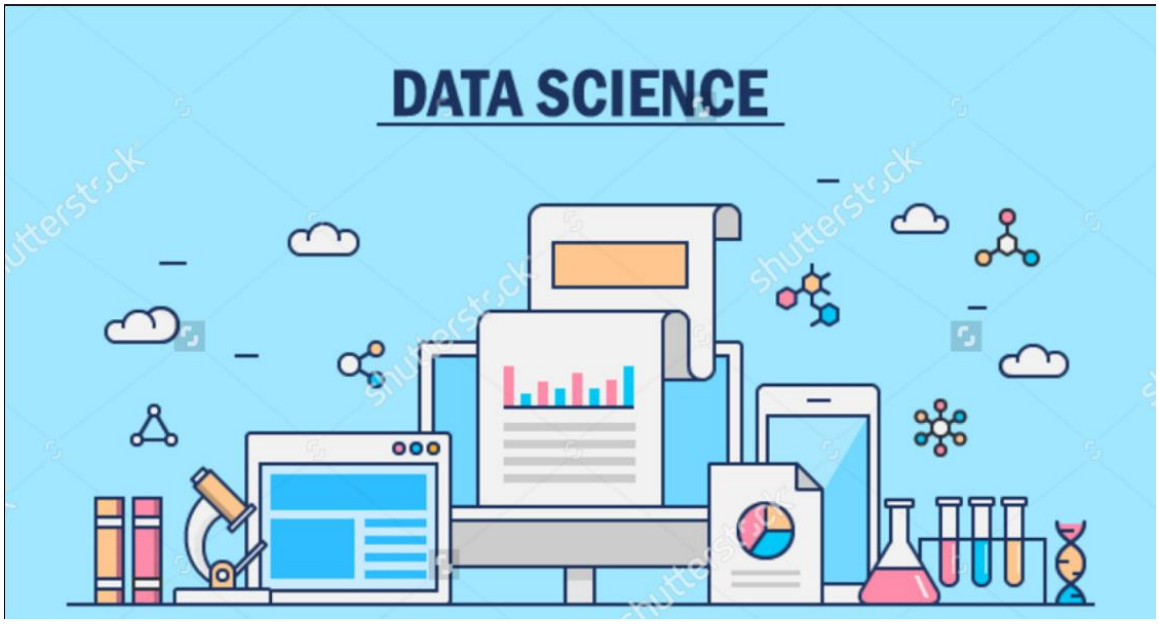


Customer churn

ANALYSIS

Chaitra Hosmani |Namita Jaggi |Noha Seddik | Shalini Varanasi
CIND119|June 2017

Contents

Summary

This report answers the research question, 'which customers are likely to churn in the future?' Data was gathered from the organization, Telus, related to their customers.

The team of data scientists then prepared the data for predictive modelling by cleaning up the data by: ensuring all attribute types were identified, calculating basic statistics for the attributes, checking for outliers, checking for the normality of attributes and selecting the attributes that will most accurately predicts the customers that will.

After data preparation, the team applied select predictive modelling algorithms to the dataset using all the attributes as the base line and then removing attributes that were identified as weak at predicting the customers likely to churn. The team identified 9 attributes that were strong predictors -- Intl Plan, VMail Plan, Intl Mins, Intl Calls, Intl Charge, CustServ Calls, Total Mins, Total Charge, Churn.

The predictive modelling techniques used were Decision Tree, Naives Bayes and PART. Decision Tree and Naives Bayes were the two models with the highest True Positive rates and the lowest False Positive rates thus were the best models for predicting the customers most likely to churn.

K-means clustering was then applied to identify the unique characteristics of customers most likely to churn. 3 unique groups were identified: 1. High customer service calls 2. High total charges 3. High international calls.

Recommendations have been provided on how to retain the customers that are most likely to churn.

