Churn No More

By: Grant Edwards

Summary

Using historical data from previous customers of SyriaTel we built a model to help predict whether customers will churn based on information pertaining to the customers account. With the data available, we were able to predict accurately predict all customers from the training data with an accuracy of ~98% and recall customers that will churn at a rate of ~87%.

Why Try to Predict Customer Churn?

Customer churning is a loss in business and revenue for SyriaTel. Being able to predict if a customer is about to churn, gives the business an opportunity to try and retain the customers business through targeted offers and support.

Business Problem

Goal: Build a model to accurately identify whether customers will drop the services of SyriaTel.

- What are the leading factors and predictors for when a customer will churn?
- Customer churn in our dataset is at ~14% of all customers and accounts for 16% of revenue ($31,567 loss of $198,146).
- Mean loss per customer of $65.36

Data

The data is historical data provided from SyriaTel. It contains information pertaining to customer accounts. Variables in the dataset include:

- Churn - this is our target variable.
- Account length - the number of months that the account has been active.
- Area code - splits the country into 3 regions.
- Voice mail plan/vmail messages - if the customer has a voicemail plan and number of messages
- Total (day/eve/night/intl) (minute/calls/charges) - how many minutes calls and the charge in dollars respectively for the different times of the day or if it was an international call.
- Customer service calls - the number of times the customer called customer service.
- There is also state and phone number which we will drop as these should both be arbitrary.

Methods

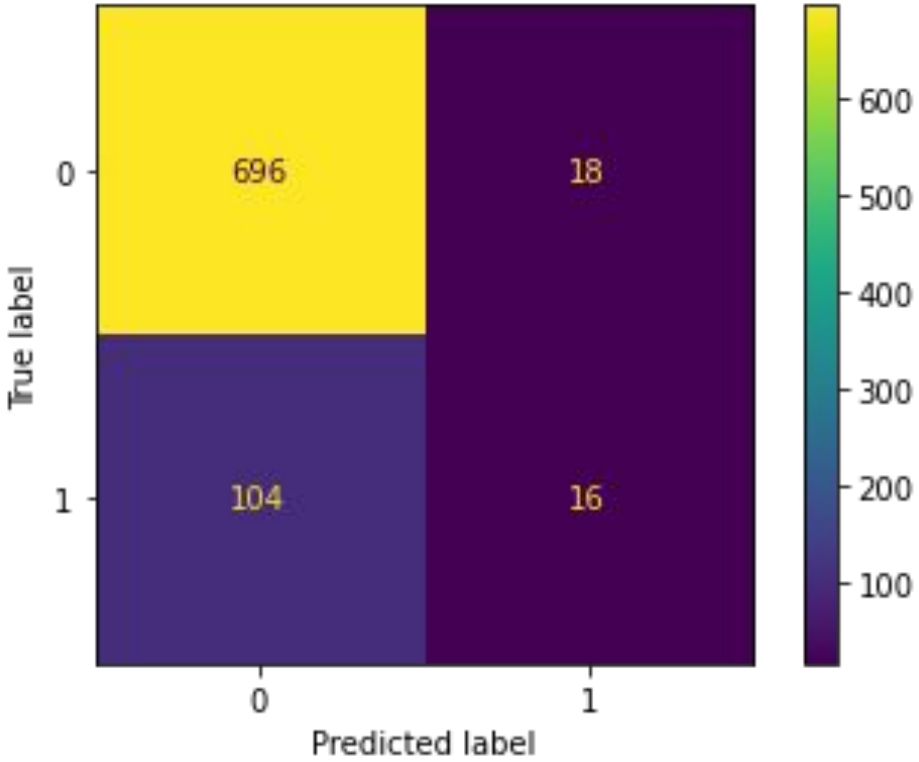
- Used Python and Pandas in a Jupyter Notebook.
- Clean and prepare data to be used in Machine Learning Modeling.
- Split our data into train and testing data.
- Use machine learning Models to make predictions on test data using training data to train the model.

