Project Title : Credit Card Default Prediction

Technologies : Machine Learning

Domain : Banking

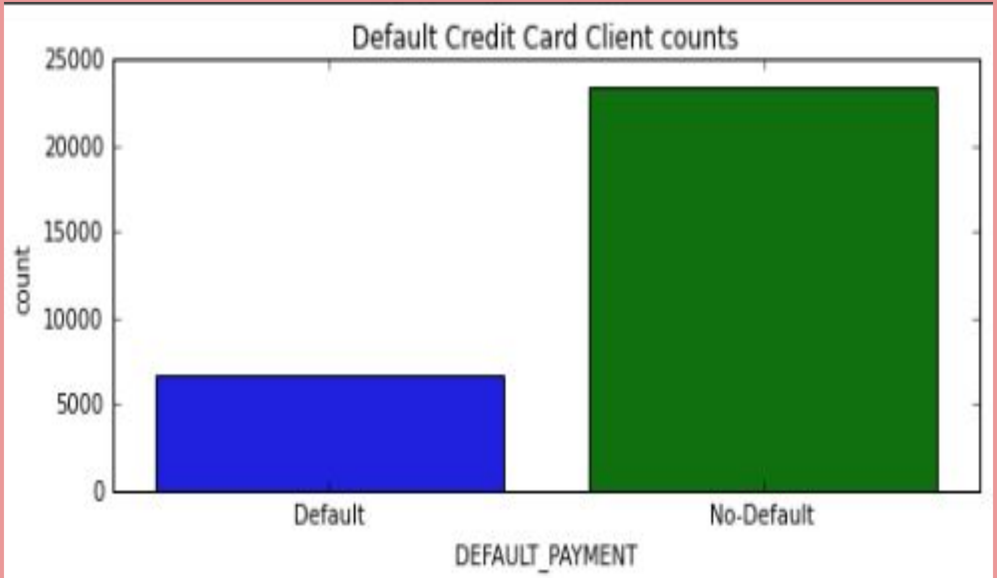
By,
SIDHARTH

PROBLEM STATEMENT

- The goal is to predict the probability of credit default based on credit card owner's characteristics and payment history.
- Owner's Characteristics :
  - Sex, Education, Marriage, Age
- Payment history :
  - Repayment status, Amount of bill statement, Amount of previous payment for 5 months.

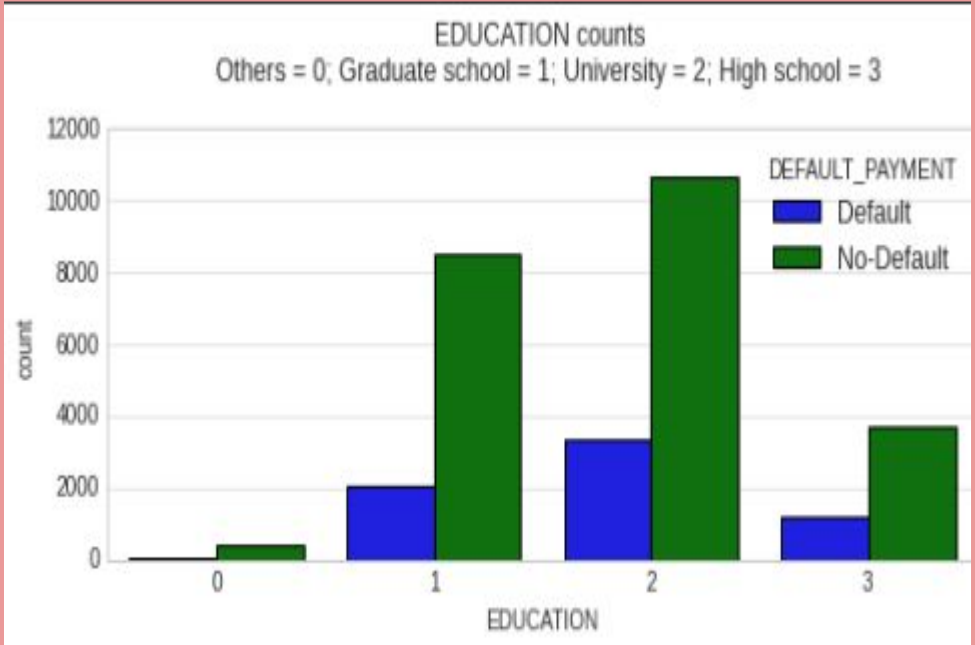
DATA EXPLORATION - Target variable

- ★ The dataset is an IMBALANCED dataset.
- ★ So we need to balance the dataset first.
- ★ I have used SMOTE oversampling technique to balance them.



