Credit Card Approval Prediction

GROUP 4 Team Members:

Krupali Shah

Nikita Shetty

Siddhant Somani

Tejaskumar Patel

Tej Kurani

Business Problem

Credit gives one access to spend with "future" money. Credit risk is loss by borrower failing to make full and timely payments. Responsible users can benefit from cash backs, reward points, purchase safety, travel convenience and others. People tend to apply for credit cards to build their credit, and in turn take out a loan with a competitive interest rate. Overall, this project of Credit approval analysis helps us in determining these various factors which contribute to approval from the lenders.

Motivation and Setting

Credit Analysis is about measuring borrower's ability to meet its debt obligations. Credit is the amount of money that a lending institute provides to an individual or entity, and in return they are expected to pay in full or certain percentage under specific time. Credit analysis is the financial analysis performed by debt-issuing entities which make decisions on whether to provide credit or not by seeing past, present, and future trends. These corporation perform analysis on entities credit history, and if not one, then financial background helps them decide on. This outcome of application varies with the features having higher weightage and the models executed as well as differs from one provider to other even though they belong to same geographical location.

In credit, a 3- digit score evaluates the creditworthiness of any issuer or entity. It suggests how likely one is to repay the debt, their past relation with it and not suggest how rich or savings one possesses. This score is a quantitative method which uses statistical models to assess creditworthiness. Factors like debt-to-income ratio, personal and financial background like credit utilization, credit history, etc. also influence the approval. Such an evaluation is essential before granting the approval decision.

Data Description

We perform our prediction models on two datasets that are connected by "customer ID". Application record dataset contains customer's personal and financial information which we would use as independent variables for prediction. And Credit card record dataset consists of customer's status of previous credit history.

There is one target variable Credit card status of users where the overdue for more than 60 days will be target risk users. Rest columns may help us determine the factors that affect debt payment which in turn can help build credit approval prediction models.

- Dataset Characteristics: Univariate
- Attribute characteristics: Categorical, Integer, Real
- Associated Tasks: Logistic Regression, KNN, Random Forest
- Records – 438558 instances; 1million instances
- Columns – 18 variables; 3 variables

The 'Status' column represented the Good or Bad borrower status which was dependent on Monthly Balance column. As shown below, we categorize them into good debt or Bad debt:

- Here, if there was No Loan for that month or Loan paid off that month or borrower is 1-29 days past due is considered under good debt while others bad debts as risky customers.
- Later, using unstack from row to column helped in getting the count of Good and Bad debts for a given customer.

And so, we transformed these columns into a single Approval status column measuring:

