Case Study 2

CAB 330

Jack Farrel – N8640866

Anton Polison – N9907637

Dennis Tran – N9937072

2017

Table of Contents

[Task 1: Data Selection and Distribution. 2](#_Toc492835509)

[Task 2: Predictive Modelling Using Decision Trees 5](#_Toc492835510)

[Task 3: Predictive Modelling Using Regression 8](#_Toc492835511)

[Task 4: Predictive Modelling Using Neural Networks 11](#_Toc492835512)

[Task 5: Comparing Predictive Models 13](#_Toc492835513)

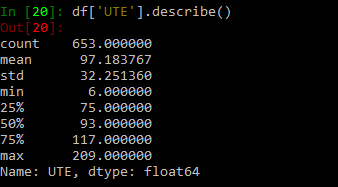
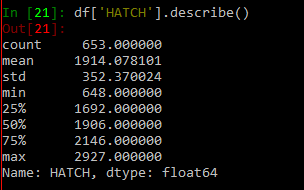
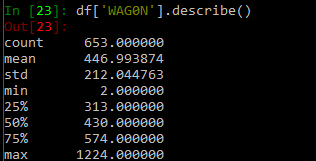
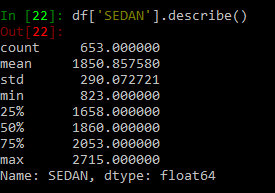
[Appendix A 16](#_Toc492835514)

# Task 1: Descriptive Data Mining - Clustering.

1. There were some changes made to the original dataset to ensure that the information from the analysis was accurate. There are several quality problems with “model\_car\_sales.csv” data set such as: missing values and redundant or irrelevant variables
   1. ***Missing Values:*** The dataset contains 22 rows which have empty values. These values were deleted as they are irrelevant for our research.
   2. ***Dropped Variables:*** In total 4 variables were dropped due to various reasons. The below table shows the dropped variables and there corressponding reasons.

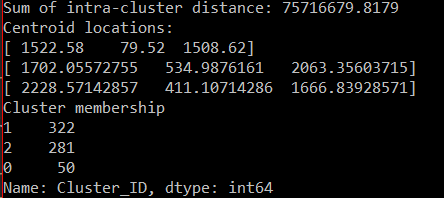
|  |  |
| --- | --- |
| *Variable* | *Reason For Drop* |
| *REPORT\_DATE* | Irrelevant. The variable has the date when the report was generated and doesn’t provide any useful information. |
| *DEALER\_CODE* | Redundancy. “LOCATION” variable provides more organised information which is at the same time easier to work with as it is numerical variable. |
| *K\_SALES\_TOT* | Irrelevant. The company want to find the minimum number of product sale segments and it is not insterested in identifying what is number of total car sales. |

1. After preprocessing and analyzing the dataset it was identified that “UTE” is clearly underperforming in sales in comparison with “HATCH”, “SEDAN” and “WAGON”.

Form the tables above we can determine that “mean” of the variables “HATCH”, “WAGON”, “SEDAN” is significantly higher than of “UTE” variable.

1. “HATCH”, “SEDAN”, “WAGON” and “LOCATION” variables were included in the analyses.
2. a)

