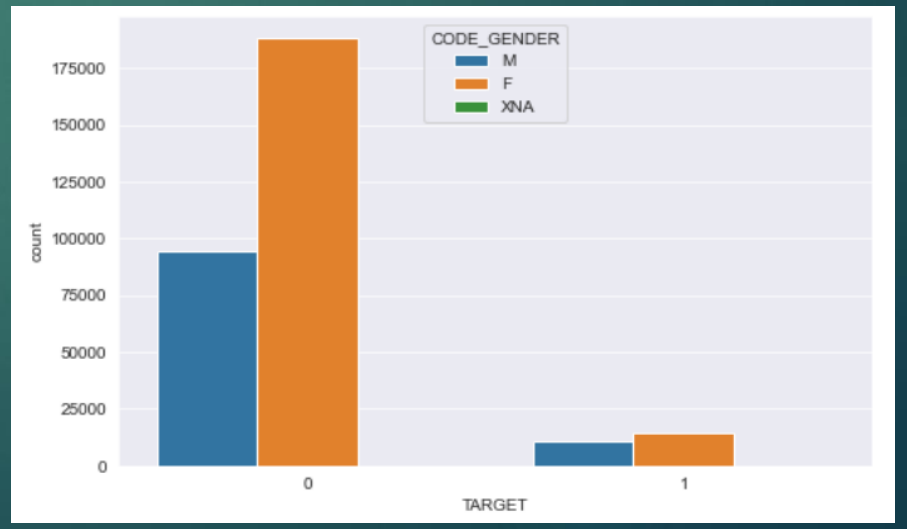
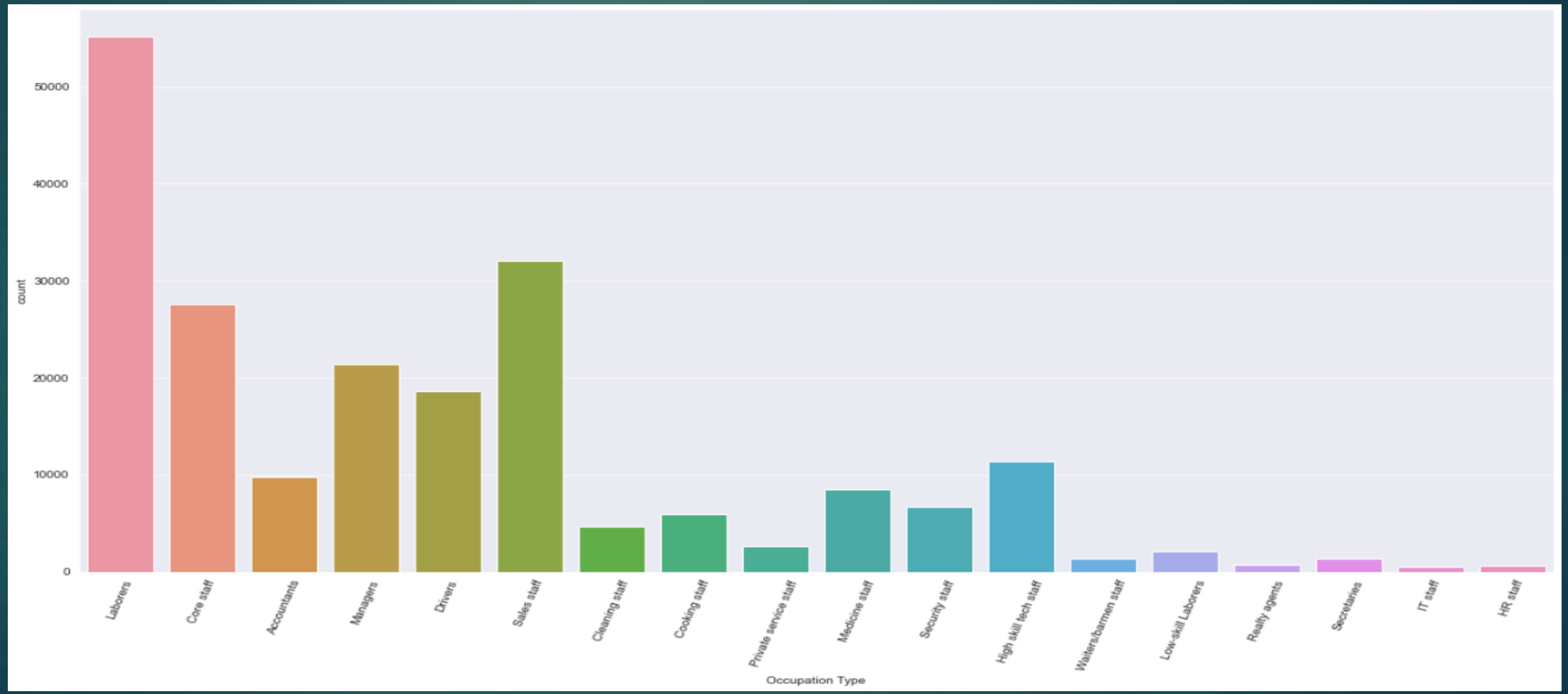