Methods

Models we will be using

- Linear Regression
- Decision Tree
- Nearest Neighbor
- Pipes - combines models
- GridSearch
- Random Forests
- Ensembles
- Bagging
- Gradient Boost
- XGBoost

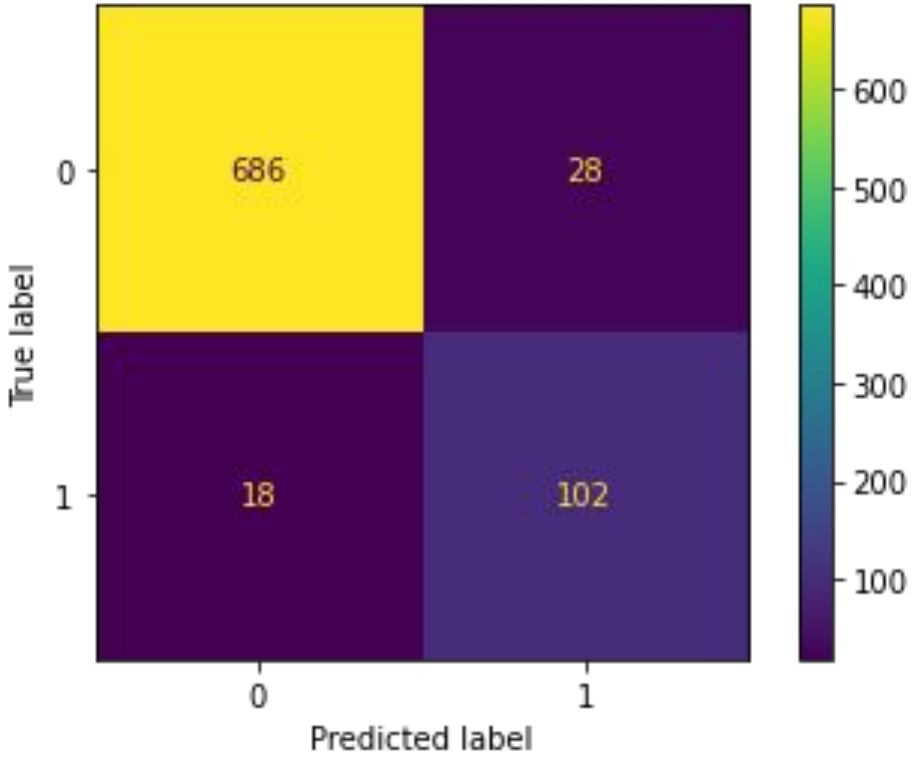
First Model - Linear Regression



Precision Score:	0.4706
Recall Score:	0.1333
Accuracy Score:	0.8537
F1-Score:	0.2078
Mean Cross validation:	0.6547

Overall this model did not do too well. Most of the customers that will churn, were predicted to not churn, and looking at the cross validation score, it would probably have been better to just guess.

Second Model - Decision Tree Classifier



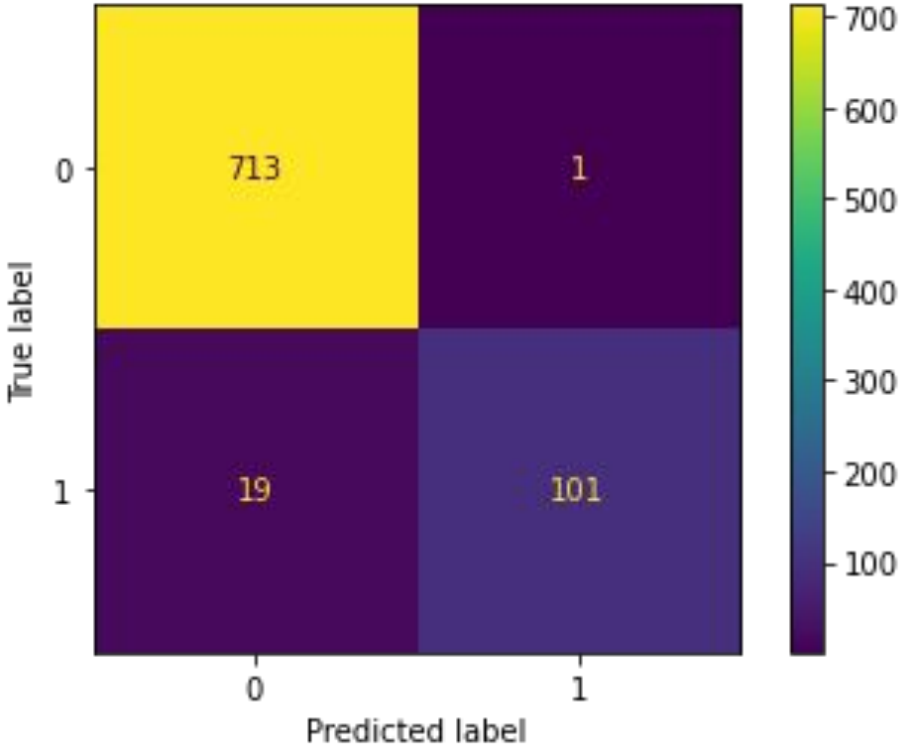
Precision Score:	0.7846
Recall Score:	0.8500
Accuracy Score:	0.9448
F1-Score:	0.8160
Mean Cross validation:	0.9616