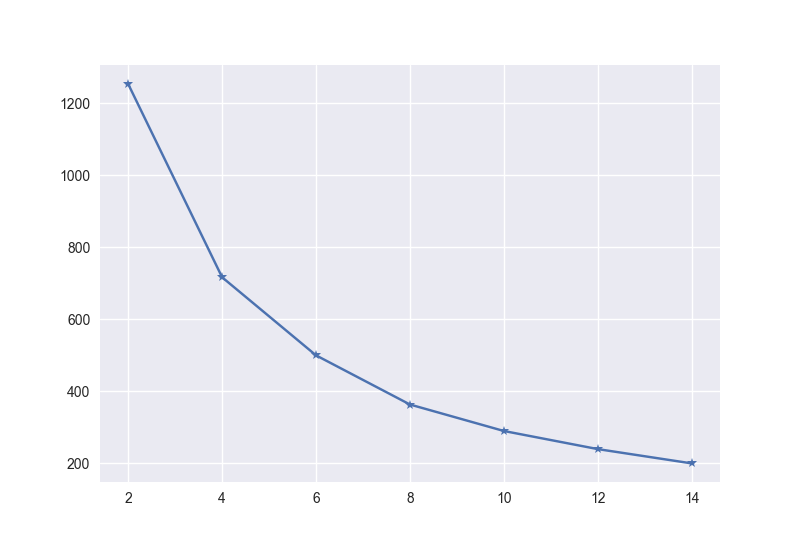


From cluster membership table we can identify that 50 records were assigned to cluster 0, 322 records to cluster 1 and 281 records to cluster 2.

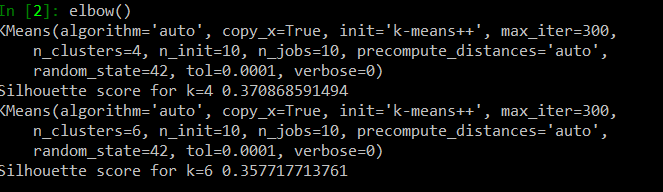
b)



1. “HATCH”, “SEDAN”, “WAGON” and “LOCATION” variables were included in the analyses.
2. Elbow method



“Silhouette Method”

”

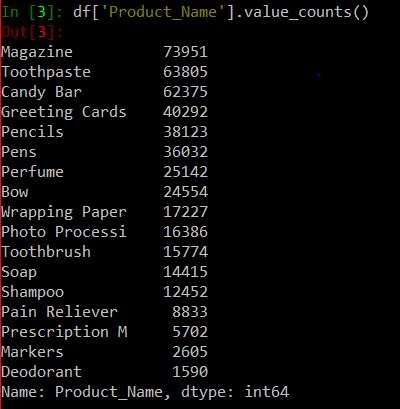
# ”

# Task 2: Descriptive Mining – Association

|  |  |
| --- | --- |
| *Variable* | *Importance* |
| *Location* | Irrelevant. The variable has no use to our research |
| *Transactin\_Date* | Irrelevant. The transaction date’s values are similar and of no use. |
| *Quantity* | Irrelevant. Each row of Quantity variable has a value of “1”. |

1. “pos\_transaction.csv” doesn’t have obvious erroneous or missing values. Some of the variables were dropped due to irrelevancy in our research.

1. “Transaction\_Id” and “Product\_Name” were 2 most valuable variable which we were using to perform association mining. “Transaction\_Id” variable include the transaction ids which group the products under particular transaction. At the same time “Product\_Name” attribute has the following values:







* 1. The table is illustrating the association rules between different items. It’s noticed that 2 resulting rules have the highest lift of 3.601370. They are “**Toothbrush => Perfume**” and vice versa “**Perfume => Toothbrush**”.
  2. Furthermore, it is identified that the highest confidence value has “**Perfume => Toothbrush**” rule with the value 0.323979.
  3. If we plot together the confidence, lift and support we get the following scenario. The customers who buy **Toothbrush** are likely to purchase **Perfume** with it. The lift of this rule is quite high which denotes strong association rules between these two products. On the other hand, the rule “**Perfume => Toothbrush**” has the same lift level as “**Toothbrush => Perfume**” but higher confidence level implying that if there is **Perfume** in the transaction it’s likely that **Toothbrush** will be in the transaction as well. Further we can observe interesting rules is **“Bow => Toothbrush”** with a quite a high lift level of 3.081236 but if we look closer at the confidence level it appears to be significantly lower than confidence level of other association rules with the same lift level. However, from the business prospective this rule is “interesting” as it has positive correlation (positive lift value). Respectively, there are couple of other association rules which can attract business’s attention such as **“Greeting Cards => Candy Bar, Magazine”**, **“Candy Bar => Pencils, Toothpaste”**, **“Greeting Cards => Magazine, Toothpaste”**. It’s worth to mention that confidence level for the 3 rules is very high (around 40%) meaning that if an item form the “Right\_side’ (see table above) appears in the transaction it’s 40% of the chance that items form the “Left\_side” appear in the same transaction.

