Abdullah Al Huwaishel
Ali Altamimi

# Loan Defaulter

## Abstract

The purpose of this research was to employ classification models to predict if a client is going to pay their loan on time or not in order to help banks. We worked with the information supplied by Gaurav Dutta in Kaggle. To get the best model feasible for this data, numerous methods have to be used. We used seaborn to visualize the outcome after acquiring the model.

## Data

There are 300,000 observations in the dataset, each with 24 characteristics, ten of which are categorical. Name of contract type, gender, If he/she owns a car,If he/she owns realty , and If he/she having children, Income of each client ,credit ,annuity, Income type, Education type, family status, Housing type, Birth,Id publish.

## Algorithms

1. Apply smote and standard scalar to have better data to work with.
2. Fill NA values with the mean when it’s numerical and mode when it’s categorical.
3. Converting categorical features to binary dummy variables
4. Combining particular dummies and ranges of numeric features to highlight strong signals and illogical values for the target during EDA

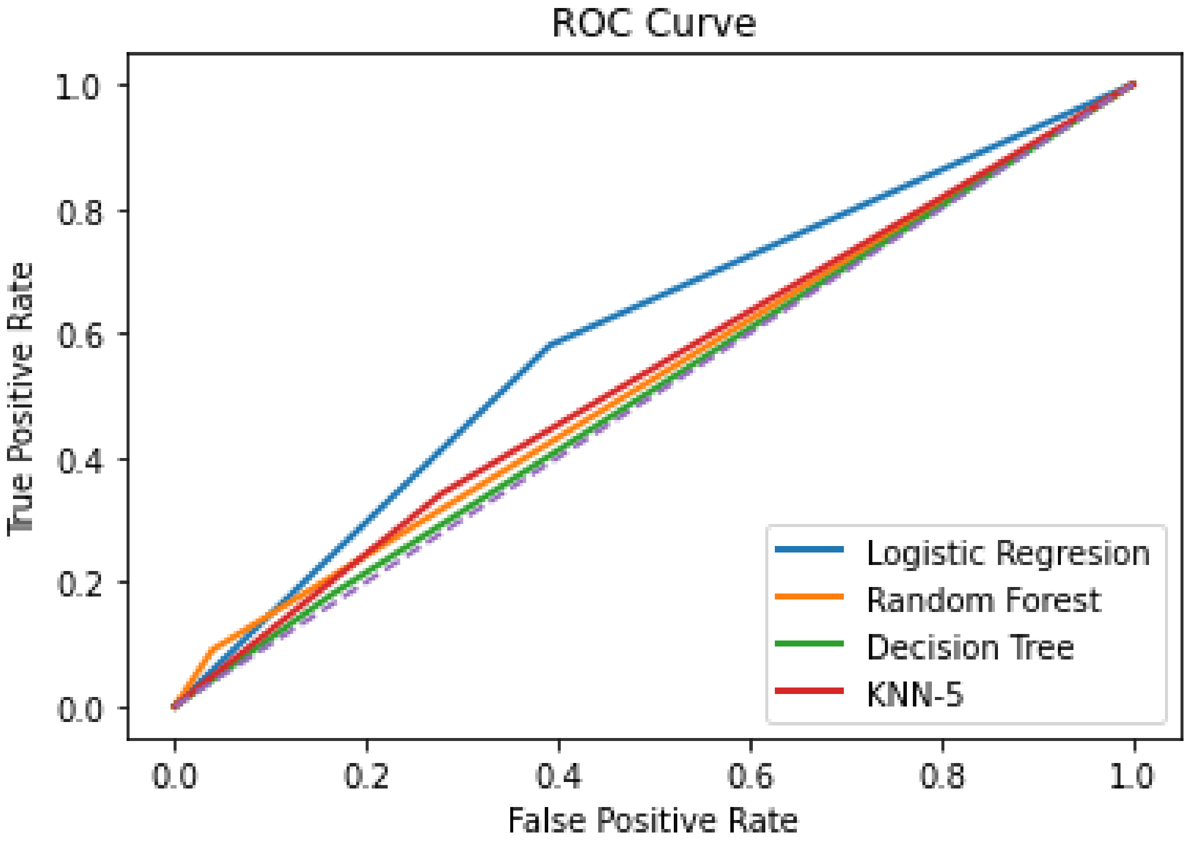
## Models

*Model Evaluation and Selection*

The dataset was divided into 80/20 for training and testing. We utilized KNN, Logistic Regression, Decision Tree and Random Forest models to train them using the splitted data. All of the scores shown below were obtained using the test split, and the actual result was compared to the anticipated result.

|  | precision | recall | f1-score |
|---|---|---|---|
| RandomForest | 0.16 | 0.10 | 0.13 |
| Logistic Regression | 0.12 | 0.63 | 0.20 |
| Decision Tree | 0.10 | 0.16 | 0.12 |
| KNeighbors | 0.10 | 0.36 | 0.16 |

We can conclude the best model for our dataset and business need is Logistic regression since the recall is high. Moreover, the Area under the curve was 0.59 for logistic regression which is higher than the other results that we got.



## Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling

- Matplotlib, Seaborn and Plotly for plotting