CIND119 Presentation
Churn Dataset

Group Name: Group_DJ0_ 1
Amarpreet Kaur, amarpreet.kaur@torontomu.ca
Eric Ding, pai.ding@torontomu.ca
Raymond Chan, r.chan@torontomu.ca

RYERSON UNIVERSITY

Executive Summary

- Problem : The telecom company wants to predict which customers will churn in the near future.
- Solution : Characterizing customers using two Predictive Models (Decision Tree & Naïve Bayes) to predict which customer may leave.
- Tools used: Python and SAS

Executive Summary

- Conclusion: Decision Tree model with feature selection provides best prediction accuracy (Model accuracy 92.32%);
- The most important predictors of customer churn are (based on both DT / NB model):
  - Day Mins: The number of minutes the customer used the service during daytime
  - CustServ Calls: The number of calls to customer support service
  - Int'l Plan: whether the customer has international calling plan

RYERSON UNIVERSITY

Workload Distribution

Member Name	List of Tasks Performed
Amarpreet	• Data Preparation: Analyze the distribution of numeric attributes (normal or other); Plot histograms; handle missing values or transform any attributes; • Predictive Modelling: Decision Tree & Naïve Bayes models in Python • Conclusion and Recommendation: (Shared Equally - We reached conclusion & wrote recommendation together) • Presentation: Predictive Modeling - Python
Eric Ding	• Data Preparation: Correlated attributes; Elimination of attributes (subjective decision or an objective decision) • Predictive Modelling: Decision Tree & Naïve Bayes models in SAS • Conclusion and Recommendation: (Shared Equally - We reached conclusion & wrote recommendation together) • Presentation: Predictive Modeling- SAS; Conclusion & Backup Slides
Raymond Chan	• Data Preparation: attribute type; descriptive analysis • Predictive Modelling: Decision Tree & Naïve Bayes models in Python • Conclusion and Recommendation: (Shared Equally - We reached conclusion & wrote recommendation together) • Presentation: Introduction & Data Preparation

Data Preparation

Step 1 : Look at the attribute type

Column	Explanation	Variable Type	Data Type
State	Customer’s state	Categorical (Nominal)	object
Account Length	Integer number showing the duration of activity for customer account	Quantitative (Continuous)	int64
Area Code	Area code of customer	Categorical (Nominal)	int64
Phone Number	Phone number of customer	Categorical (Nominal)	object
Inter Plan	Binary indicator showing whether the customer has international calling plan	Categorical/ Binary (yes, no)	object
VoiceMail Plan	Indicator of voice mail plan	Categorical/ Binary (yes, no)	object
No of Vmail Mesgs	The number of voicemail messages	Quantitative (Discrete)	int64
Total Day Min	The number of minutes the customer used the service during day time	Quantitative (Continuous)	float64
Total Day calls	Discrete attribute indicating the total number of calls during day time	Quantitative (Discrete)	int64
Total Day Charge	Charges for using the service during day time	Quantitative (Continuous)	float64

Data Preparation

Step 1 : Look at the attribute type

Column	Explanation	Variable Type	Data Type
Total Evening Min	The number of minutes the customer used the service during evening time	Quantitative (Continuous)	float64
Total Evening Calls	The number of calls during evening time	Quantitative (Discrete)	int64
Total Evening Charge	Charges for using the service during evening time	Quantitative (Continuous)	float64
Total Night Minutes	Number of minutes the customer used the service during night time	Quantitative (Continuous)	float64
Total Night Calls	The number of calls during night time	Quantitative (Discrete)	int64
Total Night Charge	Charges for using the service during night time	Quantitative (Continuous)	float64
Total Int Min	Number of minutes the customer used the service to make international calls	Quantitative (Continuous)	float64
Total Int Calls	The number of international calls	Quantitative (Discrete)	int64
Total Int Charge	Charges for international calls	Quantitative (Continuous)	float64
No of Calls Customer Service	The number of calls to customer support service	Quantitative (Discrete)	int64
Churn	Class attribute with binary values (True for churn and False for not churn)	Categorical/ Binary (TRUE, FALSE)	object

