



# CUSTOMER CHURN PROJECT

**CHAITRA HOSMANI**

**NAMITA JAGGI**

**NOHA SEDDIK**

**SHALINI VARANASI**

CIND119 | JUNE 2017



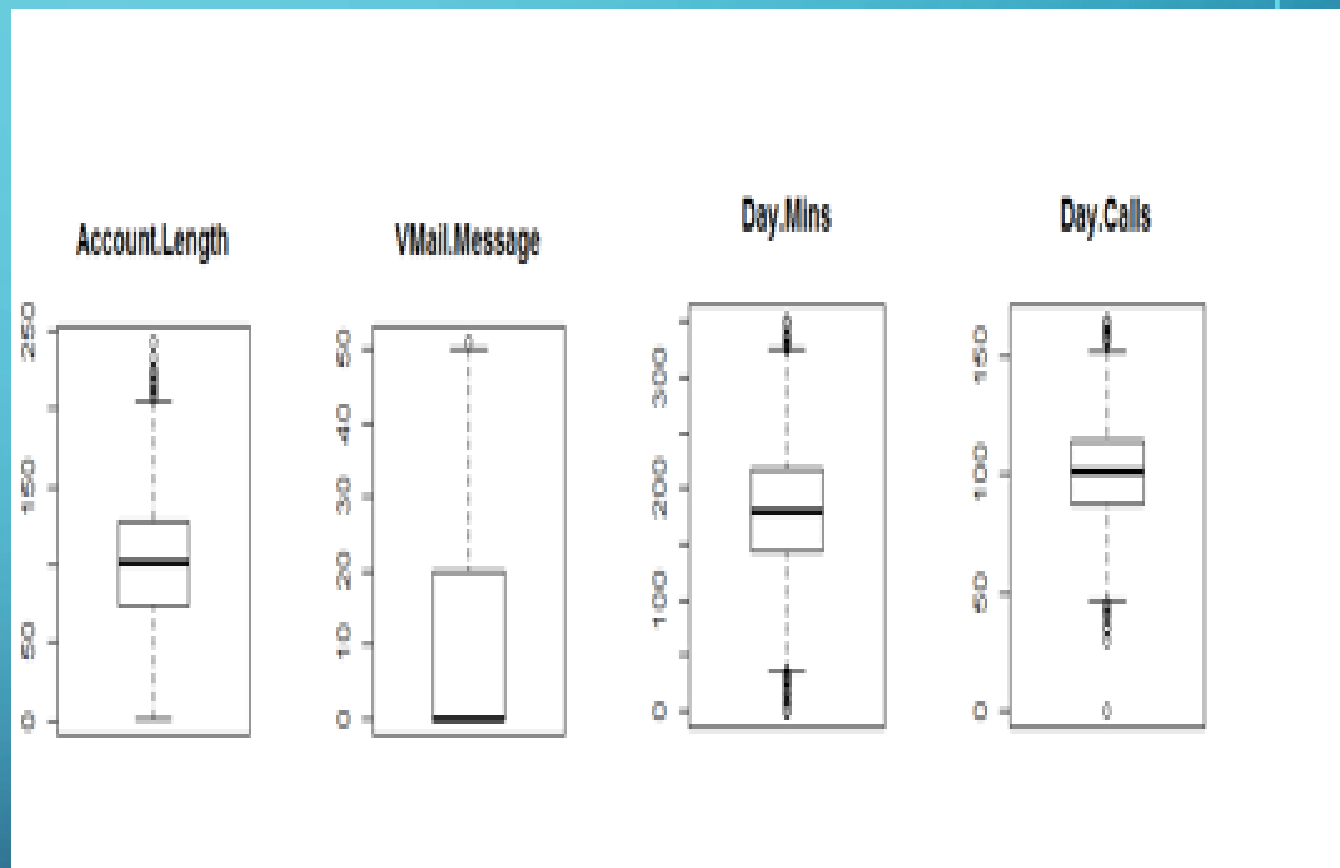
The background is a blue gradient with faint concentric circles. White circuit-like lines with circular nodes are positioned in the corners: top-left, top-right, bottom-left, and bottom-right.