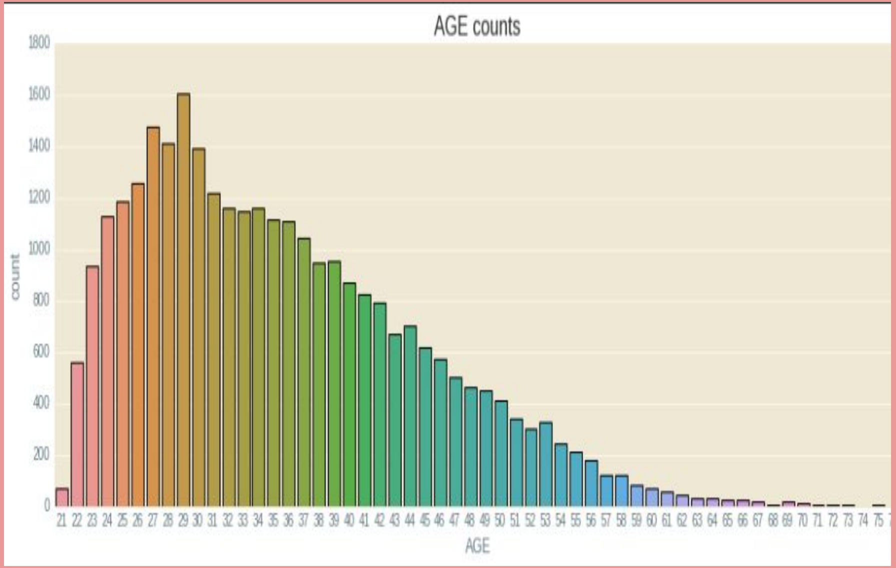
DATA EXPLORATION - Features (categorical)

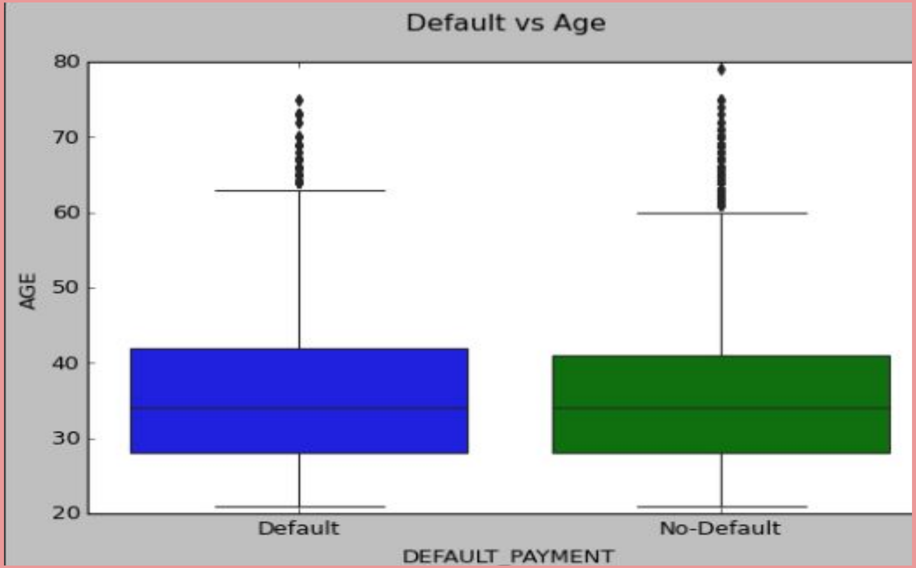
- **University & High school** graduated persons use more credit card and also default more.
- **Singles** use credit cards more than that of **Married.**


EDUCATION counts
Others = 0; Graduate school = 1; University = 2; High school = 3

DATA EXPLORATION - Features (continuous)

- ★ **Age** group between **30 to 40** use Credits more.
- ★ Hence default is also more in this age group.