Data Preparation

Step 2 : Stat summary of numerical columns

index	count	mean	std	min	25%	50%	75%	max
Account Length	3333.0	101.0648065	39.82210593	1.00	74.00	101.00	127.00	243.00
No of Vmail Mesgs	3333.0	8.099009901	13.68836537	0.00	0.00	0.00	20.00	51.00
Total Day Min	3333.0	179.7750975	54.4673892	0.00	143.70	179.40	216.40	350.80
Total Day calls	3333.0	100.4356436	20.06908421	0.00	87.00	101.00	114.00	165.00
Total Day Charge	3333.0	30.56230723	9.259434554	0.00	24.43	30.50	36.79	59.64
Total Evening Min	3333.0	200.980348	50.71384443	0.00	166.60	201.40	235.30	363.70
Total Evening Calls	3333.0	100.1143114	19.92262529	0.00	87.00	100.00	114.00	170.00
Total Evening Charge	3333.0	17.08354035	4.310667643	0.00	14.16	17.12	20.00	30.91
Total Night Minutes	3333.0	200.8720372	50.57384701	23.20	167.00	201.20	235.30	395.00
Total Night Calls	3333.0	100.1077108	19.56860935	33.00	87.00	100.00	113.00	175.00
Total Night Charge	3333.0	9.039324932	2.275872838	1.04	7.52	9.05	10.59	17.77
Total Int Min	3333.0	10.23729373	2.791839548	0.00	8.50	10.30	12.10	20.00
Total Int Calls	3333.0	4.479447945	2.461214271	0.00	3.00	4.00	6.00	20.00
Total Int Charge	3333.0	2.764581458	0.753772613	0.00	2.30	2.78	3.27	5.40
No of Calls Customer Service	3333.0	1.562856286	1.315491045	0.00	1.00	1.00	2.00	9.00

Data Preparation

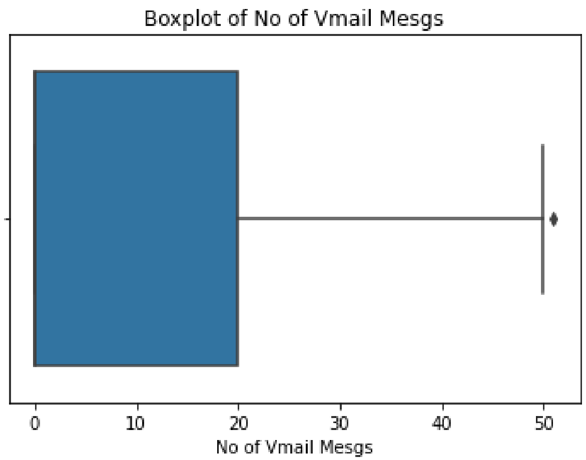
Step 3 : Outliers & Missing Values

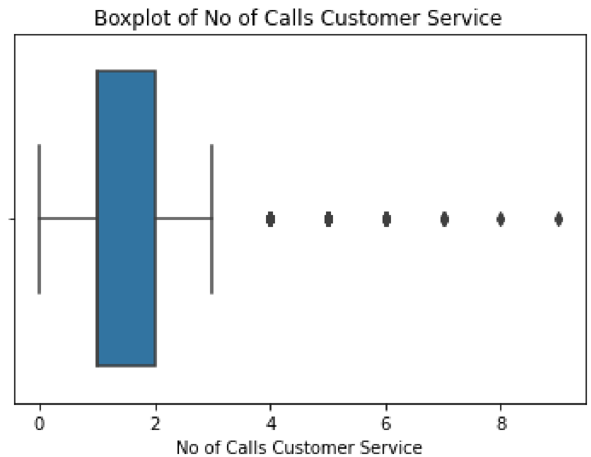
- No Missing Values detected

```
State                              0
Account Length                     0
Area Code                          0
Phone Number                       0
Inter Plan                         0
VoiceMail Plan                     0
No of Vmail Mesgs                  0
Total Day Min                      0
Total Day calls                    0
Total Day Charge                   0
Total Evening Min                  0
Total Evening Calls                0
Total Evening Charge               0
Total Night Minutes                0
Total Night Calls                  0
Total Night Charge                 0
Total Int Min                      0
Total Int Calls                    0
Total Int Charge                   0
No of Calls Customer Service       0
Churn                              0
```

