CUSTOMER CHURN PREDICTION MODEL AND ANALYSIS IN TELECOMMUNICATIONS

Sanaz Eghbali, sanaz.eghbali@ryerson.ca

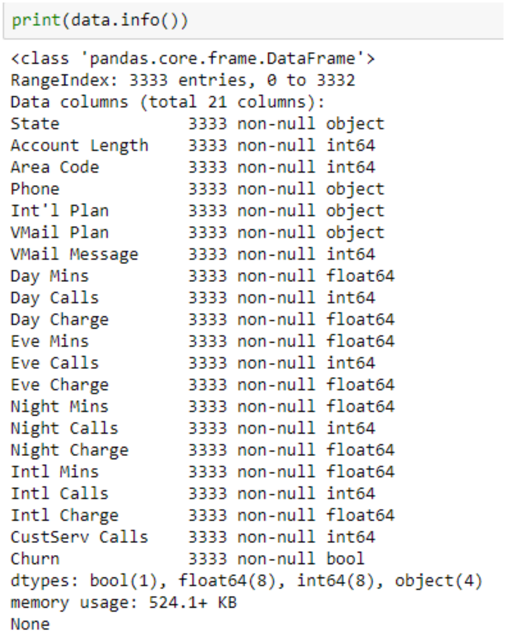
Contents

Summary

Customer churn rate is a critical performance indicator for subscription-based businesses like the telecommunications industry. The telecommunications company in this report would like to know in advance which of their customers will churn in the near future and would like recommendations for strategies to reduce the likelihood that these customers churn. Statistical analysis and data mining techniques were employed in conjunction with the dataset provided by the company to build three different predictive models for customer churn, using the Naïve Bayes, J48 Decision Tree, and Random Forest algorithms. These models were then tested against the same test dataset, yielding accuracies of 82.7%, 93.9%, and 96.9% respectively. The model generated by the Random Forest algorithm was ultimately selected based on the criteria of having the lowest False Positive Rate, 0.7% compared to 19.3% and 1.6% for the Naïve Bayes and J48 Decision Tree models respectively. In addition to classification, the dataset was divided into clusters of relevant customer attributes and behaviours and association rules were generated to determine the characteristics of customers likely to churn. The analysis of the predictive model shows that customers likely to churn fall into three main clusters based on the overall accumulation of charges and calls to customer service. The recommendations include the company responding proactively to offer retention packages to each cluster of customers before they contact customer service, with each package custom-tailored to increase the evening minutes, day minutes, and international or voicemail plan availability.

Data Preparation

I started our analysis by understanding the dataset by looking at the attribute types, attribute summaries, and checking for any missing values. This was done by opening the CSV file in both MS Excel and Weka. As shown in FIGURE 1; I observed 21 attributes with two types: categorical and numerical, and no missing values. FIGURE 2 demonstrates the histograms for attributes' distribution. For better data visualization I decided to discretize Account Length, Area Code, Number of voicemail messages, Day/Evening/Night/Int calls, and Number of Customer Service Calls. I could also see that Phone Number is a unique identifier for every entry and will not allow any generalization, so it can be dropped from our analysis.



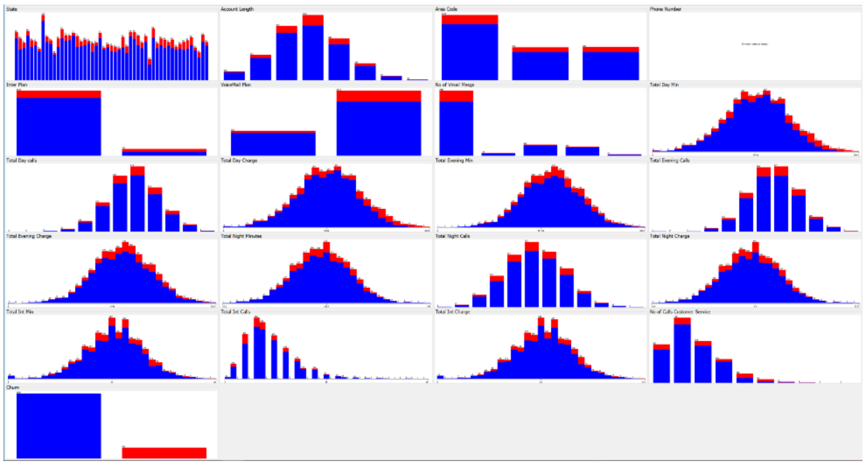
```
print(data.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3333 entries, 0 to 3332
Data columns (total 21 columns):
State             3333 non-null object
Account Length    3333 non-null int64
Area Code         3333 non-null int64
Phone             3333 non-null object
Int'l Plan        3333 non-null object
VMail Plan        3333 non-null object
VMail Message     3333 non-null int64
Day Mins          3333 non-null float64
Day Calls         3333 non-null int64
Day Charge        3333 non-null float64
Eve Mins          3333 non-null float64
Eve Calls         3333 non-null int64
Eve Charge        3333 non-null float64
Night Mins        3333 non-null float64
Night Calls       3333 non-null int64
Night Charge      3333 non-null float64
Intl Mins         3333 non-null float64
Intl Calls        3333 non-null int64
Intl Charge       3333 non-null float64
CustServ Calls    3333 non-null int64
Churn             3333 non-null bool
dtypes: bool(1), float64(8), int64(8), object(4)
memory usage: 524.1+ KB
None
```