Data Scientists

MEMBER NAME	LIST OF TASKS PERFORMED
Chaitra Hosmani	Decision tree/Selection of Attributes/K-means Clustering/Presentation
Noha Seddik	Data preparation/Naïve Bayes /Reporting/Presentation
Shalini Varanasi	Data preparation/Part/Presentation
Namita Jaggi	Decision tree/Selection of Attributes/Clustering/Reporting

Data preparation

Data Preparation

Attributes were classified attributes into different types and basic statistics like mean and standard deviation were calculated. Then, attributes that were most likely to predict customers that will churn were identified. Other attributes were removed or transformed for better analysis as below:

- State: Not relevant
- Area Code: Not relevant
- Phone Number: Not relevant
- Number of Voice Messages: Not relevant
- Day/Evening/Night Mins/Calls/Charge: Transformed to total Mins/Calls/Charge

Attribute Types

Listed in the table below are all the attributes in the dataset provided. The data scientists identified the type of each attribute. The status of each attribute refers to whether the attribute was a weak predictor or a strong predictor to predict the customers most likely to churn in the future.

Status 'removed' indicates that the attribute was deemed a weak predictor hence removed for the final analysis.

The team also identified that by combining the following into new attributes added up to be the totals provided the best predictions:

1. day, evening and night minutes = total minutes
2. day, evening and night calls = total calls
3. day, evening and night charges = total charges

Attributes	Type	Status
State	Categorical	Removed
Account Length	Quantitative discrete	Removed
Area Code	Categorical	Removed
Phone	Categorical	Removed
Intl Plan	Categorical	
VMail Plan	Categorical	
VMail Message	Quantitative discrete	Removed
Day Mins	Quantitative continuous	Removed -Used in Total Local Mins
Day Calls	Quantitative discrete	Removed -Used in Total Local Calls
Day Charge	Quantitative continuous	Removed -Used in Total Local Charge
Eve Mins	Quantitative continuous	Removed -Used in Total Local Min
Eve Calls	Quantitative discrete	Removed -Used in Total Local Calls
Eve Charge	Quantitative continuous	Removed -Used in Total Local Charge
Night Mins	Quantitative continuous	Removed -Used in Total Local Min
Night Calls	Quantitative discrete	Removed -Used in Total Local Calls
Night Charge	Quantitative continuous	Removed -Used in Total Local Charge
Intl Mins	Quantitative continuous	
Intl Calls	Quantitative discrete	
Intl Charge	Quantitative continuous	
Cust-Service Calls	Quantitative discrete	
Total Local Mins	Quantitative continuous	New
Total Local Calls	Quantitative discrete	New / Removed
Total Local Charge	Quantitative continuous	New
Churn?	Categorical (Class attribute)	

