Enterprise Database Technologies

Spring, 2017

ITT DUBLIN

Institute of Technology Tallaght

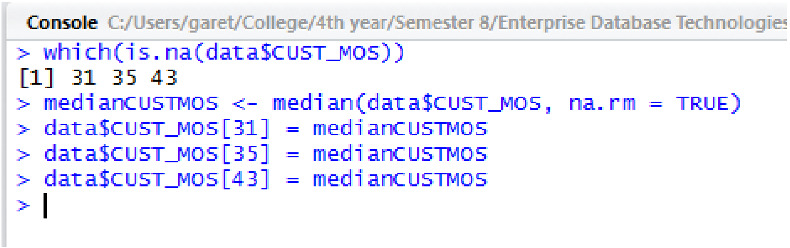
Institiúid Teicneolaíochta Tamhlacht

- Data Understanding
- Data Exploration
- Data Mining

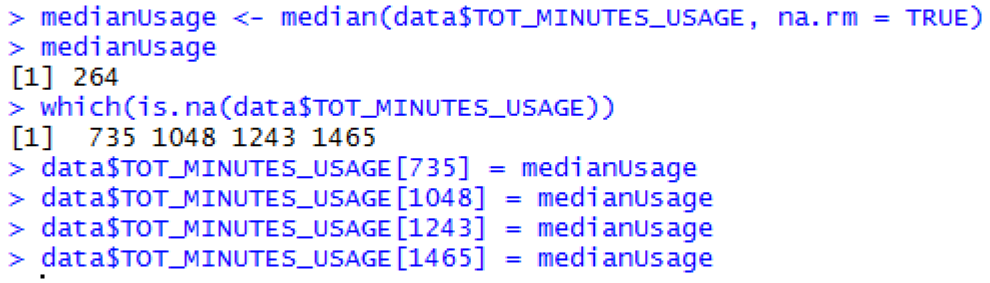
By Gareth Quirke X00108966

Part 1.

Preforming data pre-processing methods on the data set. Neglecting to do this process before the data modelling process would likely lead to models that are unreliable and results/findings that are inaccurate. To do this we will replace null values (NA) for the median of that column. Here I searched for missing values and replace them correctly. Some of the columns have no missing values. I also had to include missing string values in my search to ensure that the data mining will be effective.

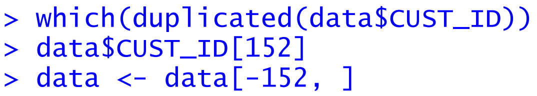

Console C:/Users/garet/College/4th year/Semester 8/Enterprise Database Technologies

```
> which(is.na(data$CUST_MOS))
[1] 31 35 43
> medianCUSTMOS <- median(data$CUST_MOS, na.rm = TRUE)
> data$CUST_MOS[31] = medianCUSTMOS
> data$CUST_MOS[35] = medianCUSTMOS
> data$CUST_MOS[43] = medianCUSTMOS
> 
```



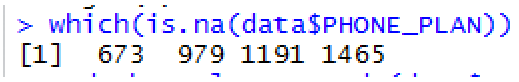
```
> medianUsage <- median(data$TOT_MINUTES_USAGE, na.rm = TRUE)
> medianUsage
[1] 264
> which(is.na(data$TOT_MINUTES_USAGE))
[1]  735 1048 1243 1465
> data$TOT_MINUTES_USAGE[735] = medianUsage
> data$TOT_MINUTES_USAGE[1048] = medianUsage
> data$TOT_MINUTES_USAGE[1243] = medianUsage
> data$TOT_MINUTES_USAGE[1465] = medianUsage
```

This Process was repeated for all of the columns in the data set. Another phase of question one was to check for duplicate records which I found in the correct column to check, Customer ID. I found one record which was a duplicate and removed it from the dataset.

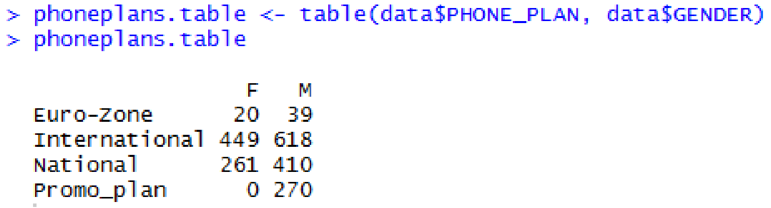


```
> which(duplicated(data$CUST_ID))
> data$CUST_ID[152]
> data <- data[-152, ]
```

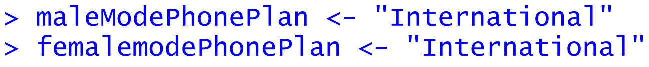
In this part of the assignment I found the columns that had NULL value for the categorical data value of which phone plan each customer had. As you can see 4 of the records did not exist. To determine the most common plan for each gender I created a table, this broke down the stats neatly and from here I Identified which value to place in this empty rows. I again repeated this same process for education which had some missing values also.



```
> which(is.na(data$PHONE_PLAN))
[1]  673  979 1191 1465
```



```
> phoneplans.table <- table(data$PHONE_PLAN, data$GENDER)
> phoneplans.table

                F   M
  Euro-Zone    20  39
  International 449 618
  National     261 410
  Promo_plan     0 270
```



```
> maleModePhonePlan <- "International"
> femalemodePhonePlan <- "International"
```

Part 2.

In this part, we create 3 categories for Income, I used this command.

```
data$INCOMEBRACKET<-cut(data$INCOME, c(0,38000,88000,321000))
```

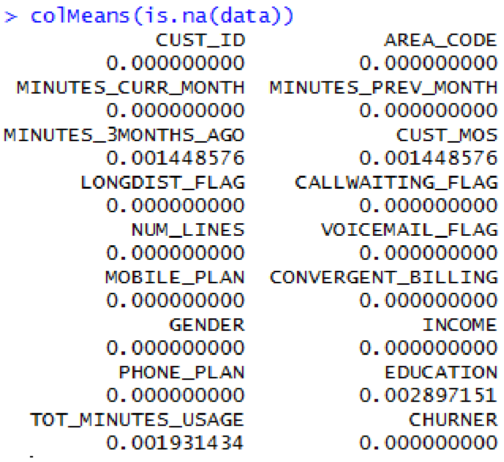
Part 3.

