Analysis on Dataset for Customer Churn

# Members

Steven Connolly, sconnolly@ryerson.ca   
Sangwoo Chun, sangwoo.chun@ryerson.ca

**Summary:**

**Business Problem**

Customer churn (or customer attrition) is the loss of a customer for a company. The problem our team is faced with is trying to predict the probability of churn for a company. Based on the dataset provided by the company, as data scientists,our goal is to build a predictive model that will allow us to predict which customers are likely to churn. By using statistical methods of analysis and forecasting, the company will be able to leverage this information to gain a much more accurate insight to their KPIs as well as to conduct other analysis such as predicting monthly or annual recurring revenue.

In order to conduct exploratory data analysis and modelling, we used tools such as Python, SciKit-Learn, Naïve Bayes, Decision Trees and Random Forest. During our modelling process, class imbalance was found and corrected using SMOTE. It appears that Decision Tree and Random Forests provided the best results.

In the end, we believe that the Decision Tree is the most suitable classifier to use as it is a supervised machine learning method. We achieved an accuracy (weighted) of 95%. Our accuracy in predicting TRUE (will churn) was 81% after removing attributes that either had high correlation with one another or had no bearing on customer churn.

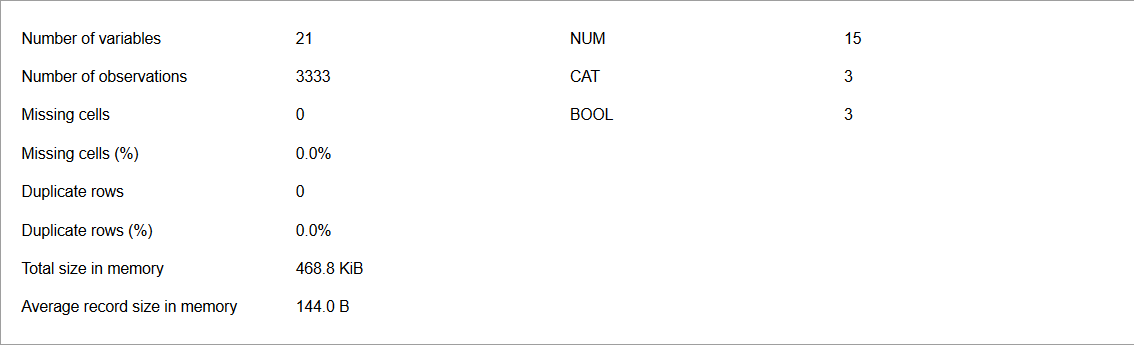
Based on our analysis, our recommendation for the company is to offer voicemail plans, better international plans, and cheaper daytime charges. Also, customer service should be improved and made more accessible as customers who never contact customer service are more likely to churn.

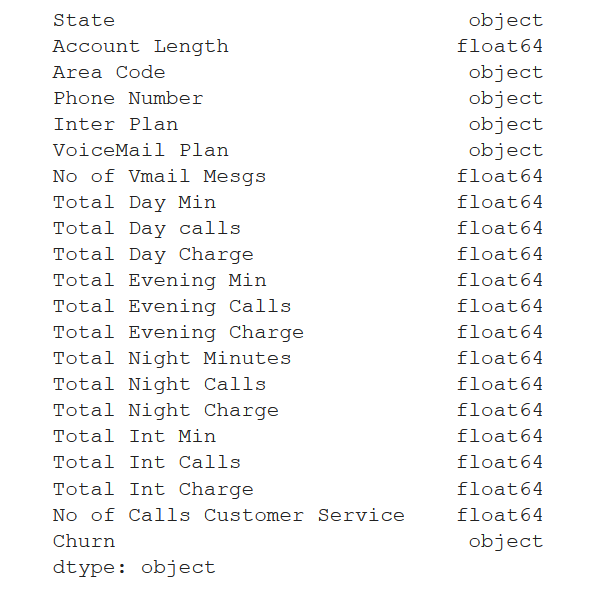
# Workload Distribution

|  |  |
| --- | --- |
| Member Name | List of Tasks Performed |
| Steven Connolly | Selecting dataset, data preparation, predictive modeling, post-predictive analysis, conclusion and recommendations, write-up |
| Sangwoo Chun | Selecting dataset, data preparation, predictive modeling, post-predictive analysis, conclusion and recommendations, write-up |

