SyriaTel Churn Data

By Eduardo Osorio

Objective

...

Our objective today is to accurately predict when a customer is about to leave SyriaTel for another provider.
We will be looking at individual customer account data to see if we can find any correlation to help us determine the likelihood of a customer churn.

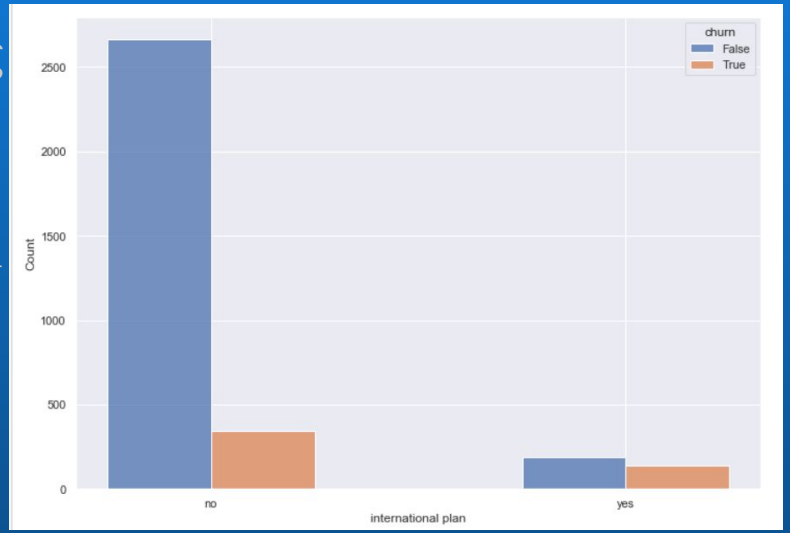
Data used:

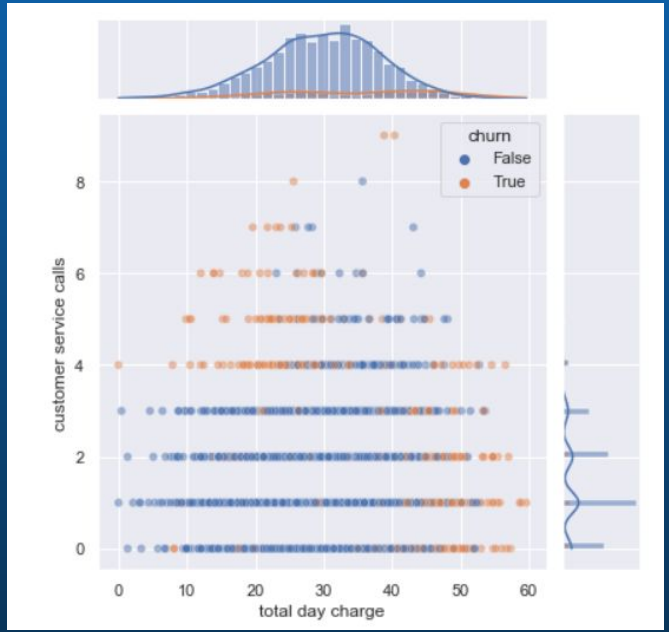
The data used is SyriaTel's own customer data. The data set provided an insight into certain individual account details like:

- If a customer is enrolled in a voicemail plan or if its enrolled in an international plan.
- Their call amount during the day, evening, and nights.
- The total amount of minutes used during the day, evening and nights.
- The amount of customers service calls.

EDA

Some interesting findings during the EDA phase was that customers tend to leave at a higher rates if they don't have an international plan





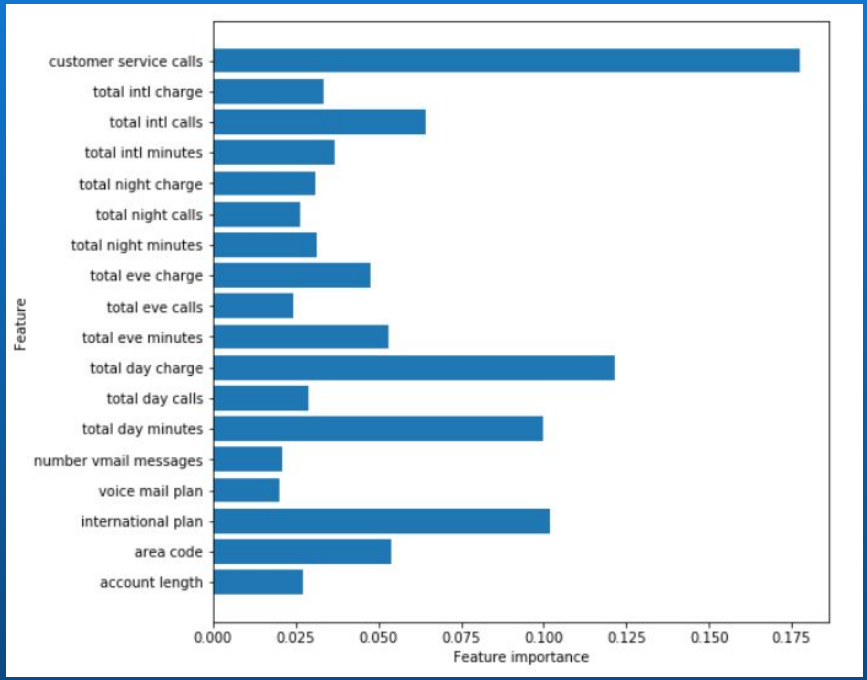
Customers that call customer service over 4 or have a total day charge over 45 tend to leave at higher rates.