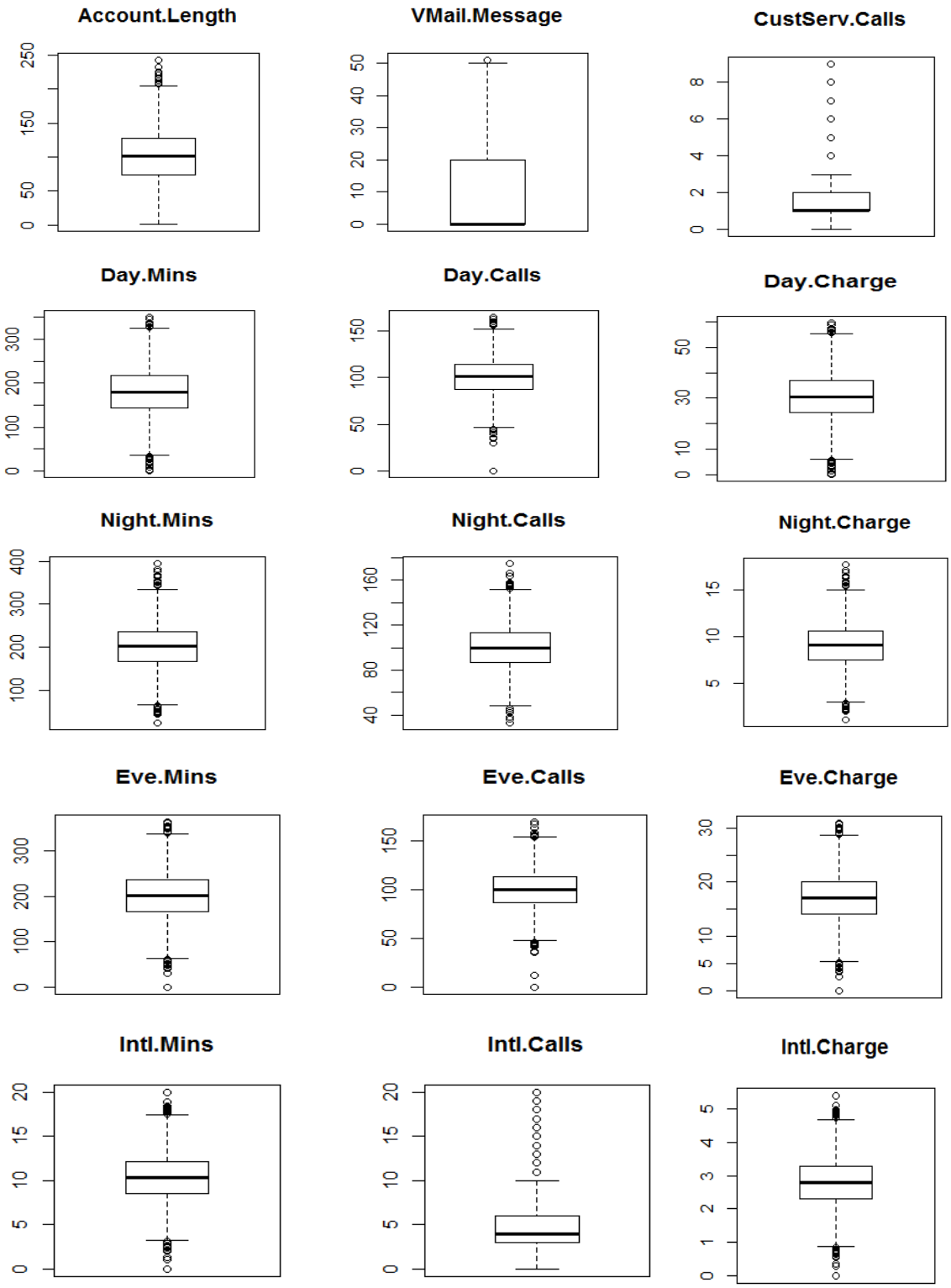
Basic Statistics for Attributes

Basic statistics were calculated to better understand quantitative attributes:

Attributes	Mean	STD	Min	Max
State	…….	…….	…….	…….
Account Length	101.0648	39.82211	1	243
Area Code	…….	…….	…….	…….
Phone	…….	…….	…….	…….
Intl Plan	…….	…….	…….	…….
VMail Plan	…….	…….	…….	…….
VMail Message	8.09901	13.68837	0	51
Day Mins	179.7751	54.46739	0	350.8
Day Calls	100.4356	20.06908	0	165
Day Charge	30.56231	9.259435	0	59.64
Eve Mins	200.9803	50.71384	0	363.7
Eve Calls	100.1143	19.92263	0	170
Eve Charge	17.08354	4.310668	0	30.91
Night Mins	200.872	50.57385	23.2	395
Night Calls	100.1077	19.56861	33	175
Night Charge	9.039325	2.275873	1.04	17.77
Intl Mins	10.23729	2.79184	0	20
Intl Calls	4.479448	2.461214	0	20
Intl Charge	2.764581	0.753773	0	5.4
CustServ Calls	1.562856	1.315491	0	9
Total Local Mins				
Total Local Calls				
Total Local Charge				
Churn?				

Check for Outliers

Box-whisker plot charts were used to identify the outliers as shown below:

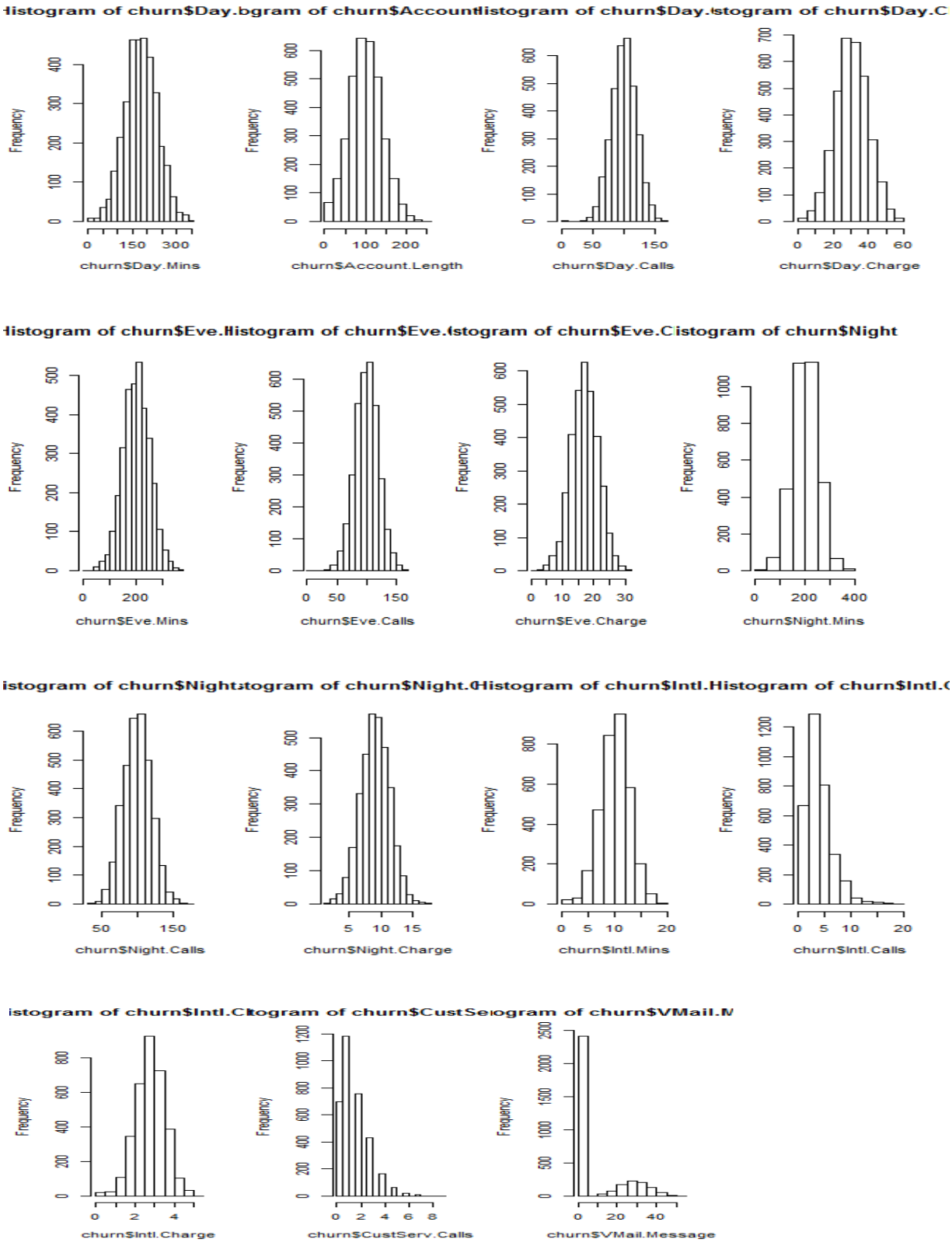


Outlier determination

Attribute	Outliers
outVMailM	51
outDayMin	332.9 337.4 326.5 350.8 335.5 30.9 34.0 334.3 346.8 12.5 25.9 0.0 0.0 19.5 329.8 7.9 328.1 27.0 17.6 326.3 345.3 2.6 7.8 18.9 29.9
outDayCall	158 163 36 40 158 165 30 42 0 45 0 45 160 156 35 42 158 157 45 44 44 44 40
outDayCharge	56.59 57.36 55.51 59.64 57.04 5.25 5.78 56.83 58.96 2.13 4.40 0.00 0.00 3.32 56.07 1.34 55.78 4.59 2.99 55.47 58.70 0.44 1.33 3.21 5.08
outEveMin	61.9 348.5 351.6 31.2 350.5 42.2 347.3 58.9 43.9 52.9 42.5 60.8 58.6 56.0 48.1 60.0 350.9 49.2 339.9 361.8 354.2 363.7 0.0 341.3
outEveCall	164 46 168 42 37 12 157 155 45 36 156 46 44 155 46 43 0 155 159 170
outEveCharge	5.26 29.62 29.89 2.65 29.79 3.59 29.52 5.01 3.73 4.50 3.61 5.17 4.98 4.76 4.09 5.10 29.83 4.18 28.89 30.75 30.11 30.91 0.00 29.01
outNightMin	57.5 354.9 349.2 345.8 45.0 342.8 364.3 63.3 54.5 50.1 43.7 349.7 352.5 23.2 63.6 381.9 377.5 367.7 56.6 54.0 64.2 344.3 395.0 350.2 50.1 53.3 352.2 364.9 61.4 47.4
outNightCall	46 42 44 42 153 175 154 158 155 157 157 154 153 166 33 155 156 38 36 156 164 153
outNightCharge	2.59 15.97 15.71 15.56 2.03 15.43 16.39 2.85 2.45 2.25 1.97 15.74 15.86 1.04 2.86 17.19 16.99 16.55 2.55 2.43 2.89 15.49 17.77 15.76 2.25 2.40 15.85 16.42 2.76 2.13
outIntMin	20.0 0.0 17.6 2.7 18.9 0.0 18.0 2.0 0.0 18.2 0.0 0.0 1.3 0.0 0.0 0.0 2.2 18.0 0.0 17.9 0.0 18.4 2.0 17.8 2.9 3.1 17.6 2.6 0.0 0.0 18.2 0.0 18.0 1.1 0.0 18.3 0.0 0.0 2.1 2.9 2.1 2.4 2.5 0.0 0.0 17.8
outIntCall	19 15 11 12 13 11 12 11 13 12 11 11 18 11 12 13 12 12 11 15 13 15 11 11 14 13 11 13 13 12 11 14 15 18 12 13 11 14 11 12 14 15 12 11 16 11 11 11 11 15 11 14 11 11 12 13 11 11 16 13 11 13 11 15 11 12 13 18 12 12 12 11 13 11 13 14 20 17
outIntCharge	5.40 0.00 4.75 0.73 5.10 0.00 4.86 0.54 0.00 4.73 4.73 4.91 0.00 0.00 0.35 0.00 0.00 0.00 0.59 4.86 0.00 4.83 0.00 4.97 0.54 4.81 0.78 0.84 4.75 0.70 0.00 0.00 4.91 0.00 4.86 0.30 0.00 4.94 0.00 0.00 0.57 0.78 4.73 0.57 0.65 0.68 0.00 0.00 4.81.0 17.6 326.3 345.3 2.6 7.8 18.9
