BDA 564

TERM PROJECT REPORT

CHURN PREDICTION

IN

TELECOM INDUSTRY

FERAY ECE TOPCU

MEF University

ISTANBUL

2018

1. INTRODUCTION

Customer churn is one of the mounting issues of today's rapidly growing and competitive telecom sector. Customer churn can be defined as customers that stop doing business with a company or a service. It is also referred as loss of clients or customers. One industry in which churn rates are particularly useful is the telecommunications industry, because most customers have multiple options. Churn in the terms of telecommunication industry are the customers leaving the current company and moving to another telecom company. In addition, the focus of the telecom sector has shifted from acquiring new customer to retaining existing customers because of the associate high cost. Thus, companies want to prevent them to leave.

1.1. DATASET

The telecom dataset supplied by Orange Company and It is available on the URL:

https://bigml.com/user/francisco/gallery/dataset/5163ad540c0b5e5b22000383

The dataset has 3333 rows and 20 columns. Explanation of columns as follow:

Column Name	Description
State	the US state in which the customer resides, indicated by a two-letter abbreviation; for example, OH or NJ
Account Length	the number of days that this account has been active
Area Code	the three-digit area code of the corresponding customer's phone number
Int'l Plan	whether the customer has an international calling plan: yes/no
VMail Plan	whether the customer has a voice mail feature: yes/no
VMail Message	presumably the average number of voice mail messages per month
Day Mins	the total number of calling minutes used during the day
Day Calls	the total number of calls placed during the day
Day Charge	the billed cost of daytime calls
Eve Mins, Eve Calls, Eve Charge	the billed cost for calls placed during the evening
Night Mins, Night Calls, Night Charg	the billed cost for calls placed during nighttime
Intl Mins, Intl Calls, Intl Charge	the billed cost for international calls
CustServ Calls	the number of calls placed to Customer Service
Churn	whether the customer left the service: true/false

Figure 1.1.1: Explanation of Columns of The Dataset

2. METHODOLOGY

Microsoft Azure Machine Learning Studio is selected for churn prediction as tool. Microsoft Azure Machine Learning Studio is a collaborative, drag-and-drop tool you can use to build, test, and deploy predictive analytics solutions on your data.

As summary of methodology; firstly, dataset is examined. After that, data manipulation is applied. Then, SMOTE module is used to handle with imbalanced dataset. After SMOTE, feature selection methods are tried with different methods. In final, different classification algorithms are executed and they are tuned to make prediction better.

2.1. Dataset & Data Manipulation & Feature Selection

Figure 2.1.1 shows the pipeline to prepare input dataset for models.

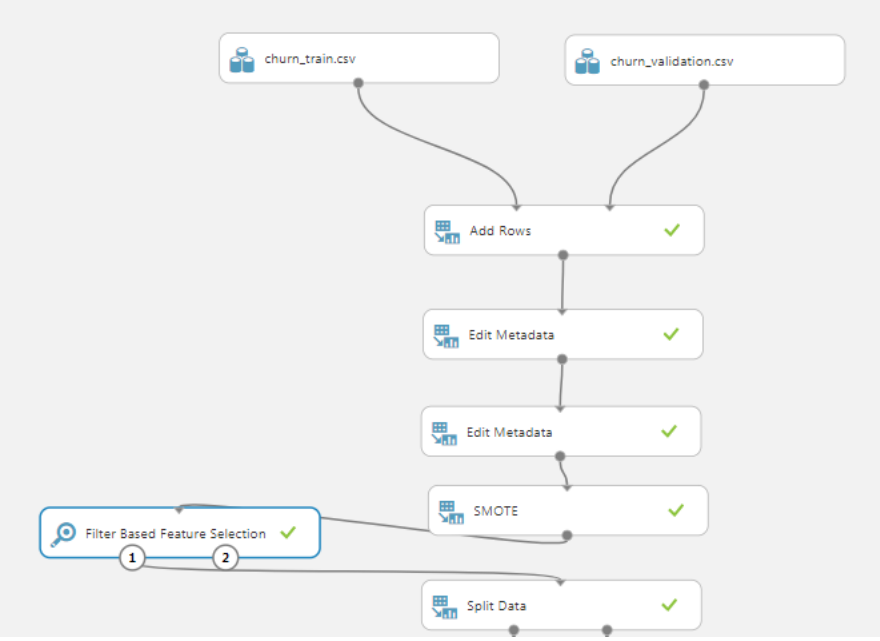

Figure 2.1.1: Pipeline for Data Manipulation and Feature Selection

Firstly, the columns of the dataset are examined in detail with visualize method of "Add Rows" module. Churn column is the target variable and it has two unique values "true" and "false". When target variable is checked, it is clearly seen that class are imbalanced. It means that one class is dominated to other class due to its number of observations.