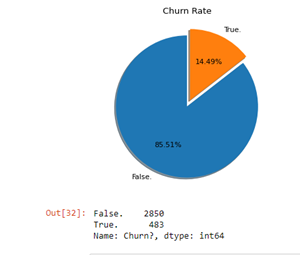
# Data Preparation

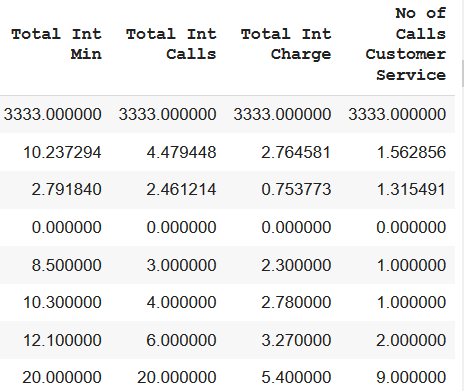
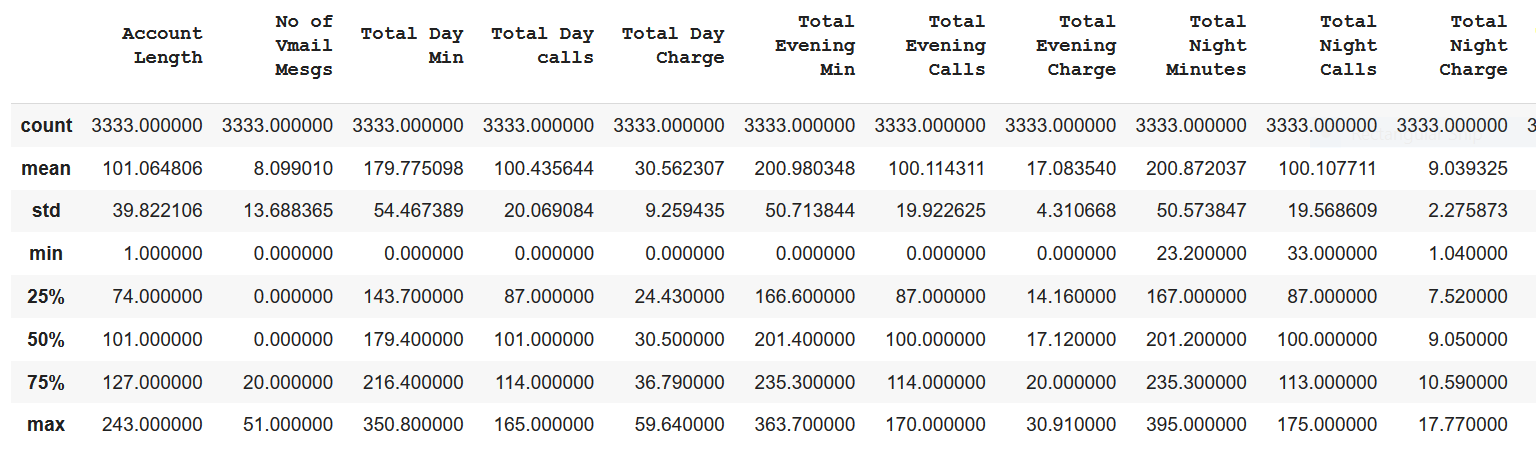
**Customer Churn Data Set (See Appendix for Data Dictionary)**

The data set that is being analyzed consists of 21 attributes which includes, numeric, categorical, and binary class attribute indicating whether a customer churned or not. There were no duplicate rows or missing values found. 



There are 15 numerical, 5 categorical attributes and one class in this dataset. Out of 3333 records, 2850 were classified as ‘FALSE’ and 483 were classified as ‘TRUE’. This class is imbalanced and may impact the accuracy of the dataset.



The numerical attributes were analyzed in an attempt to find data inconsistencies. There weren’t any notable outliers so none were removed. The max/min, mean and std deviation were found for all numerical attributes to detect any outliers.

**Analyzing the Distribution of Numeric Attributes**

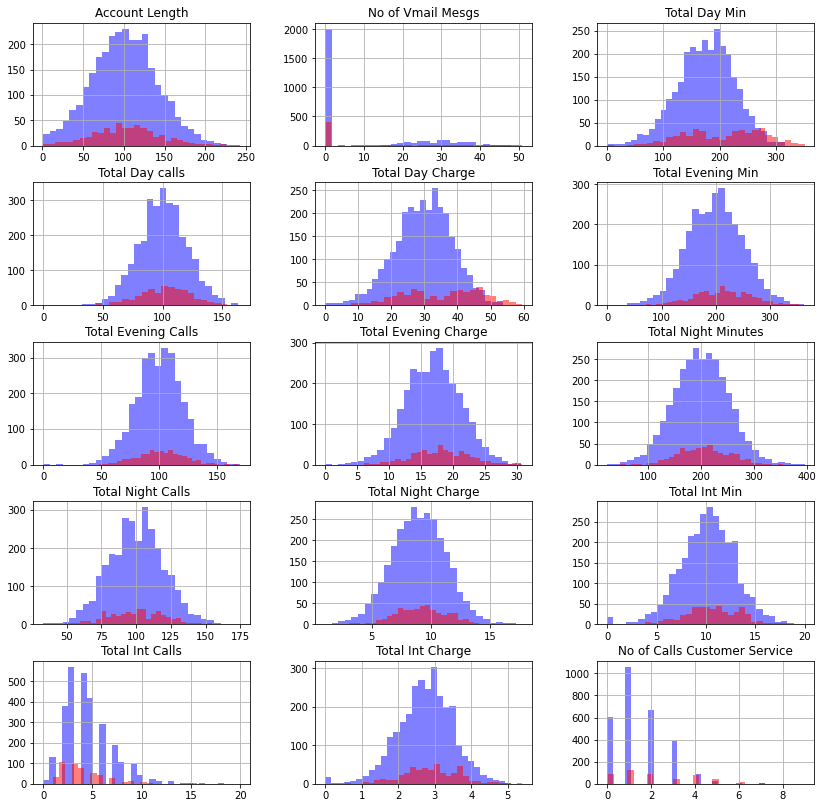
The histogram of all numeric attributes were compared to the churn class to find irregularities and potential correlations. Based on the figure below, most numerical attributes are normally distributed with the exceptions being Total International Calls, No of Calls to Customer Service, and No of VoiceMail Messages. As seen above, International Calls and Customer Service Calls fields are right-tailed. It also appears that the majority of customers make 0 voicemail messages. Looking at the two right-tailed histograms, (International Calls and Customer Service Calls), it can be seen that the frequency is exponentially decreasing.