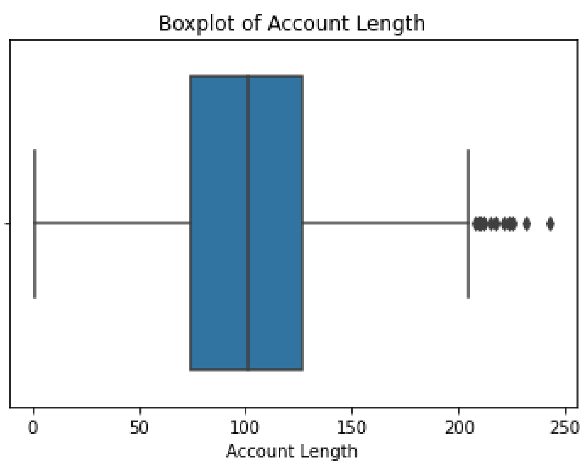
RYERSON UNIVERSITY

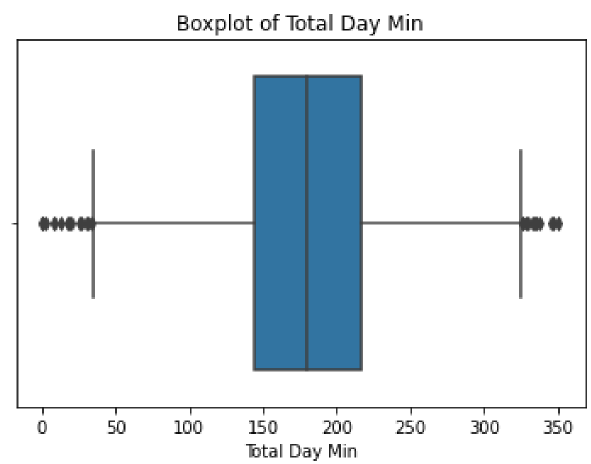
Data Preparation
Step 3 : Outliers & Missing Values