Predictor Variables

- Gender
- Education
- Total Minutes Usage
- Mobile Plan
- Convergent Billing

These variables will help determine factors that contribute to the churn rate of the company.

By running this command I found the percentage of missing values for every column. I used this to find part B for each of my predictor variables.



```
> colMeans(is.na(data))
             CUST_ID           AREA_CODE
         0.000000000         0.000000000
  MINUTES_CURR_MONTH  MINUTES_PREV_MONTH
         0.000000000         0.000000000
MINUTES_3MONTHS_AGO            CUST_MOS
         0.001448576         0.001448576
       LONGDIST_FLAG    CALLWAITING_FLAG
         0.000000000         0.000000000
           NUM_LINES      VOICEMAIL_FLAG
         0.000000000         0.000000000
         MOBILE_PLAN  CONVERGENT_BILLING
         0.000000000         0.000000000
              GENDER              INCOME
         0.000000000         0.000000000
          PHONE_PLAN           EDUCATION
         0.000000000         0.002897151
  TOT_MINUTES_USAGE             CHURNER
         0.001931434         0.000000000
```

Gender

a) Nominal

b) 0%

C) NA

D) NA

E) `ggplot(data, aes(x = data$CHURNER, fill = data$GENDER)) + geom_bar(position = "fill")`

From looking at this histogram I can see that Male members of the dataset lean slightly on the churn rate. This can impact the churn rate as there a higher count of males in the entire data set.

Education

A) Ordinal

B) 0.0028971%

C) NA

D) NA

E)
```
ggplot(data, aes(x = data$CHURNER, fill = data$EDUCATION)) + geom_bar(p
osition = "fill")
```

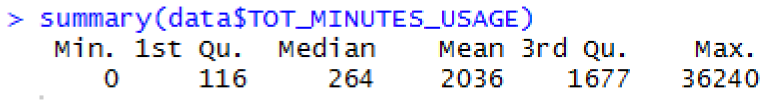
High levels of education (Bachelors and Masters) have a high level of churn rate.

Total Minutes Usage

A) Numeric

B) 0.001931%

C)



```
> summary(data$TOT_MINUTES_USAGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0     116     264    2036    1677   36240 
```

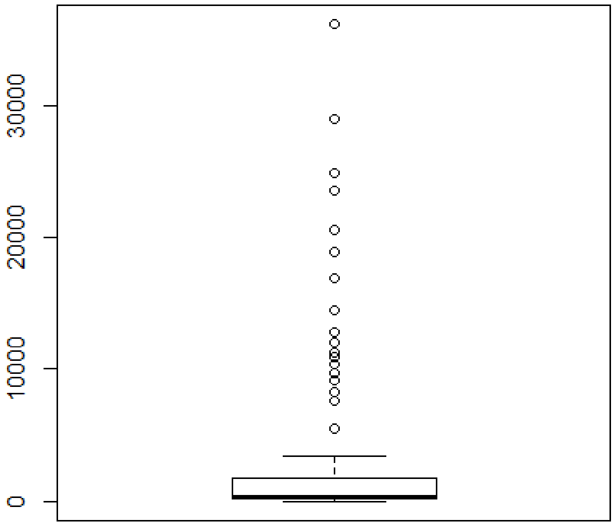
D) Positive

E) Higher total minutes tends to have a tendency to be on the yes side of the churn rate.

F) Skewness = (3 * (mean - median)) / standard deviation

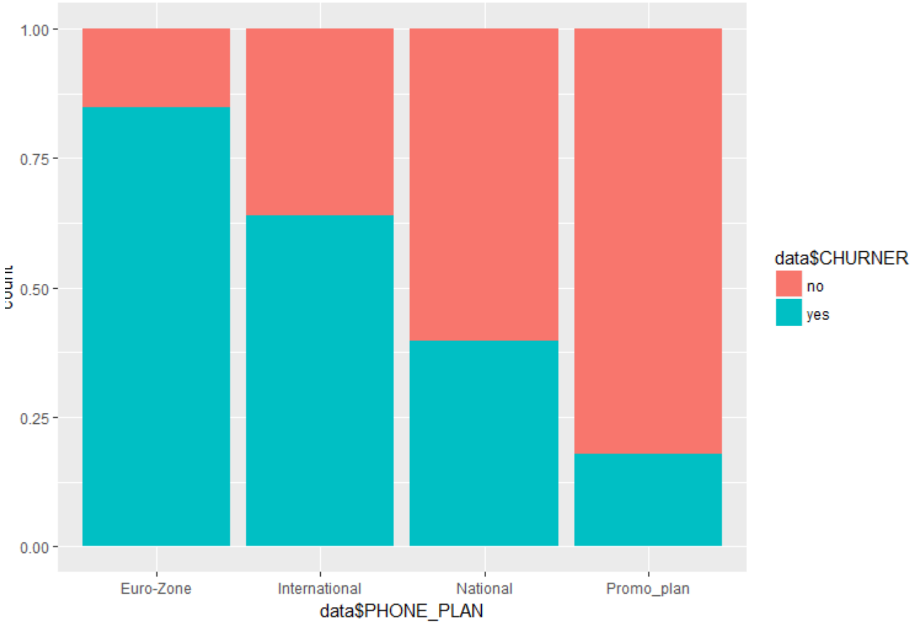
= 4.234245 (Positively Skewed)

G) I used boxplot to identify outliers for this data set. It made it easy to find the outliers. As we can see from the graph below there are a few high outliers whose overall usage is extremely high in comparison to the majority of the data which ranges between less than 10,000 and less than 20,000.