Early round analysis shows that Total Day Mins, Total Day Charge, number of calls to customer service, and number of voicemail messages can be strong signals for customer churn.

It’s important to note that Total Day Min and Total Day Charge are correlated variables. That said, we can see that customers who have greater than average day time phone minutes have a higher tendency to churn.

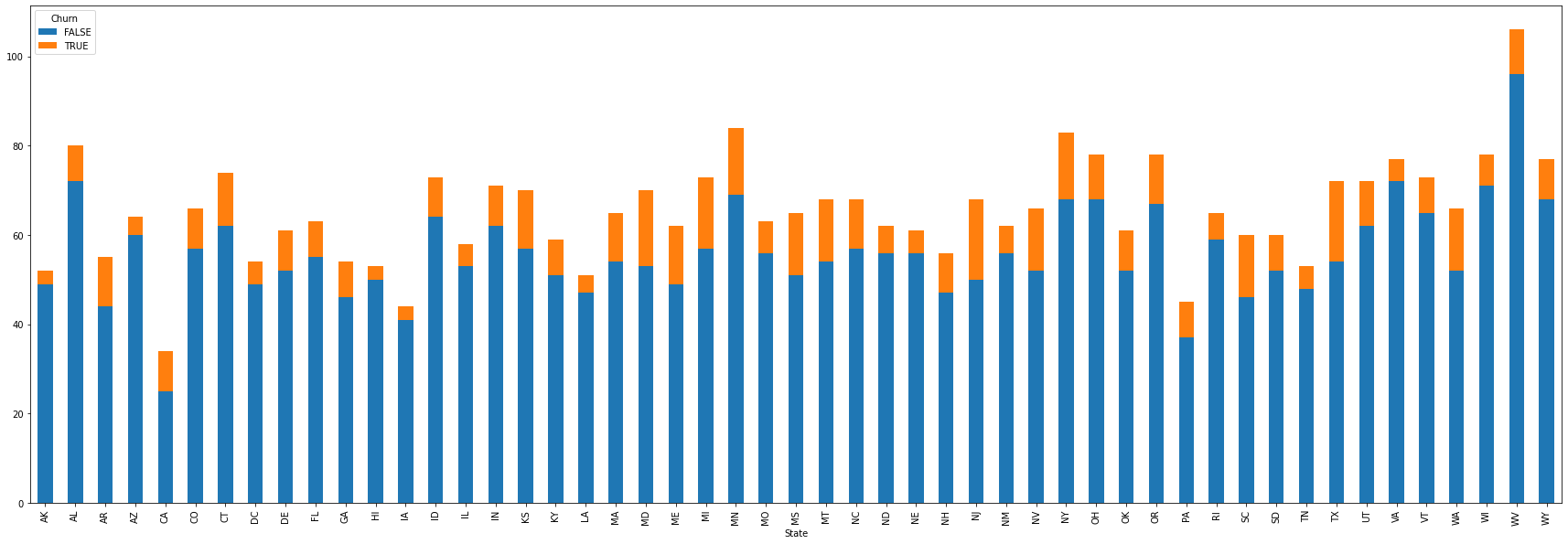
For customer service calls, there’s almost an even distribution of churners regardless of the number of times they call customer service (including zero). This might suggest that the company has a weak customer service structure to accommodate any issues.

For the number of voicemail messages, there’s an overwhelming number of churners among those who have zero voicemails. This could suggest that this is an unnecessary plan for churners.