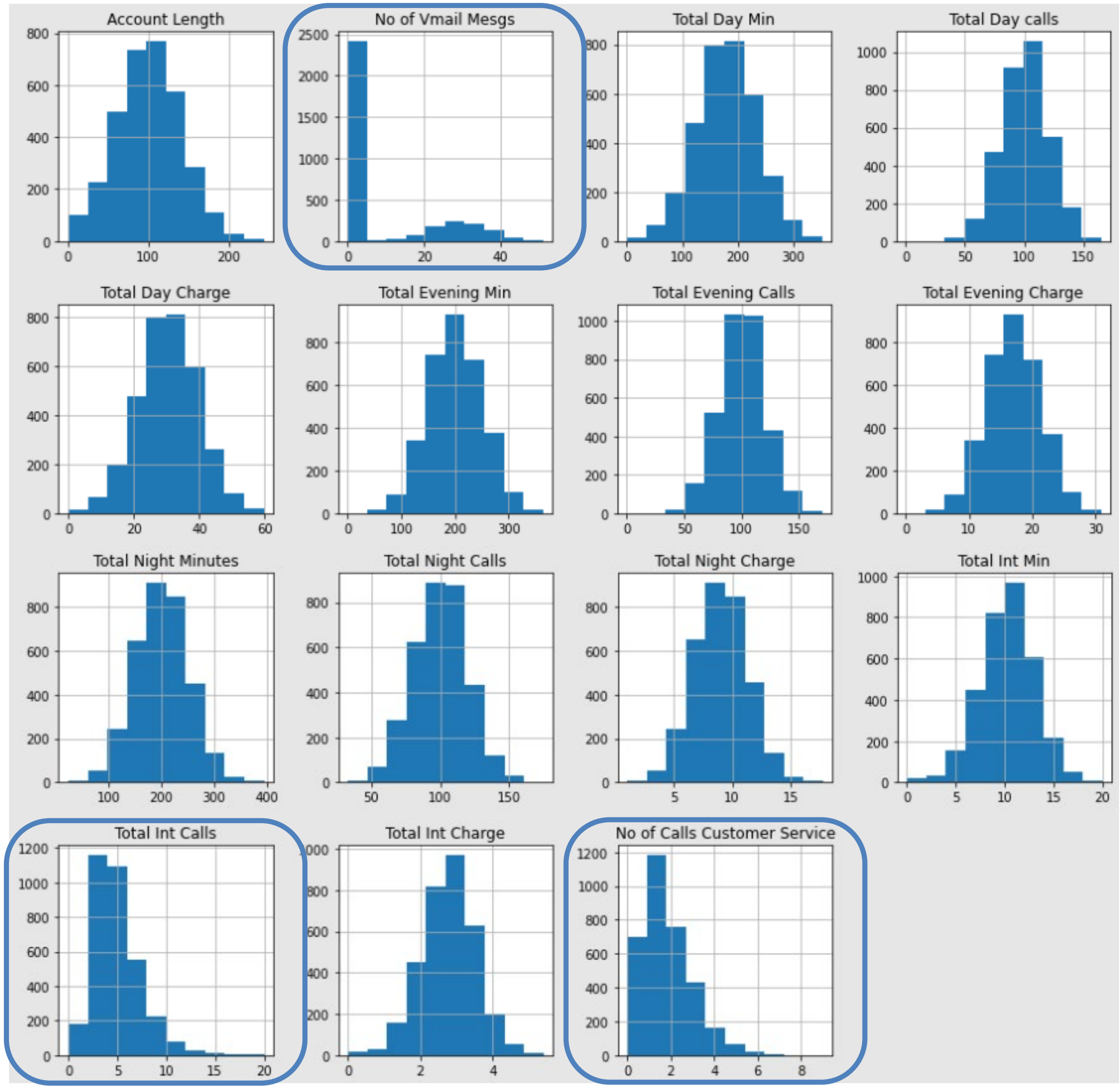
We take the data as they are





Data Preparation

Step 4: Distribution Visualization

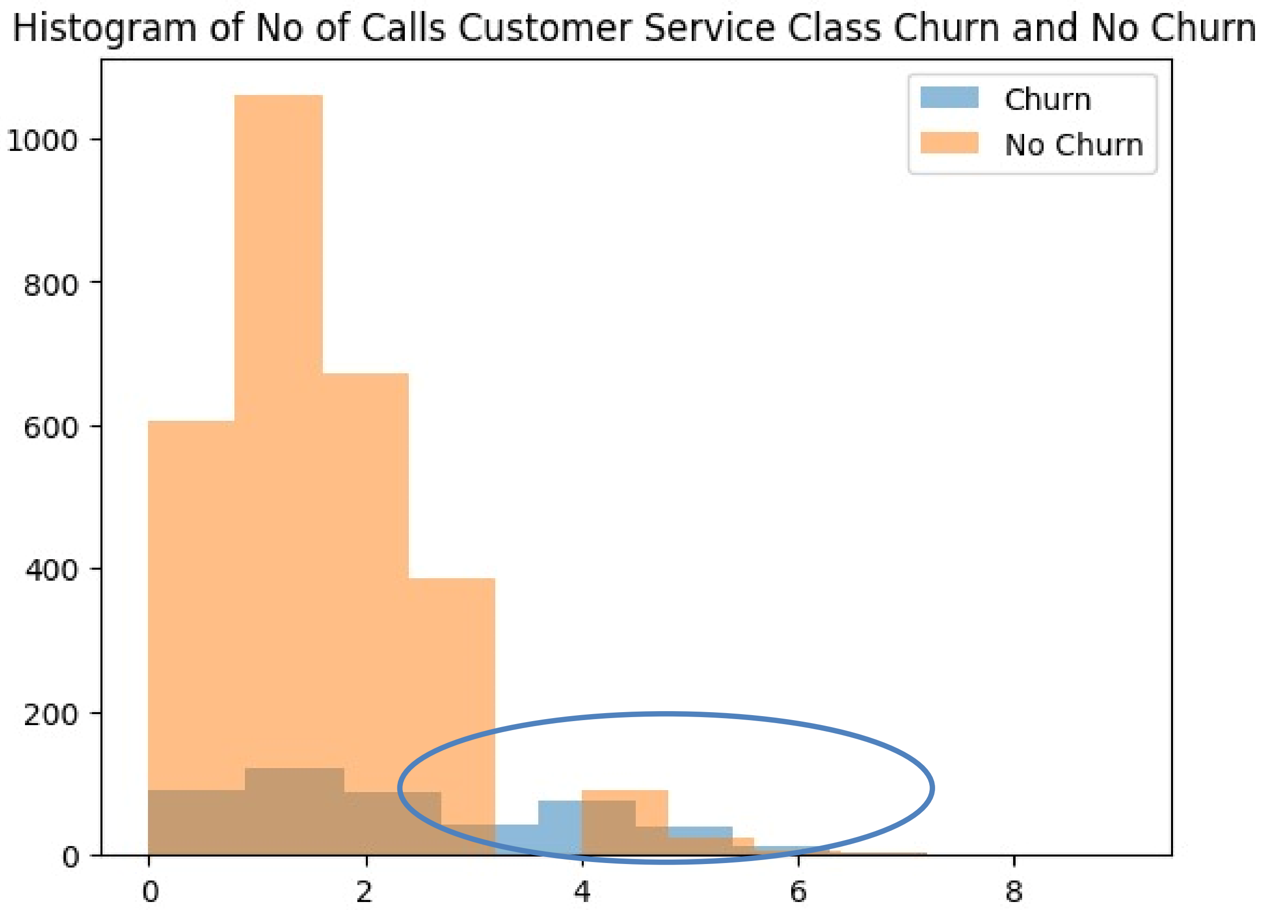


Vmail Mesgs has a different dist