FIGURE 1. Data types and summary

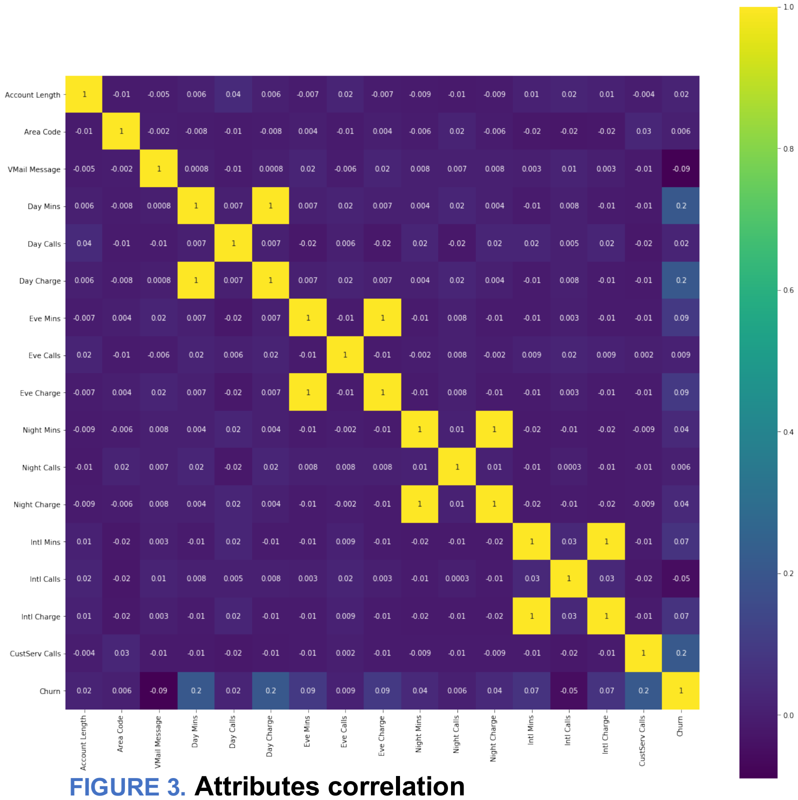
	Account Length	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge
count	3333	3333	3333	3333	3333	3333	3333	3333	3333	3333	3333	3333	3333
mean	101.0	179.7	100.4	30.56	200.9	100.1	17.08	200.8	100.1	9.04	10.24	4.48	2.76
std	39.82	54.47	20.07	9.26	50.71	19.92	4.31	50.57	19.57	2.28	2.79	2.46	0.75
min	1	0	0	0	0	0	0	23.2	33	1.04	0	0	0
25%	74	143.7	87	24.43	166.6	87	14.16	167	87	7.52	8.5	3	2.3
50%	101	179.4	101	30.5	201.4	100	17.12	201.2	100	9.05	10.3	4	2.78
75%	127	216.4	114	36.79	235.3	114	20	235.3	113	10.59	12.1	6	3.27
max	243	350.8	165	59.64	363.7	170	30.91	395	175	17.77	20	20	5.4

FIGURE 2. Attributes distribution



When I looked at the correlation of attributes as shown in FIGURE 3, I found high correlations between several attributes and Churn, our class attribute for the dataset. This high correlation of explanatory variables is an indication of multicollinearity and since all Calls and Mins are subsets of Charges, I dropped all Calls and Mins and used Charges instead. Similarly, Voicemail Plan was chosen over Number of Voicemail messages and Area Code was chosen over State. As one of our goals is to create a prediction model for customers likely to churn, I also investigated the percentage of instances where Churn was TRUE. This was found to be only 14.5%, as shown in FIGURE 4. Though a relatively high churn rate by modern industry standards (Canadian telecommunications companies routinely boast monthly churn rates of about 1-2%) and an obvious problem for the company in this instance, this low percentage is an indication of an imbalanced dataset and makes churned customers the minority class by a fair margin. To overcome this issue for the purpose of generating a predictive algorithm, I used the resample filter in Weka to simulate the Oversampling technique, which duplicates random records from the minority class. This oversampling allowed us to have more instances which reflected the minority class in the data sample used to train and test the different predictive algorithms selected.

```
weka.filters.supervised.instance.Resample -B 1.0 -Z 171.3
```