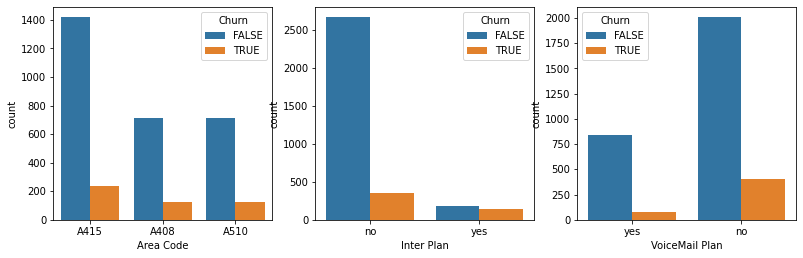


**Analyzing the Distribution of Numeric Attributes**

There are 4 categorical attributes that must also be analyzed (State, Area Code, International Plan, and Voicemail Plan).

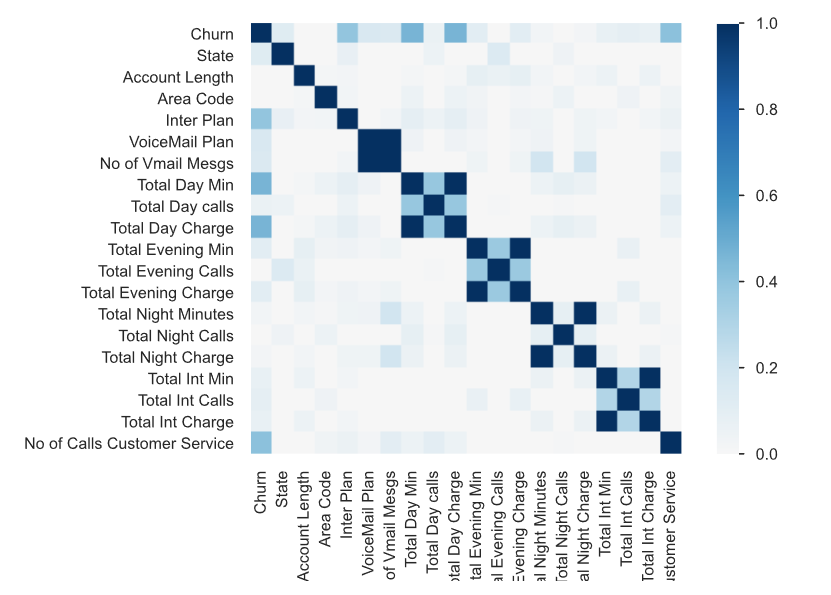


It would appear that the STATE the customer lives in has no bearing on customer churn. As a result this attribute isn’t needed. Also, phone numbers are all unique and even if correlation was discovered, it wouldn’t be accurate and should be removed.



Area Code has no bearing on customer churn and should be removed to improve the model. Inter Plan and VoiceMail Plan have significance. A customer is more likely to churn if they have an Inter Plan. A customer is also more likely to churn if they don’t have a VoiceMail Plan. Most customers have no Inter or Voicemail Plan.

**Correlation of Attributes**

****

There are several attributes that are highly correlated with one another. They will be removed and the model tested later.

# Predictive Modeling/Classification

**Data Split Strategy**

The data was split into a 25% test and 75% training set with random sampling. The training and test set were then compared against one another.

**Applying Classification Algorithm**

Three models were chosen to perform the classification algorithm: Naïve Bayes, Decision Tree and Random Forest. The criteria for weighing results was the accuracy of the TRUE (churned) more so than the FALSE (not churned). We are looking for a good FN/TN rate.

***Naïve Bayes***

The Naïve Bayes Classifier is based on conditional probabilities. It will consider each attributes of a record independently when attempting to determine the class of that record. Thus, rather than developing an archetype of attributes for a given class and then testing each record against the collection, Naïve Bayes will evaluate each attribute separately. The underlying theory of Naïve Bayes is that each attribute contributes independently toward the class of a sample record. There is a correlation between the differing attributes so Naïve Bayes will perform poorly.

***Decision Tree***

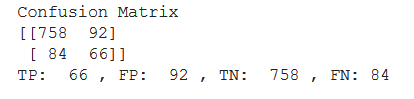
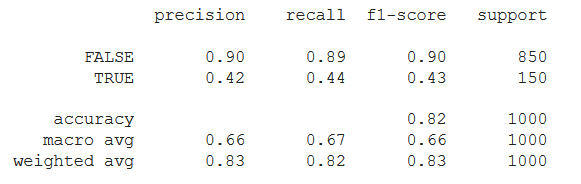
6 objects per leaf was used. This prunes the tree after it has been built and attempts to increase accuracy. It was found that 6 provided the best results.