If Good debt > Bad debt, then Approval status is set to 1

else if Bad debt >= Good debt, then Approval is set to 0

Application Record

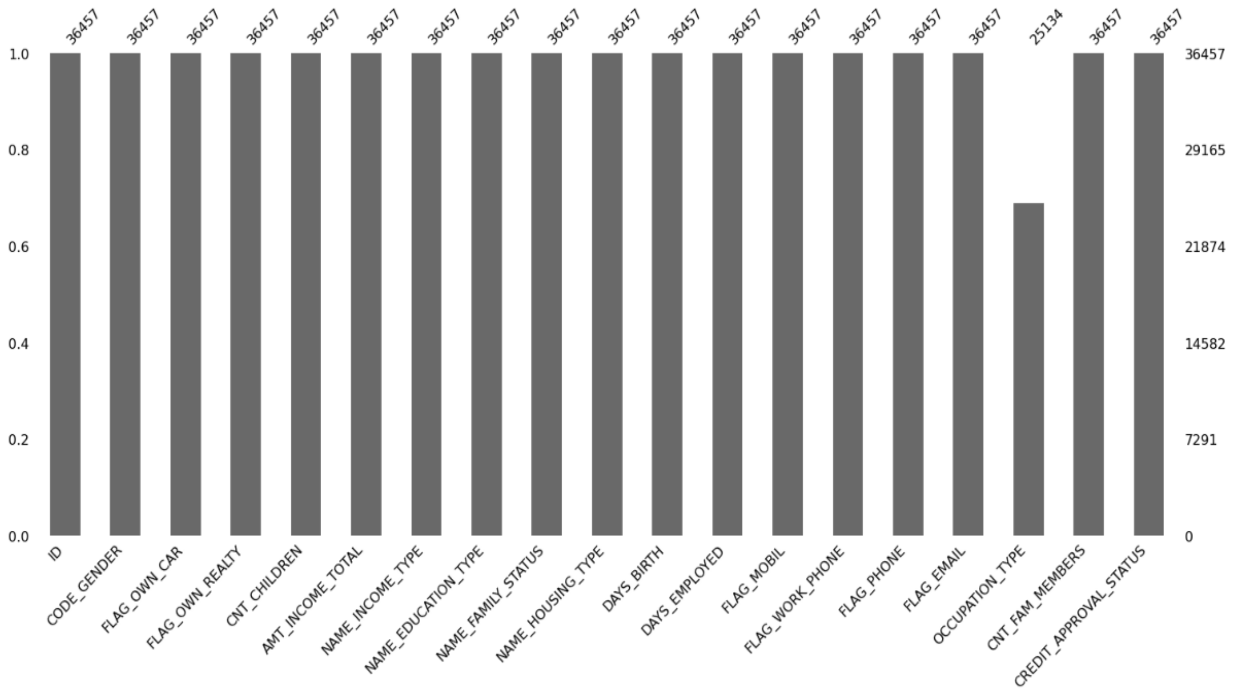
Feature Name	**Explanation**
ID	Client number
CODE_GENDER	Gender
FLAG_OWN_REALTY	Is there a property
CNT_CHILDREN	Number of children
AMT_INCOME_TOTAL	Annual income
NAME_INCOME_TYPE	Income category
NAME_EDUCATION_TYPE	Education level
NAME_FAMILY_STATUS	Marital status
NAME_HOUSING_TYPE	Way of living
DAYS_EMPLOYED	Start date of employment
FLAG_MOBIL	Is there a mobile phone
FLAG_WORK_PHONE	Is there a work phone
FLAG_PHONE	Is there a phone
FLAG_EMAIL	Is there an email
OCCUPATION_TYPE	Occupation
CNT_FAM_MEMBERS	Family size

Credit Record

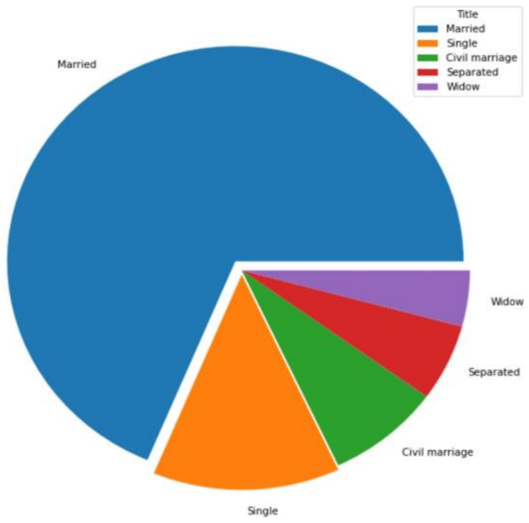
Feature Name	**Explanation**
ID	Client number
MONTHS_BALANCE	Record month
STATUS	Status