	Account Length	Area Code	VMail Message	Day Mins	Day Calls	Day Charge	Eve Mins	Eve Calls	Eve Charge	Night Mins	Night Calls	Night Charge	Intl Mins	Intl Calls	Intl Charge	CustServ Calls	Churn
Account Length	1	-0.01	-0.005	0.006	0.04	0.006	-0.007	0.02	-0.007	-0.009	-0.01	-0.009	0.01	0.02	0.01	-0.004	0.02
Area Code	-0.01	1	-0.002	-0.008	-0.01	-0.008	0.004	-0.01	0.004	-0.006	0.02	-0.006	-0.02	-0.02	-0.02	0.03	0.006
VMail Message	-0.005	-0.002	1	0.0008	-0.01	0.0008	0.02	-0.006	0.02	0.008	0.007	0.008	0.003	0.01	0.003	-0.01	-0.09
Day Mins	0.006	-0.008	0.0008	1	0.007	1	0.007	0.02	0.007	0.004	0.02	0.004	-0.01	0.008	-0.01	-0.01	0.2
Day Calls	0.04	-0.01	-0.01	0.007	1	0.007	-0.02	0.006	-0.02	0.02	-0.02	0.02	0.02	0.005	0.02	-0.02	0.02
Day Charge	0.006	-0.008	0.0008	1	0.007	1	0.007	0.02	0.007	0.004	0.02	0.004	-0.01	0.008	-0.01	-0.01	0.2
Eve Mins	-0.007	0.004	0.02	0.007	-0.02	0.007	1	-0.01	1	-0.01	0.008	-0.01	-0.01	0.003	-0.01	-0.01	0.09
Eve Calls	0.02	-0.01	-0.006	0.02	0.006	0.02	-0.01	1	-0.01	-0.002	0.008	-0.002	0.009	0.02	0.009	0.002	0.009
Eve Charge	-0.007	0.004	0.02	0.007	-0.02	0.007	1	-0.01	1	-0.01	0.008	-0.01	-0.01	0.003	-0.01	-0.01	0.09
Night Mins	-0.009	-0.006	0.008	0.004	0.02	0.004	-0.01	-0.002	-0.01	1	0.01	1	-0.02	-0.01	-0.02	-0.009	0.04
Night Calls	-0.01	0.02	0.007	0.02	-0.02	0.02	0.008	0.008	0.008	0.01	1	0.01	-0.01	0.0003	-0.01	-0.01	0.006
Night Charge	-0.009	-0.006	0.008	0.004	0.02	0.004	-0.01	-0.002	-0.01	1	0.01	1	-0.02	-0.01	-0.02	-0.009	0.04
Intl Mins	0.01	-0.02	0.003	-0.01	0.02	-0.01	-0.01	0.009	-0.01	-0.02	-0.01	-0.02	1	0.03	1	-0.01	0.07
Intl Calls	0.02	-0.02	0.01	0.008	0.005	0.008	0.003	0.02	0.003	-0.01	0.0003	-0.01	0.03	1	0.03	-0.02	-0.05
Intl Charge	0.01	-0.02	0.003	-0.01	0.02	-0.01	-0.01	0.009	-0.01	-0.02	-0.01	-0.02	1	0.03	1	-0.01	0.07
CustServ Calls	-0.004	0.03	-0.01	-0.01	-0.02	-0.01	-0.01	0.002	-0.01	-0.009	-0.01	-0.009	-0.01	-0.02	-0.01	1	0.2
Churn	0.02	0.006	-0.09	0.2	0.02	0.2	0.09	0.009	0.09	0.04	0.006	0.04	0.07	-0.05	0.07	0.2	1

FIGURE 3. **Attributes correlation**

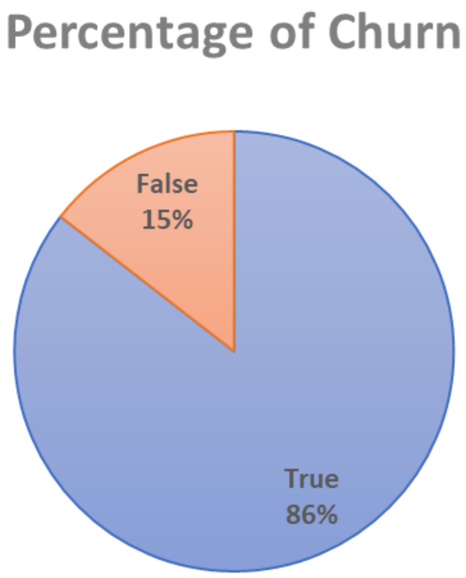


FIGURE 4. **Churn percentage**

Our next step was the selection of the appropriate attributes for our analysis. Attribute selection is vital for selecting the most relevant features of a dataset. In Weka, I used the Correlation Attributes Ranking Filter and Information Gain techniques to aid in attribute selection. In our investigation of the churn dataset, I selected only the top 5 features, each having high ranking values in the results of both techniques. Two machine learning methods were used for attribute selection to acquire the most relevant attributes; Information Gain entropy was used in decreasing order and Correlation Attributes Ranking Filter technique was used for selecting a subset of relevant features. The attribute ranking was employed to identify the factors and hidden patterns in data that are the main reasons of customers churning. The ranking values of Information Gain and Correlation Attributes Ranking Filter are shown in TABLE1 below. Since International Plan and International Charge functionally represent the same feature, Inter Plan was chosen due to its higher correlation ranking.