1. a.From the table above, we can notice 4 association rules describing items which are bought together with “Candy Bar”. They are **“Candy Bar => Pencils, Toothpaste”**, **“Candy Bar => Magazine**, **Greeting Cards”**, **“Candy Bar => Magazine, Toothpaste”**, **“Candy Bar => Toothpaste, Greeting Cards”**.

# 

# Task 3: Predictive Modelling Using Regression

1. In preparation for regression, apply transformation method(s) to the variable(s) that need it. List the variables that needed it.
   1. All methods and variables have been transformed according to the specifications laid out in Task 1 where data pre-processing was described. This includes the changing of categorical variables to binary, and fixing data set issues such as missing values
2. Build a regression model using the default regression method with all inputs. Once you did it, build another one and tune it using GridSearchCV.
   1. All 53 features from the data set are included in the regression model
   2. Based on the Feature importance metric, the following five variables were considered most valuable:

|  |  |
| --- | --- |
| *Variable* | *Importance* |
| *AGE* | -0.665997663715 |
| *GENDER (GENDER\_F)* | 0.390662546399 |
| *GENDER (GENDER\_U)* | -0.399017685825 |
| *AFFL (*FFL\_21.0) | 0.376235146978 |
| *AFFL (*FFL\_22.0) | 0.307160310202 |

* 1. The model shows no sign of overfitting due to the small difference between the train and test accuracies.
  2. To build the regression model with GridSearchCV, the parameter ‘cv’ with a value of 10 eas used for cross validation. *‘LogisticRegression’* was used for the *‘estimator’*, with the *‘n\_jobs’* parameter being set to -1 to allow the model to use as much processing as possible.
  3. **Train accuracy:** 0.80901259964  
     **Test accuracy:** 0.81205939703

1. Build another regression model using the subset of inputs selected by RFE and selection by model methods.
   1. Variables Used in RFE regression model:

|  |  |  |  |
| --- | --- | --- | --- |
| AGE | GENDER\_M | AFFL\_15.0 | REGION\_Midlands |
| GENDER\_F | AFFL\_14.0 | CLASS\_Tin | AFFL\_9.0 |
| AFFL\_6.0 | AFFL\_4.0 | AFFL\_8.0 | CLASS\_Silver |
| AFFL\_5.0 | LTIME | REGION\_South East | REGION\_South West |
| AFFL\_7.0 | NGROUP\_F | NGROUP\_B | REGION\_Scottish |

* 1. Based on the Feature importance metric, the following five variables were considered most valuable:

|  |  |
| --- | --- |
| *Variable* | *Importance* |
| *AGE* | 0.545747237844 |
| *GENDER (GENDER\_F)* | 0.237567223101 |
| *AFFL (*FFL\_6.0) | 0.0356209181906 |
| *AFFL (*FFL\_5.0) | 0.376235146978 |
| *AFFL (*FFL\_7.0) | 0.034188828339 |

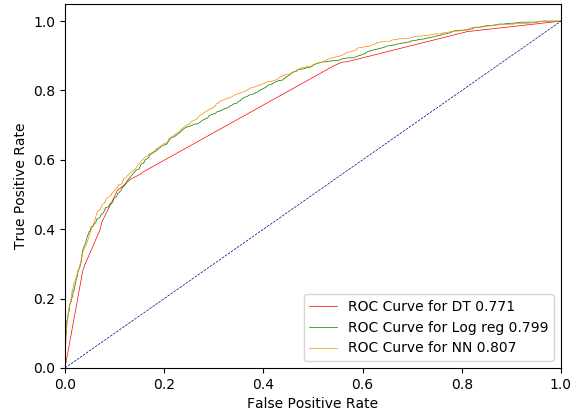
* 1. There is no signs of over fitting in this model due to the closeness of test and train accuracies
  2. **Train accuracy:** 0.812483929031  
     **Test accuracy:** 0.812959352032