***Random Forest***

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

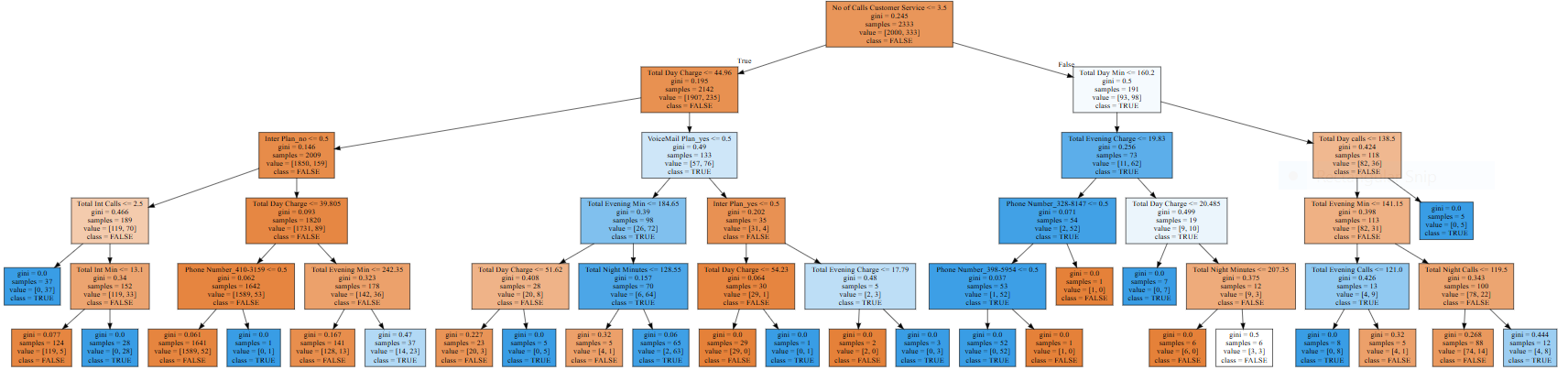
**Summary of Modeled Results**

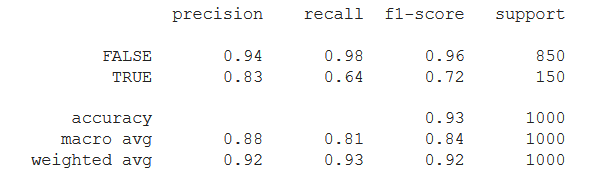
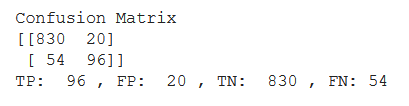
**Naïve Bayes**

. 

Naïve Bayes performed adequately with the original dataset. Recall on True is only 44% and must be improved.

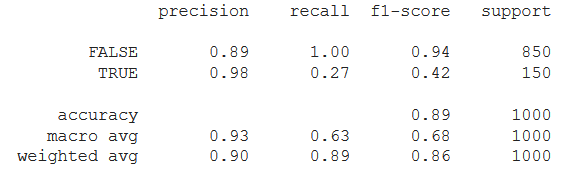
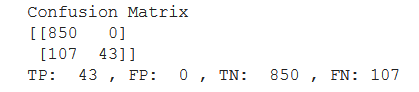
**Decision Tree**

****

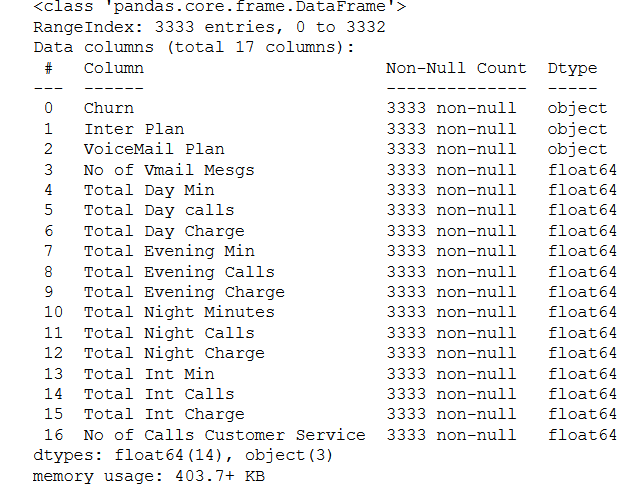


Decision Trees performed well with greater scores. Still far too many false negatives. Type2 errors are still too high.

**Random Forest**

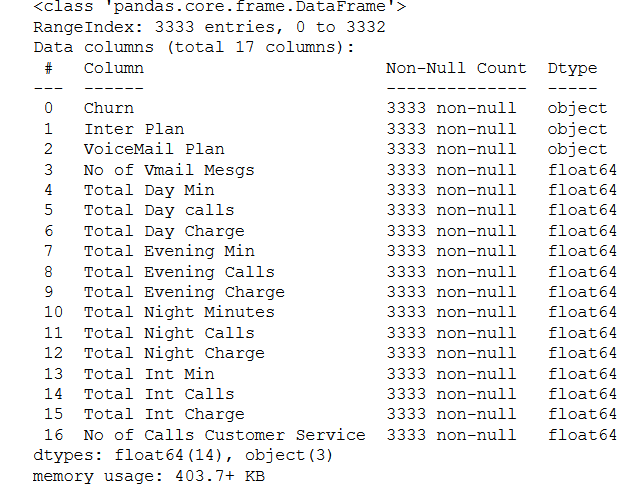


Random Forest did poorly in predicting Churn. The class imbalance seems to be affecting it. This was corrected and some attributes were removed manually.