Attributes	**Correlation Attributes Ranking Values**	**Information Gain Ranking Values**
Account Length	0.01654	0
Area Code	0.00514	0.0000383
Inter Plan	0.25985	0.0368789
VoiceMail Plan	0.10215	0.0082165
No of Vmail Mesgs	0.08973	0.0082165
Total Day Charge	0.20515	0.0773975
Total Evening Charge	0.09279	0.0054209
Total Night Charge	0.0355	0
Total Int Charge	0.06826	0.0067401
No of Calls Customer Service	0.20875	0.0500934

TABLE 1. **Attributes selection**

Finally, with our attributes of interest selected, I conducted an investigation into outliers and extreme values for each. Using the InterquartileRange filter in Weka, I found that the only variable in the dataset that contained outliers or extreme values was the Number of Customer Service Calls. This makes sense as the histogram for the distribution of this variable is clearly skewed as can be seen clearly in the final frame of FIGURE 2. Though the higher number of customer service calls might be considered outliers and therefore could be considered to have an outsized influence on the results of the analysis, I also had to consider that this was one of the variables with the highest correlation with customer churn. By eliminating the instances with the highest values in the Number of Customer Service Calls, I would also inadvertently eliminate many of our instances where the value of Churn was TRUE. Since this dataset is already imbalanced, I did not want to move such a large chunk of out minority class instances if I could help it. I decided to run our predictive model generation twice, once on the data with outliers removed and once with them left in. All three algorithms selected, Naïve Bayes, J48 Decision Tree, and Random Forest, were better able to identify customers likely to churn based on the dataset with these Customer Service Call outliers left in.

Predictive Modeling (Classification)

As far as our class attribute is concerned, there are two types of customers in our dataset, the non-churn customers and churn customers. The proposed models target churn customers and attempts to identify the reasons behind their classification. Three machine learning techniques were used for classifying customers' data to assess which of the algorithms best classifies the customers into the churn and non-churn categories. The detailed accuracy of each method is shown in TABLE 2 below.

TABLE 2. **Accuracy of various algorithms**

Method	TP Rate	FP Rate	Precision	Recall	F-measure	ROC area
Naïve Bayes	82%	18%	82%	82%	82%	86%
Decision Tree	97%	3%	96%	97%	96%	95%
Random Forest	98.5%	1.5%	98.5%	98.5%	98.5%	99%

Naïve Bayes:

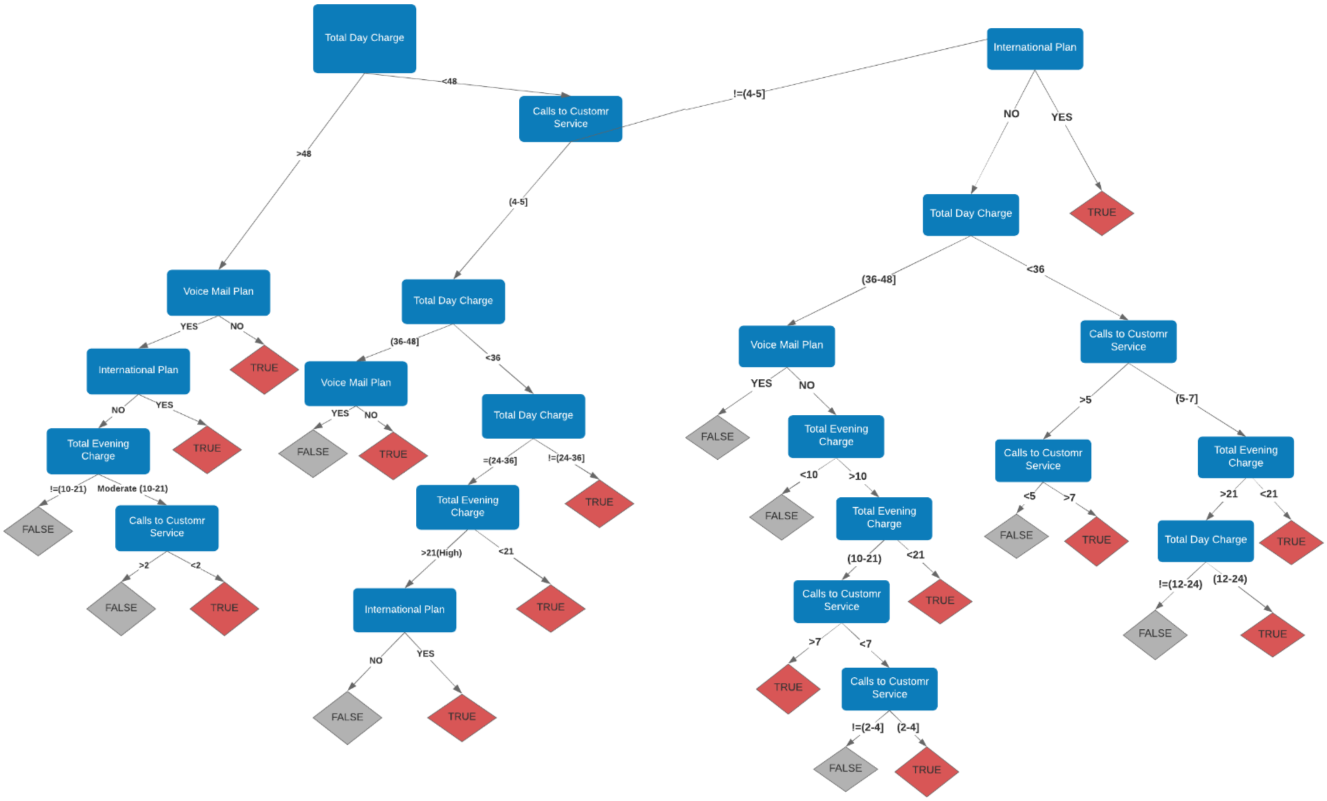
The Naïve Bayes classification algorithm was performed and evaluated using 10-fold cross validation on the entire dataset. The accuracy of the algorithm was found to be 82% but the Recall was also only 82%. This means that the possibility of a type II error was still quite high using this approach. Almost 20% of the time, a customer with a high probability of churn is incorrectly classified by the model as not likely to churn. Since I am interested primarily in correctly classifying customers likely to churn, and since this rate is comparatively high compared to the other resulting models, this model would likely not be best suited to our purposes.