outaccount	208 215 209 224 243 217 210 212 232 225 225 224 212 210 217 209 221 209

Check for normality

Histograms were used to check the normality of the attributes. Below are the results:



Attribute selection (by Weka)

The Weka Attribute Selection tool was used to understand which attributes would be the best predictors of customers that were most likely to churn. Below are the results:

Attribute evaluator	Attribute selected	Excluded attribute	
Best first + cfsubseteval	Phone Number, Inter Plan, Total Day Min, No of Calls Customer Service		
Ranker-Correlation		State, total day call, account length, phone no, total evening call, total night call, area code	Based on correlation ranking
Ranker-gain ratio attribute eval		Account, total night call, night min, day call, evening call, night charges	Based on gain ratio ranking
Ranker-information gain		Account, night mint, day call, evening call, night call, night charges	Information gain ranking

Imbalanced class distribution

For churn FALSE it is 2850 and TRUE is 483. The given data set is imbalance data set.

Predictive Modeling/Classification

Predictive Modeling/Classification

The dataset was run in Weka with different algorithms and using different methods to divide the data into training and test set as the data scientists paid close attention to the TP and FP rates. The optimal model will have the highest TP rate and lowest FP rate. Following are the results:

NOTE: TP – TRUE PSOTIVE, FP – FALSE POSITIVE

Decision Tree

All 21 attributes of the original dataset were used to obtain the decision tree results. The TP rates were used as a baseline.

	All Attributes					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.981	0.352	0.994	0.463	0.98	0.36
Correctly classified instances	93.28%		92.76%		93.07%	

8 attributes were then removed (State, account, area, phone, day call, even call, night call])