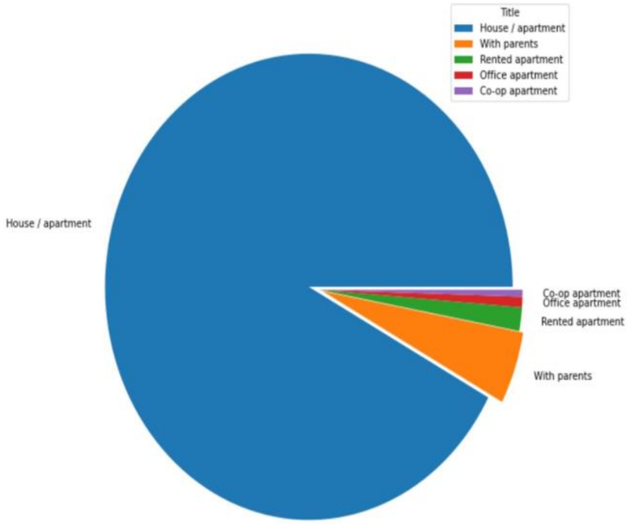
Feature Engineering

- Inner Joined both the tables based on customer ID, which reduced the size of dataset
- Performed Exploratory Data Analysis
- Dropped columns with No correlation
- Performed One hot encoding for columns like INCOME_TYPE, EDUCATION_TYPE, HOUSING_TYPE AND FAMILY_STATUS, and converted them into numerical data.
- Using timedelta() function, analyzed age and years of employment for each customer
- Visualized null values present across columns
- Replaced the null values of Occupation type to Others for better representation

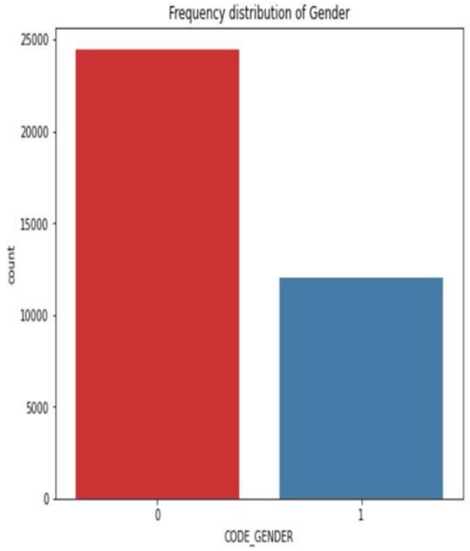


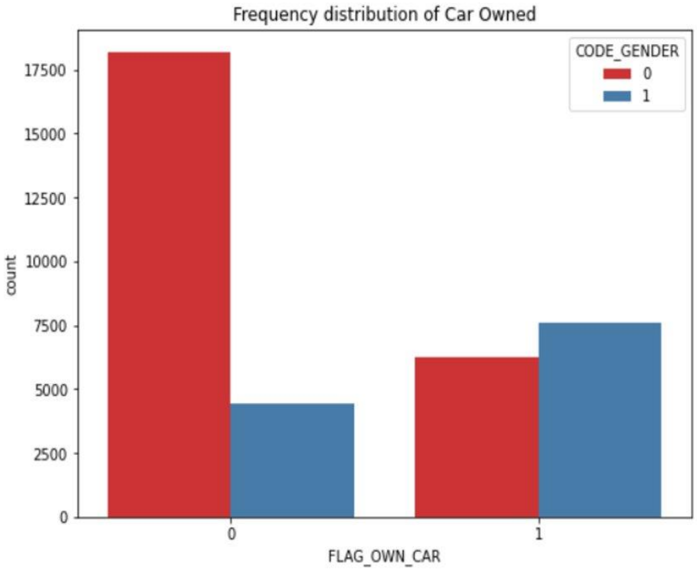
Below most of our data consists of married people followed by single and majority own house/apartment.



