Department of Mathematics
University of Calabria

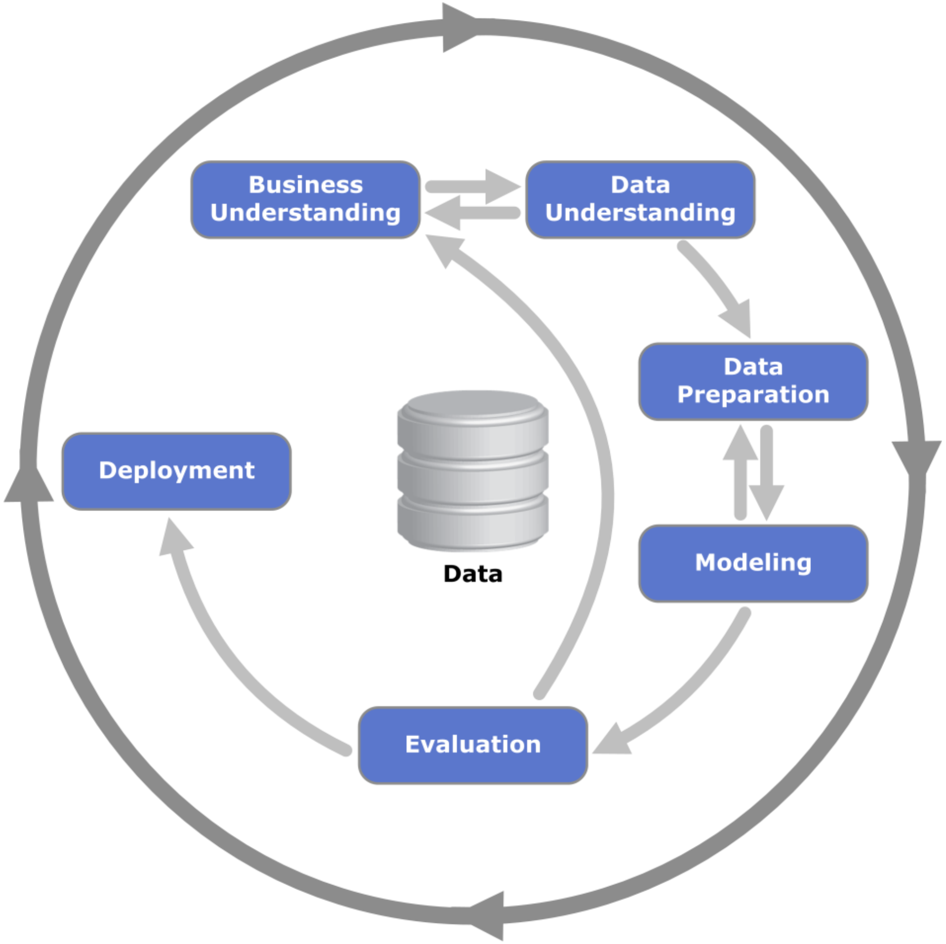
Data Warehouse and Data Mining

Module II – Data Mining

Study case: Churn Analysis

Ph.D. Ettore Ritacco

The Knowledge Discovery Process (CRISP-DM)



Business Understanding

- We are interested in modeling the customer attrition of a US mobile network operator (mobile phone company)
  - Customer attrition, also known as customer churn, is a business term describing the rate at which customers leave or cease paying for a product or service.
  - It's a critical figure in many businesses, as it's often the case that acquiring new customers is a lot more costly than retaining existing ones.

Data Understanding and Manipulation

- The dataset:
  - contains 3333 tuples
  - has 21 attributes
  - is unbalanced (w.r.t. the class attribute):
    - 2850 tuples are labeled as False. (85.5%)
    - 483 tuples are labeled True. (14.5%)

Data Understanding and Manipulation

- Data Schema (1/3):
  - State: categorical, for the 50 states and the District of Columbia
  - Account length: integer-valued, how long account has been active
  - Area code: categorical, regions within the states
  - Phone number: essentially a surrogate for customer ID
  - International Plan: dichotomous categorical, yes or no
  - VoiceMail Plan: dichotomous categorical, yes or no
  - Number of voice mail messages: integer-valued

Data Understanding and Manipulation

- Data Schema (2/3):
  - Total day minutes: continuous, minutes customer used service during the day
  - Total day calls: integer-valued
  - Total day charge: continuous, perhaps based on foregoing two variables
  - Total evening minutes: continuous, minutes customer used service during the evening
  - Total evening calls: integer-valued
  - Total evening charge: continuous, perhaps based on foregoing two variables
  - Total night minutes: continuous, minutes customer used service during the night