Decision Tree (J48):

The Decision Tree classification algorithm was the second method used to construct a predictive model for classifying customers likely to churn. Using 10-fold Cross-Validation, a predictive model was constructed that resulted in a fairly high accuracy of 96% and an impressive recall of 99%. Furthermore, if I look into ROC Area and see that this model was able to distinguish 95% of the dataset correctly between customer churn and non-churn.

In order to build a clean Decision Tree in Weka, the numeric attributes were discretized into intervals, 10-fold Cross-Validation binary splits set to true. The tree generated using this approach can be visualized as seen in FIGURE 5 below.

FIGURE 5. **Decision Tree (J48)**

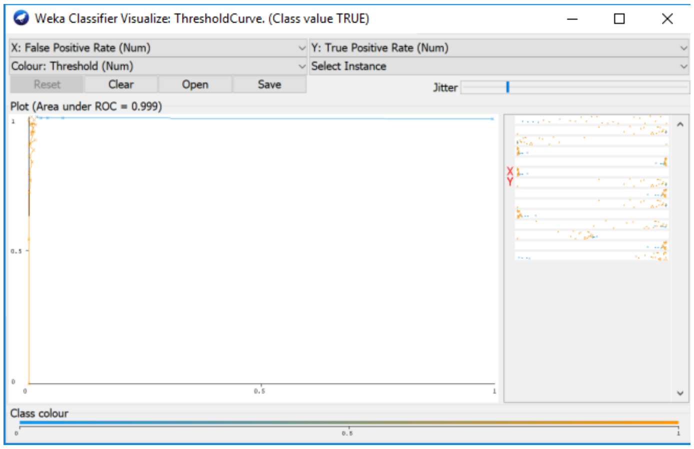


Random Forest:

Finally, a Random Forest classification algorithm was used to create a third predictive model. This model achieved the highest accuracy of the three at 98% with the 10-fold Cross-Validation. The sensitivity rate concluded that 100% of churn customers were correctly classified and the specificity rate concluded that 96% of non-churn customers were accurately classified. ROC Area was at its highest of the three models at 99%, as was F-measure at 98%.

Random Forest can handle nonlinear data efficiently and performs better if correlated features exist in the data. This method generates a series of decision trees using multiple algorithms, then averages the predicted results of each tree to obtain the overall prediction for each instance. In this instance the multiple, random decision trees of a Random Forest allowed for a more precise prediction model and produced more accurate predictive results.

FIGURE 6. Random Forest Threshold curve



Weka Classifier Visualize: ThresholdCurve. (Class value TRUE)

X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Plot (Area under ROC = 0.999)

I see that I fall under the excellent band ROC= .999 for the current model.

Ultimately, in order to compare and select the best classification model for our purposes, I needed to test each of our predictive models against the same test set. For this, I ran each of the models against the original dataset, without the oversampling present in the data used to construct the models. This was done in order to ensure I had not selected a model that had been designed to overfit our training sample data. TABLE 3 below represents the findings of the accuracy measures for each model using the same test set.

TABLE 3. Accuracy of models with test set

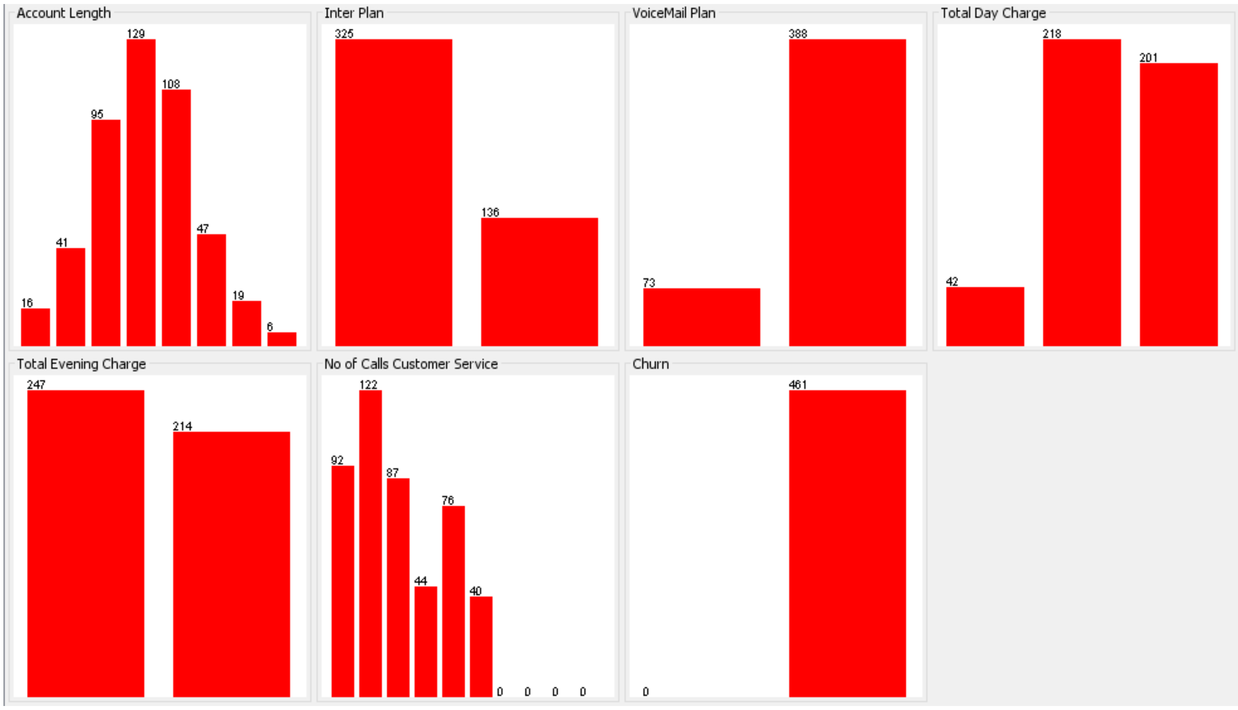
Method	TP Rate	FP Rate	Precision	Recall	F-measure	ROC area
Naïve Bayes	82.7%	19.3%	88.7%	82.7%	84.5%	86.1%
Decision Tree	93.9%	1.6%	95.6%	93.9%	94.3%	97.5%
Random Forest	96.9%	0.7%	97.5%	96.9%	97.1%	99.9%