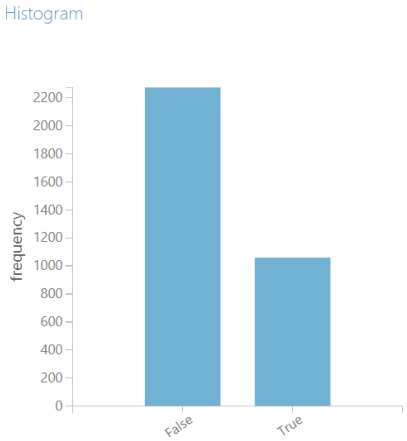

Figure 2.1.2: Histogram of Churn Column

In addition, although distribution of all features seems normal, the distribution of "Number vmail messages" is worth considering.

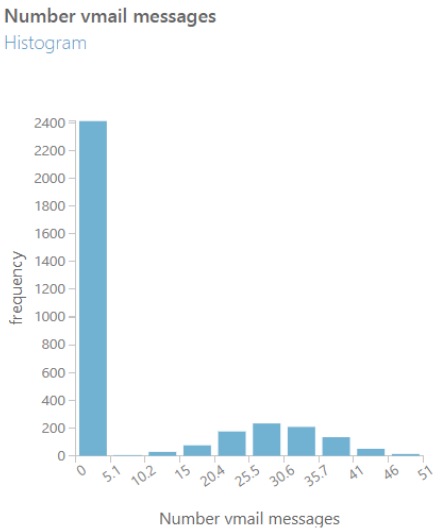

Figure 2.1.3: Histogram of Number of Vmail Messages

The last result of dataset examination is categorical columns which are stored as numerical. First “Edit Metadata” module converts State, International plan, Voice mail plan and Area code columns into categorical. Second “Edit Metadata” module labels the “Churn” column as target variable.