Data Understanding and Manipulation

- Data Schema (3/3):
  - Total night calls: integer-valued
  - Total night charge: continuous, perhaps based on foregoing two variables
  - Total international minutes: continuous, minutes customer used service to make international calls
  - Total international calls: integer-valued
  - Total international charge: continuous, perhaps based on foregoing two variables
  - Number of calls to customer service: integer-valued
  - **Churn?: dichotomous categorical, yes or no, CLASS ATTRIBUTE**

Data Understanding and Manipulation

- Attribute correlations
  - There are groups of 3 attributes that potentially may exhibit linear correlation:
    - Minutes
    - Calls
    - Charge

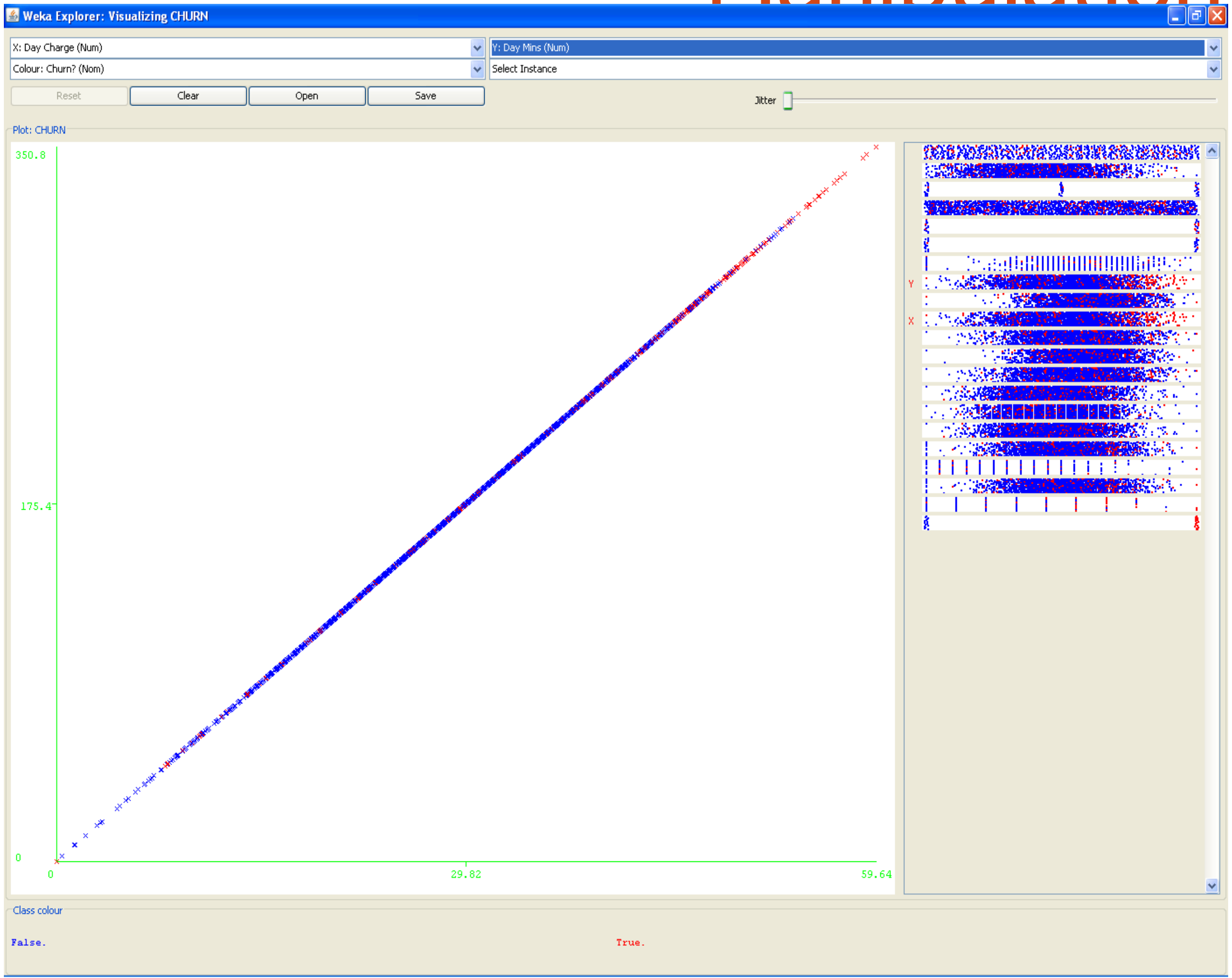
Data Understanding and Manipulation

e.g.:

day charge

and

day mins



Weka Explorer: Visualizing CHURN

X: Day Charge (Num)