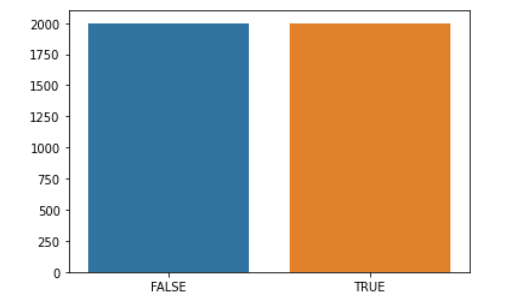
**

**Eliminating Attributes for Dataset #2**

The attributes such as State, Area Code and Phone Number and Account Length were manually eliminated. This is because these attributes are not significant contributors to customer churn.

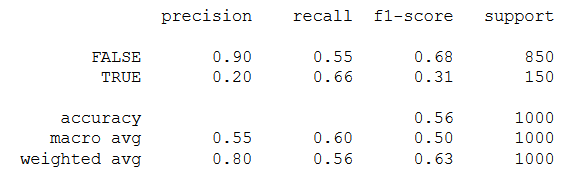
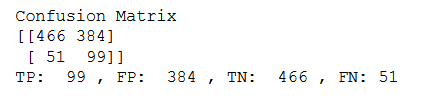


The class imbalance was then corrected using SMOTE.



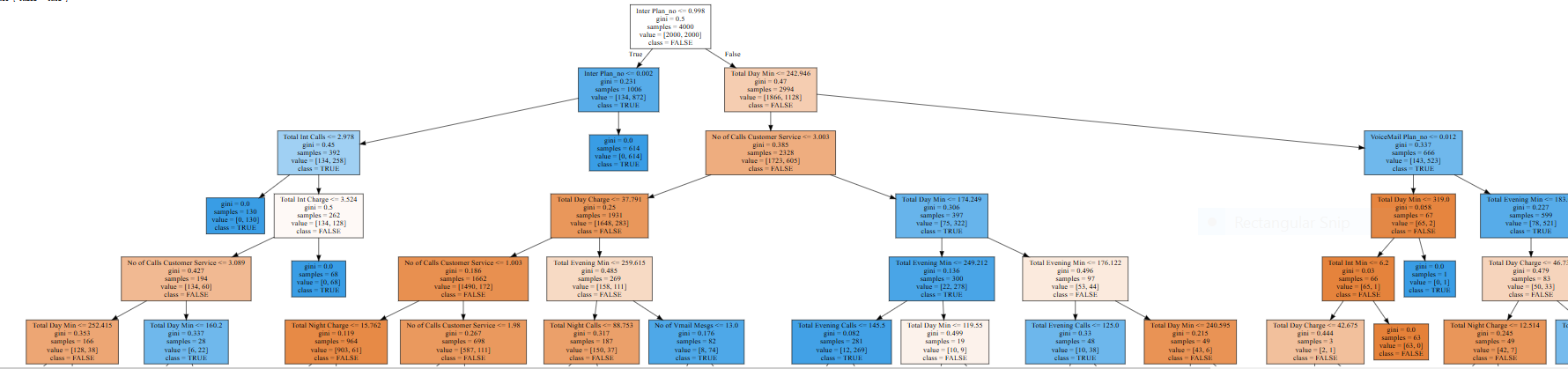
SMOTE was used on the training set ONLY while the test set remained unchanged. The 3 models were tested again to see if accuracy on type2 was improved.

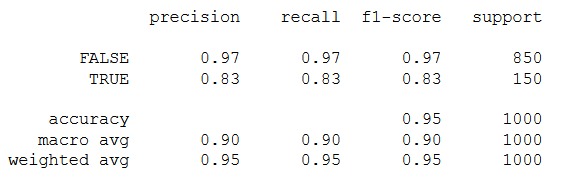
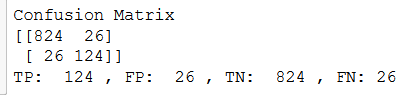
**Naïve Bayes on Dataset #2**

****

There are too many numerical attributes. The attributes are correlated so Naïve Bayes performed poorly and will not be used going forward. When SMOTE was applied to the training set and not the test set, it had a negative effect on Naïve Bayes.

