Customer Relationship Management

-BY Team Noobs
Aparna Wani
Divyanshu Torlikonda
Ravindra Manne
Vaishnavi M R

Problem Statement

A French Telecom company Orange wants to predict the propensity of customers to switch providers (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling). They want to do this from some data which they have collected and want to use this to build a model that can predict the customers' churn, appetency, and up-selling probability.

Approach

Step 1

Preprocessing and Cleaning

Step 2

Exploratory Data Analysis

Step 3

Modeling

Step 4

Analysis

Step 5

Interpretation and Conclusion

Preprocessing and Cleaning

Before Preprocessing:

- Total Features:230, out of which 190 are Numerical and 40 are Categorical.
- 156 features has >70% null values.

After Preprocessing:

- Dropped features with>70% null values.
- Total Features:75, out of which 32 are Numerical and 40 are Categorical.
- For categorical features with missing values less than 30% are imputed with mode.
- For numerical features with missing values less than 30% are imputed with median
- Filtered out the categorical columns with more than 10 unique values and one hot encoded the remaining categorical columns.
- Performed standardization using Normalization.

Exploratory Data Analysis

- Using countplots, we deduce that the response variable is highly imbalanced.
- Univariate data analysis:

  1)Independent analysis of the numerical data and categorical data.

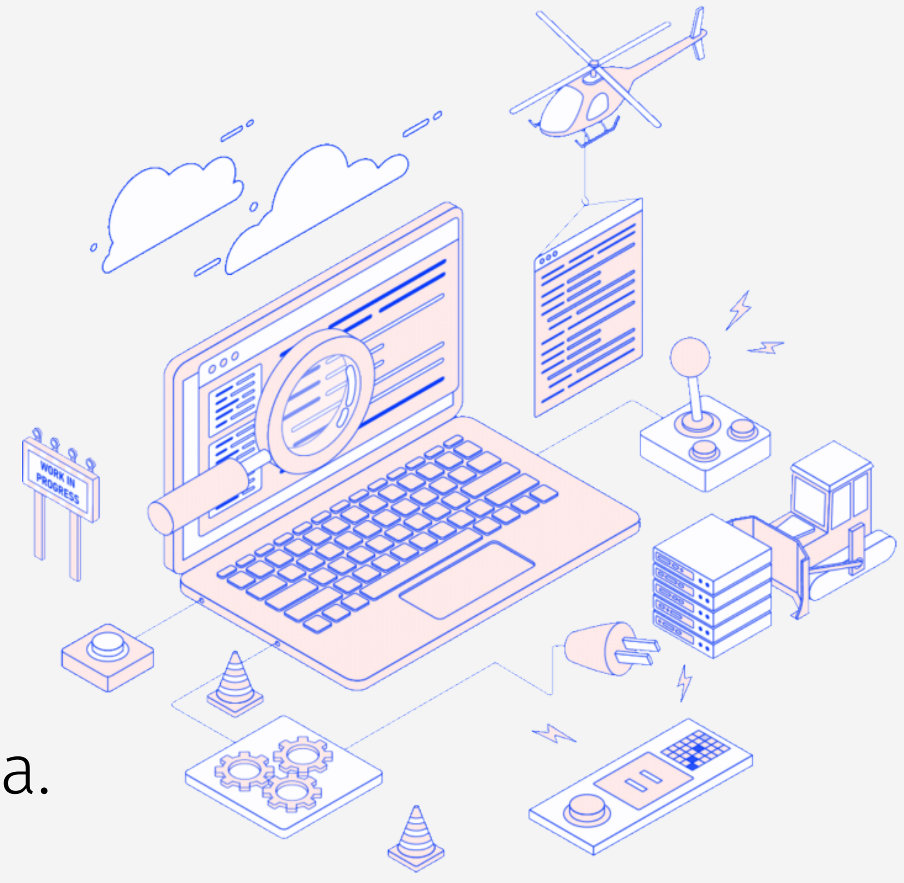
  2)Distribution of the numerical data shows that these variables are highly skewed.

  3)Determining the presence of many outliers, thereby determining which scaling metric to use.
- Standardscaler() does not help us to eliminate the skewness or variance of the datasets. So, we used Normalization
- Multivariate data analysis :

  1)Checking for variables that are highly correlated and removing some of the to reduce multi-collinearity.
- Categorical data analysis;

  1)The categorical predictors are highly imbalanced as well.
- After EDA there are 77 columns a total of which 10 are categorical and 67 are numerical.

