Classification Project Write-up

Predicting Customer Churn for Telco

Abstract

Project goal: The objective of the project was to arrive at a predictive classification model for identifying customers who are most likely to churn. It is typically cheaper to retain an existing customer than to acquire a new customer. Hence, I prioritized Recall as my main model performance evaluation metric.

Design

I was interested in applying machine learning to address one of the most frequently seen topics in marketing: customer churn.

Data

The data comes from a customer churn dataset for a fictional telecommunications company, Telco, with 7043 observations and 33 quantitative and qualitative features from IBM. I have also utilized the six-level urban-rural classification scheme for U.S. counties and county-equivalent entities from National Center for Health Statistics (NCHS).

Algorithms

Feature Engineering

1. Mapped cities to Large Metro, Medium/Small Metro and Rural classifications taken from NCHS.
2. Converting categorical features to binary dummy variables
3. Grouped particular dummies such as Churn Reason, Core & Premium Bundles and Customer Tenure.
4. Added calculated features such as Number of Add-ons, Media Streaming and Auto-Payment Preference counts.

Models

Random Forest, XGBoost and Logistic regression were used in my model. My base model was Random Forest, which I used for feature selection for the following models. After the base

model, I upsampled my target class, and used GridSearchCV to arrive at the best combination of parameters, such as class weight, n_estimators and max depth of trees. I also selected the best decision threshold after testing each model.

Model Evaluation and Selection

The entire dataset was split into 70/30 train vs. test. All results are reported based on the test performance.

I was optimizing my models for Recall, but also tested different class weights to optimize F1. Testing different decision thresholds also finetuned AUC for each model.

Final Logistic Regression Model (29 features):

```
Train Performance - Logistic Regression
---------------------------------
Model lr Predictions: AUC 0.81 | Accuracy 0.81 | Recall 0.84 | Precision 0
.74 | F1 0.78

Test Performance - Logistic Regression
---------------------------------
Predictions: AUC 0.78 | Accuracy 0.76 | Recall 0.83 | Precision 0.54 | F1
0.66
```

Tools

- Data cleaning and analysis: Pandas, Numpy
- Modeling: scikit-learn, xgboost, imbalanced-learn
- Visualizations: Matplotlib, Seaborn, Tableau

Communication

Presentation slides.