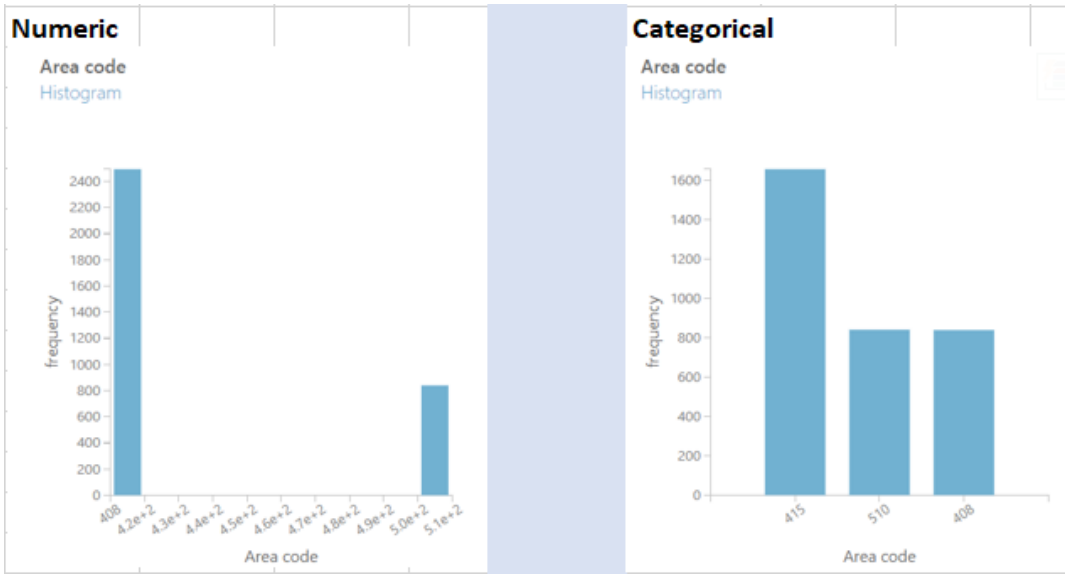

Figure 2.1.4: Histogram of Area Code Before and After Conversion

After data manipulation, “SMOTE” module is used to handle imbalances dataset. SMOTE stands for *Synthetic Minority Oversampling Technique*. The algorithm takes samples of the feature space for each target class and its nearest neighbors and generates new examples that combine features of the target case with features of its neighbors. It does not change the number of majority cases. SMOTE needs two parameters; 200 is given for *SMOTE percentage* and 2 is given for *number of nearest neighbors* for this project after trial of different values. After SMOTE module, dataset has 5451 rows and 20 columns and Figure 2.1.5 shows the distribution of Churn -target- column after SMOTE.

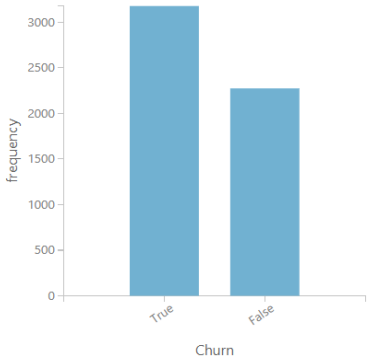

Figure 2.1.5: Histogram of Churn after Oversampling

After data manipulation, “Filter Based Feature Selection” module is applied to extract the most valuable features for prediction. This module needs two parameters, Chi-Squared is selected as *Feature Scoring Method* and 10 is given as *Number of Desired Features.* This module returns ten important features as shown on Figure 2.1.6.

Figure 2.1.6: Significant Features for Churn Prediction

Before modeling, dataset is randomized split as train and test. The proportion of splitting is 0.8 and train dataset has 4361 and test dataset has 1090 rows.

2.2. Models for Prediction

Two class decision forest and two class decision forest algorithms are used for modeling part. “Tune Model Hyperparameters” module is executed for each algorithm to make better prediction and improve evaluation metrics.

2.2.1. Two Class Decision Forest

The pipeline for two class decision forest can be examined on Figure 2.2.1.1 below.

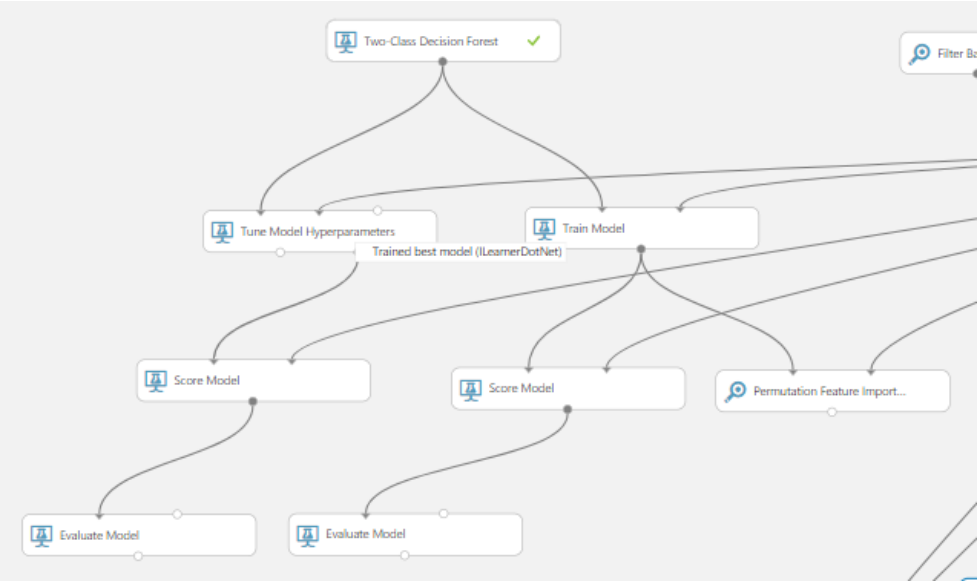

Figure 2.2.1.1: Pipeline of Two Class Decision Forest

Firstly; Two-Class Decision Forest algorithm is trained with train dataset with below parameters.

