Project 3: Classification

Introduction

For our third project at the Metis Data Science Bootcamp we explored machine learning and classification problems. For my project I decided to investigate telecom customers who are at risk of changing companies (churn). As with most projects, there are endless options for the focus and depth of the investigation. I made the conscious choice to avoid hours of data wrangling in hopes that I would have time to focus on the modeling process and solidifying my understanding of classification. To that end I chose a dataset from Kaggle with roughly 7,000 entries for telecom customers.

For any classifier we need to consider the effectiveness of the model at correctly predicting the desired outcome. It is rare that a model will correctly identify the desired class so we also need to evaluate the errors. The predictions of any classifier can categorized as True Positive, True Negative, False Positive and False Negative. Prioritizing these four categories is highly dependent on the business case. As a first pass I used recall (true positive rate) as my primary scorer.

Exploratory Data Analysis

The first step in any of these projects is exploratory data analysis. Each entry contained 20 pieces of information (features) about each customer along with a field indicating if they churned. It's worth noting that most of the features were yes/no binary type features indicating which services the customer subscribed to. It's also worth noting that the dataset was slightly imbalanced with 27% of the customer's churning. Finally, when I looked at the breakdown of churn I learned that 88% of the customers who switched phone providers were on month to month contracts.

Model Identification

After completing the EDA I started the evaluation of the different classifiers. I chose to evaluate 6 different classifiers: Logistic Regression, K-Nearest Neighbors, Random Forest, Support Vector Classifiers, Naïve Bayes and a Hybrid Classifier. In order to address the slight class imbalance for the churn customers I decided to utilize Systematic Minority Oversampling known as SMOTE. For each classifier candidate I configured a pipeline of SMOTE and the classifier. The pipeline was passed to a grid search tool along with a list of values to test for each of the hyperparameters for classifier step. The output of each search was checked to make sure the optimal parameters were not at an end of the search space. Once a set of parameters were identified the classifier was retrained with the test data and re-scored.

In creating the hybrid classifier mentioned above, I tried to utilize the correlation between month to month contracts and churn. I divided the sample group into two smaller groups: month to month contracts and long term contracts. Then for each of these smaller groups I ran a grid search for each of the classifier candidates mentioned above. The best model for each of

these sub-groups was then merged into a single predictor. The output was scored using recall as with the other models.

Results

The scores for each of the models are shown in Table 1.

	Metrics	
Models Tested	**Recall**	**ROC AUC Score**
Logistic Regression	0.76	0.83
KNN Classifier	0.75	0.81
Random Forest Classifier	0.86	0.82
Support Vector Classifier	0.78	0.82
Naïve Bayes Classifier	0.78	0.77
Hybrid Model	0.70	0.77

While many of the models performed similarly, I chose the Logistic Regression model primarily because the model is interpretable, i.e. we are able to easily see the relation between changes in inputs and the prediction.

I then applied this model to data held out for final testing. The ROC AUC score was similar to the initial trials but the Recall Score was significantly lower for the test dataset. This may be due to overfitting the training data.

Going Forward

As I mentioned earlier, the metric selected for evaluating models is highly dependent on the business case for your application. In order to truly optimize the model for a specific business case I believe a custom score needs to be developed to account for the specific costs associated with my business case. For example, for telecom customer churn there are several factors that should be considered when optimizing model parameters: customer monthly spend, promotional cost per customer, and expected promotion acceptance rate for each class. With each of these factors incorporated into a custom score a true optimal model could be identified and evaluated.