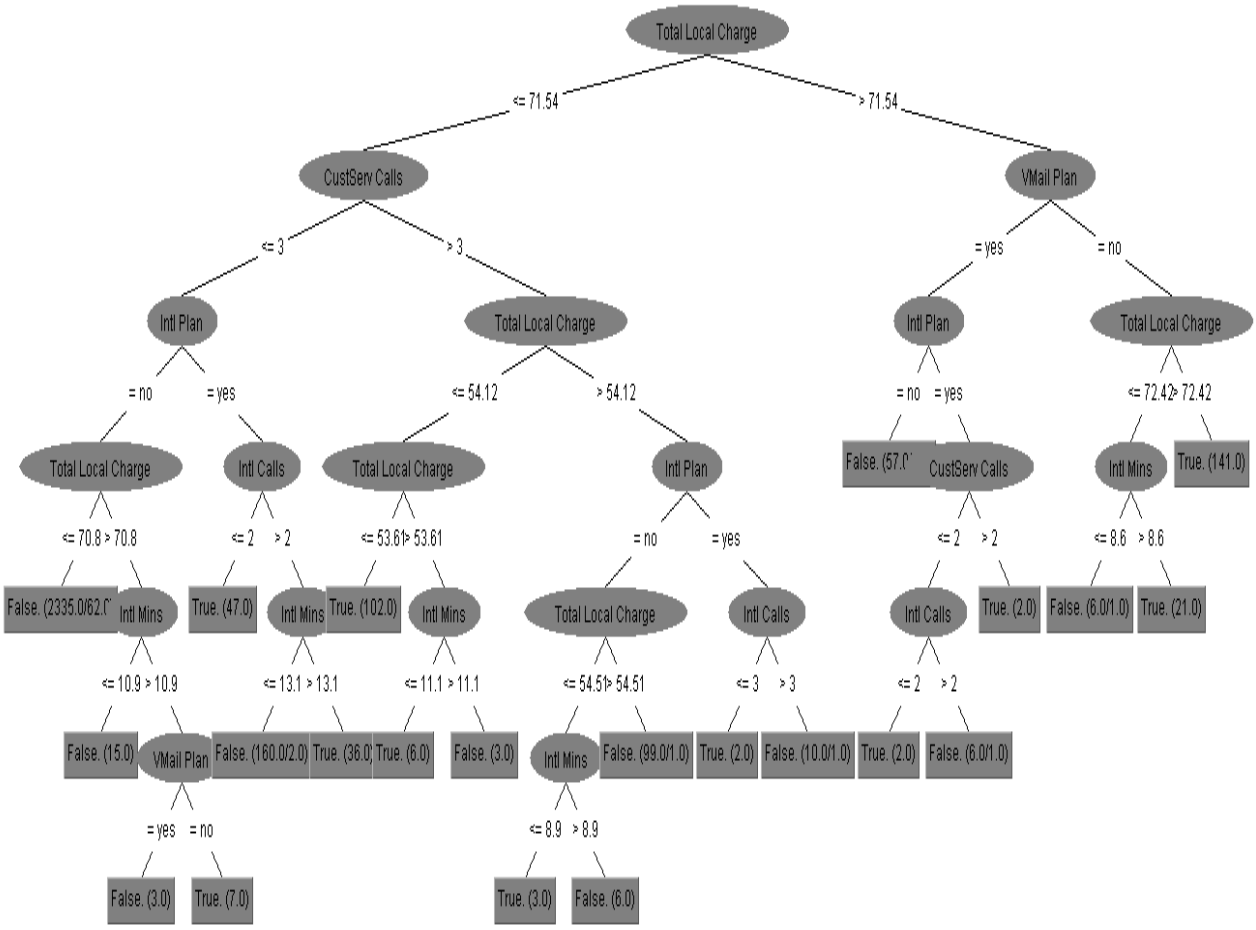
	14 Attributes (remove [State, account, area, phone, day call, even call, night call])					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.978	0.296	0.994	0.366	0.977	0.286
Correctly classified instances	93.82%		94.17%		93.91%	

The below results show that the dataset with 9 attributes (Int plan, vmail plan, int min, int calls, int charge, Cust calls, total min, Total charge, churn) with outliers has the highest TP rates and lowest FP rates thus is the best predictor (within the decision tree algorithm) for customers that are most likely to churn.