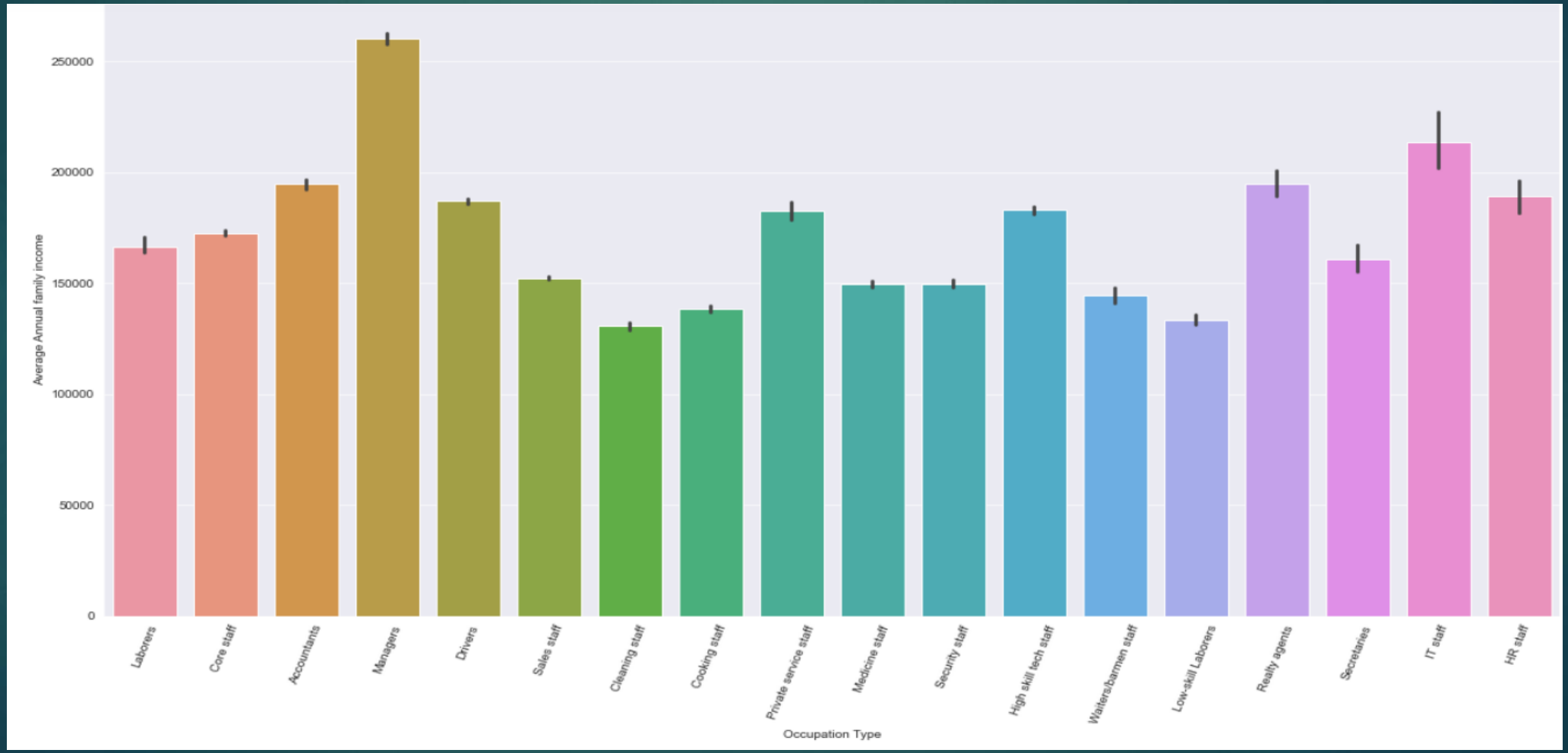
Figure 2:



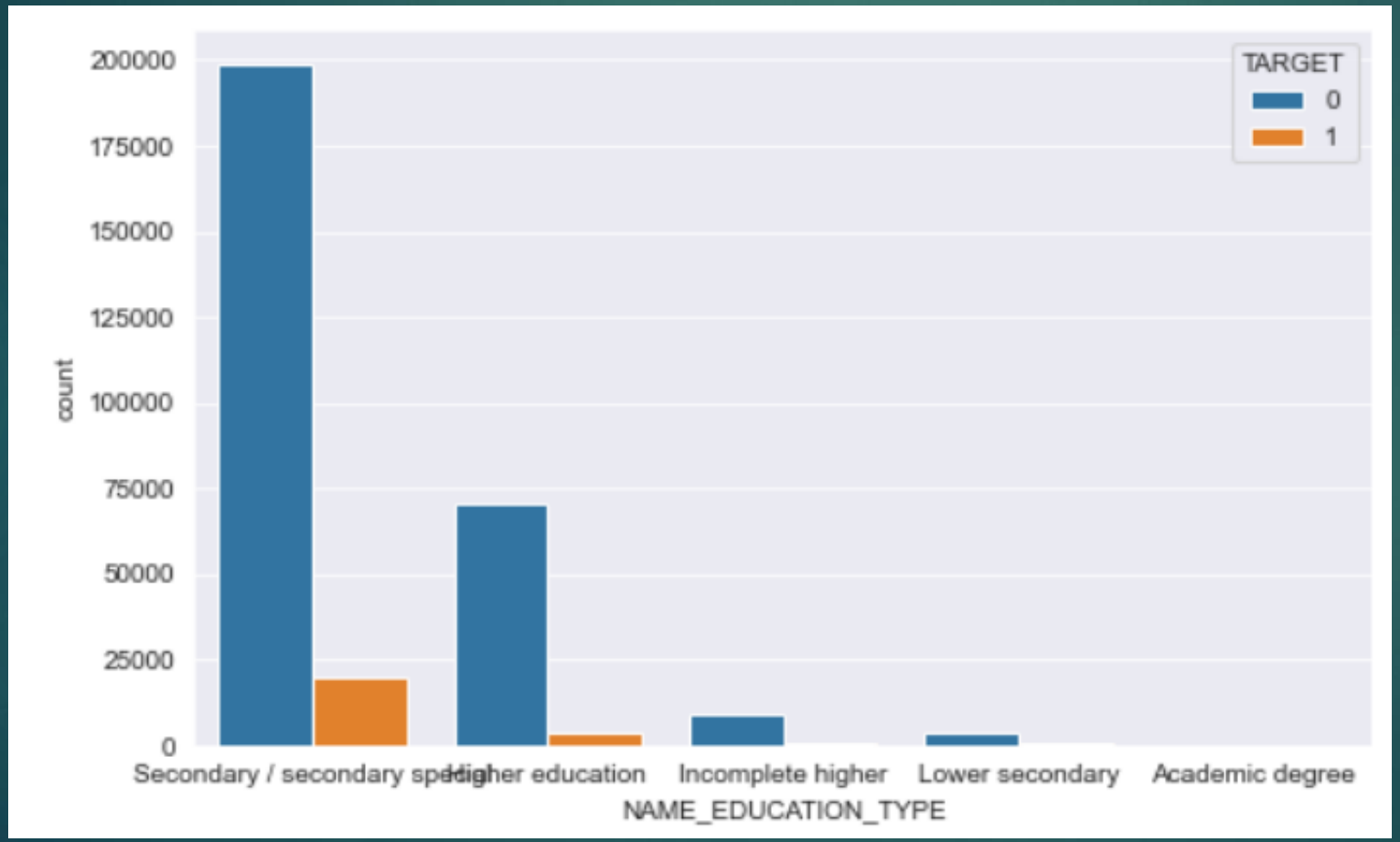
- Most of the clients barrower is laborers.
- And the least of the clients are IT and HR staff.