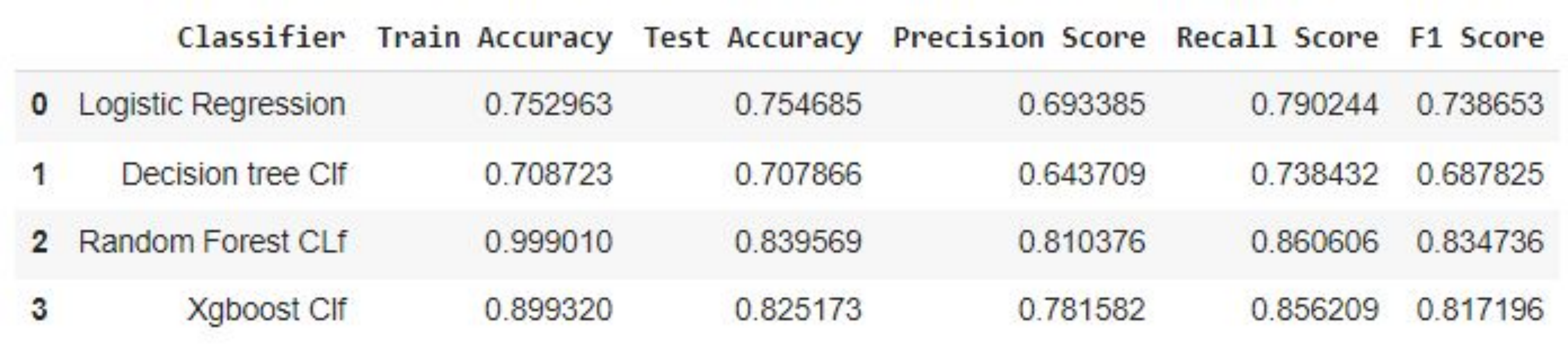
DATA CLEANING & FEATURE ENGINEERING

- Renaming the columns for better understandings (in Payments features)
- Consolidating the ambiguous values (in Education, Marriage features)
- One hot encoding (Education, Marriage, Pay)
- Label encoding (Sex)
- Dropping unnecessary features

MODELLING

1. Logistic Regression
2. Decision tree
3. Random Forest
4. XGBoost

MODEL TESTING

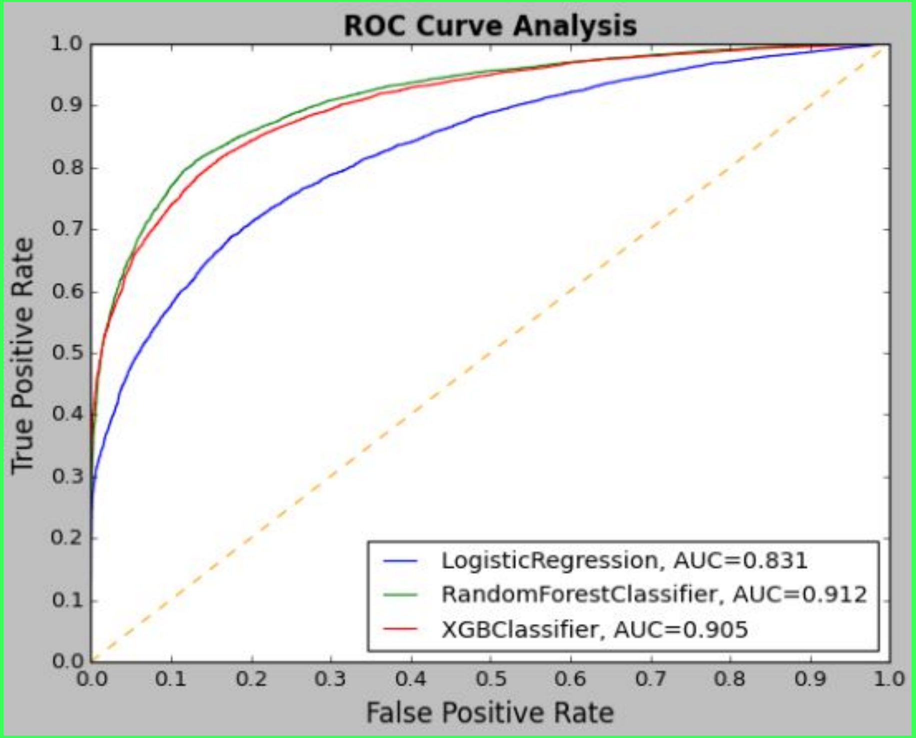


	Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
0	Logistic Regression	0.752963	0.754685	0.693385	0.790244	0.738653
1	Decision tree Clf	0.708723	0.707866	0.643709	0.738432	0.687825
2	Random Forest CLf	0.999010	0.839569	0.810376	0.860606	0.834736
3	Xgboost Clf	0.899320	0.825173	0.781582	0.856209	0.817196

- **Random forest** model, even though it is overfitting, gives the **highest** F1-score (mean of Precision & Recall).
- **Decision tree** model performs **poorly** on this dataset.

MODEL TESTING (AUC_ROC curve)

- ➔ Random Forest gives the best score of **91%** followed by XGBoost (**90%**).
- ➔ Hence we can conclude that ***Random Forest is the best ML model*** for this dataset.



IMPROVEMENTS

- We can further increase the accuracy of the model using:
  - More Quality data
  - Better fine-tuning of hyperparameters
- Thus, Defaulters can be predicted in advance and help company reduce the losses.

THANKYOU