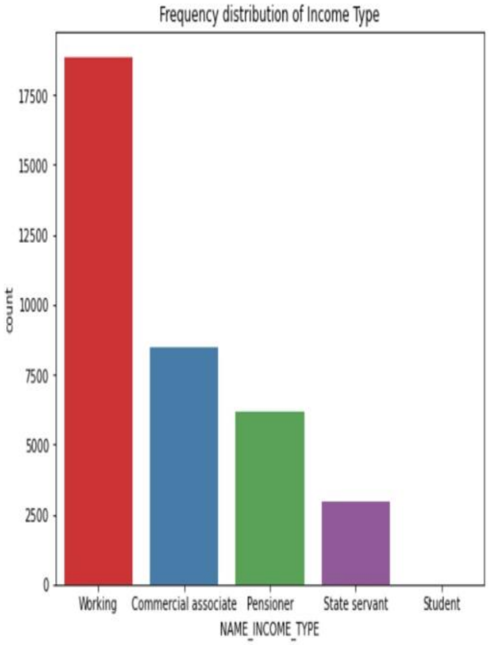


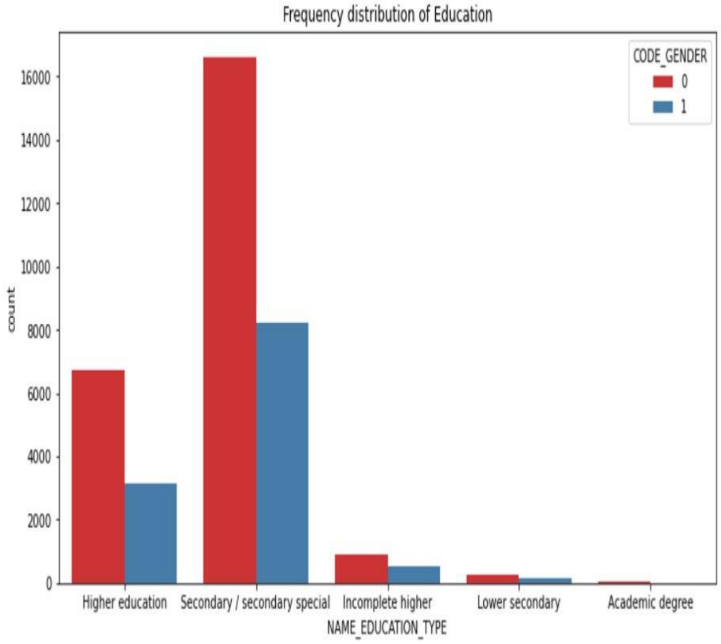
Majority of gender is men/women and based on gender distribution majority of men own car as compared to female.



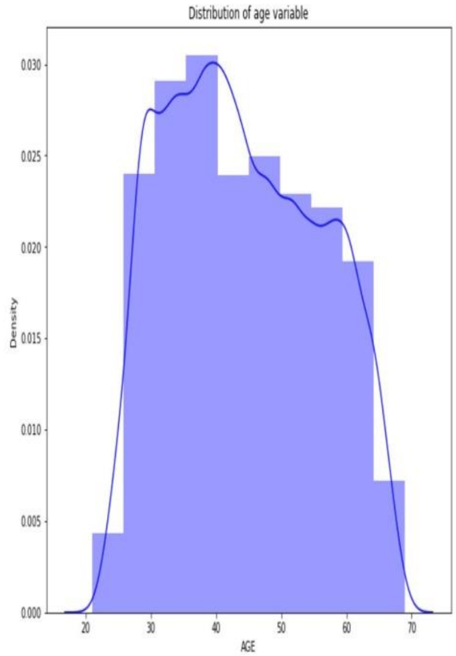


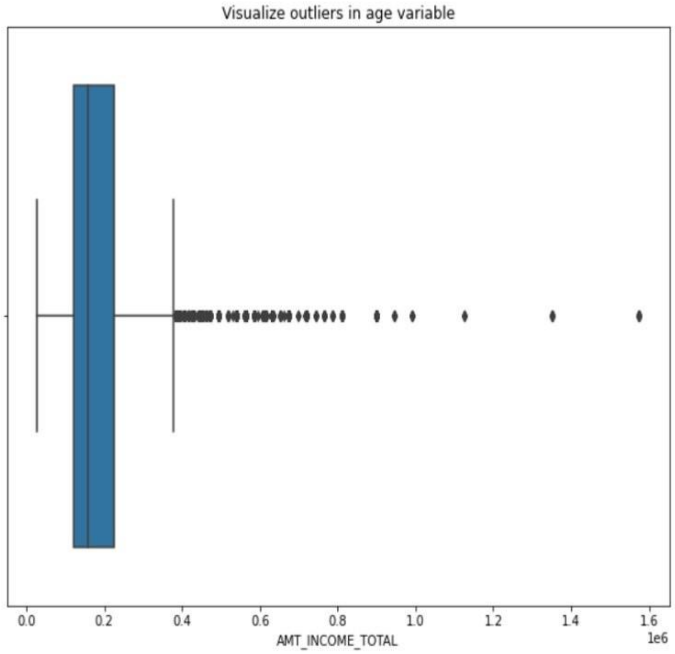
Majority of our dataset is working class & majority of people have studied till secondary education followed by higher education