No of Customer Service Calls has a skewed Dist

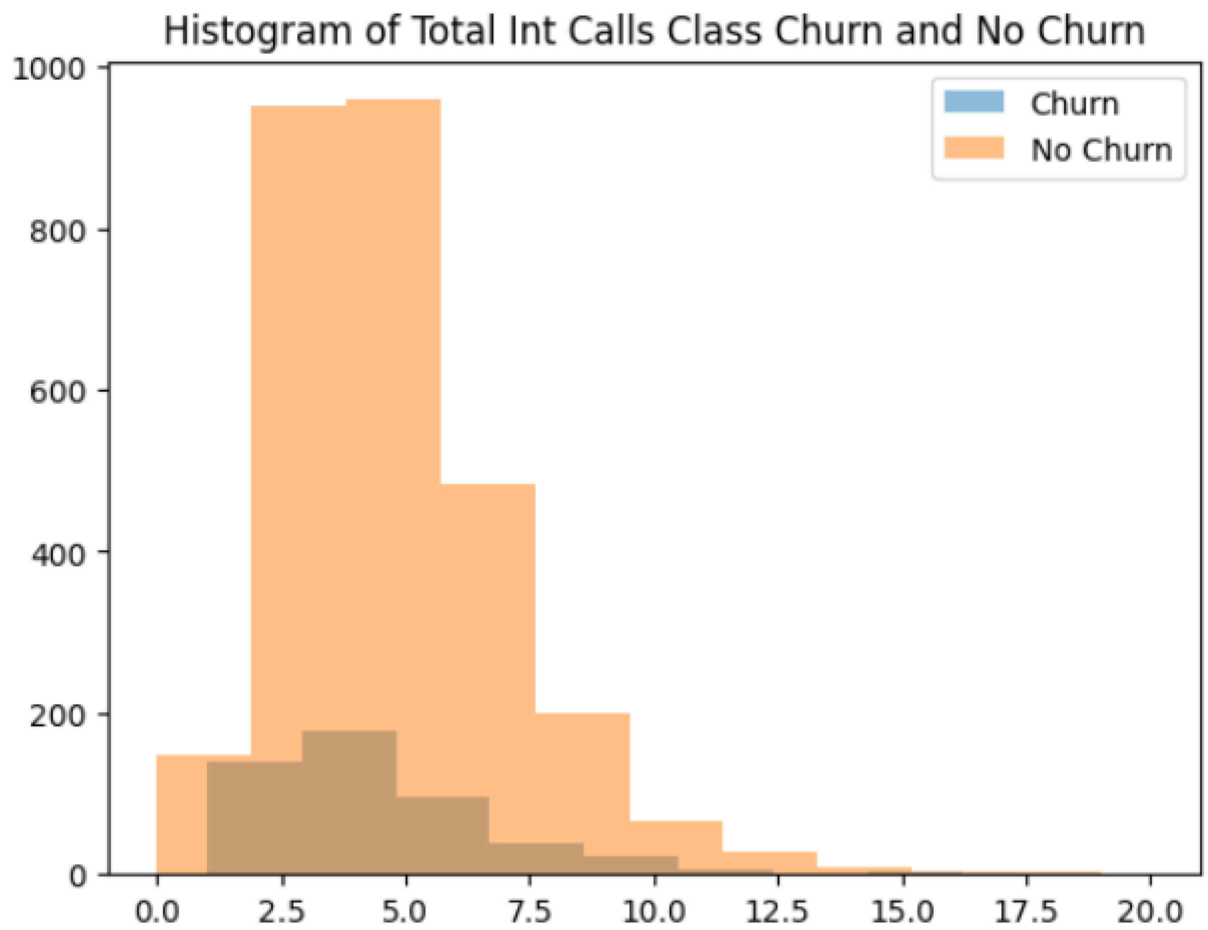
Data Preparation
Step 4: Distribution Visualization

Service Calls seems relevant (fatter tail on the right)