Y: Day Mins (Num)

Colour: Churn? (Nom)

Select Instance

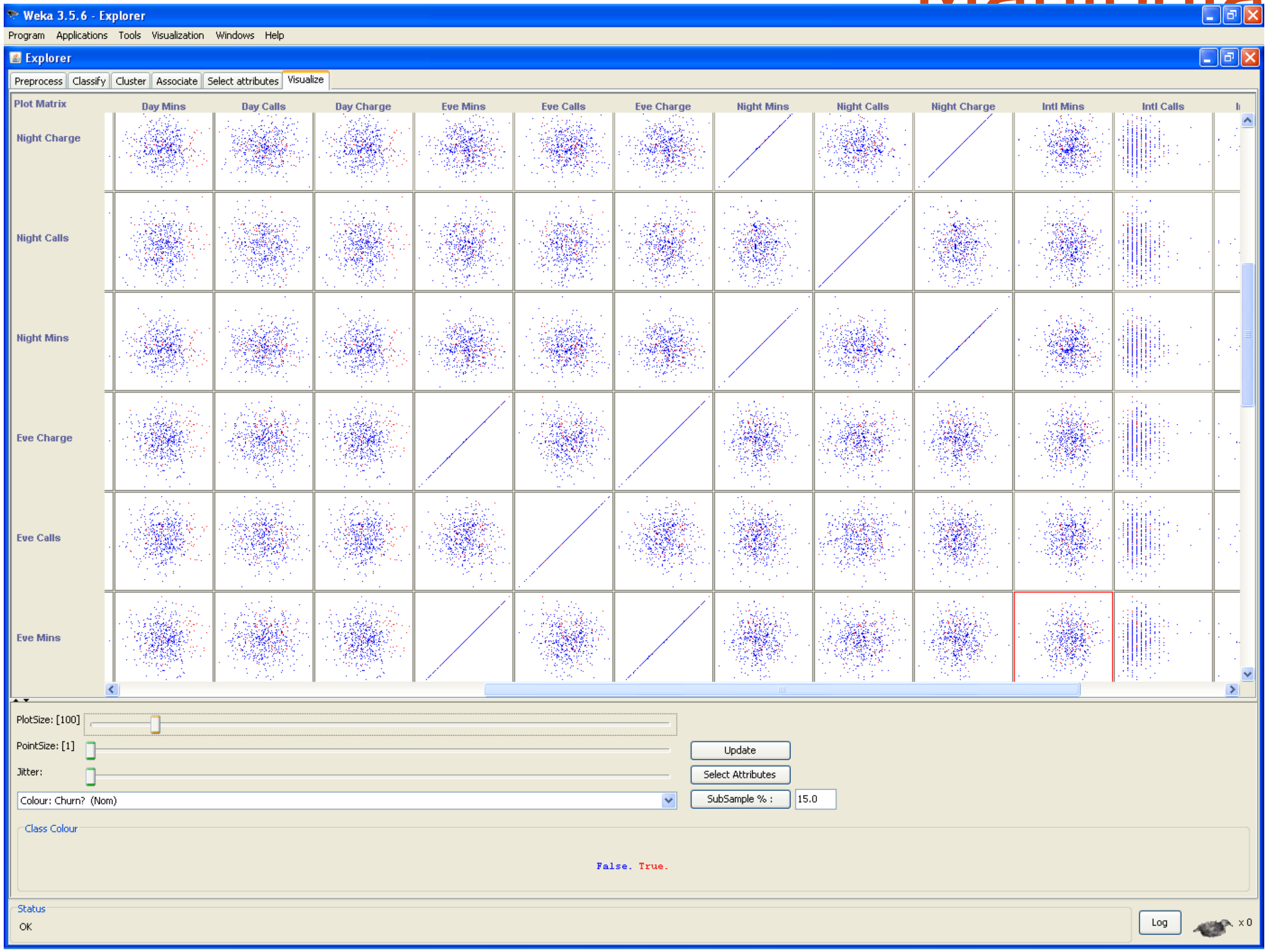
Plot: CHURN

Class colour

False.

True.