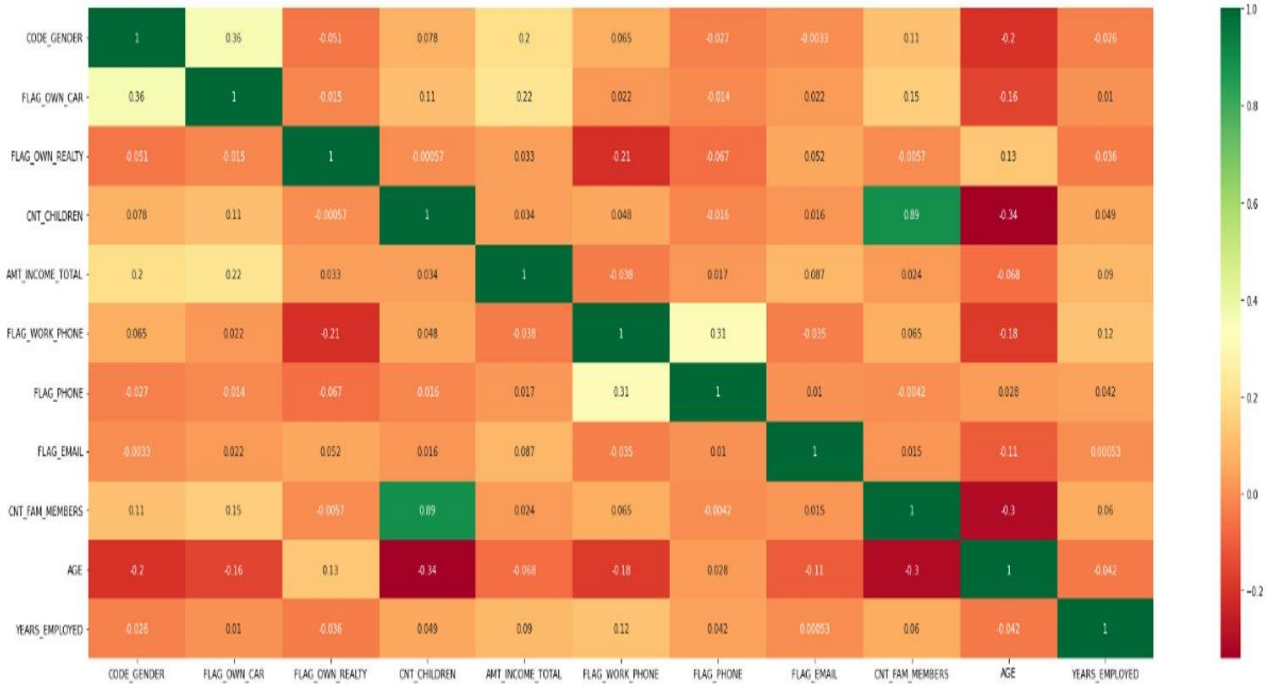


The age group is between 30-45 and we have also built a box plot to understand age variables with respect to total income.





Correlation matrix displays the heatmap of correlation amongst each of the other variables affecting the approval prediction.