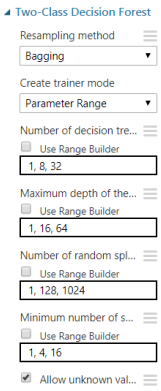

Figure 2.2.1.2: Initial Parameters of Two-Class Decision Forest

After training; Two-Class Decision Forest algorithm is applied without tuning. The model is scored with “Score Model” module. In final; “Evaluate Model” and “Permutation Feature Importance” modules are used to interpret the evaluation metrics.

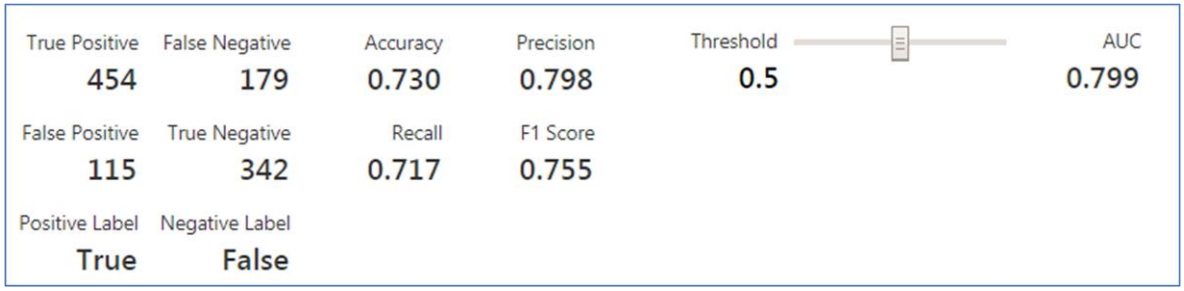

Figure 2.2.1.3: Results for Two Class Decision Forest before Tuning

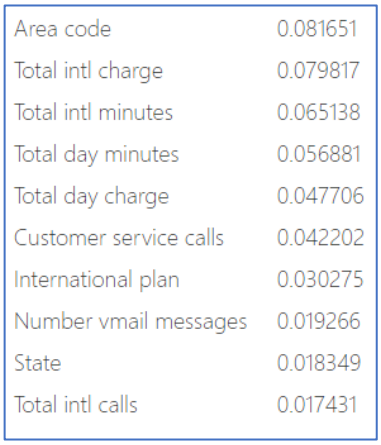

Figure 2.2.1.4:Feature Importance based on Two Class Decision Forest before Tuning

According to Figure 2.2.1.3; accuracy is 0.73 for this model but although imbalance is handled with SMOTE, accuracy is still not a reliable metric for prediction on this dataset. F-Score and AUC are better metrics to compare models with each other.

F-Score and AUC can be improved with tuning the model. Thus, “Tune Model Hyperparameters” module is executed with Two-Class Decision Forest and train dataset. “Tune Model Hyperparameters” module is executed with below settings:

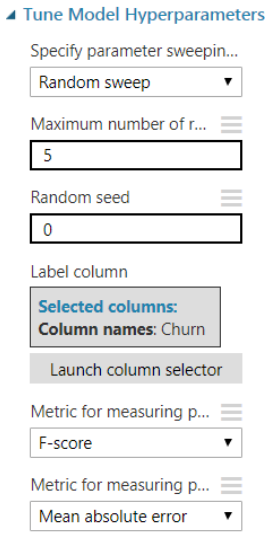

Figure 2.2.1.5: Initial parameters for Tune Model Hyperparameters

Figure 2.2.1.6 and Figure 2.2.1.7 shows the result of two class decision forest after tuning. F1-Score and AUC are increased by tuning. In addition; the order of feature importance is changed after tuning.