Post-predictive Analysis

In order to determine which types of customers the company would benefit the most from targeting with retention strategies, I had to group the customers into similar clusters using the data on hand. Customer clusters are therefore used to partition the customers' data into groups based on their behavior information and their relationships. Additionally, I am only interested in the customers that churn, as this is the group I wish to target with retention strategies, so I partitioned the data by removing all the instances where Churn was FALSE. Then, I used the K-means technique on these remaining churn-only customers to segment the data into different groups, as shown in FIGURE 7 and used real-valued data to find a relationship and reveal patterns in the data which belong to one class.

Although not highly correlated to finding patterns in customer churn originally, to better understand the real-world business problem of when customers are leaving, I added the Account Length attribute to our cluster analysis. With 8 bins representing the month of the account, I can see that 70% of churn happens between 3-5 months and 62% of churn happens within 0-2 calls to customer service

FIGURE 7. **Churn attribute distribution**



I tested the K-means technique with 5 cluster numbers ranging from 2-6, as shown FIGURE 8. After careful observation of the Within cluster sum of squared errors, I decided to use K as 3 for the k-means algorithm to segment the data into three groups. The results are displayed in TABLE 4 below.

FIGURE 8. **K-means cluster analysis**

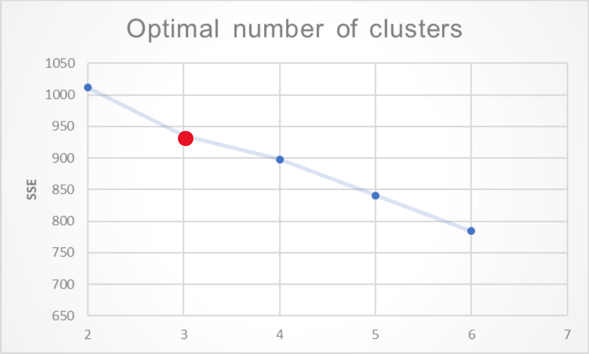


TABLE 4. **K-means cluster centroids**

Attribute	Cluster 1	Cluster 2	Cluster3
Account Length (months)	4	3	5
Inter Plan	No	Yes	No
Voicemail Plan	No	No	No
Total Day Charge	>40	(20-40]	(20-40]
Total Evening Charge	>18	>18	<18
No of Calls Customer Service	1	1	4

FIGURE 9 shows that of our resulting clusters, cluster 1 and cluster 3 have the highest rates of customer churn at 53% and 26% respectively. These two clusters are more valuable to the company to maximize the profit by retaining them. It’s interesting to see that the customers in both of these clusters have no Voice mail or International Plan. I know that 70% of churn happens between 3-5 months and it drops to 15% after 6-7 months. I also know that 29% of customers are unlikely to churn if they have a Voicemail plan; this is perhaps an opportunity to offer one-or more months of complementary Voicemail or International plan service within the 3rd month of the customer’s plan. Although this might not be effective due to other reasons customers choose to select these features at the outset of setting up their accounts that are not accounted for in this dataset, such as the need to keep in touch with family or business contacts in other countries or time zones, this could also prove to be an effective way to retain these customers.

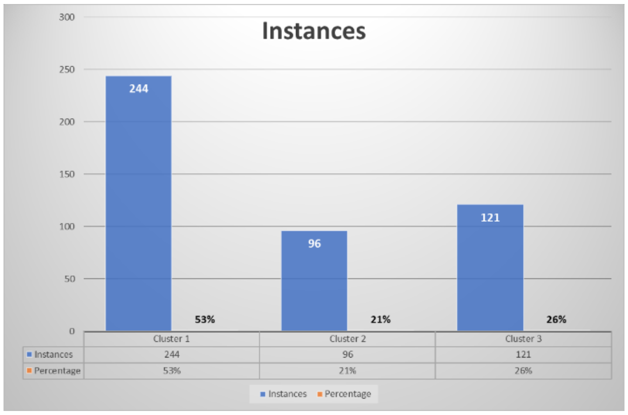


FIGURE 9. **Cluster instance analysis**

In order to get a better visualization of our clusters behaviour and characteristics I segmented Day Charges by Low, Medium, High and Evening Charges by Low and High, shown in FIGURE 10. I then segmented our clusters' probability of churn based on the customers' spending in FIGURE 11.