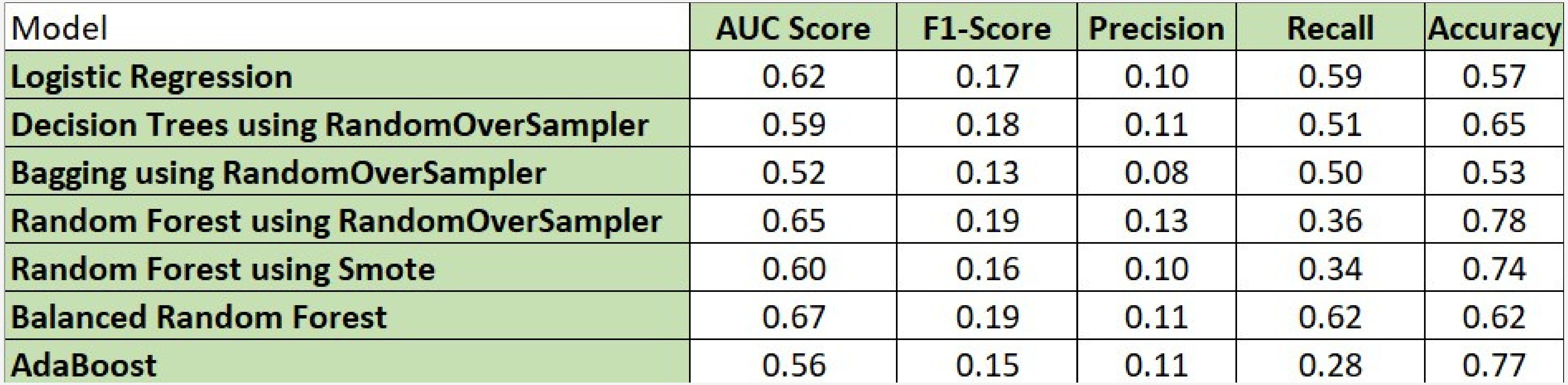
Modelling



- Split the dataset using ***train_test_split().***
- Upsampled the data using ***RandomOverSampler()*** to handle the imbalanced data.
- Use ***GridSearchCV*** to get the best estimator.
- Fit the best estimators on upsampled trained data for the following models:

  1.Logistic Regression 2.Decision Trees 3.Bagging 4.Random Forest 5.Random Forest using Smote 6.Balanced Random Forest 7.AdaBoost
- Predicted on the test data and built ***confusion matrices*** and calculated the ***classification metrics*** for each model.
- Plotted ROC-AUC curves for each model.
- Compared these metrics to get an idea of which model gives best results for our dataset.

Results



Model	AUC Score	F1-Score	Precision	Recall	Accuracy
Logistic Regression	0.62	0.17	0.10	0.59	0.57
Decision Trees using RandomOverSampler	0.59	0.18	0.11	0.51	0.65
Bagging using RandomOverSampler	0.52	0.13	0.08	0.50	0.53
Random Forest using RandomOverSampler	0.65	0.19	0.13	0.36	0.78
Random Forest using Smote	0.60	0.16	0.10	0.34	0.74
Balanced Random Forest	0.67	0.19	0.11	0.62	0.62
AdaBoost	0.56	0.15	0.11	0.28	0.77

Conclusion

- The data is highly imbalanced with lots of missing values and outliers.
- Due to the lack of documentation on the data, couldn't relate to the data set from the domain perspective.
- Observed Multi-Collinearity in the data.
- Balanced Random Forest model performs best on this data.
- As expected,Tree-based Models performed well compared to Logistic Regression.
- Due to constraints on computational power couldn't explore the power of GridSearchCV.
- There is a scope for improvement in the performance of the model.

Future Scope

- Work on improving the model performance by -
  1. Treating the outliers
  2. Identifying the multi-collinearity between the categorical variables
  3. Exploring GridSearchCV with full potential on highly configured VMs
  4. Explore feature selection methods Information Gain and Permutation Importance
- Extend the model to predict Appetency and Upselling.
- Learning to deal with large datasets by experimenting with them.

Thank You