Data Understanding and Manipulation



PlotSize: [100]

PointSize: [1]

Jitter:

Colour: Churn? (Nom)

SubSample % : 15.0

Status

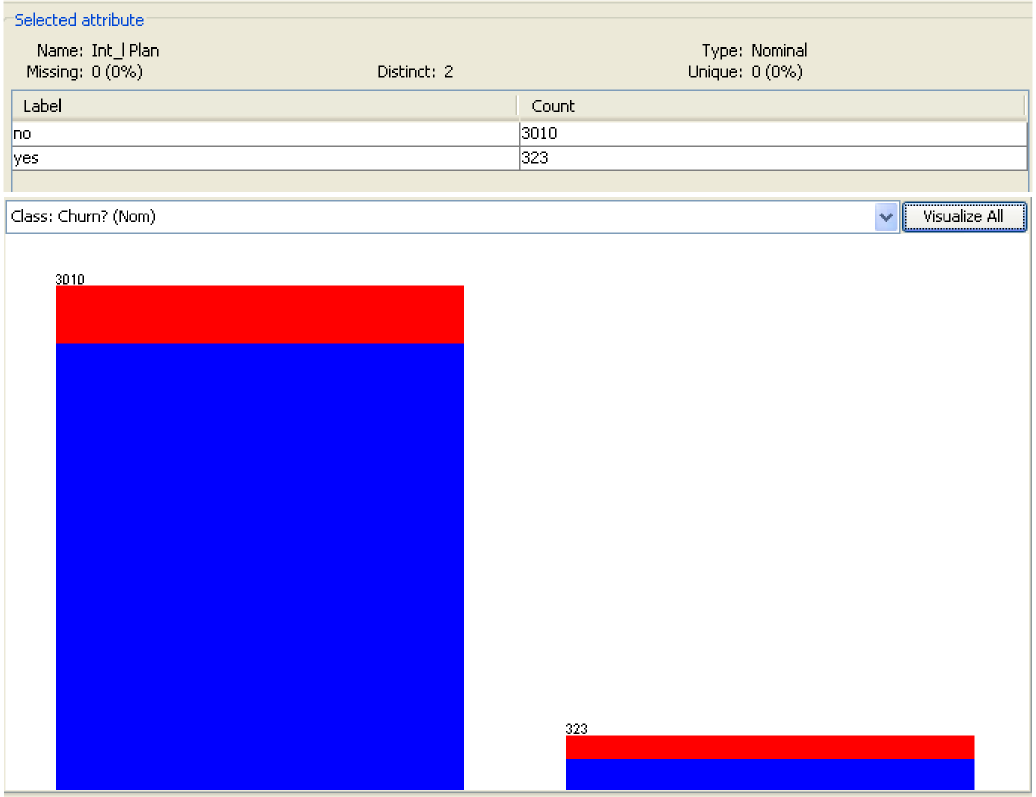
OK

Data Understanding and Manipulation

- Exploratory Analysis

Data Understanding and Manipulation

- International plan:



Selected attribute

Name: Int_l Plan
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

Label	Count
no	3010
yes	323

Class: Churn? (Nom)