# DATA PREPARATION

# ATTRIBUTES

Attributes	Type	Status
State	Categorical	Removed
Account Length	Quantitative discrete	
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
Total Local Charge	Quantitative continuous	New
Churn?	Categorical (Class attribute)	

# ANY OUTLIERS?



# BASIC STATS FOR ATTRIBUTES

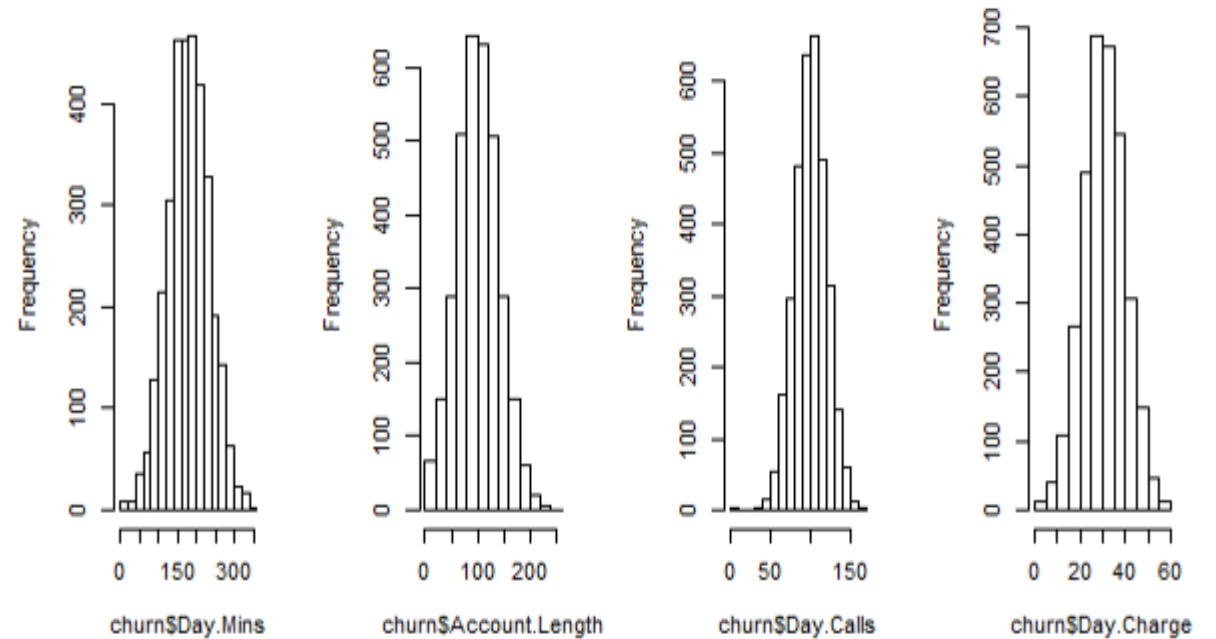


DATA  
PREP

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?				

## NORMALITY OF ATTRIBUTES

histogram of churn\$Day.Mins    bgram of churn\$Account.Length    histogram of churn\$Day.Calls    istogram of churn\$Day.Charge



# ATTRIBUTE SELECTION

DATA  
PREP

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

**Weka Explorer**

Preprocess Classify Cluster Associate **Select attributes** Visualize

**Attribute Evaluator**

Choose **CfsSubsetEval -P 1 -E 1**

**Search Method**

Choose **BestFirst -D 1 -N 5**

**Attribute Selection Mode**

☒ Use full training set  
☐ Cross-validation Folds   
Seed

(Nom) Churn? ▼

Start Stop

Result list (right-click for options)

**Attribute selection output**

The background is a blue gradient with decorative white circuit-like lines in the corners. The lines consist of straight segments and small circles, resembling a stylized electronic circuit or data flow diagram.