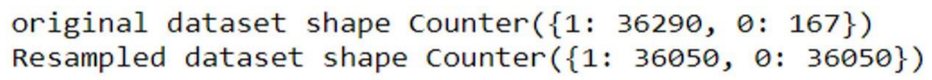


Feature Scaling performed in below columns:

FAMILY_STATUS				EDUCATION_TYPE	
Married	1			Secondary / secondary special	1
Single / not married	1			Secondary	1
Single	2			Higher education	2
Civil marriage	3			Incomplete higher	3
Separated	4			Lower secondary	4
Widow	5			Academic degree	5

		HOUSING_TYPE			
		House / apartment	1		
		With parents	2		
		Municipal apartment	3		
		Rented apartment	4		
		Office apartment	5		
		Co-op apartment	6		

Dealing with Unbalanced Data:



```
original dataset shape Counter({1: 36290, 0: 167})
Resampled dataset shape Counter({1: 36050, 0: 36050})
```

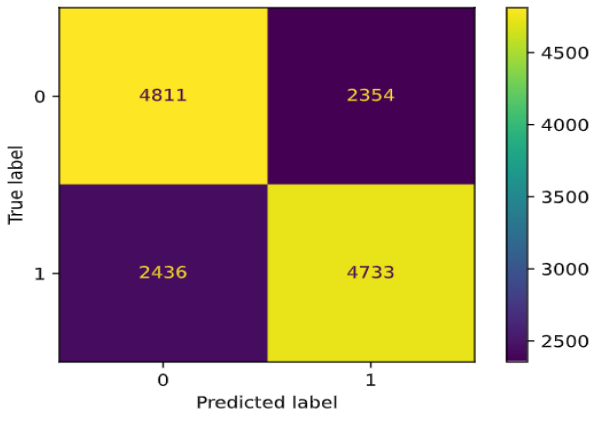
- We have 36K data for credit card approval whereas we have only 167 rows for rejection.
- Hence Up sampling the data to remove biasness.

MODELS

1. **Logistic Regression:**

   Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. By applying Logistic Regression on our dataset, we got an accuracy score of 66.5%.

   Confusion Matrix:



True Negative: 4811

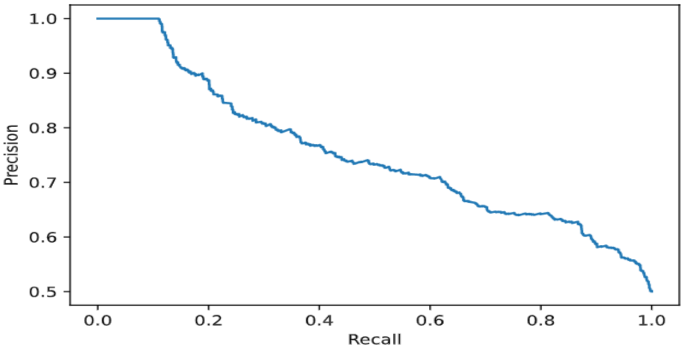
False Positive: 2354

False Negative: 2436

True Positive: 4733

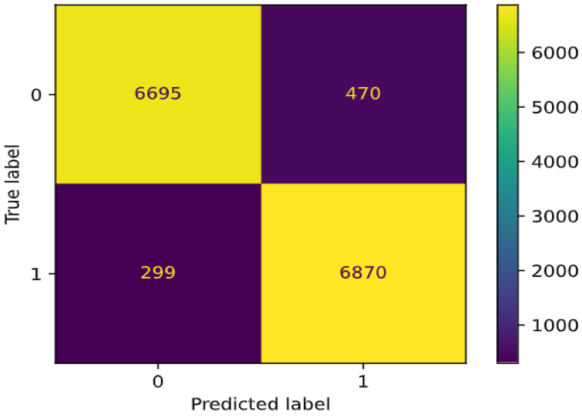
Precision and Recall Curve:

From Precision Recall curve we see that the model has High Recall and constant or decreasing Precision. It is because of the higher False Positive in our model. Therefore, the F1 Score will be better unbiased performance measure, which is 0.66.



2. **K Nearest Neighbors:**

   The k-nearest neighbors (KNN) algorithm is a data classification method for estimating the likelihood that a data point will become a member of one group, or another based on what group the data points nearest to it belong to. Using KNN on our dataset we got the accuracy score of 94.2%. Using Grid Search CV, we got the best K parameter = 3. By using K = 3 the accuracy score for our model is 94.6%.