1. Using the comparison statistics, which of the regression models appears to be better? Is there any difference between two models (i.e. one with selected variables and another with all variables)? Explain why those changes may have happened.
   1. Both models display similar accuracy results, however when comparing the difference in the accuracies in both models it is clear that the RFE tuned model is the best. This model was also tuned with GridSearchCV which measures a variables importance. This helps the model select which features are best suited to be used within the analysis. The RFE model has its advantage as it removes features that are not useful for the analysis. In this case the model chose to use 20 features, rather than the full 53 in the standard model.
2. From the better model, can you identify which customers to target? Can you provide some descriptive summary of those customers?
   1. In regression, interpreting the model requires looking at the coefficients of the features presented. AGE being the most important has a negative coefficient, from this we can say that a buyer who is younger is more likely to buy organic products. Looking at GENDER\_F, which is the second most important feature we can see that the positive value indicates that a woman is more likely then a man to purchase an organic product.   
      When looking at AFFL features it can be seen that any affluence level higher than 20 has a positive coefficient. These is an indication that customers with a high affluence level is more likely to buy an organic product than one with a low to mid-level. Only customers from the Midlands region show a positive coefficient for the region feature, with tin level loyalty members also displaying a positive coefficient.  
      Therefore, based on the RFE model it can be concluded that a young female with a tin loyalty card, a high affluence level and lives in the Midlands region is the most likely to buy the organic products.

# Task 4: Predictive Modelling Using Neural Networks

1. Build a Neural Network model using the default setting.
   1. Multi-Layer Perception Classifier (MLPClassifier)
   2. The default value of 200 iterations was required to train this model
   3. This model did show signs of overfitting, this was seen through the train and test set accuracies. The significantly larger train set accuracy shows that this model was overfitting
   4. The model did achieve convergence
   5. **Train accuracy:** 0.86603660366  
      **Test accuracy:** 0.788455357717
2. Refine this network by tuning it with GridSearchCV. What are the parameters used? Explain your decision. Report the trained model.
   1. Multi-Layer Perception Classifier (MLPClassifier)
   2. 1000 iterations were required to train this model to ensure convergence was achieved.
   3. This model did not show signs of overfitting as the previous one did. The train and test set accuracies were very similar in this instance indicating no over fitting occurred
   4. The model did achieve convergence
   5. **Train accuracy**: 0.809912573926  
      **Test accuracy:** 0.813109344533
3. Build another Neural Network model with inputs selected from RFE with regression (use the best model generated in Task 3) and selection with decision tree
   1. Feature Selection decreases both training and testing accuracy slightly, and the values are further apart, showing a slightly worse model. This model has reduced the features down from 52 to nearly half at 24, with only 9 important features chosen. These are:
      1. AGE
      2. GENDER\_F
      3. AFFL\_14.0
      4. AFFL\_13.0
      5. GENDER\_U
      6. AFFL\_7.0
      7. AFFL\_6.0
      8. GENDER\_M
      9. AFFL\_12.0
   2. **Train accuracy:** 0.787670352276.   
      **Test accuracy:** 0.797660116994
   3. Once again 1000 iterations were used to ensure convergence was achieved.
   4. No Overfitting was seen is this model due to the similarity in train and test accuracy.
   5. The model converges and allows for the best possible model
   6. Both train and test accuracy have improved, as well as showing less difference between them. This shows the model is improved through the use of GridSearch CV. These values are seen below: **Train accuracy:** 0.811326819234  
      **Test accuracy:** 0.812809359532
4. Using the comparison methods, which of the models (i.e. one with selected variables and another with all variables) appears to be better? From the better model, can you identify which customers to target? Can you provide some descriptive summary of those customers?
   1. Through the use of comparison methods, its clear that the RFE with GridSearchCV is the best Neural Network model as it shows a better fit than the others. Although the model is very accurate, interpreting neural network models are quite difficult. Because of this it is hard to use this model to provide a description of the type of customer that would buy organic products