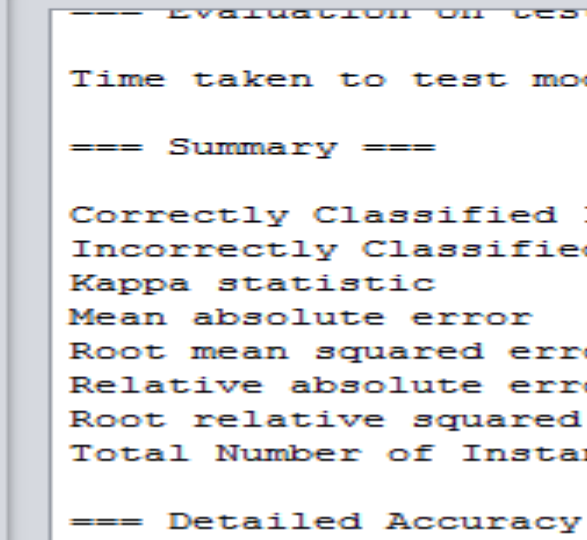
# **PREDICTIVE MODELLING / CLASSIFICATION**



A word cloud visualization of data science and machine learning terminology. The words are arranged in a circular pattern, with 'PREDICTIVE' and 'ANALYTICS' being the largest and most central. Other prominent words include 'REGRESSION', 'DATA', 'MODEL', 'CUSTOMER', 'VARIABLE', 'SERIES', 'CLASSIFICATION', 'LEARNING', 'STATISTICAL', 'USED', 'FUNCTIONS', 'METHODS', 'EXAMPLES', 'DEPENDENT', 'INDEPENDENT', 'VARIABLES', 'TECHNIQUES', 'DISTRIBUTION', 'RELATIONSHIP', 'FINANCIAL', 'PERSON', 'MODELS', 'RISK', 'LOGIT', 'SURVIVAL', 'POSSIBLE', 'CUSTOMER', 'NETWORK', 'MACHINE', 'DIFFERENTIAL', 'INTEGRATION', 'SERIES', 'CLASSIFICATION', 'ANALYSIS', 'FRUIT', 'MULTINOMIAL', 'ORDINAL', 'CARDINAL', 'RATIOS', 'PERCENTAGES', 'PROBABILITIES', 'STATISTICS', 'STATISTICAL', 'USED', 'FUNCTIONS', 'METHODS', 'EXAMPLES', 'DEPENDENT', 'INDEPENDENT', 'VARIABLES', 'TECHNIQUES', 'DISTRIBUTION', 'RELATIONSHIP', 'FINANCIAL', 'PERSON', 'MODELS', 'RISK', 'LOGIT', 'SURVIVAL', 'POSSIBLE', 'CUSTOMER', 'NETWORK', 'MACHINE', 'DIFFERENTIAL', 'INTEGRATION', 'SERIES', 'CLASSIFICATION', 'ANALYSIS', 'FRUIT', 'MULTINOMIAL', 'ORDINAL', 'CARDINAL', 'RATIOS', 'PERCENTAGES', 'PROBABILITIES', 'STATISTICS'.

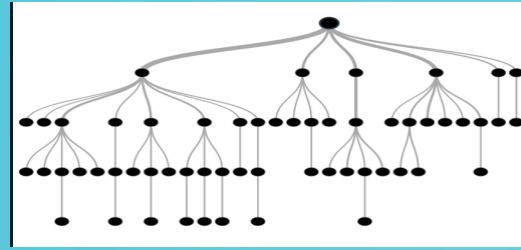
- 10 fold cross validation
- 3 fold cross validation
- Percentage split(66%)

