Springboard Capstone Project 1:
Predicting Telecom Churn Customers

Albert Chiu
December 2018

Machine Learning: In-Depth Analysis

The goal of this capstone project is to take the dataset of telecom customers and build a model that predicts churning. The predictive model would be making a categorical prediction (whether customers churn or not) and the dataset already has labels that indicate whether each observation was a churn/non-churn customer; because of this, I elected to undertake this problem with a supervised classification approach.

As a first step, I split the dataset into training and test sets (80:20 split) and applied several classification algorithms (k-NN, Lasso Regression, Ridge Regression and Random Forest) to the training set. Hyperparameter tuning was applied to each of these classifiers to optimize results. When examining the classification report results for these classifiers, they performed moderately well (all classifiers had a churn recall score of approx. 50%, except k-NN, which had a 36% churn recall score) but there was room for improvement.

From the initial exploratory data analysis, I learned that there is a data imbalance in the dataset; approx. 74% of the customers are classified as non-churn. Because of this, we cannot use a simple accuracy score to measure the performance of the classifiers. A business would be most concerned with correctly identifying as many of actual churn customers as it can so that it can take the proper actions to retain these customers. Therefore, I used churn customer recall as the primary metric to determine model performance.

To compensate for the data imbalance and improve classifier performance, oversampling and undersampling techniques were applied to the data set. Oversampling techniques (SMOTE and random oversampler) and undersampling techniques (Tomek links and random undersampler) were combined with Random Forest and Logistic Regression classifiers. While a slight performance improvement was seen on Random Forest when it was paired with random undersampler, the biggest performance gains were seen in Logistic Regression pairing with SMOTE, random oversampler and random undersampler. Below are the classification report results:

Classification report results for non-churn customers:

	Precision	Recall	F-score	Support
k-NN	0.7872	0.9404	0.857	1007
Lasso Reg	0.8249	0.9027	0.862	1007
Ridge Reg	0.8223	0.9007	0.8597	1007
Random Forest	0.8116	0.9067	0.8565	1007
Random Forest w/ Random Over-Sampler	0.8204	0.8709	0.8449	1007
Random Forest w/ SMOTE	0.7982	0.8918	0.8424	1007
Log Reg w/ Random Over-Sampler	0.912	0.7408	0.8175	1007
Log Reg w/ SMOTE	0.9096	0.7398	0.816	1007
Random Forest w/ Random Under-Sampler	0.8708	0.7498	0.8058	1007
Random Forest w/ Tomek Links	0.8313	0.8858	0.8577	1007
Log Reg w/ Random Under-Sampler	0.9101	0.7239	0.8064	1007
Log Reg w/ Tomek Links	0.8427	0.862	0.8522	1007

Classification report results for churn customers:

	Precision	Recall	F-score	Support
k-NN	0.7059	0.36	0.4768	400
Lasso Reg	0.6787	0.5175	0.5872	400
Ridge Reg	0.6711	0.51	0.5795	400
Random Forest	0.6667	0.47	0.5513	400
Random Forest w/ Random Over-Sampler	0.6154	0.52	0.5637	400
Random Forest w/ SMOTE	0.6135	0.4325	0.5073	400
Log Reg w/ Random Over-Sampler	0.5569	0.82	0.6633	400
Log Reg w/ SMOTE	0.5544	0.815	0.6599	400
Random Forest w/ Random Under-Sampler	0.5333	0.72	0.6128	400
Random Forest w/ Tomek Links	0.6557	0.5475	0.5967	400
Log Reg w/ Random Under-Sampler	0.5413	0.82	0.6521	400
Log Reg w/ Tomek Links	0.6313	0.595	0.6126	400

There appears to be an inverse correlation between churn customer recall and non-churn customer recall as well as churn customer recall and churn customer precision. While the top performing classifiers have a lower churn customer precision value, it is an acceptable trade off. It would be better to identify more of the actual churning customers at the expense of slightly more incorrect churn predictions rather than failing to identify a large percentage of churning customers.