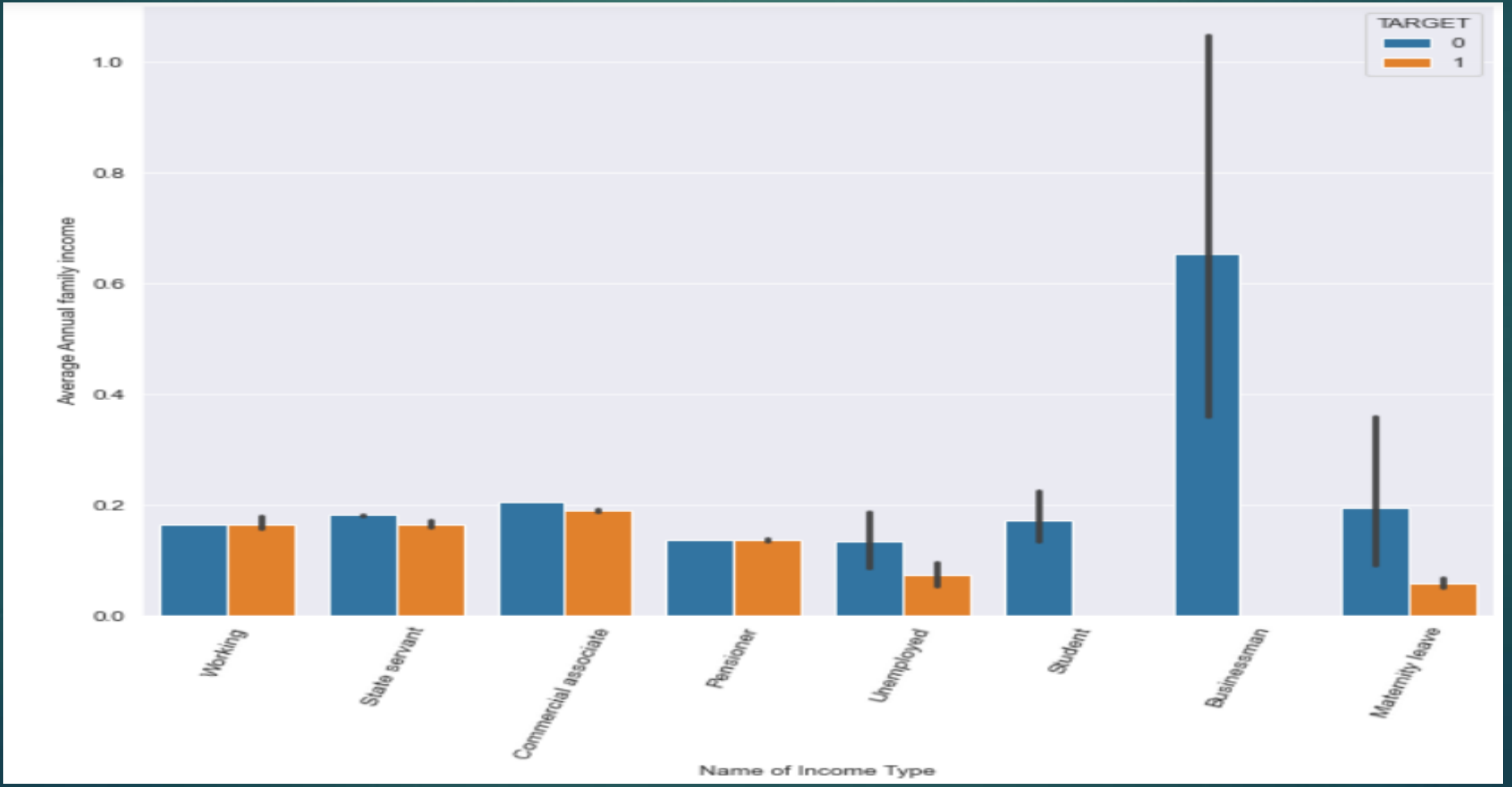
- Based on the annual family income Managers are economically stable and Cleaning staff are least stable



