Propensity of Contract Renewals

Himanshu Shekhar
CS229: Machine Learning, Stanford University

Goal

Maximize the renewal conversion of the contracts by prioritizing the set of contracts that have lower likelihood of renewal

Method

All learning methods considered for use for this project falls under the category of supervised learning algorithms. Along with the accuracy of the model, I would look at other important metrics because as you increase your sensitivity (true positives) and can identify more cases with a certain condition, you also sacrifice accuracy on identifying those without the condition (specificity). Thus we will look at sensitivity and specificity:

Specificity= TN/(TN+FP) , Sensitivity=TP/(TP+FN)

In order to combine both sensitivity and specificity, we will use the G-mean:

Gmean = sqrt(sensitivity*specificity)

We will train several learning models to find the best prediction, and to establish performance, we will train each model with the first 70% of the loans and test the trained models on the last 30% of the loans in our dataset.

Data Preparation

Raw data

- Largest networking company(dataset)
- 45 features

Filtered data

- Filter out non-descriptive fields such as contract ID, product key
- Remove contracts for which at least one field is missing

Expanded data

- Convert categorical fields into multiple boolean fields
- Conduct TF-IDF on case notes (conversation between customers & service provider) to identify relevant keywords

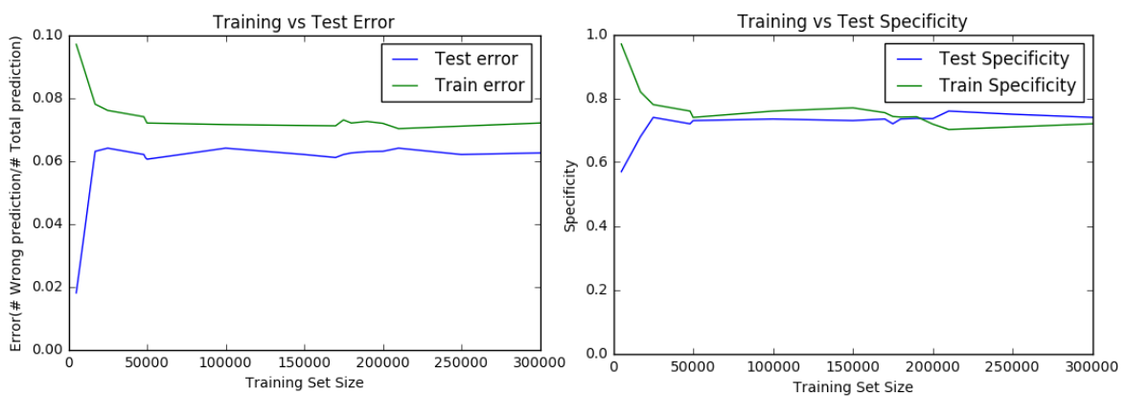
NOTE: To label the dataset, I classified any contract that expired as negative(0) examples, while I classified any contract that renewed as positive (1) examples

Models

Logistic Regression

Started modeling with logistic regression with Newton's method. I trial-trained with a logistic classification on all features and got rid of all the factors/variables where the p value is greater than .05 (i.e. 95% wald confidence interval) and also computed VIF to ensure there is no multicollinearity. Training and test converges to an optimal solution within 11 Newton iterations.

From the diagnostic of bias vs variance, we see that the model still has high bias, but renewal prediction may benefit from better feature selections:



Based on the ablative analysis on all available features, we filtered out features that decreased the test specificity, and without hurting overall test accuracy, we managed to bump specificity from 74.8% to 76.1% for our logistic regression model with better feature selection.

Accuracy	Precision	Sensitivity	Specificity	G-mean
91.4%	95.1%	94.2%	76.1%	84.6%

SVM

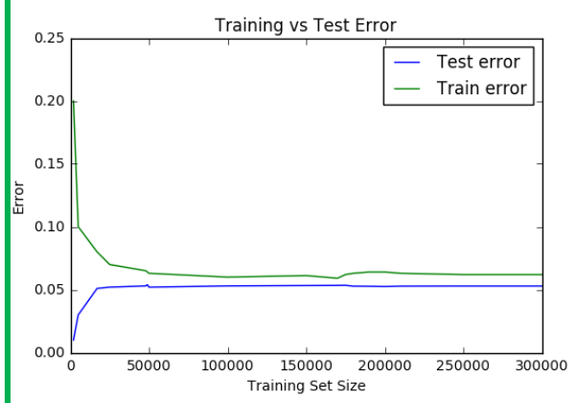
I used L1 regularization (soft margin SVM). For training the data points, the model is the result of the optimization:

$$min_{\gamma,w,b}\frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$
$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, ..., m$$

The performance of an SVM model depends on the kernel used, the parameters of the kernel, and the soft margin parameter C. I ran SVM with various kernels:

Kernel	Accuracy	Precision	Sensitivity	Specificity	G-mean
Linear	92.0%	95.6%	95.3%	79.0%	86.7%
Polynomial	92.0%	93.5%	96.6%	60.8%	76.6%
RBF	91.8%	92.9%	97.0%	56.5%	74.0%
Sigmoid	89.8%	90.8%	97.2%	40.0%	62.3%

Using a linear kernel, we see that the model has a high bias:



Including in additional features extracted through text mining led to a 0.5% increase in specificity and 0.7% increase in G-mean. We then iterated on various values of C but ultimately C=1 performed the best.

Accuracy	Precision	Sensitivity	Specificity	G-mean
92.7%	95.9%	95.3%	79.0%	86.7%

Results

I found that SVM has the highest specificity (79%), and thus renewal rate prediction is best with SVM. If we apply the model and prioritize the set of contracts which has lower likelihood of renewals with the right incentives identified from the model like bundled offers, appropriate discount, dedicated account manages among others, we can increase the renewal rate by 47%.