# Task 5: Comparing Predictive Models

1. Use the comparison methods to compare the best decision tree model, the best regression model, and the best neural network model.
   1. Findings:
      1. **ROC:** The RFE based neural network produces the best ROC score, with a value of 0. 0.808288934649, with RFE regression showing the second best with 0.798668160168 and DT in third with a score of 0.771157411543. This means on a carried discrimination threshold, the RFE neural network performs best. This is however marginal and can be seen more clearly in the below graph:  
         
      2. **Accuracy Score:** Based on accuracy score, the RFE regression shows a slight advantage over the neural network with a score of 0.812959352032 vs. 0.812509374531, with the decision tree a little bit behind with a score of 0.797810109495. Although the accuracy scores suggest that the regression is marginally better, it important to note that ratio of the target classes isn't relatively equal. This means that an accuracy score isn’t the best method for comparing models in this instance.
      3. **Classification Report:**
         1. Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1score | support |
| 0 | 0.82 | 0.93 | 0.87 | 5015 |
| 1 | 0.65 | 0.4 | 0.49 | 1652 |
| Avg/total | 0.78 | 0.8 | 0.78 | 6667 |

The above classification report shows the decision tree is quite effective at predicting positive observations when determining a customer has not purchased organics, and also has a good rate in which positive predictions are correct. However, when it comes to predicting and success at predicting buyers of organic the model is not as effective

* + - 1. Regression Model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1score | support |
| 0 | 0.83 | 0.95 | 0.88 | 5015 |
| 1 | 0.72 | 0.4 | 0.52 | 1652 |
| Avg/total | 0.8 | 0.81 | 0.79 | 6667 |

Shows the same positives as the DT model however slightly improved, however, there is a significant improvement in making correct predictions for Buyers

* + - 1. Neural Network:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Precision | Recall | F1score | support |
| 0 | 0.83 | 0.95 | 0.88 | 5015 |
| 1 | 0.72 | 0.4 | 51 | 1652 |
| Avg/total | 0.8 | 0.81 | 0.79 | 6667 |

Displays same traits as regression

* 1. Both the decision tree and regression models agree on a few traits. Those being that younger woman are more likely to buy organic products. However, they differ on values of affluence with the decision tree indicating lower affluence levels than the regression model. Regression provides more detail in that it also shows loyalty level and location characteristics.  
     Due to the neural network being difficult to interpret it is hard to say with the features it shows is similar or not.

1. Finally, based on all models and analysis, is there a particular model you will use in decision making? Justify your choice. How the outcome of this study can be used by decision makers?
   1. Based on ROC and Classification tables, it appears regression and neural network models are the best at making predictions. However, when considering factors such as interpretability DTs hold a significant advantage, with regression also much better than NN. When considering speed, a NN lags behind with its slow speed with regression and DT holding an advantage here. Adaptability gives NN Advantage with the ability to use online training if the model needs to slowly adapt to new data trends. DTs is the worst for this. Therefore, based on ROC and classification reports and other factors RFE regression would be the best model.  
        
      The outcome can be used by decision markers through more targeted advertising. With a known demographic, they can market with more specificity. Along with this information, if other purchasing habits of the same demographic are studied, possible deals could be created. This could also be used to determine where in the store the products are displayed.
2. Can you summarise positives and negatives of each predictive modeling method based on this analysis?

|  |  |  |
| --- | --- | --- |
|  | POSITIVE | NEGATIVE |
| DT | * Reasonable training time * Fast application * Easy to interpret * Easy to implement | * Not suitable for large dataset * Difficult to visualize * Too simple |
| REGRESSION | * Simple * Fast application * Relatively easy to interpret | * Cannot handle large number of features * Some features can’t be interpreted in a linear model * Very sensitive to skewed data |
| NN | * Adaptable * Can handle lots of features | * Very Slow * Using lots of computing power * Hard to interpret * Hard to implement |

* 1. The positives and negatives of the models are summed up in the below table:

# Appendix A