Data Preparation
Step 4: Distribution Visualization

Total Int Calls seems not very relevant



Data Preparation

Step 5: Check attribute correlation

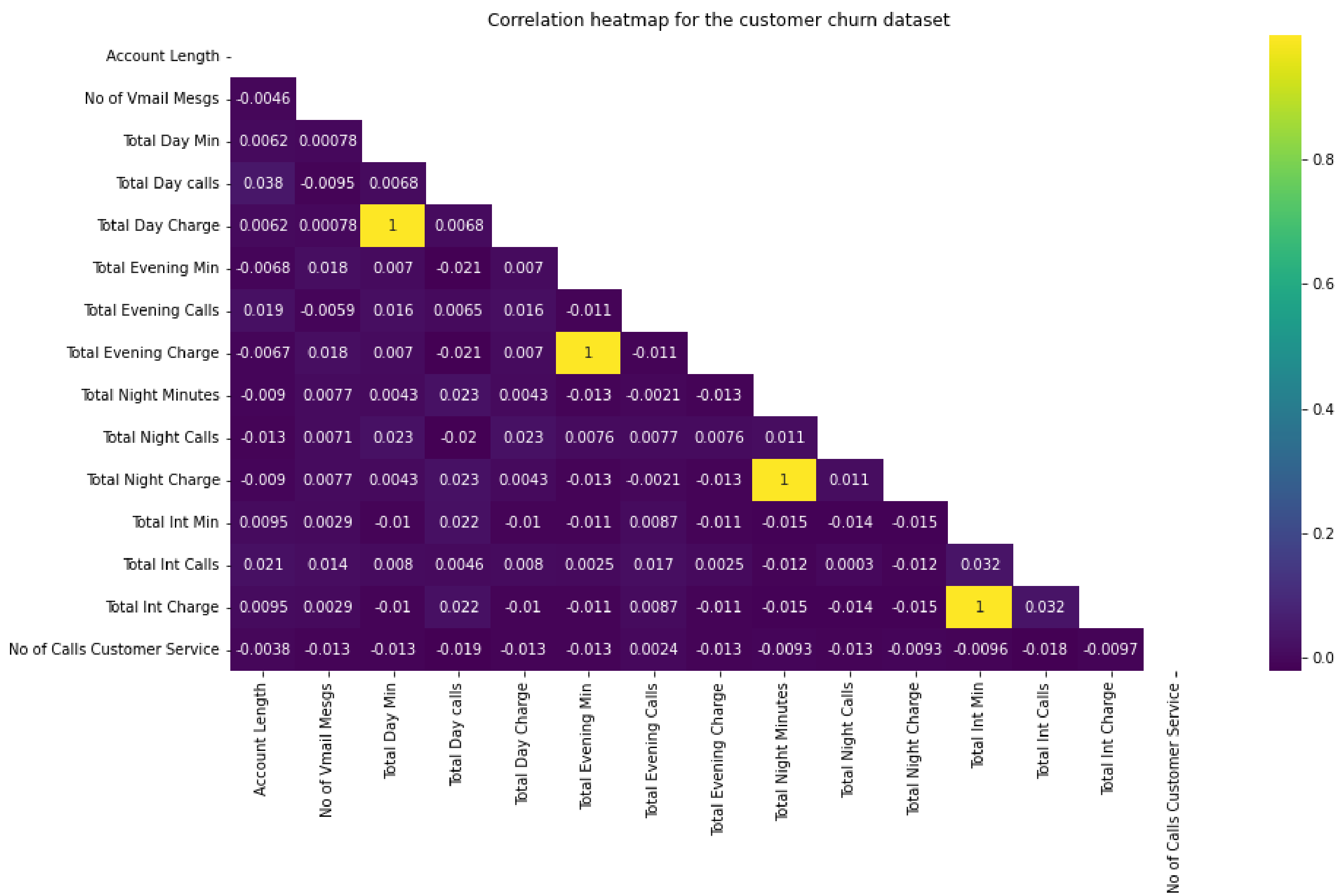
Four pairs of highly correlated attributes:

Total Day Charge – Total Day Min,

Total Evening Charge – Total Evening Min,

Total Night Charge – Total Night Minutes,

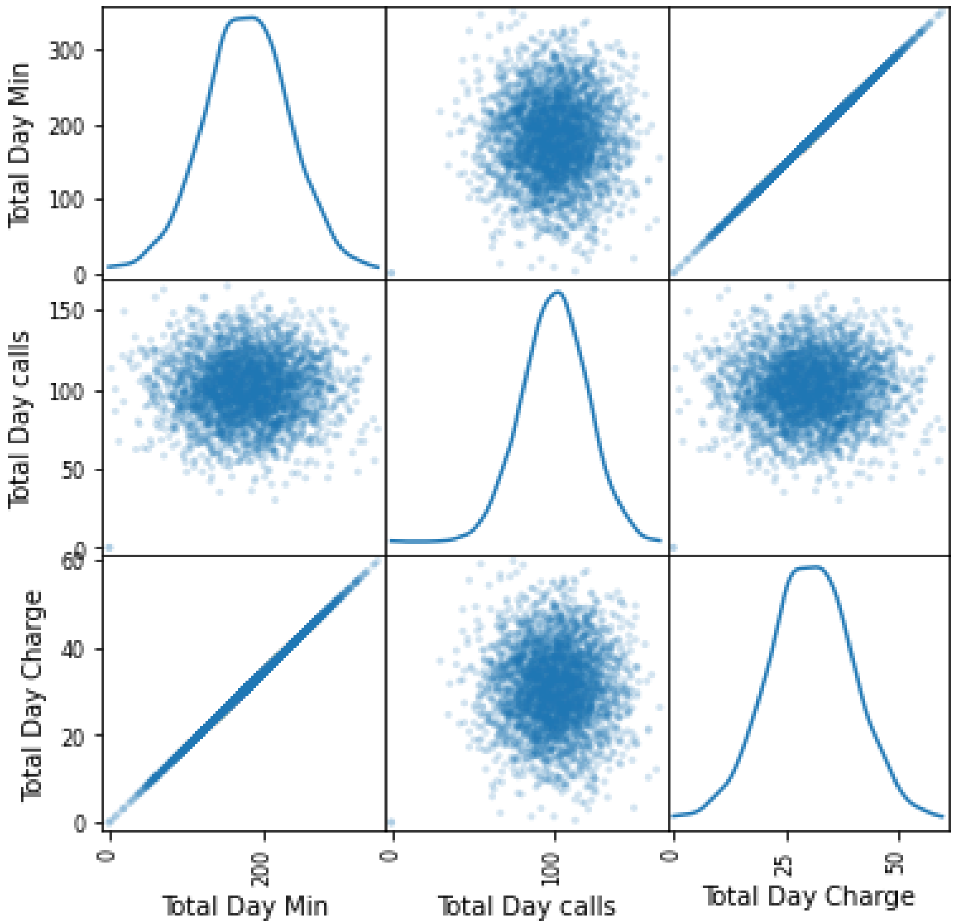
Total Intl Charge - Total Int Min.



RYERSON UNIVERSITY

Data Preparation

Step 5: Check attribute correlation



Note: Day calls & Day mins are not correlated

Data Preparation

Step 6: Attribute Selection (for Comparison)

- Subjective judgement:
  - Drop State, Area Code, Phone number from the dataset due to irrelevancy
- Objective judgement:
  - Out of the 4 pairs of highly correlated attributes, we keep the Mins set and drop the Charges set
  - Drop Phone number & State due to high cardinality
- After the removal, there is 13 (predictors) + 1 (class var) left.

Predictive Modeling - SAS

- Evaluation method:
  - train-test set split; 70/30 split
- Results (best model = Selected DT)

Model	Accuracy	Precision of predicting Not Churn	Precision of predicting Churn
Baseline decision tree	92.21%	94.92%	82.61%
Baseline Naïve Bayes	78.40%	95.67%	79.09%
Decision tree on selected features	92.32% [Best]	95.24%	83.44% [Best]
Naïve Bayes on selected features	84.79%	93.60%	64.39%

We care about "Churn=True" cases more

Confusion Matrix - DT Model

Without Selection

With Selection

(Both without Pruning)

Confusion Matrices (Without Selection)

	Actual	Predicted		Error Rate
		False.	True.	
Training	False.	2013	1	0.0005
	True.	41	302	0.1195
Validation	False.	803	33	0.0395
	True.	43	97	0.3071

Confusion Matrices (With Selection)

	Actual	Predicted		Error Rate
		False.	True.	
Training	False.	2014	0	0.0000
	True.	40	303	0.1166
Validation	False.	801	35	0.0419
	True.	40	100	0.2857

We also care about "Predicted – F / Actual – T" cases; Should be minimized

Confusion Matrix - NB Model

Without Selection

Confusion Matrix			
	predictedchurn		
Churn?	False.	True.	Total
False.	2231	619	2850
	66.94	18.57	85.51
True.	101	382	483
	3.03	11.46	14.49
Total	2332	1001	3333
	69.97	30.03	100

With Selection

Confusion Matrix			
	predictedchurn		
Churn?	False.	True.	Total
False.	2515	335	2850
	75.46	10.05	85.51
True.	172	311	483
	5.16	9.33	14.49
Total	2687	646	3333
	80.62	19.38	100