This was a massive improvement over the base model. It has greatly increased the number of accurate predictions on customers that will churn.

Parameters:

- Max tree depth: 6
- Minimum samples per split: 12

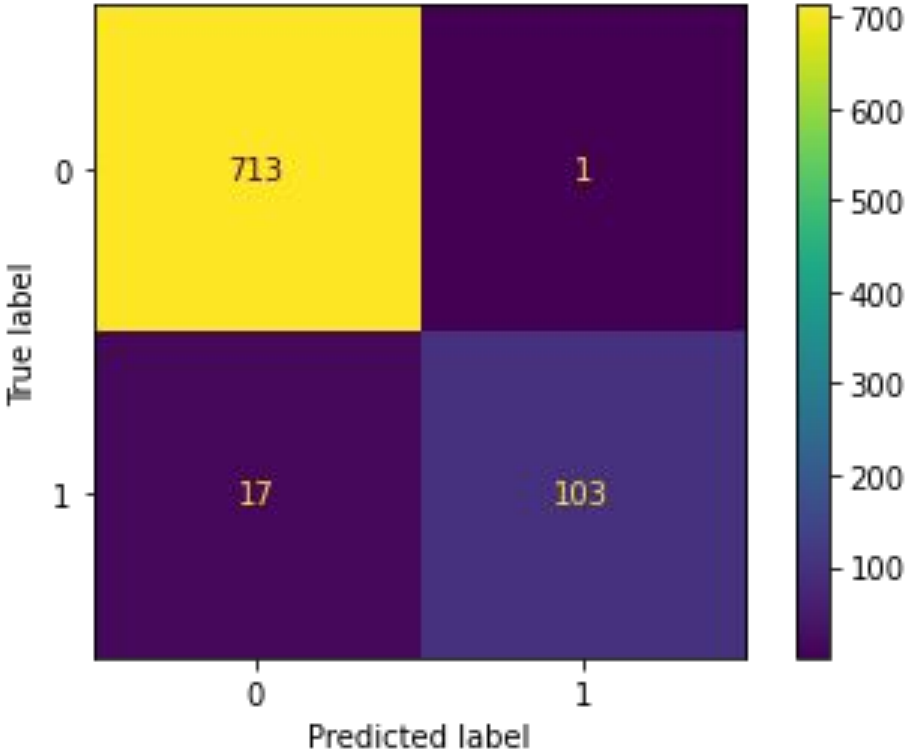
Decision Tree Classifier - Improved



Precision Score:	0.9902
Recall Score:	0.8417
Accuracy Score:	0.9760
F1-Score:	0.9099
Mean Cross validation:	0.9724

Overall an improvement on the base decision tree. However, while there was an improvement on the accuracy, there was a slight decline in recall, which is more important for this model.

