AFA Company

Abstract

The goal of this project is to use classification models to predict customers who are going to churn from a Telecom company called AFA. Predicting future churn rate will assist the AFA company to identify and improve areas where customer service is lacking. In addition to gaining a better understanding of future expected revenue.

Design

The project uses data of AFA company to build the classification models contains several features such as customer ID, gender, monthly charges, payment method...etc. We used a Random Forest, Decision Tree, KNN, Ada Boost, Naive Bayes, Logistic Regression and SVM to predict the customer churn.

Data

The dataset contains information around 8000 customers of a telecom company. There are numeric columns and categorical columns. There are several types of contracts and payment methods.

Algorithms

Feature Engineering

1-Separating numeric and categorical columns.
2-Converting categorical features to binary dummy variables.
3-Combining dummies and ranges of numeric features to gain clear insights during EDA.
4-Calculating correlation between features and Churn label.
5-Creating several data visualizations to build an understanding of the properties of data features.

Model Evaluation and Selection

The dataset was split into 80/20 train vs. test, and all scores reported below were calculated with 10-fold cross validation:

- Accuracy - Test: 80.37 %
- Accuracy - Train: 80.3 %
- Precision: 61.36
- Recall: 48.21
- F1 Score: 54.0

Validation set:

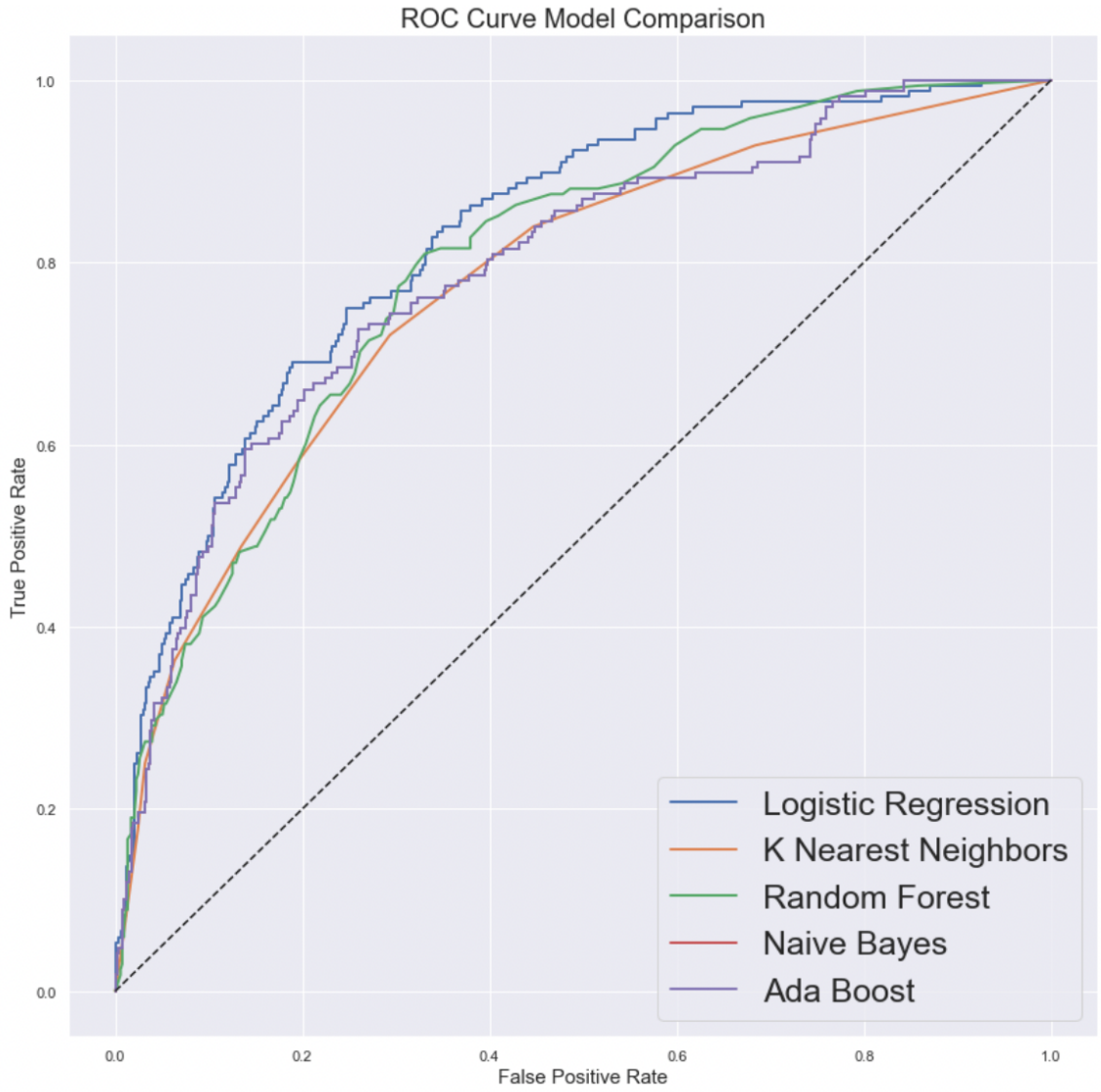
- Accuracy: 0.81
- F1: 0.74 macro avg, 0.80 weighted avg.
- Precision: 0.77 macro avg, 0.80 weighted avg.
- Recall: 0.72 macro avg, 0.81 weighted avg.

Tools

Numpy and Pandas for data manipulation
Scikit-learn for modeling
Matplotlib and Seaborn for plotting

Communication

	model	Best_score	Best_parameters
0	svm	0.745778	{'kernel': 'rbf', 'C': 12}
1	random-forest	0.790933	{'n_estimators': 15, 'max_depth': 13}
2	logisticregression	0.802489	{'C': 6}
3	decision_tree	0.784889	{'max_depth': 5}


ROC Curve Model Comparison
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
K Nearest Neighbors
Random Forest
Naive Bayes
Ada Boost