Customer Churn Detection

Utpalraj Kemprai(DS1),Aryan Joshi(CS1)

March 2024

1 Introduction

The customer churn dataset contains customer interactions with an online retail store. Our goal was to build two classifiers to predict churn using AdaBoost and Random Forest.

2 Exploratory Data Analysis

We carried out our analysis in the following manner:

1. We first checked the datatypes of all columns in the data.
2. We checked for missing values and found there were no missing values in the data
3. We checked for outliers by standardising all the numeric columns and then plotting boxplot diagrams for them. We found no outliers (all values lied inside outlier fences)
4. We plotted the correlation Heat map for numeric columns and there was very little correlation among the numeric features in the data.
5. We checked if the data is imbalanced. Our dataset had 526 churn labels out of 1000. So the data was not imbalanced.

3 Data Preprocessing

- We encode the non-numeric and non-boolean columns using One Hot Encoding.
- We splitted the data into train and test data (80% for Training and 20% for testing). We dropped the CustomerID in the training data as it was clearly independent of Churning.

4 Training the model for classification

4.1 AdaBoost Classifier

4.1.1 Approach

- We used a Decision Tree Classifier with max depth 2 as a base model for our Adaboost Classifier.
- We use SAMME algorithm as Decision Tree typically gives good classification estimates but poor probability estimates.
- We use the Gini index as our criterion.
- We performed Hyperparametric Tuning for our AdaBoost classifier on the number of estimators and the learning rate with 3 fold cross validation.

4.1.2 Results

1. The best parameters found were learning rate= 2, no. of estimators= 175
2. The best score(accuracy) on cross validation was 0.535
3. The accuracy on the training data was 0.614
4. The accuracy on the test data was 0.600
5. The confusion matrix for the test data is $\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 30 & 64 \\ 16 & 90 \end{pmatrix}$

   where N = The customer churns and P = The Customer does not churn

4.2 Random Forest Classifier

4.2.1 Approach

- We performed Hyperparametric tuning on the Random Forest classifier on the no. of decision trees,max features,max depth, ccp_alpha (Cost Complexity Pruning) and oob_score (to use or not use).
- We use Gini index as our criterion.

4.2.2 Results

1. The best parameters found were ccp_alpha= 0, max_depth = 2, max_features = 'log2', n_estimators = 30, oob_score= True
2. The best accuracy on cross validation was 0.532
3. The accuracy on training data was 0.6
4. The accuracy on test data was 0.53

5. The confusion matrix for the test data is $\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} = \begin{pmatrix} 9 & 85 \\ 9 & 97 \end{pmatrix}$ where N = The customer churns and P = The Customer does not churn

5 Comparing the results

In our case of Churn classifier we are typically interested in knowing which customer will churn and which will not. Our interest typically is not finding all customers that will churn or those who will not churn. Also as the test data is not imbalanced (proportion of churn and not churn about the same), we care equally about both True Positives and True Negatives. So accuracy is a suitable metric for comparing their performance.

AdaBoost had an accuracy of 0.6 on test data while Random Forest had 0.53. Also from the two confusion matrices above we can see that both classifiers have a high recall (similar to each other) for churn. So for Churn Classification AdaBooost is a more suitable model.