Cecilia Bell 6.11.2021

Classification Project Write-Up

Predicting Telco Customer Churn

Abstract

Reducing customer churn can drastically improve a company's revenue. The goal of this project was to predict customer churn for a telecommunications company. I worked with a dataset provided by IBM on the customer attributes of a fictional company called Telco. After feature engineering, handling class imbalance, and tuning hyperparameters, the chosen logistic regression model achieved a promising F2 score of 0.921 on the test data. Unsurprisingly, the satisfaction score proved to be the most important feature.

Design

The goal of this project was to predict customer churn for Telco, a fictional telecommunications company. Assuming that the company wanted to run a targeted promotional campaign at predicted-churn customers, and that the cost of the promotional campaign would be small compared to the lost revenue from churn, I chose to evaluate the model on an F2 score: an F-beta score that gives greater weight to recall than precision.

Data

I used two datasets provided by IBM on a telecommunication company's customer attributes, which can be found here and here. After merging and cleaning the data, and handling categorical variables, the set used for modelling contained 7043 rows and 35 features. Each row represented one customer. Features that showed strong correlation with the target included satisfaction score, contract type, and tenure.

Algorithms

- Feature Engineering: missing values were imputed, binary categorical features were label encoded, other categorical features were converted to dummy variables. Custom features were also created and cross validated.
- Modelling: logistic regression, K-nearest neighbors, Random Forest, Extra Trees, Ada Boost and Gradient Boost classifiers were tested, before the logistic regression model was selected on the basis of its F2 performance. Feature coefficients guided the choice of variables to be included in the model. Class imbalance was handled by cross validating class weights as well as probability thresholds. Hyperparameters were tuned with GridSearchCV.
- Model evaluation: the entire dataset was split into 80/20 stratified train/test sets. All scores and predictions were calculated with 5 KFolds. Predictions on the test set were only made at the very end. The official metric chosen, as explained above, was an F2 score.

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling and statistics
- Matplotlib and Seaborn for plotting
- Tableau for presentation visualizations

Communication

Model scores, as well as Tableau plots visualizing correlations and unexpected findings, are displayed in a PowerPoint, which will be presented to the Metis class and shared on GitHub.