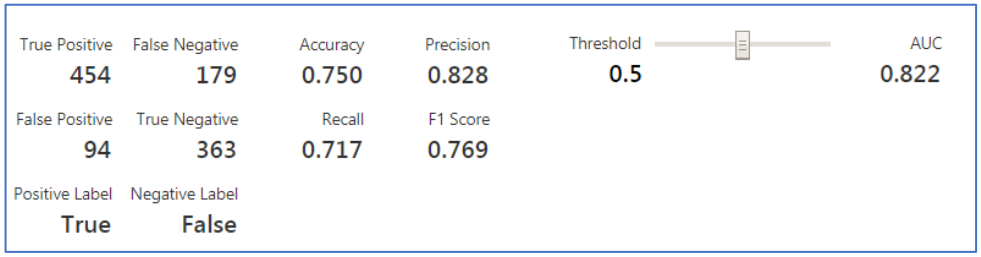

Figure 2.2.1.6: Results for Two Class Decision Forest after Tuning

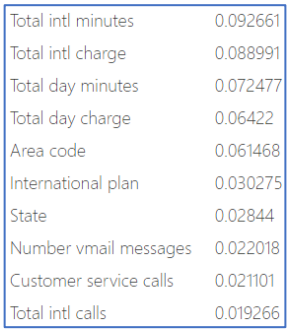

Figure 2.2.1.7: Feature Importance based on Two Class Decision Forest after Tuning

2.2.2. Two Class Decision Jungle

The pipeline for two class decision forest can be examined on Figure 2.2.2.1 below.

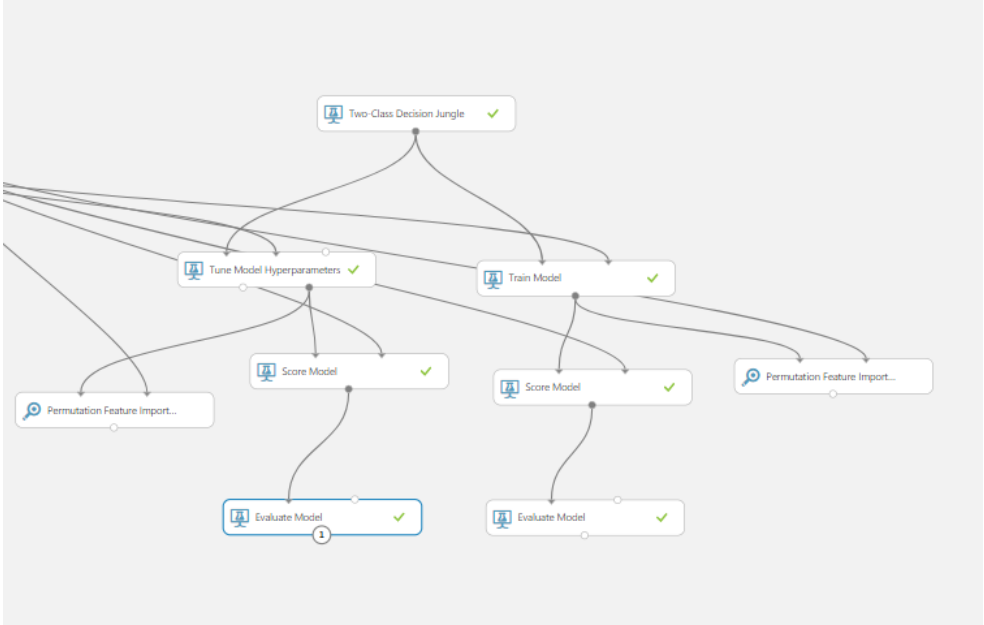

Figure 2.2.2.1: Pipeline of Two Class Decision Jungle

Firstly; Two-Class Decision Jungle algorithm is trained with below parameters and train dataset.

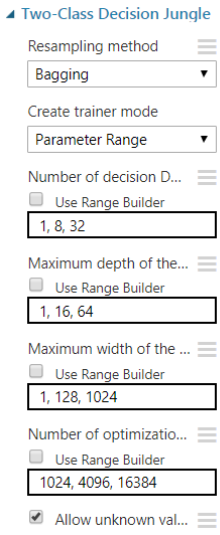

Figure 2.2.2.2: Initial Parameters of Two-Class Decision Jungle

After training; Two-Class Decision Jungle algorithm is applied without tuning. The model is scored with “Score Model” module to see predicted values in detail. In final; “Evaluate Model” and “Permutation Feature Importance” modules are used to interpret the evaluation metrics.

Figure 2.2.2.3 shows the evaluation metrics for Two Class Decision Jungle algorithm before tuning. When this model is compared to Two Class Decision Forest models, this model is not better than previous Two Class Decision Forest models according to evaluation metrics such that F1 Score and AUC.