Data Understanding and Manipulation

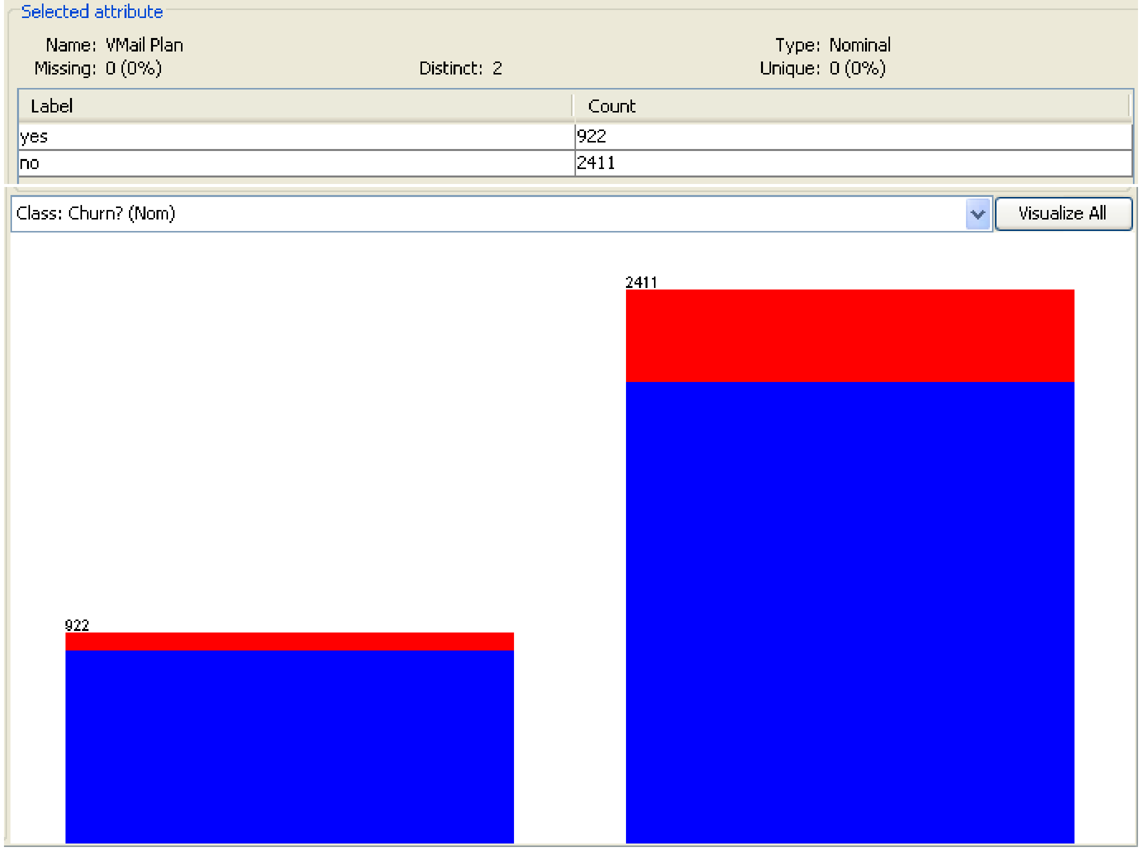
- In cross-tabulation:

International Plan		
Churn	No	Yes
False.	2664	186
True.	346	137

- The users with the international plan are 323
- The churners with this plan are 137 (42.4%).
- 137 churners (out of 483, 28.4%) have this plan.
- Then?

Data Understanding and Manipulation

- VoiceMail plan:



Selected attribute

Name: VMail Plan — Type: Nominal

Missing: 0 (0%) — Distinct: 2 — Unique: 0 (0%)

Label	Count
yes	922
no	2411

Class: Churn? (Nom)

Data Understanding and Manipulation

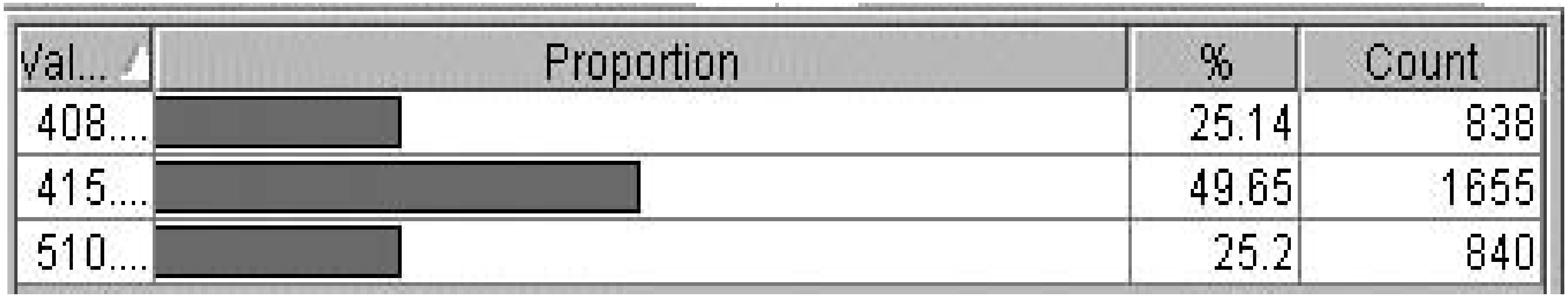
- Cross-tabulation

VoiceMail Plan		
Churn	No	Yes
False.	2008	842
True.	403	80

- The users with the international plan are 922
- The churners with this plan are 80 (8.7%).
- 137 churners (out of 483, 16.6%) have this plan.
- Then?