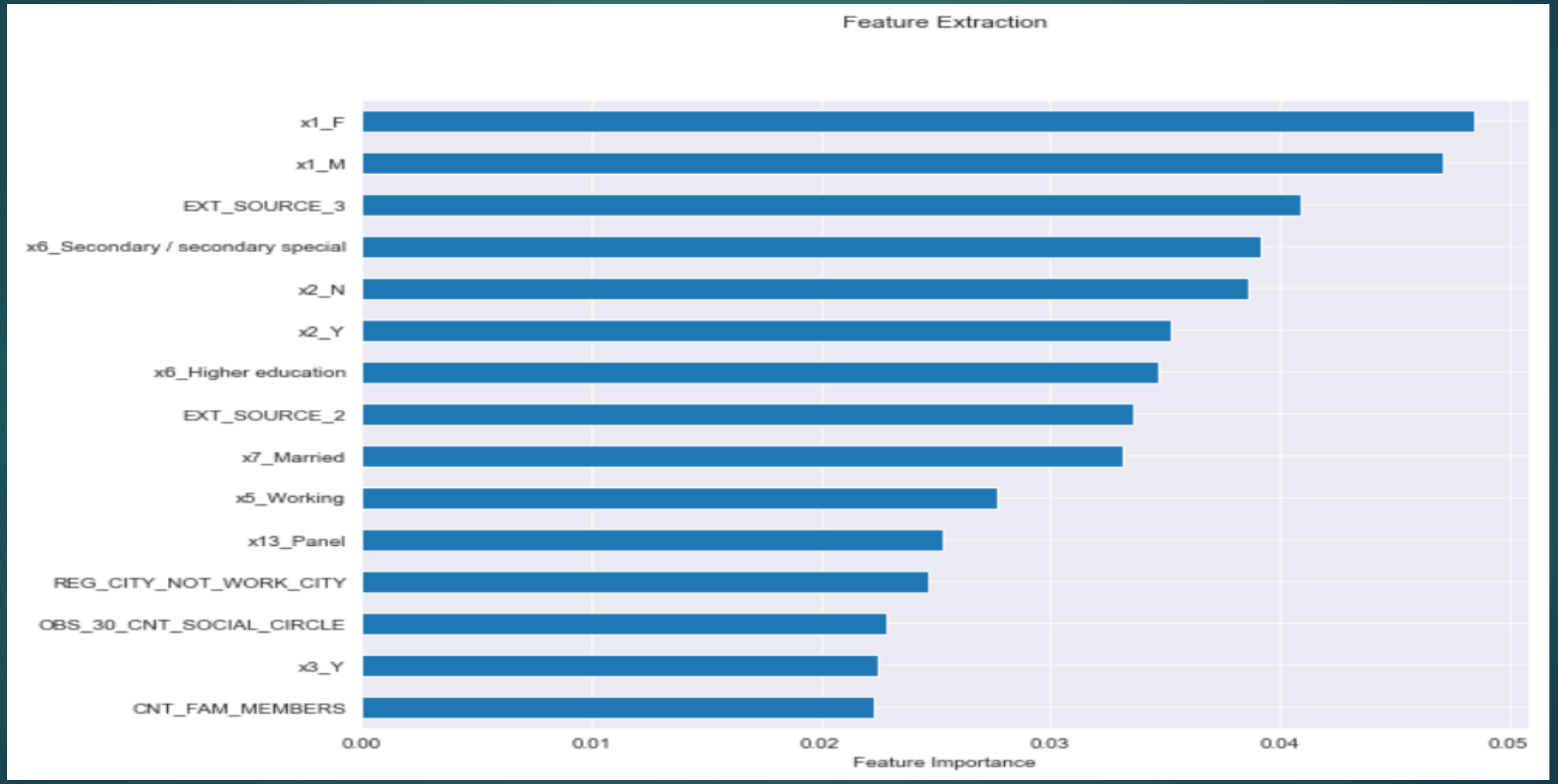
- Based on the education type there are more loans taken by those with secondary education



- Businessman's having the highest income amount and less income amount is unemployed. And Businessman are likely repay the loan.



- Feature Selection :



Further Improvement Area

- Here we used Missing values imputation by mean and mode.
- For this problem we can predict the null values, may be that gives the best result.
- Checking the outliers and Removing the outliers .
- Some more Exploratory data analysis for much more insight.
- For model building if use hyper parameter tuning we might get good accuracy.
- Using feature selection method getting which columns is most important to predict the values.
- For this project used only machine learning algorithms, we can use Deep Learning also.

How Helpful For Business

- Machine learning is a technology that helps businesses effectively gain insights from raw data .
- Machine learning specifically machine learning algorithm can be used to iteratively learn from a given data set, understand patterns, behaviors, etc.,
- Machine learning can help companies convert their data into value-adding insights.
- Humans cannot evaluate information and run many potential scenarios at the size and speed required to take the best course of action.
- ML will let you analyze the data related to past outcomes and interpret them. Therefore, based on the new and different data you will be able make better predictions of outcomes .
- For sanctioning loan may be take more time by human but ML Model it will take less time to predict

THANK YOU