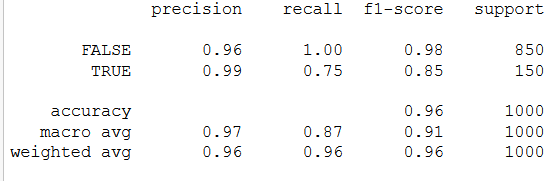
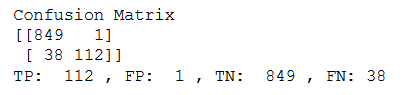
**Decision Tree on Dataset #2**

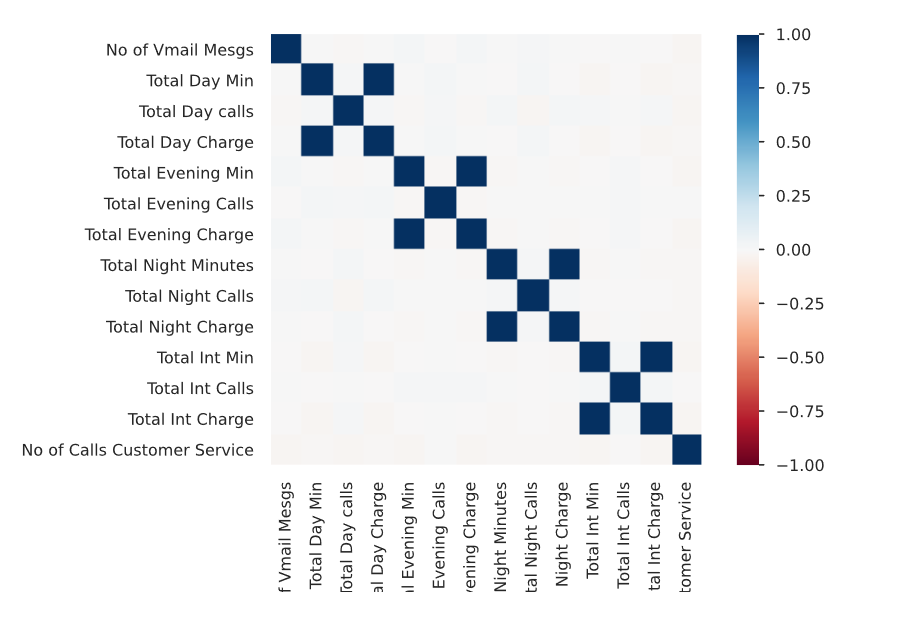
****



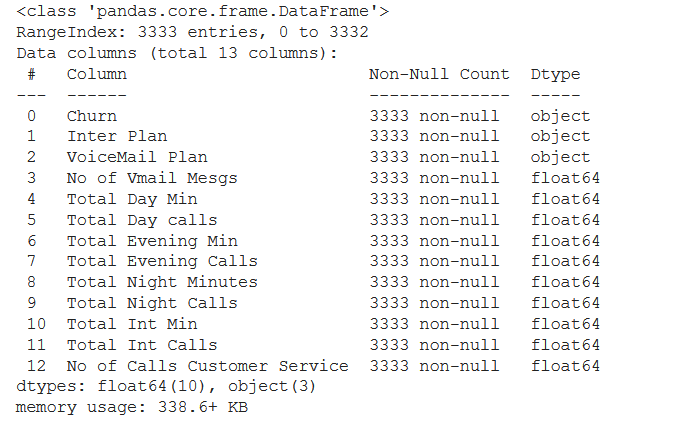
Decision Tree performed really well here and it is our best model. Accuracy for Churn is much higher. This is a good data model.