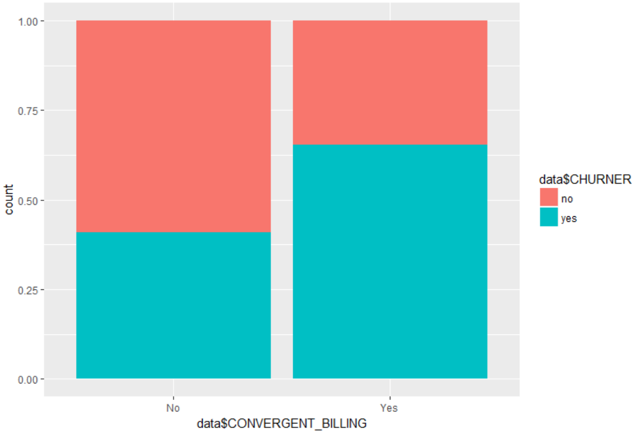
Mobile Plan

A) Ordinal
B) 0%
C) NA
D) NA
E)
```
ggplot(data, aes(x = data$PHONE_PLAN, fill = data$CHURNER)) + geom_bar(
position = "fill")
```



From this graph, I could tell that the promotional plan had a very low number in the churn rate. Both Euro-Zone and International had high levels of churn rates. This may be because of the level of competition given the scale of the market in these areas.

Convergent Billing
A) Nominal
B) 0%
C) NA
D) NA
E)
```
ggplot(data, aes(x = data$CONVERGENT_BILLING, fill = data$CHURNER)) + g
eom_bar(position = "fill")
```



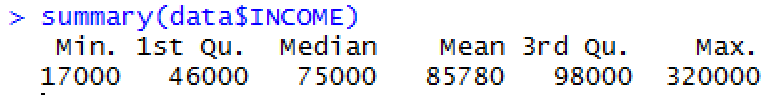
Convergent billing led to a higher churn rate than no type of this billing.

Income

A) Numeric

B) 0%

C)



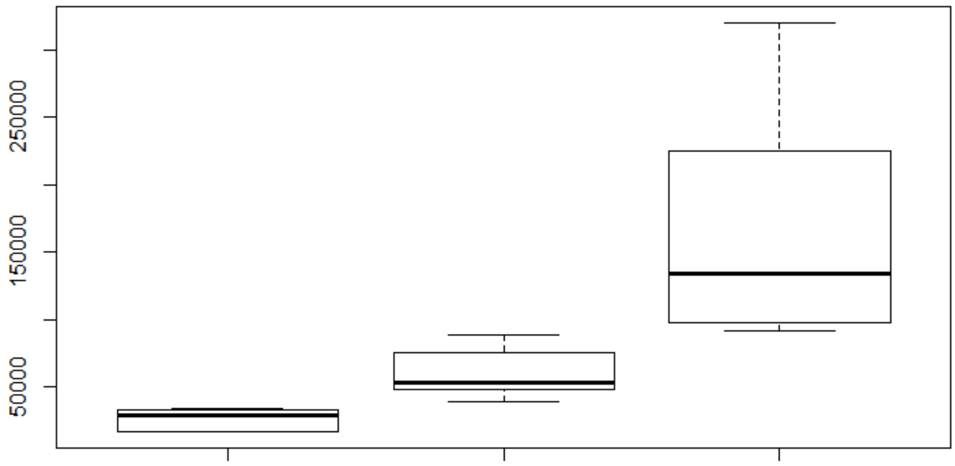
```
> summary(data$INCOME)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  17000   46000   75000   85780   98000  320000 
```

D) Evenly

E)

F) Positively Skewed (2.117082)

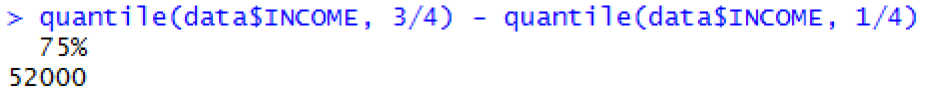
G)



There are skewed income records in the data set. They are skewed positively as I can tell from the graph above.

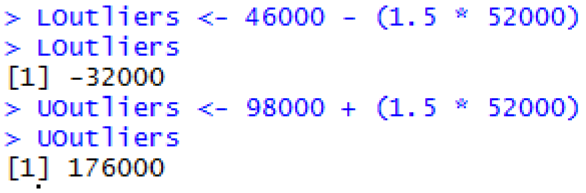
Part 4.

I will use the Interquartile range method and Z-Score standardization methods to identify outliers based on my inspections made on the numerical predictor variable – income, in step 2. First I find the IQR.

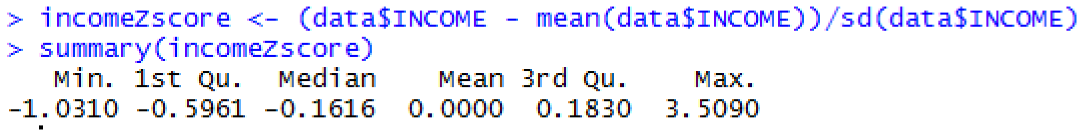


```
> quantile(data$INCOME, 3/4) - quantile(data$INCOME, 1/4)
  75%
52000
```

Now my upper outlier bound and my lower outlier bounds are as follows. This shows that there are outliers in the upper region as some income figures are well above 176,000. There are no lower outliers in this set. I did the same for Z-score standardization and found the same results, outliers in the upper or positive section of the data set.



```
> Loutliers <- 46000 - (1.5 * 52000)
> Loutliers
[1] -32000
> uOutliers <- 98000 + (1.5 * 52000)
> uOutliers
[1] 176000
```

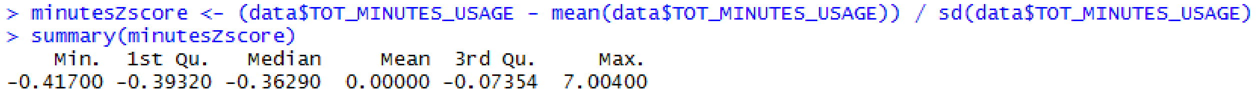


```
> incomeZscore <- (data$INCOME - mean(data$INCOME))/sd(data$INCOME)
> summary(incomeZscore)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.0310 -0.5961 -0.1616  0.0000  0.1830  3.5090
```

Part 5.