FIGURE 10. Charges churn rate

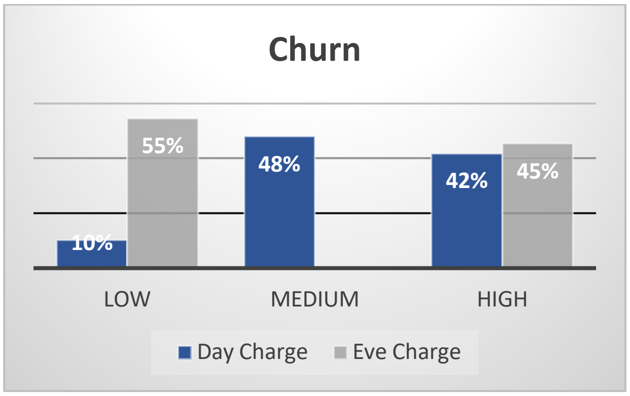
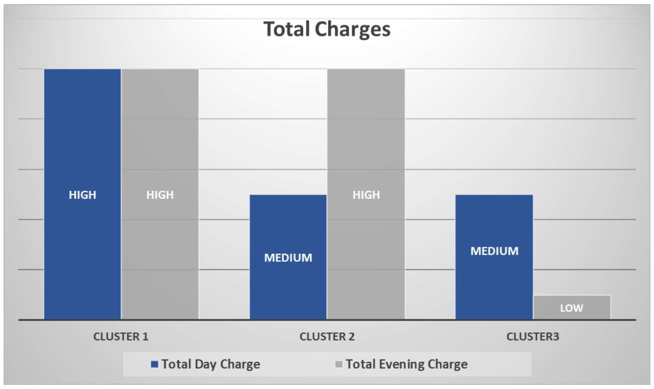
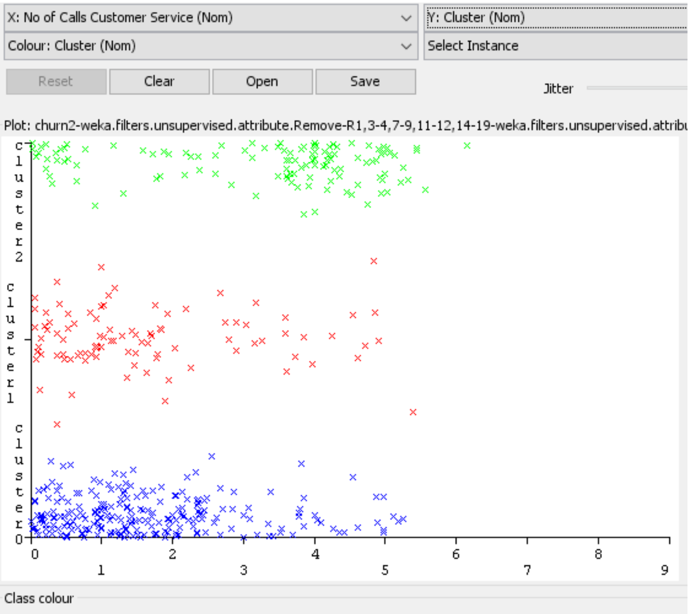


FIGURE 11. Clusters probability of Churn



To explore the number of customer service calls I visualized k-means clusters in FIGURE 12. As I can see 62% (0-2 calls) of churn lies within cluster 1 (Blue) and 24% (4-5 calls) within cluster 3 (Green).

FIGURE 12. Customer Service calls clusters



X: No of Calls Customer Service (Nom)

Y: Cluster (Nom)

Colour: Cluster (Nom)

Select Instance

Plot: churn2-weka.filters.unsupervised.attribute.Remove-R1,3-4,7-9,11-12,14-19-weka.filters.unsupervised.attribu

Next, I created customer profiles based on the behaviour they exhibited through the k-means algorithm to create retention policies for our recommendations, shown in TABLE 5.

TABLE 5. **Customer profiles**

Customer Profile	Churn Probability	Cluster
VIP • 4 months • 1 Customer service call	HIGH	1
AVERAGE-VALUE • 5 MONTHS • 4 Customer service calls	HIGH	3
HIGH-VALUE • 3 months • International Plan • 1 Customer service call	MEDIUM	2

Even though, cluster 1 and 3 have a high probability of churn, cluster 2 represents high-value customers, which is vital to the company's revenue stream, and efforts should also be made to retain this class. In order to create patterns based on our observations and finalize our recommendations I applied association rules with multiple support and confidence values using the Apriori algorithm in Weka. TABLE 6 shows the outcomes of the most frequent and logical patterns.