**Random Forest on Dataset #2**

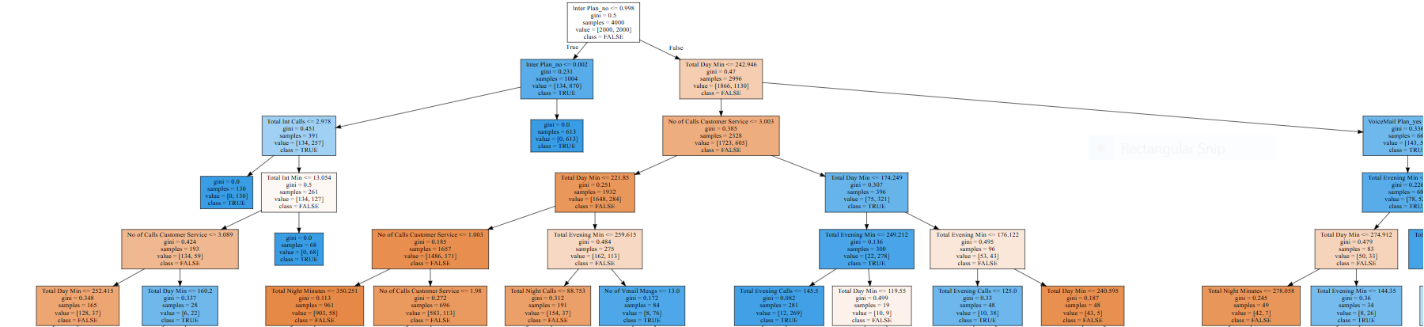


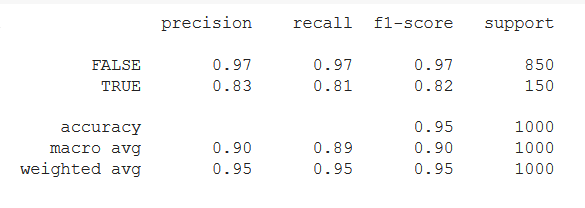
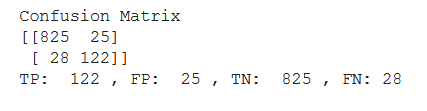
Random Forest is much improved as well, but accuracy for churn is lower than Decision Tree.

There are still some highly correlated attributes. They might interfere with the model and will be removed.

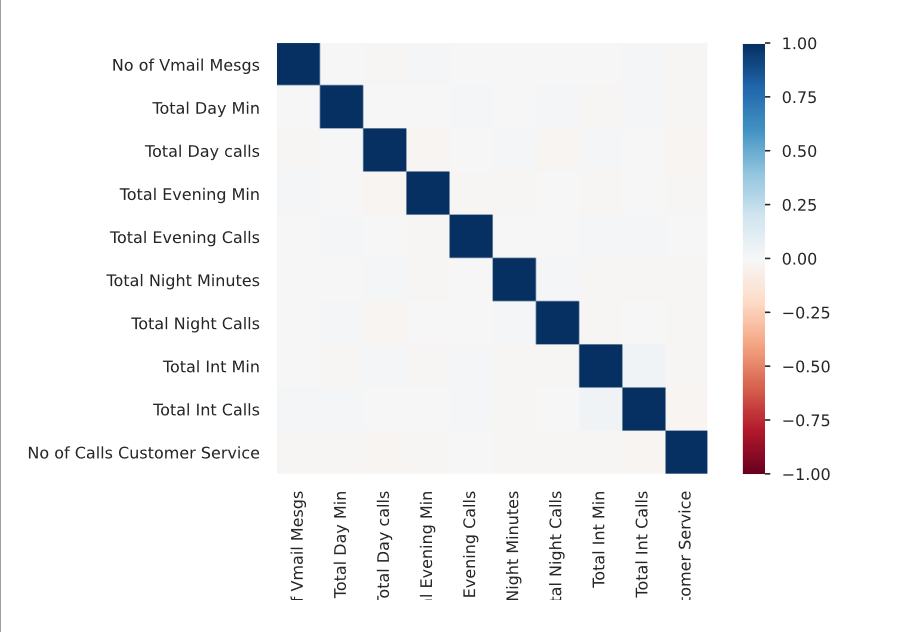


**Decision Tree for Dataset #3**

****

****

Accuracy for True (Churn) is slightly reduced but we have more confidence in this model with the removal of highly correlated features.



# Post-prediction Analysis

The 5 most important attributes appear to be International Plan, No of Calls to Customer Service, Total Day Min, Voicemail Plan. Therefore, it can be guessed that the eliminated attributes have less churn prediction capabilities and are not likely linked to the risk of customer churning. The Decision Tree model performed well. Other attributes were removed in an attempt to improve accuracy but these were the best results.

# Conclusions and Recommendations

Based on the results from our EDA and predictive modelling analysis, the 10 attributes chosen for the model are the best predictors for churn.

Based on our correlation analysis between the class variable and the 10 signal variables, customers were likely to churn when the number of customer service calls was low.  There was also a higher churn rate observed with higher day charge.

Our recommendation is that the company should encourage their customers to consider signing up for a voicemail plan as customers without Voicemail are more likely to leave. It is also recommended that the Inter Plan is improved as it appears to be causing churn.

Our recommendation to the company is to encourage its customers to use the company’s voicemail feature. Whether the company decides to include that as part of their plan or lower the cost of the voicemail plan is up to the company’s discretion. Another addon plan feature to consider is improving the company’s international plans. Whether it’s the quality of the company’s connection or the price of international plan, we recommend the company to review their addon plan structure.

The company should also invest in improving their customer service department. Our study shows that there were many customers who churned without even making a single call to customer service. This could be due to the lack of customer service presence or the quality of the customer service was poor.

In the analysis, it was discovered that if customers do call customer service it is typically 1 time. Another suggestion can be that the first time a customer contacts customer service, they can be offered a new/discounted voicemail plan or international plan in hopes of encouraging them to stay. Day charges could also be reduced. In order to find out other possible concerns of customers, the calls to customer service can be recorded for the purpose of finding out other concerns of customers that may be causing them to churn.