Random Forest Model



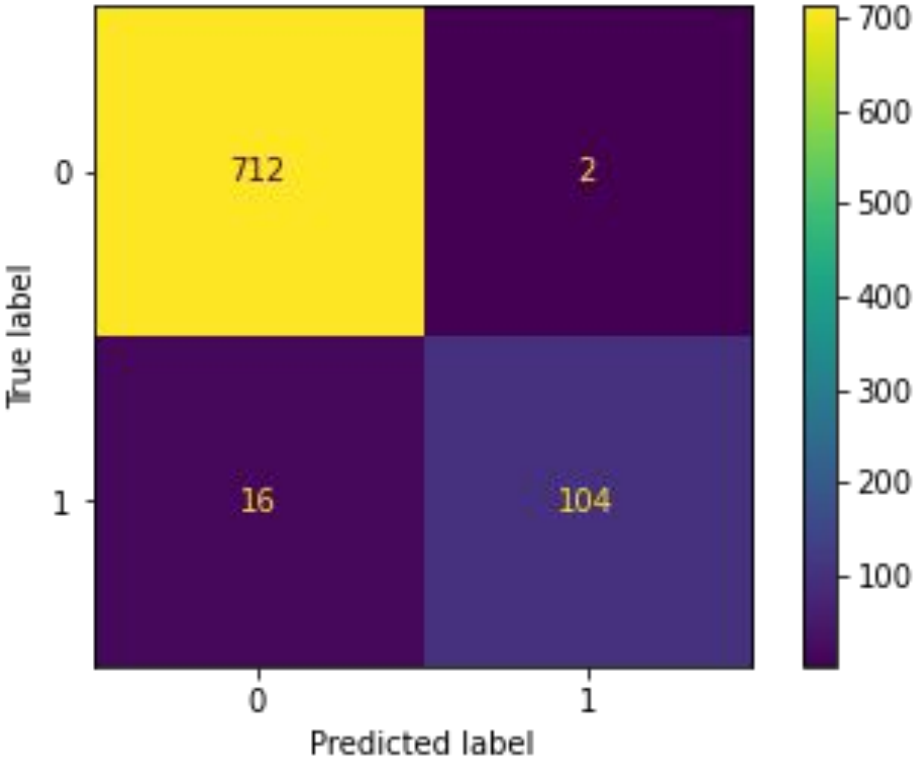
Precision Score:	0.9904
Recall Score:	0.8583
Accuracy Score:	0.9784
F1-Score:	0.9196
Mean Cross validation:	0.9700

Random Forest improved our model on the decision tree, largely reducing the number of false positives predicted.

Parameters:

- # of estimators: 20
- Max depth: 5
- Max features: 10

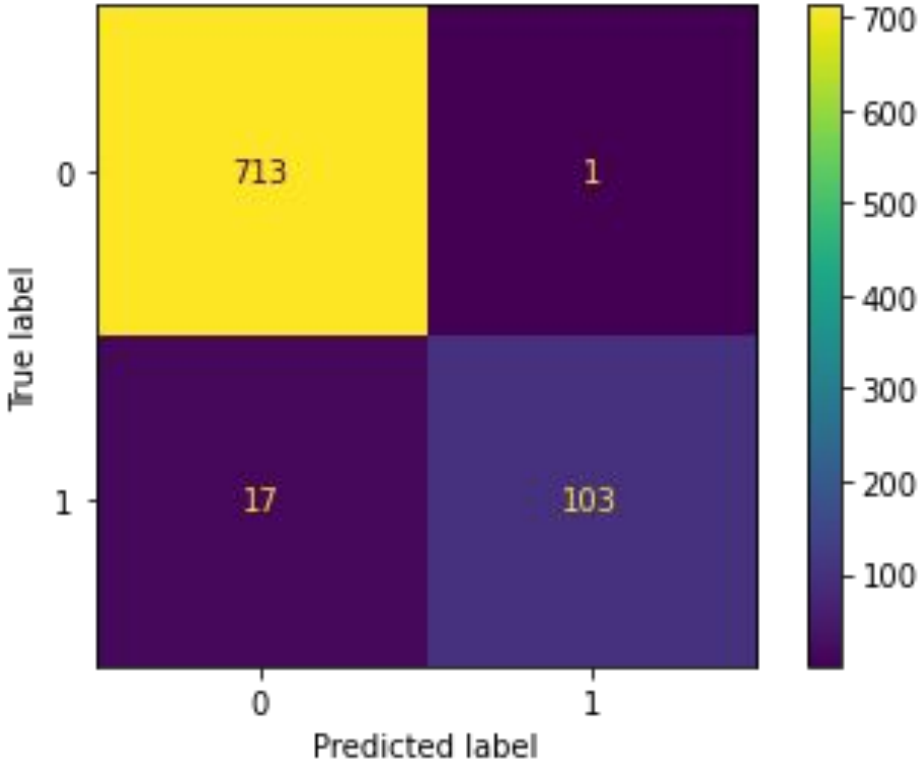
Gradient Boost - The top performing model