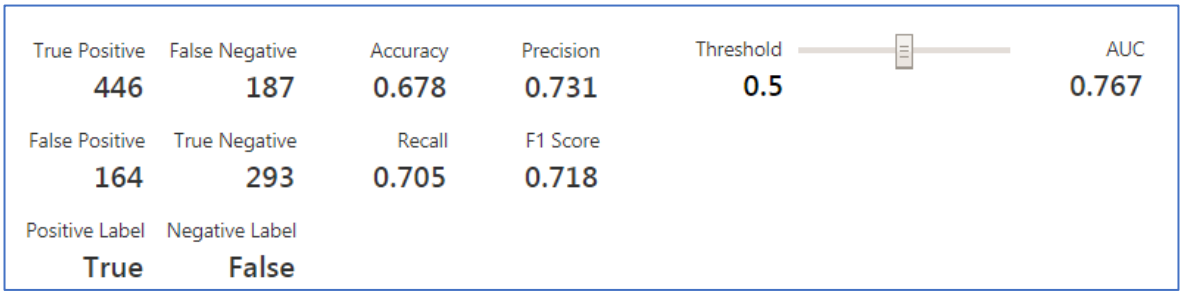

Figure 2.2.2.3: Results for Two Class Decision Jungle before Tuning

When Figure 2.2.2.4 is examined, the first two features are on same order with Two Class Decision Forest's feature importance result.

Area code	0.058716
Total day charge	0.017431
International plan	0.011927
Total intl minutes	0.010092
State	0.009174
Total intl charge	0.008257
Customer service calls	0.007339
Number vmail messages	0.005505
Total day minutes	0.004587
Total intl calls	-0.002752

Figure 2.2.2.4:Feature Importance based on Two Class Decision Jungle before Tuning

“Tune Model Hyperparameters” module is executed with Two-Class Decision Jungle and train dataset to improve the model. “Tune Model Hyperparameters” module is executed with same settings with previous module.

As shown on Figure 2.2.2.5; Two Class Decision Jungle after tuning is a better model than previous version of two class decision jungle based on the evaluation metrics. F1 Score and AUC are increased but when this model is compared to tuned Two Class Decision Forest, tuned Two Class Decision Forest is better model for prediction.

True Positive	False Negative	Accuracy	Precision	Threshold	AUC
441	192	0.720	0.796	0.5	0.800
False Positive	**True Negative**	**Recall**	**F1 Score**		
113	344	0.697	0.743		
Positive Label	**Negative Label**				
True	False				

Figure 2.2.2.5: Results for Two Class Decision Jungle after Tuning

Area code	0.051376
State	0.043119
Total day minutes	0.03945
Total intl charge	0.03945
Customer service calls	0.027523
Total intl minutes	0.023853
International plan	0.022936
Total day charge	0.017431
Number vmail messages	0.008257
Total intl calls	-0.001835

Figure 2.2.2.6: Feature Importance based on Two Class Decision Jungle after Tuning

3. CONCLUSION

Churn in the terms of telecommunication industry are the customers leaving the current company and moving to another telecom company. Thus, the dataset supplied by Orange Company is used to predict churn rate on Azure Machine Learning Studio. The dataset has 3333 rows and 20 columns.

First, because of imbalance dataset, “SMOTE” module is used for oversampling while protecting number of majority cases. After some data manipulation such that labeling categorical columns and target variable, “Filter Based Feature Selection” module is used to select significant features to generate more accurate model. The selected columns are used as input features for classification algorithms.

After preparation of dataset, Two Class Decision Forest and Two Class Decision Jungle algorithms are used to predict. Although F1 Score and AUC values are good for each algorithm, the models need to be tuned. Therefore, “Tune Model Hyperparameters” module is used to tune the algorithms and make better predictions. In summary, tuned Two Class Decision Forest model is the best choice according to its evaluation metrics to predict which customer is tending to be churn. This model has 0.75 accuracy, 0.77 F1 Score and 0.82 AUC.

In addition, the experiment of this project is available on Azure Machine Learning Studio:

https://gallery.cortanaintelligence.com/Experiment/BDA-564-TermProject