Case 3 – Churn Exploration

Nathan Gardner

Dr. Petrie

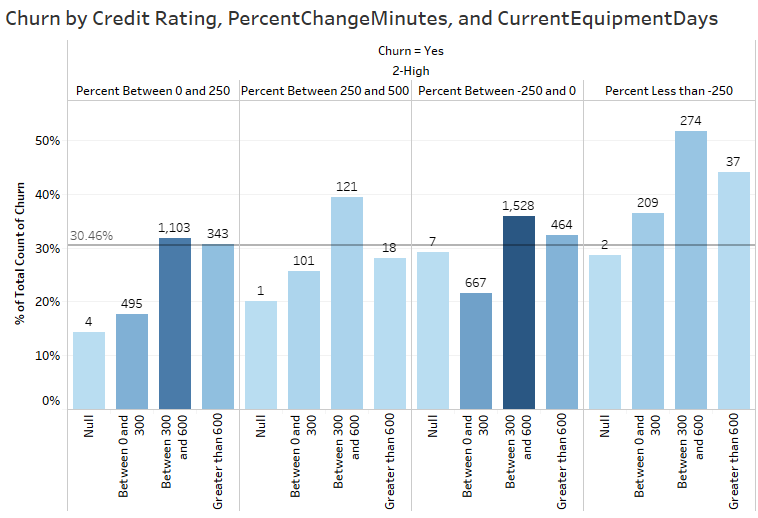
## Executive Summary

This analysis was performed with three goals in mind: To develop a model to predict whether a customer will churn, to identify the most important characteristics of churners, and to create a proactive churn management program. The data used in this analysis was the cell2celltrain data, which contains almost 70,000 rows with 51 variables (after cleaning). A potential limitation of this data is that churn is a relatively rare event (2-4%), so predicting it is hard without enough data. To combat this, more customers who churned were put in the data used to build the model.

Of the 980 individuals who had the highest predicted probabilities, the model correctly identified 52.1% of them compared with the overall 28.5% churn rate for a total difference of 23.6%.

The top 4 most important single drivers of churn are the number of days a customer has used their current equipment, the percent change in number of minutes used over the past 4 months, the average minutes used over the past 4 months, and the average monthly revenue collected over the past 4 months (CurrentEquipmentDays, PercChangeMinutes, MonthlyMinutes, MonthlyRevenue, respectively).

When combining variables, there was only one combination that had a churn rate above 50%. As seen in Figure 1, when PercChangeMinutes was less than -250, CurrentEquipmentDays was between 300 and 600, and CreditRating was High, there was a churn rate of 51.7%

Since there is limited resources and customers are likely to accept an offer if given one, offering only customers who are most likely to churn is extremely important to conserve resources. The model can identify customers with the highest chance of churning, and to start out, Cell2Cell can offer the top 10% of customers who are most likely to churn a discount. The model should be run every 30 days (in accordance with how churn is measured), and if results are promising, expanded to encompass the top 20% of churners. Progress should be constantly monitored to scale up or scale back program based on results.

Figure

## Technical Summary

### Creating the best model

I ran through many kinds of models to see which performed the best. This was determined by measuring the AUC. I tested a regularized logistic regression model, several gradient boosted models, a K-nearest neighbors model, a neural network, a support vector machine, a random forest, and an extreme gradient boosted model. Without tweaking much, the boosted models (extreme gradient boost and gradient boost models) greatly outperformed the other models in terms of AUC. Consequently, I began tweaking those two models.

I first tweaked the number of trees for the GBM, trying out 100, 500, 1000, and 1500 trees. I increased the interaction depth to 5 and 6 and decreased the shrinkage to .001 and .0001. I also reduced the minimum observations in node to 7 from 10. Finally, I increased the training sample from 5,000 to 10,000 rows. This resulted in an increase from .62 to .641. When tested on the holdout sample, this model received a lift of 1.75 and identified 50% of churners correctly.

For the extreme gradient boosted tree, I kept the training sample at 10,000. I used an eta (learning rate) of .01 and tried many different numbers of trees (250, 500, 1000, 2,500, and 5,000). I tried many different gammas, but it did not seem to affect the AUC much. The best model had a ROC of .651 and a lift of 1.83. It identified 52% of the customers with the highest predicted churn probability correctly

### A screenshot of a cell phone Description automatically generatedHow GBM models relationships

Figure

After training a model with the GBM function and obtaining the variable importance, CurrentEquipmentDays had the highest relative importance (Figure 5). Figure 2 shows the jump (~10%) in churn rate once CurrentEquipmentDays reaches about 400+ days. Figure 3 overlays the churn rate with a histogram of the data. This shows that most of the values lie between 0 and 1000, with very little showing after that. The churn also remains steady once CurrentEquipmentDays reaches 400.

A close up of a logo

Description automatically generated

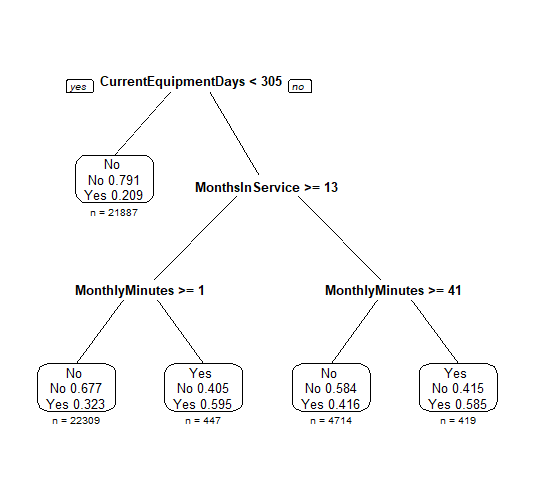
Churn Rate

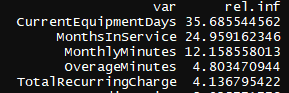
Figure

CurrentEquipmentDays Counts

### Explaining partition model

This basic partition model (Figure 4) shows how a simpler model works. It takes the most important variables (Figure 5) and finds the optimal split, or the split that predicts churn better. If the model fails to improve by what the complexity parameter is set at, it stops splitting.





Figure

Figure