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

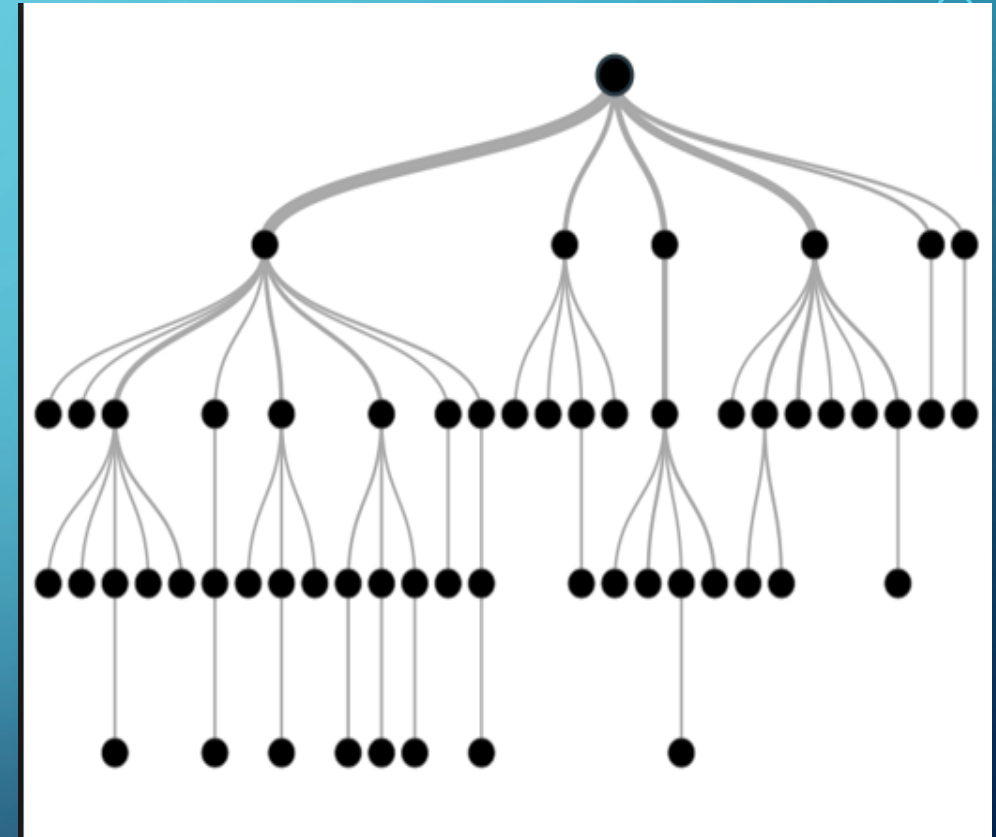




# DECISION TREE



- **First analysis using all attributes**
- **Second analysis using 14 attributes:**
  - Excluding state, account, area, phone, day call, even call, night call
- **Third analysis using 9 attributes:**
  - Int plan, vmail plan, int min, int calls, int charge, Cust calls, total min, total charge, churn
- **Forth analysis using 9 attributes without outliers**