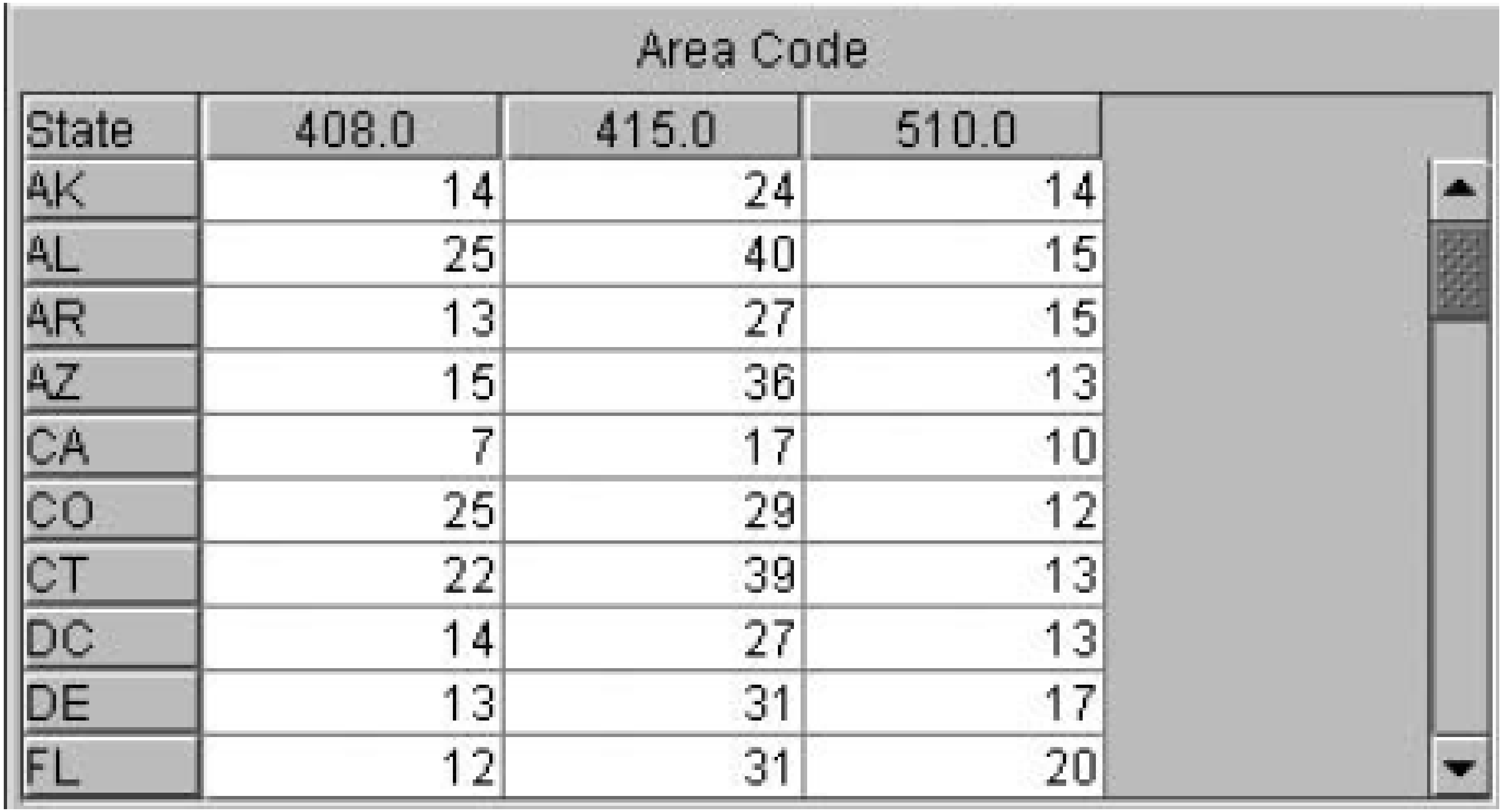
Data Understanding and Manipulation

- Anomaly Detection
  - Area code should span over all the US states, but has only 3 values
    - 408, 415, 510



Val...	Proportion	%	Count
408....		25.14	838
415....		49.65	1655
510....		25.2	840

Data Understanding and Manipulation



	Area Code		
State	408.0	415.0	510.0
AK	14	24	14
AL	25	40	15
AR	13	27	15
AZ	15	36	13
CA	7	17	10
CO	25	29	12
CT	22	39	13
DC	14	27	13
DE	13	31	17
FL	12	31	20

Data Understanding and Manipulation

- Maybe a domain expert can explain this phenomenon, or maybe there are some errors in the data

- We choose to remove this attribute

Data Understanding and Manipulation

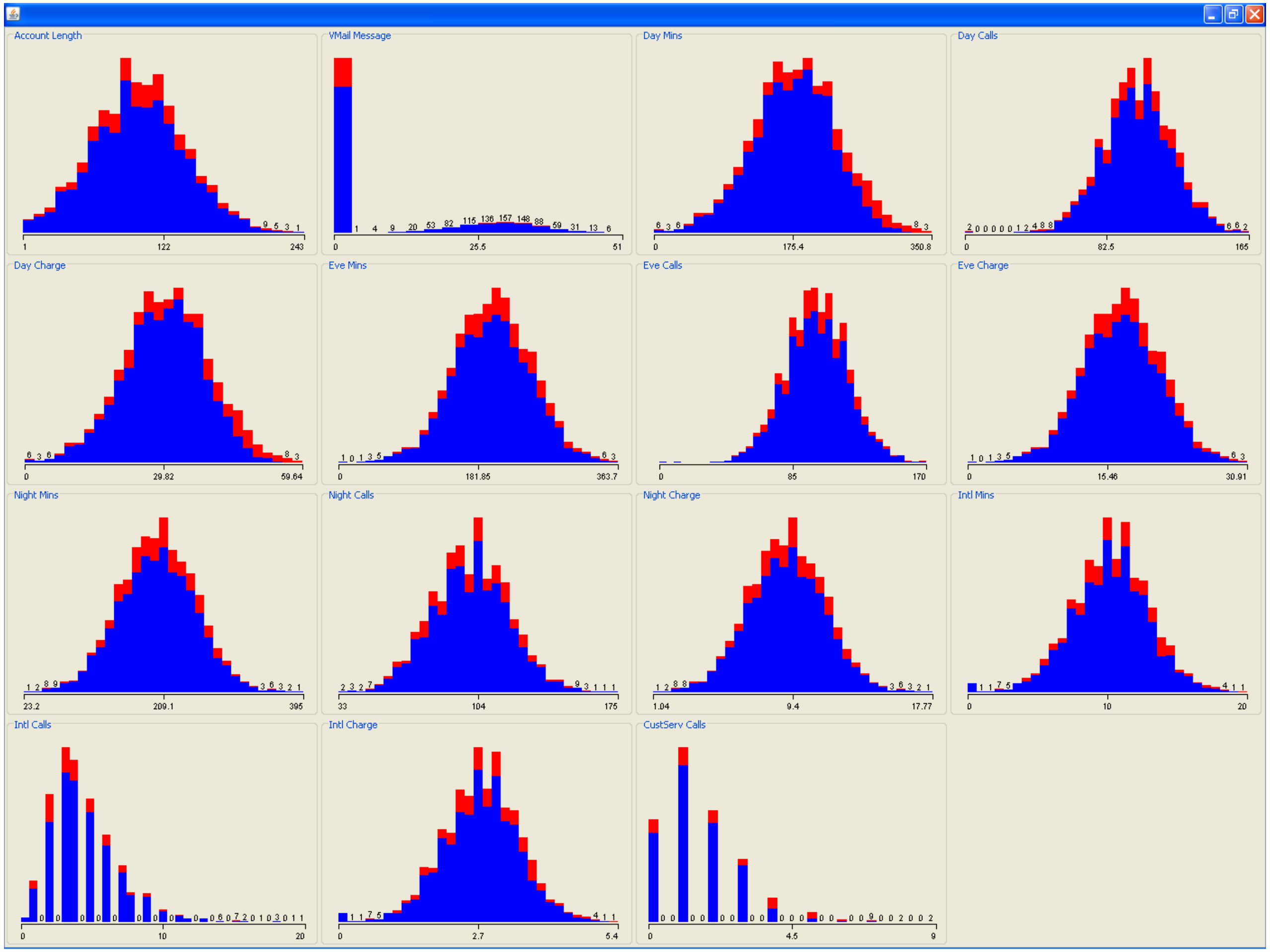
- Numerical Attributes:

	Max	Min	Avg	St.Dev.	Median
Account Length	*243*	*1*	*101,0648*	*39,81613*	*101*
VMail Message	*51*	*0*	*8,09901*	*13,68631*	*0*
Day Mins	*350,8*	*0*	*179,7751*	*54,45922*	*179,4*
Day Calls	*165*	*0*	*100,4356*	*20,06607*	*101*
Day Charge	*59,64*	*0*	*30,56231*	*9,258045*	*30,5*
Eve Mins	*363,7*	*0*	*200,9803*	*50,70624*	*201,4*
Eve Calls	*170*	*0*	*100,1143*	*19,91964*	*100*
Eve Charge	*30,91*	*0*	*17,08354*	*4,310021*	*17,12*
Night Mins	*395*	*23,2*	*200,872*	*50,56626*	*201,2*
Night Calls	*175*	*33*	*100,1077*	*19,56567*	*100*
Night Charge	*17,77*	*1,04*	*9,039325*	*2,275531*	*9,05*
Intl Mins	*20*	*0*	*10,23729*	*2,791421*	*10,3*
Intl Calls	*20*	*0*	*4,479448*	*2,460845*	*4*
Intl Charge	*5,4*	*0*	*2,764581*	*0,75366*	*2,78*
CustServ Calls	*9*	*0*	*1,562856*	*1,315294*	*1*

Data Understanding and Manipulation

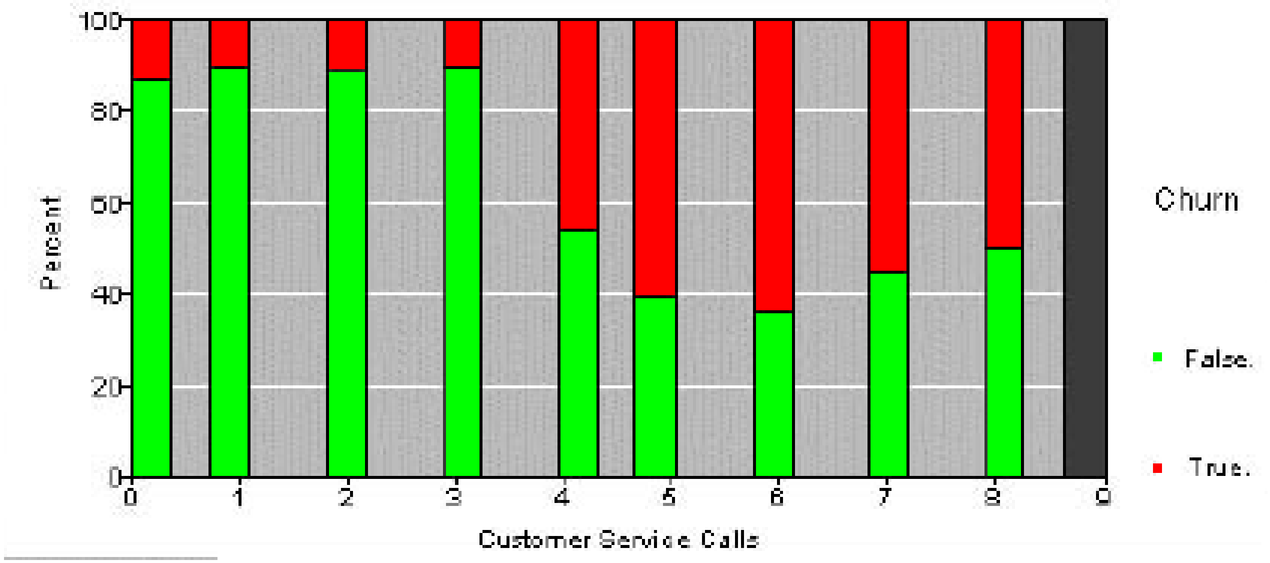
- Some attribute are symmetric:
  - Account Length, # minutes, # call, # charge
- Others are asymmetric:
  - VoiceMail message, Customer service call

Data Understanding and Manipulation



Data Understanding and Manipulation

- The CustomerServiceCalls attribute



- The attribute can be discretized with threshold 3.5, into 2 classes: *low* e *high*.

Summary

Account length	No visible relation with churn
Area code	Anomalous, removed
Phone number	ID, removed
International Plan	Good Predictor
VoiceMail Plan	Good Predictor
Number of voice mail messages	No visible relation with churn
Total day minutes	Good Predictor
Total day calls	No visible relation with churn
Total day charge	Redundant, removed
Total evening minutes	Good Predictor
Total evening calls	No visible relation with churn
Total evening charge	Redundant, removed
Total night minutes	No visible relation with churn
Total night calls	No visible relation with churn
Total night charge	Redundant, removed
Total international minutes	No visible relation with churn
Total international calls	No visible relation with churn
Total international charge	Redundant, removed
Customer service calls	Good Predictor

Modeling

- Data seem ready to be used in the modeling phase