EDA Continued:

The most relevant Features in our data according to some preliminary models are:

- The number of customer service calls.
- The total day Charge rate.
- If they have an international plan or not.

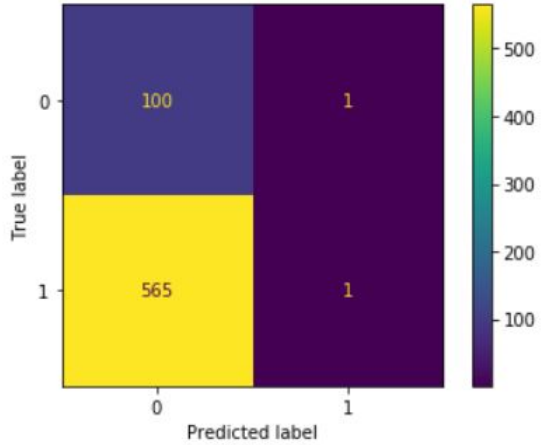
Even though there are 3,333 enteries, only 15% actually have left SyriaTel todate.
Meaning our data will have some imbalance.




```
Raw counts:

1    2850
0     483
Name: churn, dtype: int64
-----------------------------------
Normalized counts:

1    0.855086
0    0.144914
Name: churn, dtype: float64
```

Base Model:

```
In [5]: run(X, labels, RandomForestClassifier())

Training Precision:  1.0
Testing Precision:  0.9102173913043479
Training Recall:  1.0
Testing Recall:  0.8796050099709618
Training Accuracy:  1.0
Testing Accuracy:  0.9475262368815592
Training F1-Score:  1.0
Testing F1-Score:  0.8939890015575829
```



For the base model, I created a function to quickly run several vanilla models to get an idea of which ones worked the best. The two winning models were XGBoost and Random Forests. If you look at the confusion matrix for these two examples, you can clearly see the class imbalance mentioned earlier.

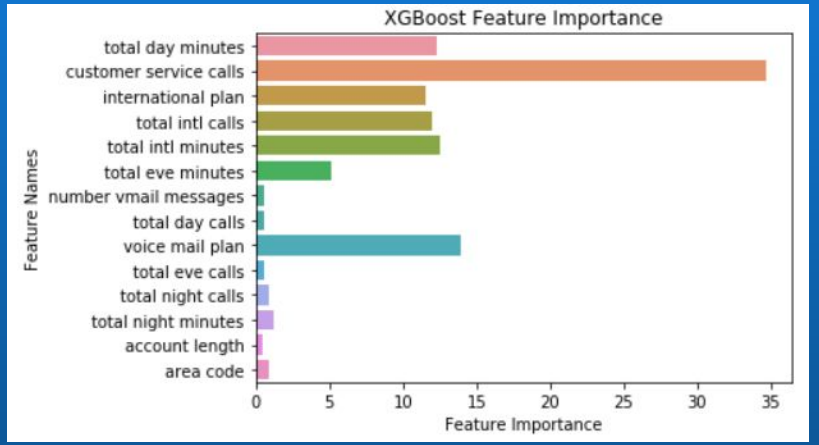
```
Training Precision:  0.07552392249901146
Testing Precision:  0.07571214392803598
Training Recall:  0.5
Testing Recall:  0.5
Training Accuracy:  0.15104784499802293
Testing Accuracy:  0.15142428785607195
Training F1-Score:  0.13122638268636208
Testing F1-Score:  0.13151041666666666
The Confusion Matrix is:
 [[101   0]
 [566   0]]
```