	9 Attributes (Int plan, vmail plan, int min, int calls, int charge, Cust calls, total min, Total charge, churn)					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.996	0.182	0.994	0.14	0.995	0.178
Correctly classified instances	96.99%		97.17%		96.97%	

Without outliers, the TP rate was lower and FP rate was higher as shown below:

	9 Attributes-Excluding outliers					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.995	0.192	0.98	0.32	0.993	0.158
Correctly classified instances	96.87%		93.61%		97.31%	


Total Local Charge
<= 71.54
> 71.54
CustServ Calls
VMail Plan
<= 3
> 3
= yes
= no
Intl Plan
Total Local Charge
Intl Plan
Total Local Charge
= no
= yes
<= 54.12
> 54.12
= no
= yes
<= 72.42
> 72.42
Total Local Charge
Intl Calls
Total Local Charge
Intl Plan
False. (57.0
CustServ Calls
Intl Mins
True. (141.0)
<= 70.8
> 70.8
<= 2
> 2
<= 53.61
> 53.61
= no
= yes
<= 2
> 2
<= 8.6
> 8.6
False. (2335.0/62.0
Intl Mins
True. (47.0)
Intl Mins
True. (102.0)
Intl Mins
Total Local Charge
Intl Calls
Intl Calls
True. (2.0)
False. (6.0/1.0)
True. (21.0)
<= 10.9
> 10.9
<= 13.1
> 13.1
<= 11.1
> 11.1
<= 54.51
> 54.51
<= 3
> 3
<= 2
> 2
False. (15.0)
VMail Plan
False. (160.0/2.0)
True. (36.0
True. (6.0)
False. (3.0)
Intl Mins
False. (99.0/1.0)
True. (2.0)
False. (10.0/1.0)
True. (2.0)
False. (6.0/1.0)
= yes
= no
<= 8.9
> 8.9
False. (3.0)
True. (7.0)
True. (3.0)
False. (6.0)

Naïve Bayes

We used the same strategies to determine whether naïve bayes was a good model for prediction. As shown below, naïve bayes consistently had a high FP rate thus is not the best algorithm to predict which customers are most likely to churn in the future.

	All Attributes					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.935	0.466	0.93	0.48	0.94	0.47
Correctly classified instances	87.66%		87.66%		87.93%	

	14 Attributes (remove [State, account, area, phone, day call, even call, night call])					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.96	0.56	0.95	0.543	0.96	0.55
Correctly classified instances	88.24%		87.91%		88.33%	

	9 Attributes (Int plan, vmail plan, int min, int calls, int charge, Cust calls, total min, Total charge, churn)					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.98	0.69	0.97	0.69	0.98	0.71
Correctly classified instances	87.91%		86.05%		87.52%	

	9 Attributes-Excluding outliers					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.97	0.6	0.96	0.56	0.97	0.64
Correctly classified instances	88.66%		89.26%		88.07%	

PART

The PART algorithm below is based on 9 Attributes as it was previously discovered that 9 attributes consistently provided the best results. As shown below, when using the PART algorithm by including outliers for the 9 chosen attributes provides the best result of the highest TP rates and lowest FP rates when compared to all the other models (including decision tree and naïve bayes).

	9 Attributes (Int plan, vmail plan, int min, int calls, int charge, Cust calls, total min, Total charge, churn)					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.996	0.164	0.995	0.14	0.995	0.188
Correctly classified instances	97.26%		97.26%		96.82%	

	9 Attributes-Excluding outliers					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.995	0.194	0.998	0.21	0.998	0.187
Correctly classified instances	96.84%		97.02%		97.13%	

OPTIMAL MODEL

The optimal models as shown by the analysis above was PART with 9 attributes (including outliers) using the 10 fold cross validation technique to train the data. The Decision Tree was a close second best model. Both models have higher True Positive rates and lower False Positive rates thus great models to predict which customers are most likely to churn in the future.

DATA SET DIVISION

The following techniques were used for data set division:

- 10 fold cross validation
- 3 fold cross validation
- Percentage split(66%)

From the analysis above it was clear that using the 10-fold cross validation provided the best results for prediction across all the algorithms used.

Post-prediction Analysis

Post-prediction Analysis

K-MEANS ALGORITHM

K-means clustering was used to segment churned customer into different categories. 3 clusters K-means clustering provided the best results.

Within cluster sum of squared errors: 268.65934735797794

Cluster 0: no,no,4.1,4,1.11,4,308,44.9,True.

Cluster 1: no,no,4.2,5,1.13,4,369,75.17,True.

Cluster 2: no,no,12.3,2,3.32,5,299,48.92,True.

