Problem approach

The model to predict whether a customer would be approved a credit card or not, was built as detailed below.

EDA and Data Preparation

Missing value imputation

From the analysis of data I found that the missing values in data were presented as '?' in each of the following columns.

1. Male
2. Age
3. Married
4. BankCustomer
5. EducationLevel
6. Ethnicity
7. ZipCode

- Since 'Age' is a continuous variable without any significant outlier, the missing values in age were replaced by its mean.
- For the remaining categorical features the missing values were replaced by the feature's mode.
- Ensured that the distribution of the features, before and after imputation remained the same.

Encoding of categorical variables

Dichotomous Categorical Variables

For Dichotomous categorical variables like PriorDefault, Employed, DriversLicense I went with binary encoding mechanism of replacing 'True' as 1 and 'False' as 0.

Ordinal Categorical Variables

Education level is the only ordinal categorical variable, however it had to be considered as a nominal variable as no information was provided on the ordering of categories as to which category represented the highest or lowest education level.

Nominal Categorical Variables

- For nominal categorical variables like Male, Married, BankCustomer, EducationLevel, Citizen, I went with one hot encoding for linear model like logistic regression.
- For non-linear models like random forest and XGBoost, since it would suffer from one hot encoding I went with mean encoding of categories.

Outlier Treatment

For a couple of categorical features removed categories that had very low frequency of occurrence and didn't significantly distinguish the target. For e.g. The variable 'Citizen' had one such category called 'p'.

Model Building

Since the distribution of target was not skewed, didn't apply any under sampling or oversampling techniques prior to model building.

Logistic Regression

Feature Encoding

The nominal features like Male, Married, BankCustomer, EducationLevel, Citizen were one hot encoded.

Feature Selection

- Built two models. The first model had almost all the features. For the second model, applied chi square test and selected only the variables having a p-value less than 0.01.
- Since the number of continuous features were already less, didn't apply any feature selection technique for continuous variables to avoid loss of information.

Model fitting, cross validation and prediction

- Took 80% of the data for training the model and 20% of the data was used as a held out sample for evaluating the model performance on unseen data.
- Ensured that the target's categories were represented in equal ratio in both train and held out sample.
- The evaluation criteria used to validate the model was AUC.

Random forest

Feature Encoding

The nominal features like Male, Married, BankCustomer, EducationLevel, Citizen were mean encoded as tree based models would suffer from encoding mechanisms like one hot encoding.

Hyper parameter tuning, model fitting, cross validation and prediction

- Similar to logistic regression took 80% of the data for training the model and 20% of the data was used as a held out sample for evaluating the model performance on unseen data.
- Since random forest has inbuilt mechanism of validating model performance on Out Of Bag data, tuned hyper parameters like number of trees in forest, depth of each tree, minimum number of samples required to split an internal node, minimum number of samples required to be at leaf node and number of features to be considered while determining the best split and validated them on both train and held out sample.
- Finally the model was fitted and obtained an AUC of 85% in train data, AUC of 89% in held out sample and f1-score of 93.7% in the actual model submitted. P.S: Random forest performed well, comparatively in both held out sample as well as in leader board.

Other approaches tried

The following approaches were also tried, but they didn't significantly increase the Area Under Curve value.

Feature Generation

- Generated interaction terms among categorical variables.
- Binning of continuous variables like age.

Model Building

- Extreme Gradient Boosting.
- Ensemble – Maximum Vote Classifier (Base classifiers used – Logistic Regression, XGBoost, Random Forest), all the classifiers were given equal weight in voting process.