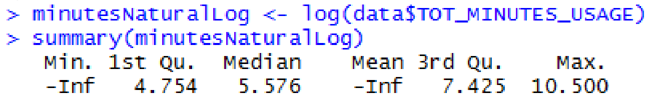
I have chosen to transform total minutes to achieve normality and determine how skewed the data is within this column. I found very skewed values in the graph produced as part of section 3 of the data mining phase for this assignment. The initial skewed values were on the positive side. The best ways to deal with this type of skewness is to use either root or log transformation of the data. I found that the data was evenly spread and showed a clear patter when using the square root transformation.

A)



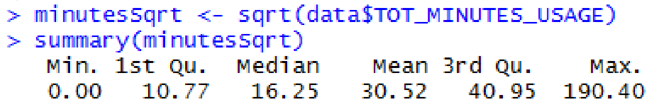
```
> minutesZscore <- (data$TOT_MINUTES_USAGE - mean(data$TOT_MINUTES_USAGE)) / sd(data$TOT_MINUTES_USAGE)
> summary(minutesZscore)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.41700 -0.39320 -0.36290  0.00000 -0.07354  7.00400
```

B)



```
> minutesNaturalLog <- log(data$TOT_MINUTES_USAGE)
> summary(minutesNaturalLog)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   -Inf   4.754   5.576    -Inf   7.425  10.500
```

C)



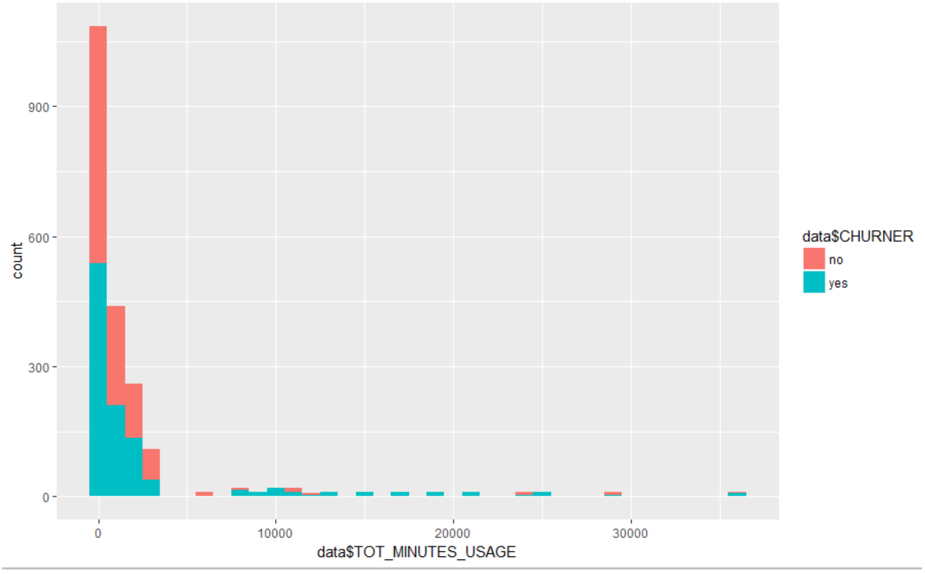
```
> minutesSqrt <- sqrt(data$TOT_MINUTES_USAGE)
> summary(minutesSqrt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   10.77   16.25   30.52   40.95  190.40
```

Part 6.

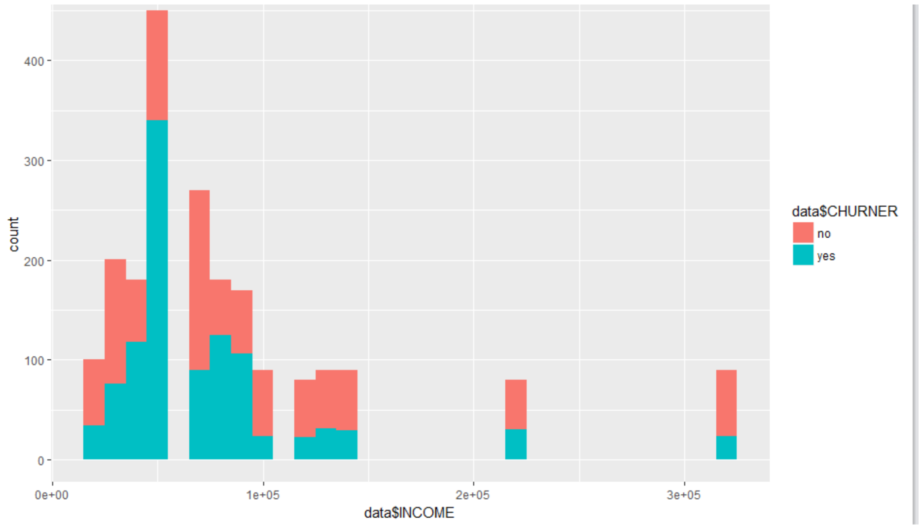
A)

For my two numerical variables, I constructed histograms with overlay of the response variable yes or no in terms of churn rate of a customer. As I've already mentioned before those who tend to have a higher demand for services are those who are most likely to be influential to the overall churn rate of the company. Those with a higher income and total minutes used overall tend to be likely key member of the churn rate.

Total Minutes Usage



Income

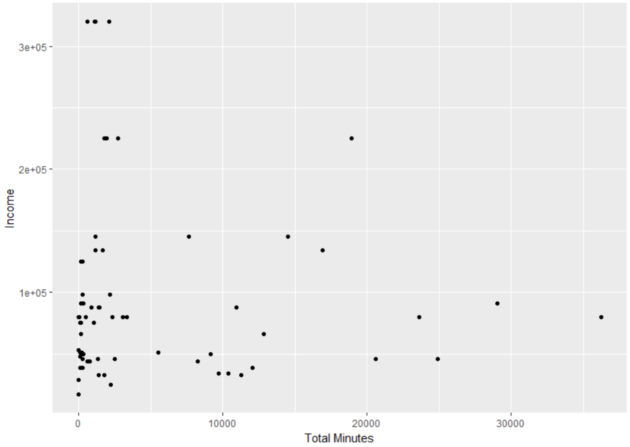


B) I believe that Income would make a significant difference in the CART data mining classification model. I feel that this variable would be prominent in the model since it would be a main decision node that really defines the data based on the level of income of the customer. As we can see from above the data is quite split particularly in the lower incomes.

Part 7.

Scatter plot for Income and Total Minutes Usage to investigate any correlations between the pair.

A)

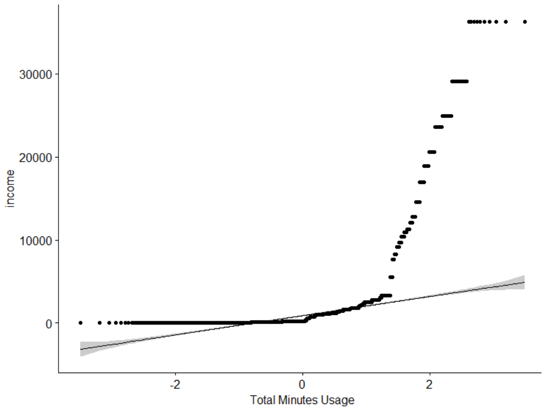


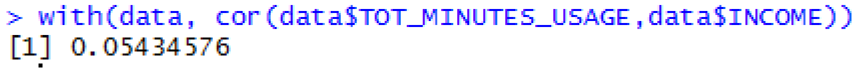
The correlation is strong in the lower areas of the data section. As the values increase the ties between the two variables decrease dramatically. This may be due to the large number of customers in between these mean values for both columns.

B)

I created a graph and used a correlation command to verify the results above. It was interesting to see the line heavily related in the lower areas of the graph. The second command computes the correlation coefficient.

```
ggqqplot(data$income, data$TOT_MINUTES_USAGE,ylab = "income", xlab = "Total Minutes Usage")
```

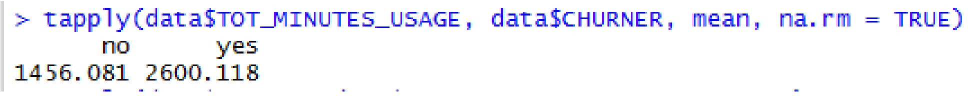




```
> with(data, cor(data$TOT_MINUTES_USAGE,data$INCOME))
[1] 0.05434576
```

C)

I find that total minutes used by a customer is highly involved in the rate of churning for this company.



```
> tapply(data$TOT_MINUTES_USAGE, data$CHURNER, mean, na.rm = TRUE)
      no      yes
1456.081 2600.118
```

As we can clearly see from above the number of records on the yes side for churning greatly exceeds the no side.

D) The two numerical values I chose Income and total minutes' usage are important so no I don't need to remove any variables. These two variables are influential in the churn rate as I have discovered up until this point.

Section 2

Weka is a data mining tool that improves decision making. The knowledge is attained through machine learning algorithms. This is a form of artificial intelligence to allow the company for example to find a set of rules that determine the most likely paths that lead to good or bad response variables. According to The Weka Workbench book published in 2016, '*pre-process a dataset, feed it into a learning scheme, and analyse the resulting classifier and its performance*'.

I will use the following schemes to analyse this data set: PART, JRip and J48. I used ZeroR as my baseline classifier as a comparison to the other learning schemes used in this section.

PART

This learning scheme creates a PART decision list. It uses divide and conquer algorithm to recursively break down a problem into sub-problems until they are simple enough to separate and solve. It builds a partial C4.5 decision tree in each iteration and makes the best leaf into a rule.

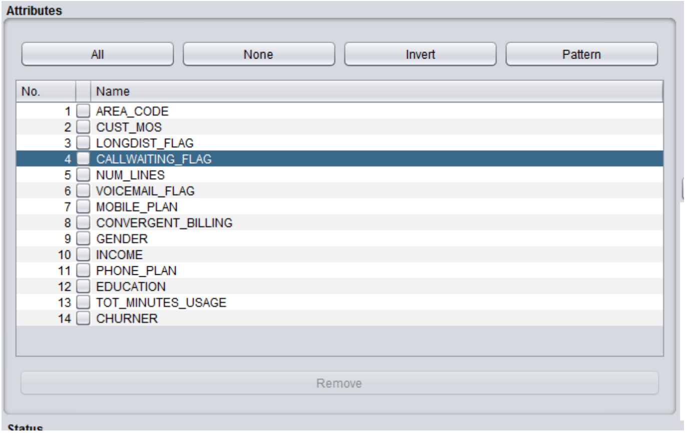
JRip

This learning scheme is based on RIPPER (repeated incremental pruning to produce error reduction). The algorithm is a rule based learner that generates a set of rules that identify classes. At each stage the rule with the highest rate of possible errors is pruned recursively until the set of rules is created with a high rate of accuracy.

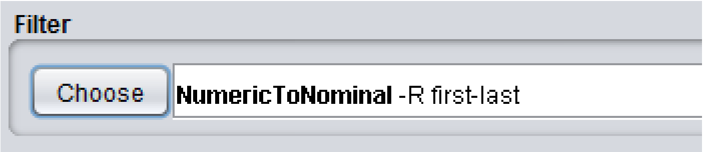
J48

This learning scheme is the implementation of the algorithm ID3 which stands for Iterative Dichotomiser 3. This is a precursor to the C4.5 algorithm. Its main characteristics are to select nodes with the highest amount of information gain. It uses information gain to help it decide which attribute goes into a decision node

I first imported the data set into Weka and removed the columns deemed to be redundant as part of the data mining task. After this I normalised all the numerical data by adding a filter to the data set.



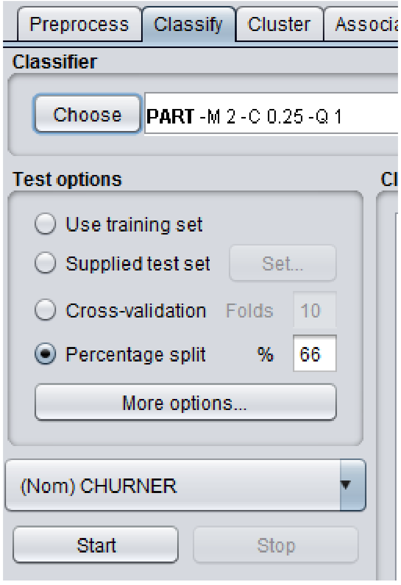
No.		Name
1	□	AREA_CODE
2	□	CUST_MOS
3	□	LONGDIST_FLAG
4	□	CALLWAITING_FLAG
5	□	NUM_LINES
6	□	VOICEMAIL_FLAG
7	□	MOBILE_PLAN
8	□	CONVERGENT_BILLING
9	□	GENDER
10	□	INCOME
11	□	PHONE_PLAN
12	□	EDUCATION
13	□	TOT_MINUTES_USAGE
14	□	CHURNER



NumericToNominal -R first-last

Using the learning schemes on the data

PART



Classifier

Choose PART -M 2 -C 0.25 -Q 1

Test options

○ Use training set

○ Supplied test set

○ Cross-validation Folds 10

● Percentage split % 66

(Nom) CHURNER

I used a percentage split for 66/34% for each of the learning algorithms used for the dataset. First, I will look at PART and determine its accuracy from the output panel after clicking start. The number of rules produced was 20.

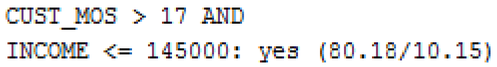
```
Number of Rules  :      20


Time taken to build model: 0.08 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.02 seconds

=== Summary ===

Correctly Classified Instances         576               81.8182 %
Incorrectly Classified Instances       128               18.1818 %
Kappa statistic                          0.6322
Mean absolute error                      0.2582
Root mean squared error                  0.3627
Relative absolute error                 51.5428 %
Root relative squared error             72.3357 %
Total Number of Instances              704
 === Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.694    0.068    0.903      0.694    0.785      0.649    0.857     0.883     yes
                 0.932    0.306    0.769      0.932    0.842      0.649    0.857     0.813     no
 Weighted Avg.   0.818    0.192    0.833      0.818    0.815      0.649    0.857     0.847
```

PART

how does the classifier determine whether a customer is a churner or not?

PART algorithm determines whether a customer is a churner or not by creating different paths to the to the set target variable which is CHURNER. It bases this of several resulting rules which lead to the highest overall accuracy of determine if the customer is a churner or not. The number of rules in this case is 20.

Do the decisions made by the classifiers make sense to you?



```
CUST_MOS > 17 AND
INCOME <= 145000: yes (80.18/10.15)
```

Yes, the decisions made by the classifier does make sense and looking at this rule it seems very sensible given the amount of data processing done in the previous section. Rules such that include for example area code and women make less sense to me. I don't think it makes much sense to generalise a customer in these areas.

What are the key predictors of churn?

Education, Gender, Income, Total minutes' usage and the number of months a customer is with the provider are key predictors that show up most commonly in this model.

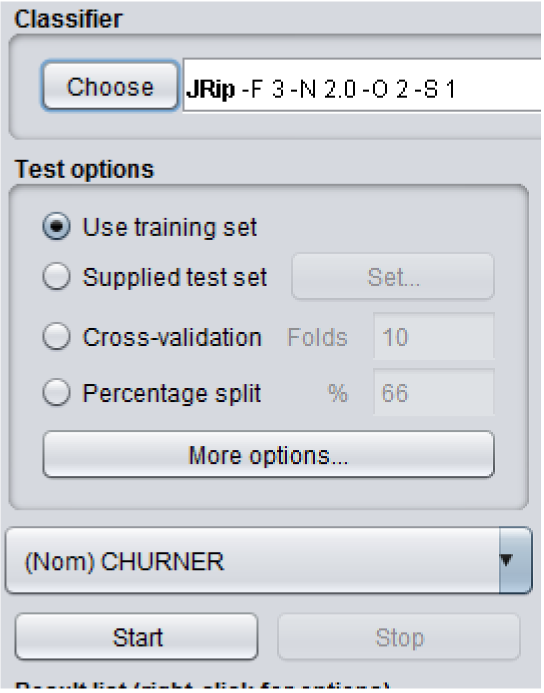
What are the significant rules\decision tree paths? Provide evidence for your assertions.

Convergent Billing = NO,
Customer months greater than 44,
Education = Masters, PHD, Bachelors,

Each of these rules had many records for each determining if the customer was a churner or not. Those with a higher degree of education tended to be a part of the churn rate being on the yes side. While convergent billing was no it also tended to have a large number of customers who didn't churn.

i) Proportion of false positives: 25
ii) Proportion of false negatives: 103
iii) Overall error rate and accuracy of model: Error rate: 18%, Accuracy rate: 81%
iv) Precision: 83%
v) True positive rate: 0.818
vi) False positive rate: 0.192
vii) ROC: 0.857 (1 is perfect, so this model is highly accurate).

JRip



Classifier

Choose JRip -F 3 -N 2.0 -O 2 -S 1

Test options

● Use training set

○ Supplied test set Set...

○ Cross-validation Folds 10

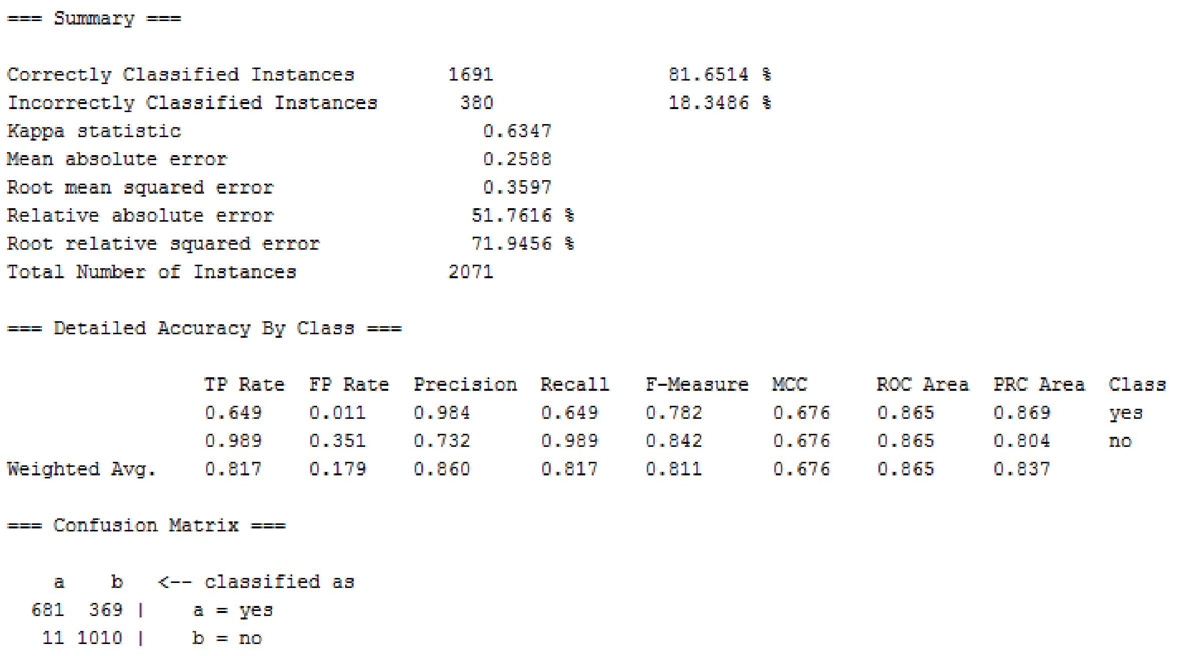
○ Percentage split % 66

More options...

(Nom) CHURNER

Start Stop

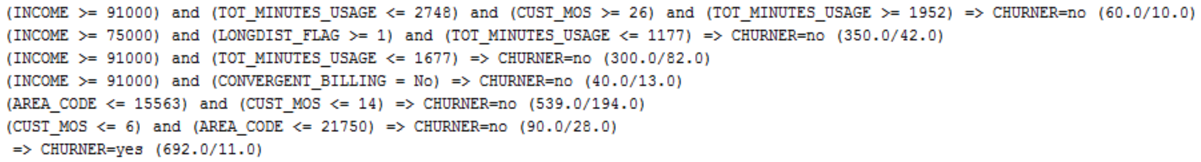
Result list (right-click for options)



```
=== Summary ===

Correctly Classified Instances        1691               81.6514 %
Incorrectly Classified Instances       380               18.3486 %
Kappa statistic                          0.6347
Mean absolute error                      0.2588
Root mean squared error                  0.3597
Relative absolute error                 51.7616 %
Root relative squared error             71.9456 %
Total Number of Instances             2071

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.649    0.011    0.984      0.649    0.782      0.676    0.865     0.869     yes
                 0.989    0.351    0.732      0.989    0.842      0.676    0.865     0.804     no
Weighted Avg.    0.817    0.179    0.860      0.817    0.811      0.676    0.865     0.837

=== Confusion Matrix ===

    a    b   <-- classified as
  681  369 |    a = yes
   11 1010 |    b = no
```

Using JRip with the training set. It created less rules than previously seen in PART, just 7. This is also highly accurate and I like that the number of rules is much lower, it is easier to determine the process of the learning scheme in this scenario.

how does the classifier determine whether a customer is a churner or not?



```
(INCOME >= 91000) and (TOT_MINUTES_USAGE <= 2748) and (CUST_MOS >= 26) and (TOT_MINUTES_USAGE >= 1952) => CHURNER=no (60.0/10.0)
(INCOME >= 75000) and (LONGDIST_FLAG >= 1) and (TOT_MINUTES_USAGE <= 1177) => CHURNER=no (350.0/42.0)
(INCOME >= 91000) and (TOT_MINUTES_USAGE <= 1677) => CHURNER=no (300.0/82.0)
(INCOME >= 91000) and (CONVERGENT_BILLING = No) => CHURNER=no (40.0/13.0)
(AREA_CODE <= 15563) and (CUST_MOS <= 14) => CHURNER=no (539.0/194.0)
(CUST_MOS <= 6) and (AREA_CODE <= 21750) => CHURNER=no (90.0/28.0)
 => CHURNER=yes (692.0/11.0)
```

The rules to determine if a customer is a churner or not is determined by these set of rules. These are a combination of rules that have the lowest error rate.

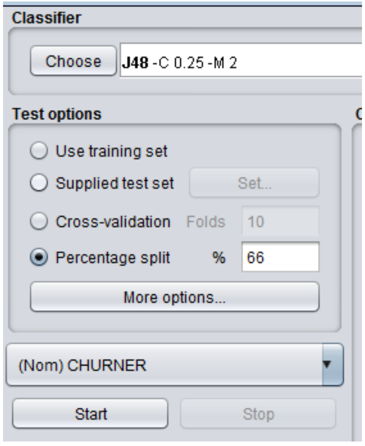
Do the decisions made by the classifiers make sense to you?
Yes, they do. It is sensible that income and total minutes' usage are common in each of the rules defined above.

What are the key predictors of churn?
Income, Total minutes' usage, number of months a customer is with the provider (CUST_MOS) are key predictors in this model.

What are the significant rules\decision tree paths? Provide evidence for your assertions.
Those living in an area code below 15563 and who are customers for less than 14 months are valuable customers as they have a very low rate of churn.

i) Proportion of false positives: 11
ii) Proportion of false negatives: 369
iii) Overall error rate and accuracy of model: Error rate is 18.3486%, Accuracy of Model: 81.65%
iv) Precision: 0.860
v) True positive rate: 0.817
vi) False positive rate: 0.179
vii) ROC: 0.865

J48



Classifier

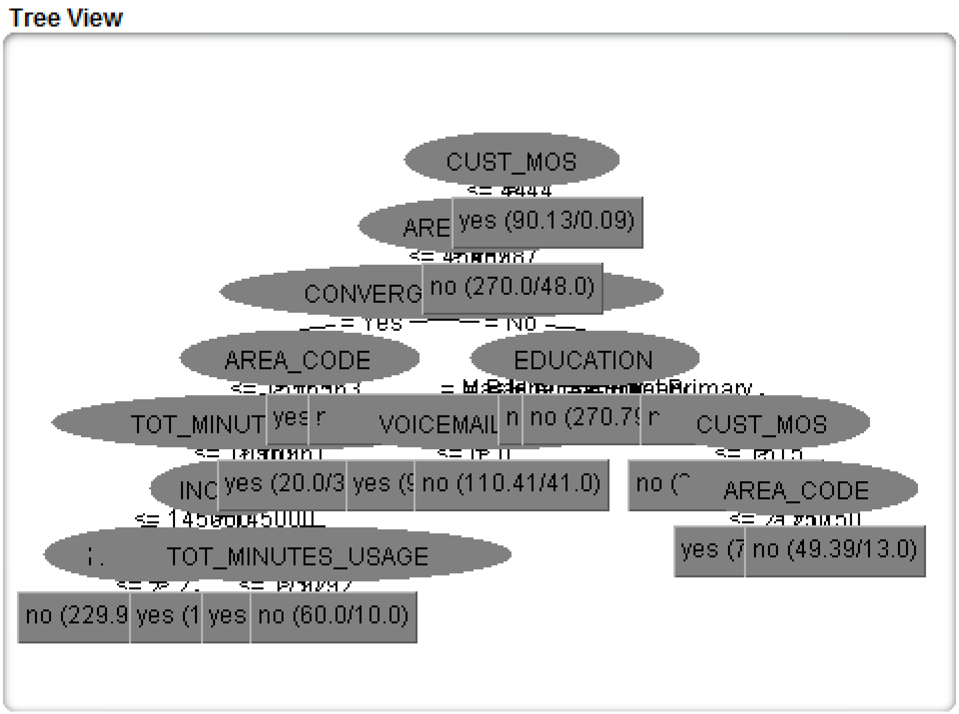
J48 -C 0.25 -M 2

Test options

○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 66

(Nom) CHURNER

J48 with the percentage split 66 – 34%. This will produce a tree with several decision nodes or leaf's. From this, I can also visualise the actual tree to get a better sense of the model produced by this algorithm.



```
=== Summary ===

Correctly Classified Instances         578               82.1023 %
Incorrectly Classified Instances       126               17.8977 %
Kappa statistic                          0.6376
Mean absolute error                      0.262
Root mean squared error                  0.3619
Relative absolute error                 52.3133 %
Root relative squared error             72.1808 %
Total Number of Instances              704

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.685    0.054    0.920      0.685    0.786      0.658    0.857     0.882     yes
                 0.946    0.315    0.766      0.946    0.846      0.658    0.857     0.801     no
Weighted Avg.    0.821    0.190    0.840      0.821    0.817      0.658    0.857     0.840

=== Confusion Matrix ===

   a   b   <-- classified as
 231 106 |   a = yes
  20 347 |   b = no
```

This learning scheme created 17 leaves and a tree with an overall size of 29. It also produced similar results in terms of accuracy to the other algorithms with an 82% rate.

how does the classifier determine whether a customer is a churner or not?

Through a series of leaves which determine if the customer is a churner or not at the root of the tree. The customer will have its attributes passed through the decision nodes and will end up at either yes or no at the root of the tree. The rules to determine if a customer is a churner or not is determined by these set of rules. These are a combination of rules that have the lowest error rate.

Do the decisions made by the classifiers make sense to you?

Yes, they do. It is sensible that income and total minutes' usage are common in the tree and make early appearances at the top of the tree. Other decisions such as separating the customer lower in the tree by level of education make sense to me also.

What are the key predictors of churn?
Income, Total minutes' usage, number of months a customer is with the provider (CUST_MOS), Education and Area Code.

What are the significant rules\decision tree paths? Provide evidence for your assertions.
Level of income is significant as it appears most often and separates the most number of nodes in the tree.

i) Proportion of false positives: 20
ii) Proportion of false negatives: 106
iii) Overall error rate and accuracy of model: Error rate is 17.8977%, Accuracy of Model: 82.1023%
iv) Precision: 0.840
v) True positive rate: 0.821
vi) False positive rate: 0.190
vii) ROC: 0.857

What is it that makes a customer churn?
Important factors that make customers churn from my analysis on this data set are level of income, the number of minutes' usage and how many months they have already been with the company as a customer. These are important to look at as some may be loyal customers and their probability of churn is very low. Those with higher education also tend to consider competition more often than a person with lower degree of education. Another important aspect I discovered was that customers with higher total minutes usage tend to become a part of the churn rate. These customers are also more valuable to the company so loosing these would be harmful to it.

Are some customers more likely to churn than others?
Yes, I have seen from my learning schemes produced in weka that customers with higher education are more often likely to churn.

How can we identify these customers before they churn?
We can identify these customers before they reach a certain point of being a long-standing customer. For example, if the customer is with the company for over 14 months it may be time to offer them a new deal to prevent them from churning. This is especially true when the customer is college educated and earns a higher income than average.