“Predicted – F / Actual – T” cases (False negative cases) are larger than DT

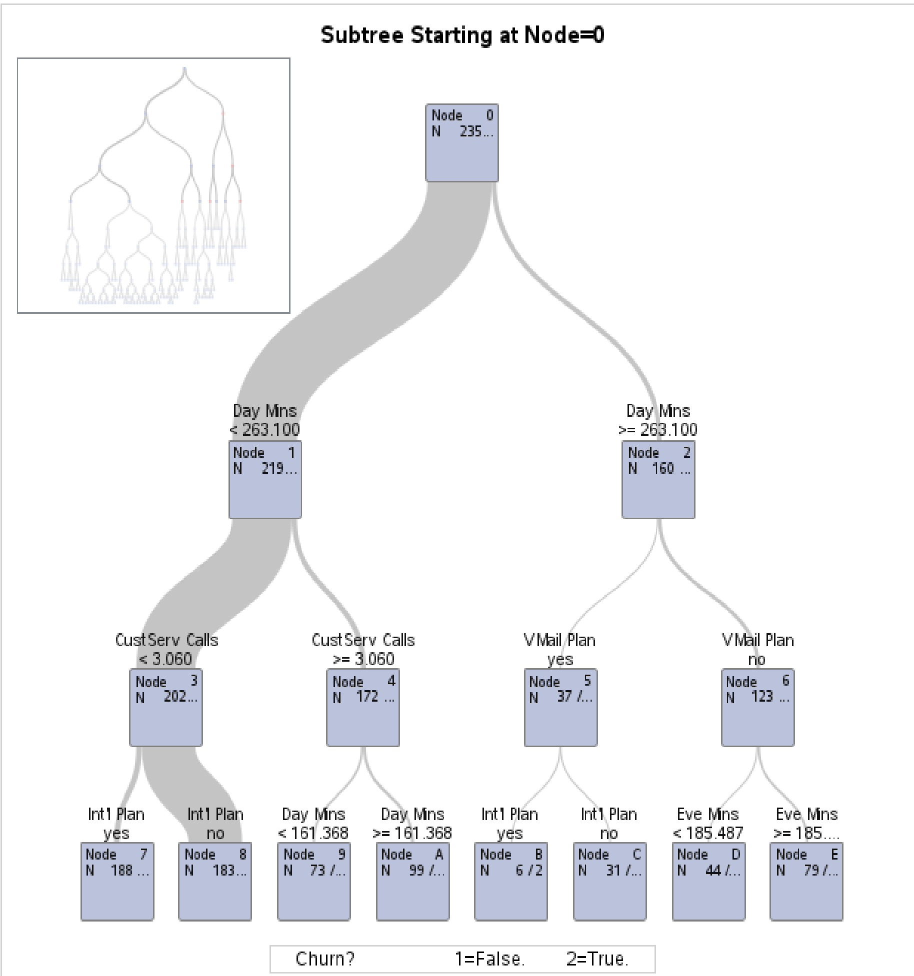
Confusion Matrix - DT Model （with Pruning）

With Selection
(with Pruning)

Confusion Matrices				
	Actual	Predicted		Error Rate
		False.	True.	
Training	False.	1995	19	0.0094
	True.	83	260	0.2420
Validation	False.	820	16	0.0191
	True.	32	108	0.2286

"Predicted – F / Actual – T" cases (False negative cases) are minimized

Best Model: DT with var selection



Variable Importance

Variable	Training Relative	Training Importance	Validation Relative	Validation Importance	Relative Ratio	Count
Day Mins	1.0000	12.1986	1.0000	6.6834	1.0000	19
CustServ Calls	0.6053	7.3840	0.8099	5.4132	1.3381	1
Int'l Plan	0.5104	6.2267	0.6831	4.5651	1.3382	3
Intl Mins	0.5533	6.7499	0.6476	4.3279	1.1703	4
VMail Plan	0.5023	6.1270	0.6144	4.1062	1.2232	5
Intl Calls	0.4983	6.0790	0.5866	3.9204	1.1771	4
Eve Mins	0.7429	9.0625	0.5737	3.8339	0.7722	15
Eve Calls	0.2319	2.8291	0.2877	1.9228	1.2405	7
Night Mins	0.4282	5.2237	0.2606	1.7418	0.6086	10
Account Length	0.3354	4.0910	0.1579	1.0555	0.4709	14
Day Calls	0.2417	2.9484	0.1111	0.7428	0.4598	5
VMail Message	0.1335	1.6279	0.1020	0.6815	0.7641	3
Night Calls	0.2077	2.5332	0.0603	0.4033	0.2906	4

Conclusion and Recommendation

- The most important predictors of customer churn are (both DT / NB model agree):
  - **Day Mins:** The number of minutes the customer used the service during daytime
  - **CustServ Calls:** The number of calls to customer support service
  - **Int'l Plan:** whether the customer has international calling plan

Conclusion and Recommendation

- Recommendations to Company / BU:
  - Heavy users (having large Day Mins or Evening Mins + have Intl Plan) tends to have higher churn rates, we should pay more attention to improve customer experience
  - Especially: When they make more customer service calls
  - Flag system (CRM Department to work on?)