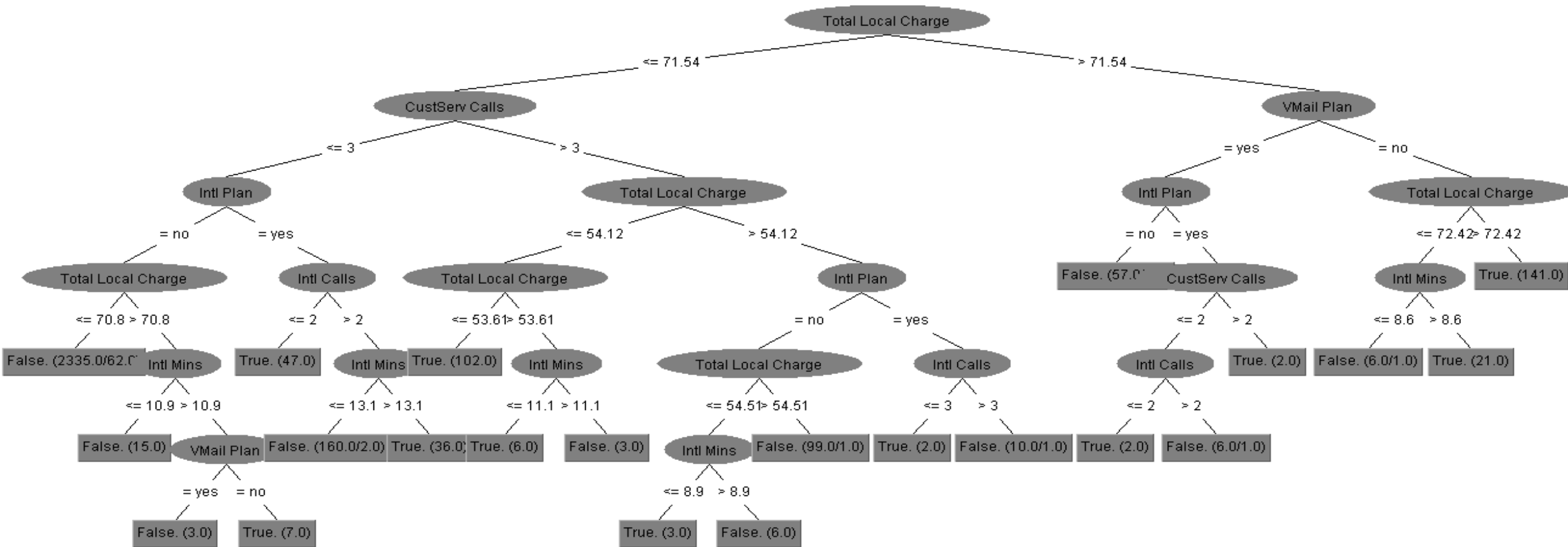


	All Attributes					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.981	0.352	0.994	0.463	0.98	0.36
Correctly classified instances	93.28%		92.76%		93.07%	

	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%	

	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%	

	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%	

[illegible]

# NAÏVE BAYES

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood:  $P(x|c)$   
 Class Prior Probability:  $P(c)$   
 Posterior Probability:  $P(c|x)$   
 Predictor Prior Probability:  $P(x)$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$



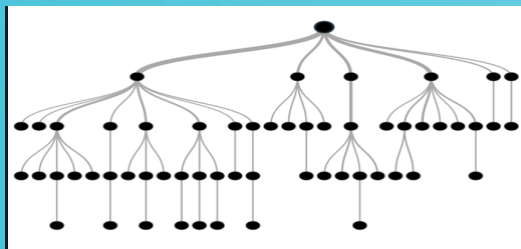
	All Attributes					
	Cross validation Folders 10		Split 66%		Cross validation Folders 3	
Class	TP	FP	TP	FP	TP	FP
FALSE	0.935	0.466	0.93	0.48	0.94	0.47
Correctly classified instances	87.66%		87.66%		87.93%	

	14 Attributes (remove[State,acount,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

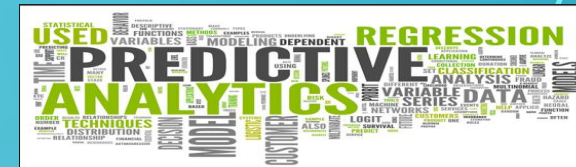


+

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood:  $P(x|c)$   
 Class Prior Probability:  $P(c)$   
 Posterior Probability:  $P(c|x)$   
 Predictor Prior Probability:  $P(x)$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$



	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%	

The background is a blue gradient with faint, stylized circuit lines in the corners. These lines are white and light blue, forming geometric patterns that resemble electronic circuitry. They are located in the top-left, top-right, bottom-left, and bottom-right corners of the image.

# POST-PREDICTIVE ANALYSIS

# POST-PREDICTION ANALYSIS

## K-means Algorithm

3 Clusters, 9 Attributes, Test mode: split 66% train, remainder test.



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



# POST-PREDICTION ANALYSIS



## K-means Algorithm

3 Clusters, 9 Attributes, Test mode: split 66% train, remainder test.

Group 1: High  
Customer service  
call

Group 2: High  
Local Cost

Group 3: High  
International and  
Local Cost

# POST-PREDICTION ANALYSIS



## K-means Algorithm

Recommendation:

Group 1: customers should get speedy access and solution to their issues.

Group 2:  
Alert customer on their usage and allow to set limit

Group 3:  
create special international package and alerts for high usage

# CONCLUSION



1. Selecting Attributes is very important to produce more accurate result in classification.
2. The best algorithm in this dataset is Decision Tree or PART as shown below:

