Relax Inc.

Relax Inc. makes productivity and project management software that's popular with both individuals and teams. Founded by several former Facebook employees, it's considered a great company to work at.

Relax would like to utilize existing customer information to predict future user adoption. Relax defines user adoption as a user who has logged in on three (or more) separate days in one (or more) seven-day periods.

Relax collects the user id and date when a user logs into the system. With that data, it was relatively easy to calculate a running total of logins for each user for the previous 7 days. With this, any user who has a count of three or higher qualifies as an adopted user.

Relax has data on 12,000 users (I found some duplicates) and for those users has collected 207,917 separate user logins. Of the 12,000 users, 1,602 (13%) are considered adopted users. This is an imbalanced dataset which in this case, will make accurate positive predictions more difficult. To combat the imbalance data, I did the following:

- For accessing and comparing classifier performance, I used ROC curves (Figure 1), the confusion matrix, and precision, recall and F1 metrics.
- Tried various classifiers

After being able to train a accurate model, I stopped but if needed, my plan was to generate synthetic samples (over-sampling) and to utilize penalized classifiers and metrics.

I trained the initial models on ~15 features. The LGBM Classifier scored the best and was the classifier I opted to use. I was also able to obtain some very good results with the KNN classifier.

I moved on to feature elimination by utilizing feature importance and LASSO (which both agreed) and ultimately resulted in three features as follows:

- Creation_Time_Delta:
    - Creation time (date) as an integer from the first recorded date.
- Last_Session_Creation_Time_Delta
    - Last session creation time (date) as an integer from the first recorded date.
- Org_ID
    - Given Org id – The organization ID if a user belongs to an organization

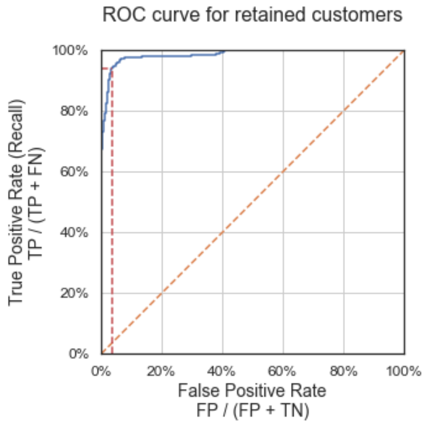
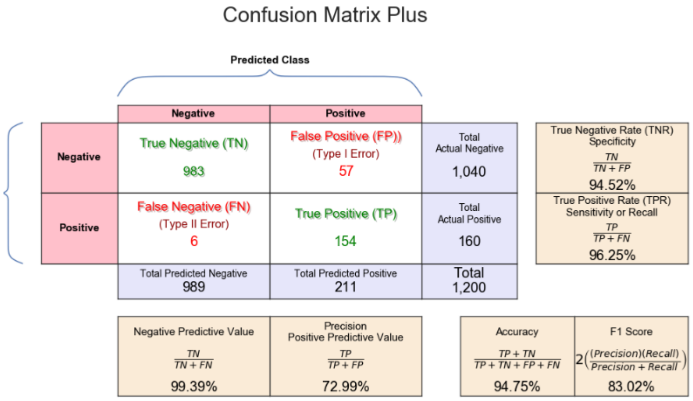


Figure 1

My initial classifier, before making my cutoff adjustment scored a True Negative Rate (TNR) of 98.56% and a True Positive Rate (TPR) of 80.21%. Normally I wouldn't look at accuracy on an imbalanced data set, but these number were so good I decided to monitor accuracy as well. The model's accuracy score was 96.11%. I then calculated a new cutoff based on the ROC Curve (Figure 1) and which resulted in a TNR of 96.47%, a TPR of 94.10% and an accuracy score of 96.16%

Finally, I retrained my model on the training and validation data and predicted on my holdout data (never seen by the model). I received a TNR of 94.52%, a TPR of 96.25% and an accuracy score of 94.75% (Figure 2).



Confusion Matrix Plus

Predicted Class

	Negative	Positive		
Negative	True Negative (TN) 983	False Positive (FP)) (Type I Error) 57	Total Actual Negative 1,040	True Negative Rate (TNR) Specificity $\frac{TN}{TN + FP}$ 94.52%
Positive	False Negative (FN) (Type II Error) 6	True Positive (TP) 154	Total Actual Positive 160	True Positive Rate (TPR) Sensitivity or Recall $\frac{TP}{TP + FN}$ 96.25%
	Total Predicted Negative 989	Total Predicted Positive 211	Total 1,200	

Negative Predictive Value $\frac{TN}{TN + FN}$ 99.39%	Precision Positive Predictive Value $\frac{TP}{TP + FP}$ 72.99%	Accuracy $\frac{TP + TN}{TP + TN + FP + FN}$ 94.75%	F1 Score $2\left(\frac{(Precision)(Recall)}{Precision + Recall}\right)$ 83.02%

Figure 2

I am concerned my model, or the data might have some leakage because of the incredible score. I couldn't find any problems, but my next step would be to reevaluate my features to ensure I am providing good, clean data with no special information. In that process, I would like to get a better understand of why these features are predicting so well and ensure we could utilize this model as a predictive tool for Relax Inc.