TABLE 6. **Association rules**

1	High Day Charge/Inter Plan=no/Voicemail Plan=no ==> Churn=TRUE 159 <conf:(1)>
2	Medium Day Charge/High Evening Charge ==> Churn=TRUE 140 <conf:(1)>
3	High Day Charge/ High Evening Charge/Voicemail Plan=no ==> Churn=TRUE 117 <conf:(0.97)>
4	High Day Charge/No of Calls Customer Service=1/Voicemail Plan=no ==> Churn=TRUE 63 <conf:(0.93)>
5	High Day Charge/No of Calls Customer Service=0/Voicemail Plan=no ==> Churn=TRUE 47 <conf:(0.98)>
6	Account Length=Month 4/High Day Charge/Voicemail Plan=no ==> Churn=TRUE 50 <conf:(0.96)>
7	Account Length=Month 3/High Day Charge/Voicemail Plan=no ==> Churn=TRUE 47 <conf:(0.96)>
8	Inter Plan=no/Medium Day Charge ==> Churn=TRUE 133 <conf:(1)>
9	Account Length=Month 5 ==> Churn=TRUE 108 <conf:(1)>

As can be seen, the probability of churn for customers without Voicemail and International plans and high value customers with voice mail and international plans is very high, so I can craft our recommendations around customers with no Voicemail or International Plan. Another pattern detected is the association between high value customers and the number of customer service calls.

Associations 2 and 3 above confirm the high churn probability for these high value customers and associations 4 and 5 confirm the high churn rate for customers service calls between 0-1. This can be interpreted as a sign to the company that it is necessary to act proactively, and make the customers an offer before the customers contact the customer service department themselves. This can be implemented in time using the pattern detected by association 6, high churn rate for high value customers during month 4. Finally, associations 6, 7, and 9 confirm our earlier observation that high churn is observed during the period of time from months of 3-5.

Conclusion and Recommendations

The customers most likely to churn can be broadly divided into three main clusters by overall charges incurred. The customers in each cluster exhibit slightly different behaviours when it comes to their tendency to contact the customer service department, what services they choose to subscribe to, and their typical account length before they churn. Because of this, each group can be dealt with using a custom-tailored approach to retention to meet the needs of these customers the best, without overspending by providing services the customers in that cluster are unlikely to require in order to be retained. Additionally, it should be noted that efforts at customer retention should likely be targeted at and designed to keep customers subscribed up until the sixth month of service, as the data shows a sharp drop off point for customer churn beyond this threshold. As such, our recommendations are aimed largely at getting existing customers to this point in their account length.

Strategies:

- Express customer service department for selected customers' profile
- Three Retention Packages tailored specifically for each individual profile

Rules:

Rule 1: If Account Length >= 4 & Day Charge >= 40 & Evening Charge >= 18 & VM =No & Inter Plan = No
THEN Class A: VIP

Rule 2: If Account Length >= 5 & Day Charge >= (20-40] & Evening Charge < 18 & Calls >= 4 & VM =No & Inter Plan = No
THEN Class B: AVG

Rule 3: If Account Length >= 3 & Day Charge >= (20-40] & Evening Charge >= 18 & Calls >= 1 & VM =No & Inter Plan = No
THEN Class C: High

1. Class A: VIP
    I. This class identifies the highest spenders or the VIP sector. These are the people who require lots of attention and love! When they contact customer service, they should transfer to the express line, so their problem is solved as soon as possible. The representatives should be notified by the system that the customer calling is classified as VIP and must respond according to the VIP script provided by the company.
    II. Month 4 has the highest churn rate for this customer group and retention efforts must be maximized during this period. Tailored packages have proven very successful for subscription-based business models and fit our criteria in this instance. The VIP Package offers high day and evening minutes, with a complementary offer of two months of either Voicemail or International Plan, customer's choice. The key here is to increase the customer lifetime value to 6 months, and as evident in our data customers are less likely to churn after 6 months.
2. Class B: AVG
    I. This classifies the average-spending customers. They've been with the company for 5 months, complain often and are unhappy with their service. Since this group are frequent complainers, and 4 customer service calls has a high probability of churn, Class B should also be recognized by the system and transferred to the express line.
    II. As shown in TABLE 6, Association 2, the probability of churn for average value customers is high. I know that 70% of churn happens between 3-5 months, so in order to increase the CLV for this class, I offer them the Discount Package. This includes discounted Evening minutes with a complementary offer of one month of either Voicemail or International Plan, customer's choice.
3. Class C: High
    I. This class has a medium probability of churn but are also part of our high-value customer sector, since the probability of churn in terms of customer service is high during 0-1 calls, they will also be transferred to the express line for a better service experience.
    II. This segment could benefit from the Business Package; High Day and Evening minutes + Voicemail or International plan at a fixed price.

The provision of these packages tailored for each identified class should help reduce churn to a great extent. The tailored nature of the solutions on offer is also an aid to cost overruns, as a blanket solution of a discount package to get customers to the 6-month mark might have. As identification of a customer’s cluster can be done easily by classifying their account data, this should be done automatically at fixed intervals and attempts to retain customers should be done proactively by a representative of the company reaching out to them instead of waiting for the customer to contact customer service to complain.

Finally, it is of interest that this dataset largely represents the situation of major Canadian telecom companies in the past decade or so, and that the strategies and recommendations derived from our analysis closely mirrors the real world tactics that were employed by most of the major companies in question over the past several years. The move away from individual minutes, a division between daytime, evening, and night time calling, and additional charges for international or long-distance calling that has occurred across the board with these major telecommunications firms are all supported here by the data analysis tools and predictive algorithms used. It is, in effect, an excellent case study in the way data can be and has been used to make informed business decisions.