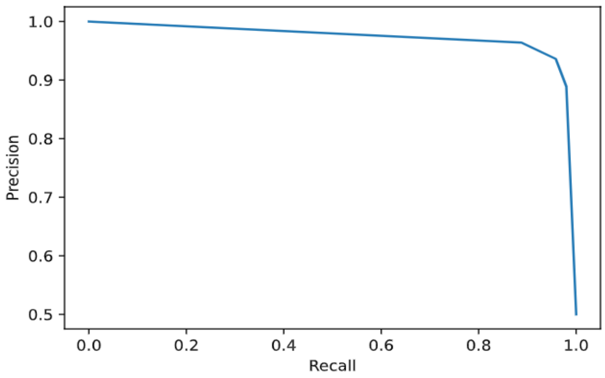
True Negative: 6695

False Positive: 470

False Negative: 299

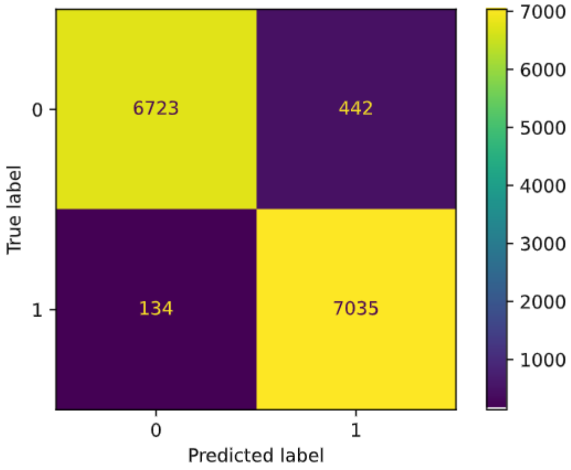
True Positive: 6870

Using this, the False Positive rate was reduced. Hence, there is little improvement in the accuracy and precision score. From the above figure, we can say that both precision and recall score are high.



3. **Random Forest:**

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification problem. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. The accuracy score for this model is 95.9%. By applying Grid Search CV, we got best n-estimator = 150.



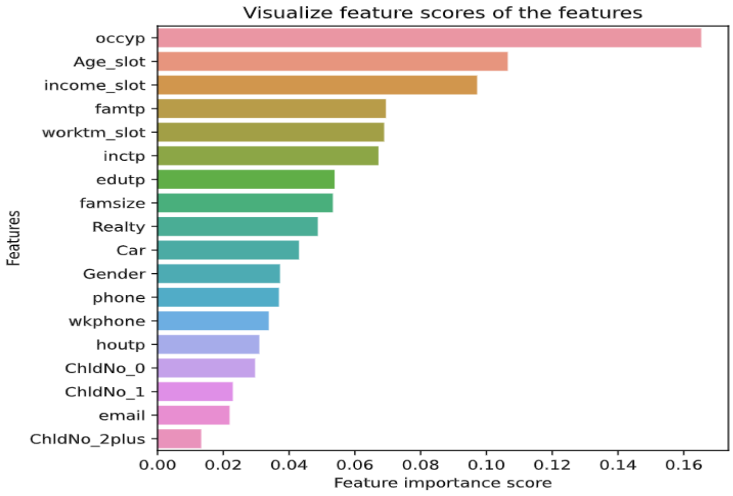
True Negative: 6723

False Positive: 442

False Negative: 134

True Positive: 7035

From the figure, we observe that variable Occupation types have the highest Feature score in Random Forest model.



Performance Evaluation and Model Comparison

Performance Metric	Logistic Regression	K Nearest Neighbor (K=3)	Random Forest
Accuracy	0.665	0.946	0.959
Precision	0.667	0.935	0.94
Recall	0.66	0.958	0.981
F1	0.664	0.946	0.96
R-Square	0.665	0.946	0.959
Adjusted R-Square	0.665	0.946	0.959
Mean Absolute Error	0.334	0.053	0.040

Conclusion

As per the above model evaluations we can conclude that KNN and Random Forest are the best fitted models for our dataset. Hyperparameter tuning helped to find the best K value and N-Estimator for our models, to improve the performance metrics. By training more data we can improve the Logistic Regression model in future.