Final cluster centroids:

Attribute	Full Data	0	1	2
Intl Plan	no	no	no	no
VMail Plan	no	no	no	no
Intl Mins	10.7733	10.8885	8.0909	**13.1289**
Intl Calls	4.1258	3.5862	4.1	4.5372
Intl Charge	2.9092	2.9403	2.1851	3.5451
CustServ Calls	2.2862	**4.4023**	1.5545	1.4298
Total Calls	303.5723	297.8046	302.6	308.6033
Total Charge	62.6373	47.2477	**68.1117**	**68.7258**
Churn	True.	True.	True.	True.

According to analysis above, the following are three groups of customers that are most likely to churn:

Group 1 – High Customer Service Calls

This group had low total charges, average international call but high customer service calls. This group churned because they did not like the customer service by the network provider.

Recommendation: Group 1 customers should get speedy access and solution to their issues. Moment the customer calls is 3 or greater, flag should be raised to quickly resolve their issues.

Group 2 – High Total Charges

This group had high local cost, less customer service call and average international call. They churned because they felt they are paying too much for the network.

Recommendation: Customer should get an alert of their usage and also they should be able to set the upper limit for call costs.

Group 3-High International Charges

This group had high local cost, low customer call and high international call cost. They churned because the international call cost was high.

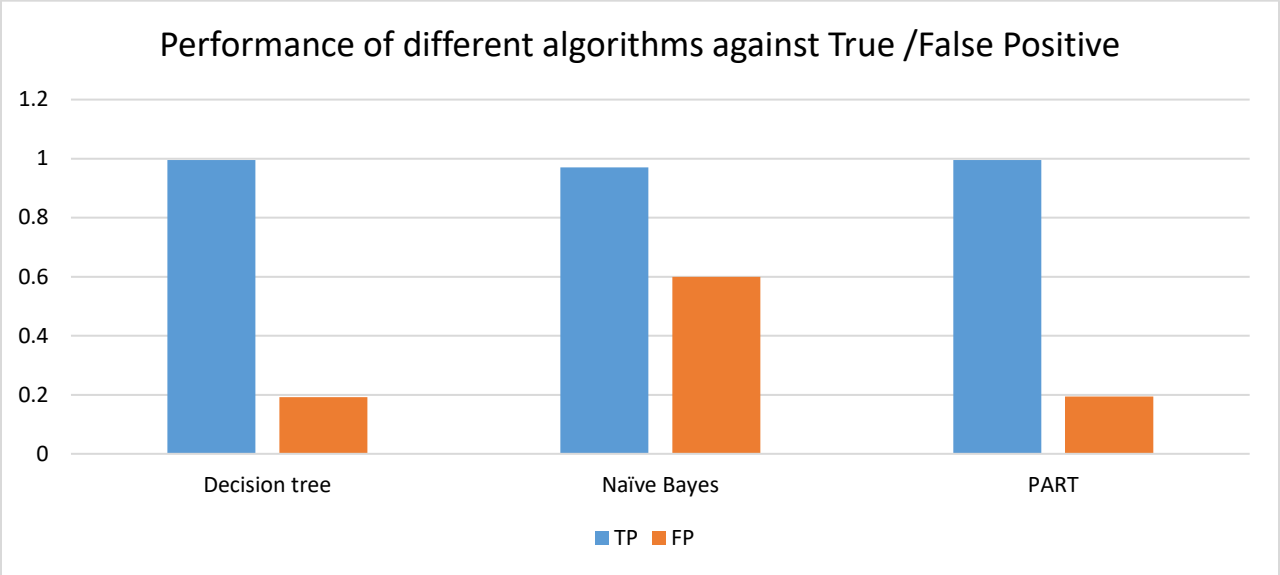
Recommendation: Should create special international package. Also alert customers their minutes and cost after each call.

Conclusion

Conclusions and Recommendations

After trying three algorithms with different attributes, below are the findings:

1. Selecting Attributes is very important to produce more accurate result in classification.
2. The best algorithm in this dataset is PART as shown below, as it had the lowest FP rates and highest TP rates



TP – TRUE POSITIVE

FP- FALSE POSITIVE

Further Recommendations

- Customer rating/survey on biannual basis
- Data plan (type), data usage
- Contract/No contract
- Type of phone offered