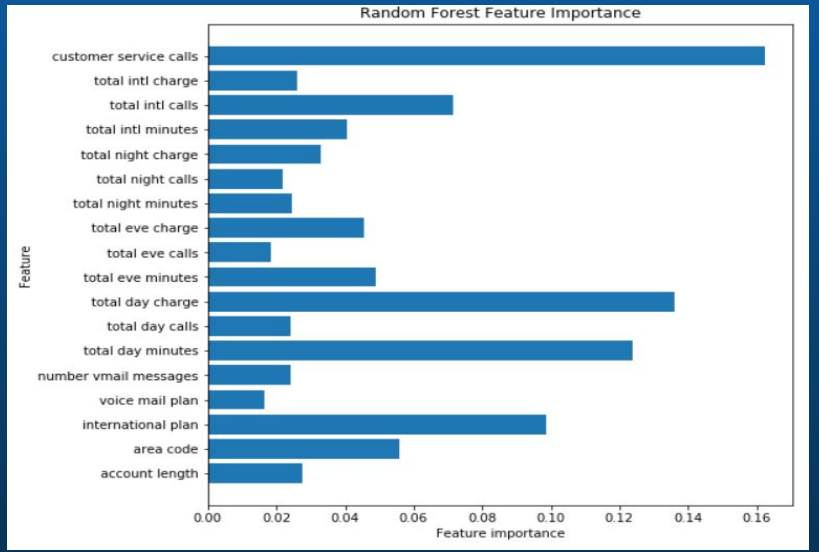
Final Model:

- The first thing we need to address from the base model is the class imbalance. I used Tomek Links to help establish well defined clusters and then proceeded to use SMOTE to generate new values inside of the minority class cluster. This helps us balance our training data to help the model interpret it better. There was no need to use SMOTE with XGBoost because it actually affected the accuracy of the model.
- In our case we want to focus on increasing our recall score because we want the most amount of True Positives and also increasing our False Positives does not negatively impact our end goal.

Final Model Continued:

The most important features for our final models are shown here. The main takeaway is that customer service calls and account plans like voice mail and international tend to have the biggest influence on the churn rate. After that total minutes used and total charge rate tend to have the second biggest impact.



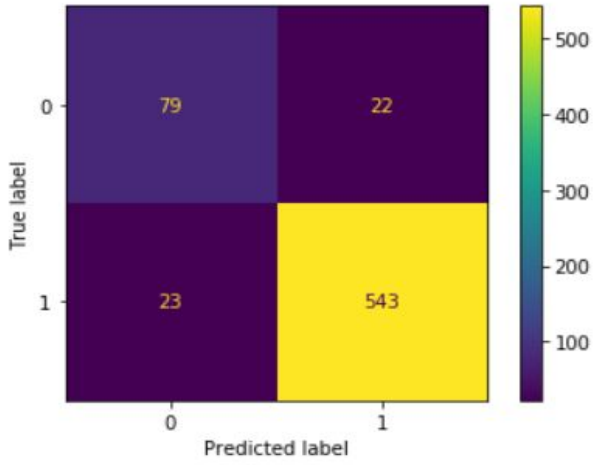


Results:

Compared to the base model, we were able to correct the class imbalance issue and increase our Recall score significantly. With Random Forest we also fixed the overfitting Tree classifiers tend to have with the training data.

Random Forests

```
Training Precision:  0.9794661915886931
Testing Precision:  0.8677858754121117
Training Recall:  0.9792993630573248
Testing Recall:  0.8707710877094776
Training Accuracy:  0.9792993630573248
Testing Accuracy:  0.9325337331334332
Training F1-Score:  0.9792975622226932
Testing F1-Score:  0.8692686623721106
```



XGBoost

```
Training Precision:  0.9723945316926821
Testing Precision:  0.9157864334048522
Training Recall:  0.8862701395103847
Testing Recall:  0.8525522163523773
Training Accuracy:  0.9640173981810992
Testing Accuracy:  0.9430284857571214
Training F1-Score:  0.9229821376281113
Testing F1-Score:  0.8802173913043478
The Confusion Matrix is:
 [[ 73  28]
 [ 10 556]]
```

summary/conclusions

In order for SyriaTel to improve their customer churn rate. They should offer:

- More plans like their international calling plan
- If a customer calls customer service more than 3 times, SyriaTel should look into maybe giving that customer more attention or better deals to prevent them from leaving.
- If they suspect a customer might be leaving soon, they could offer them a special promotional deal to hold on to the customer.

(if applicable) next steps

- With more time, I would like to explore the relationship between customer churn and time of day usage. I believe with this information SyriaTel could create certain plans to maybe offer free minutes at night or first x amount of minutes are free.
- Also I believe with more data on people leaving could also greatly increase the accuracy of the model.

Thank you

- Thank you to Yish and my cohort mates for all the help and guidance.
- Sources used: towardsdatascience.com, medium, stackoverflow, and statquest
  - https://towardsdatascience.com/a-beginners-guide-to-xgboost-87f5d4c30ed7 (This article really helped me with XGBoost)
  - https://towardsdatascience.com/how-does-xgboost-work-748bc75c58aa