Precision Score:	0.9811
Recall Score:	0.8667
Accuracy Score:	0.9784
F1-Score:	0.9204
Mean Cross validation:	0.9772

Gradient Boosting saw a very small improvement on the random forest but with the improvement on the mean cross validation shows an ever so slightly stronger model.

XGBoost

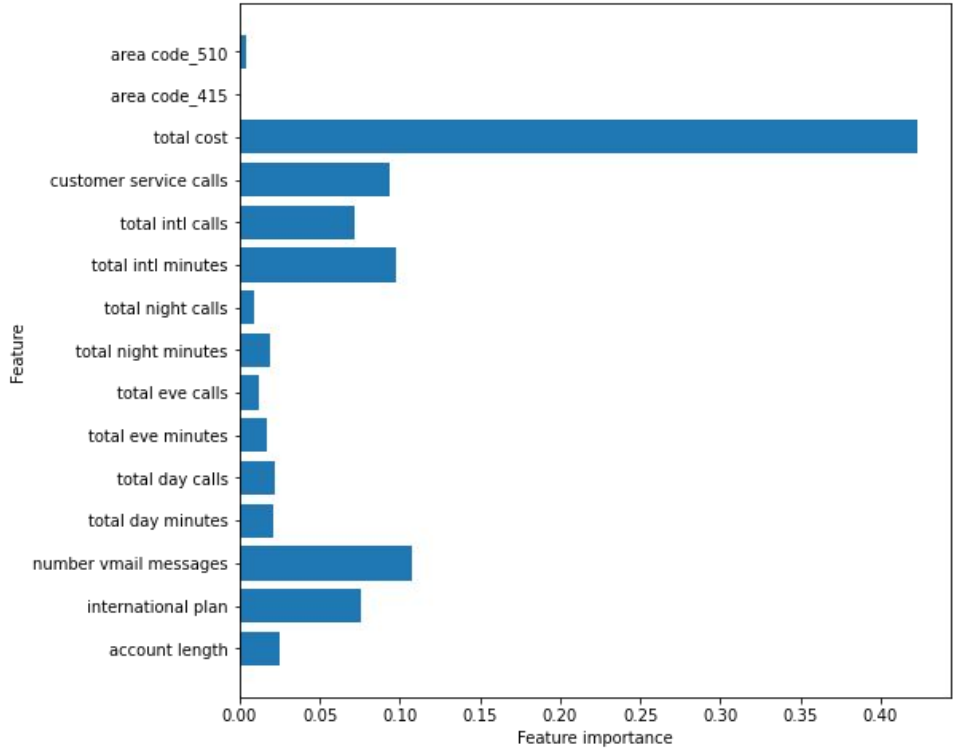


Precision Score:	0.9904
Recall Score:	0.8583
Accuracy Score:	0.9784
F1-Score:	0.9196
Mean Cross validation:	0.9736

Extreme gradient boosting is one of the leading model building tools right now. However we saw a slight decline in our model, from the gradient boost.

Features and their Importance

- Total cost has the strongest significance to customers churning.
- Variables pertaining to international call also have a stronger overall correlation to churn.



Conclusions

We were able to build a model that was able to correctly recall customers that will churn at 87% and have an overall accuracy for all customers of 98% .

- Using this model we should be able to be able to identify churning customers.
- Paired with a customer retention plan, churn could be heavily reduced.

Knowing that total cost is a major predictor for churning, as well as international calling, voicemail, and customer service calls:

- Offer a better international plan or lower cost international after a certain point.
- Lower cost voicemail plans.
- Incorporate customer retention in service calls.

Thank You!

Email: grantedwards11@gmail.com
GitHub: @gzedwards
LinkedIn: www.linkedin.com/in/grant-edwards-25206914a