JOB-A-THON - May 2021

Problem Statement

Data contains details of customers including basic demographic data like *Age, Gender, Region* etc. along with data on their association with the bank like *Vintage, Account Balance, Is Active* etc. The objective is to Predict probability of existing customers being interested in buying a credit card.

Imputing Missing Values

There are more than 10% missing values in the *Credit_Product* column in both train and test data. This is a significantly high number and hence we cannot just ignore these rows. The following are the different methods tried to impute these values:

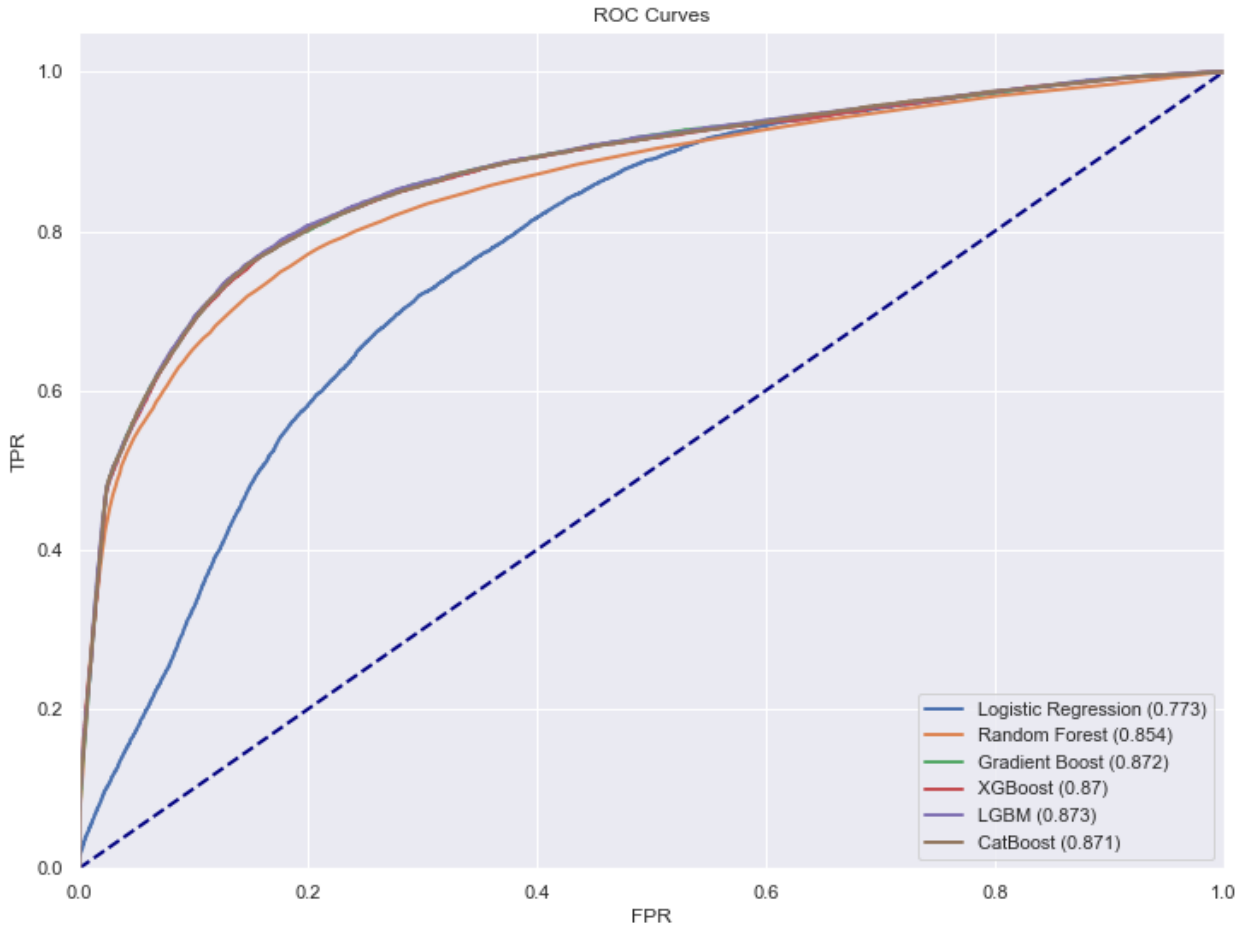
1. Replace all missing values with *"No"* which is the more frequent value.
2. Replace all missing values with *"Yes"* which is statistically more similar to the missing values in terms of the box plot against *Age* and *Vintage* variables. This gave increased the ROC AUC scores by about 10% as compared to method 1.
3. I attempted to use KNNImputer to impute missing values based on the nearest neighbors but this worsened the scores given by the model as compared to method 2
4. **I defined the missing values as a separate third category. This gave the best results and hence, I have used this for my final solution**

Data Preparation

- **All categorical variables are label encoded**
- **All numerical variables are log transformed (gave better results) and then standardized.**
- Feature selection using Recursive Feature Elimination on a base LogisticRegression method was attempted which selected 5 out of the 9 features but this worsened the final results
- Oversampling using SMOTE also did not help

Classification Models

Various Classification algorithms were fit and evaluated on train data. The following graph shows a comparison of the ROC curves:



Conclusion

LightGBM Classifier gave the best scores and hence was chosen as the final model.