**Data Mining**

**IMAT3613**

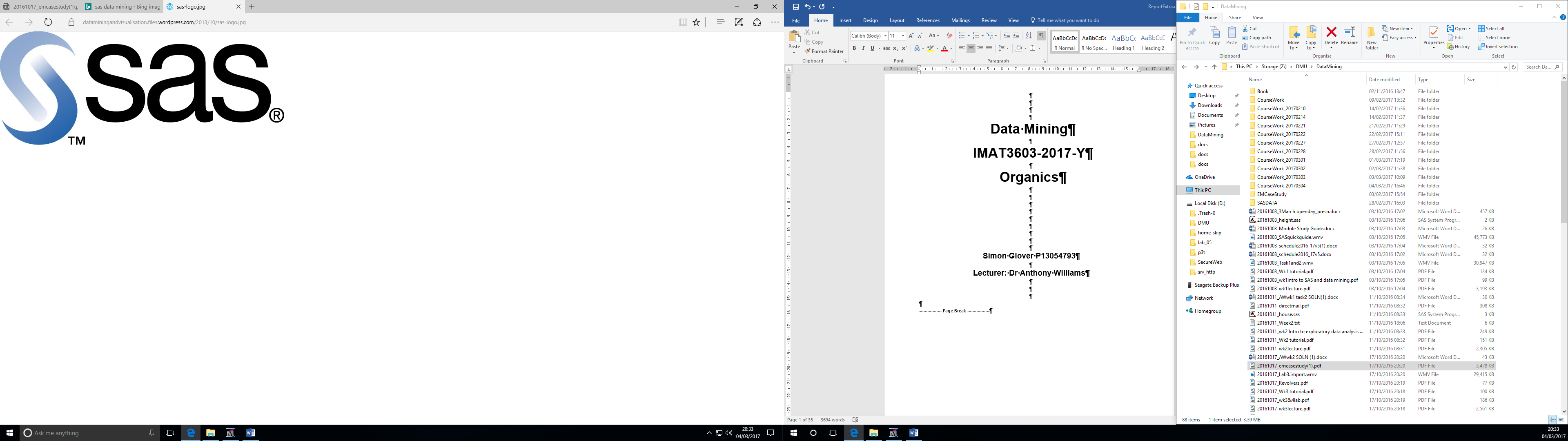
**Author**

**PNumber**

**Date**

**Organics Assignment**

(Word count)

 ****

**Lecturer: Dr Anthony Williams**

**Table of contents**

Contents

[Summary 3](#_Toc524305355)

[Business Problem 3](#_Toc524305356)

[Data Mining Representation 3](#_Toc524305357)

[Methodology - data mining approach 3](#_Toc524305358)

[Data Exploration 4](#_Toc524305359)

[Data partition creation of model sets 5](#_Toc524305360)

[Data Modelling 6](#_Toc524305361)

[Regression 6](#_Toc524305362)

[Neural Network 7](#_Toc524305363)

[Development of models 7](#_Toc524305364)

[Model performance 8](#_Toc524305365)

[Neural network architecture of best model 8](#_Toc524305366)

[Decision Tree 10](#_Toc524305367)

[Development of models 10](#_Toc524305368)

[Performance of models 10](#_Toc524305369)

[Overfitting and limitations 11](#_Toc524305370)

[Analysis of the best model 11](#_Toc524305371)

[Conclusion 12](#_Toc524305372)

[Recommendation 13](#_Toc524305373)

[References and Bibliography 13](#_Toc524305374)

[Other Resources 13](#_Toc524305375)

[Appendix 14](#_Toc524305376)

# Summary

This report presents the results of using the \_\_\_\_\_data mining framework, to build models that will help the management of a supermarket concentrate their resources on targeting customers that are most likely to purchase organic products. Three binary \_\_\_\_\_\_\_\_\_\_\_\_\_\_models were generated using \_\_\_\_\_\_\_\_ analysis, decision \_\_\_\_\_\_ and \_\_\_\_\_\_\_ \_\_\_\_\_\_ . \_\_\_\_\_and \_\_\_\_\_ were identified as the most important predictors. All three models were predictive and have a \_\_\_\_\_\_\_\_level of performance. \_\_\_\_\_ was chosen as the champion models based on performance and also the techniques ability to provide an non-technical explanation of the model with a lift value of \_\_\_\_\_times greater than selecting customers at random. Several recommendations are made for improving data quality for the next cycle of data mining.

# Business Problem

The business problem is to identify the customers that are most likely to \_\_\_\_\_ organic products in the supermarket. A data models will be built data set (organics.xls) collected during the supermarket incentive period. By identifying \_\_\_\_\_\_ who are \_\_\_\_\_\_ to purchase organic products the company will be able to target its marketing efforts more effectively which should result in more sales per marketing advertising spend.

# Data Mining Representation

The business problem, identification of customers who are \_\_\_\_\_\_ to \_\_\_\_\_\_ organic products is a type of data mining representation known as a \_\_\_\_\_ \_\_\_\_\_\_\_\_\_ problem. The most suitable target variable is \_\_\_\_\_ which is identified as a **binary variable**. The remaining variables will be given roles \_\_\_\_\_ and are assigned the default measurement levels with the exception of AFFL which has been changed from interval to ordinal.

# Methodology - data mining approach

The process to be adopted is the first flow in the virtuous cycle of data mining which has four distinctive steps:

1. Identify the business problem or opportunity.
2. Mining data to transform it into actionable information.
3. Acting on the information this is outside the remit of the brief (marketing initiative driven by the model, eg targeted marketing offer)
4. Measuring the results of the marketing initiative this is outside the remit of the brief (measure profitability of the pilot marketing study using the model).

The data mining framework used in the generation of models for steps 1 and 2 of the virtuous cycle of data mining is based upon \_\_\_\_\_, Sample, Explore, Modify, Model and Assess. In the brief the data set has already been provide, so there is no need to sample or collect the data. However limitations in data collection may become apparent in the data mining process, these are discussed in the recommendations section.

Explore:

Explore the raw data as provided. This will result in a brief overview of the variables so that they can be classified as qualitative data (binary, nominal, ordinal), or quantitative data (discrete, interval) with a brief description.

Establish a Target variable (which variable can be used to establish the required results).

Modify:

Assess the data quality (what changes can be made to the data classifications/levels, model roles). If necessary modify and transform the data, impute missing values, transformations to normalise distributions of heavily skewed data.

Models:

Create models of data (regression, neural networks, decision trees) to identify patterns relationships and parameters that can predict the target variable. Each model representation has their own advantages and disadvantages.

Assess:

Asses the model performance in order to identify the champion model and investigate their limitations. If necessary make changes to the data set, model tuning through a second cycle of data mining to improve on the previous model results. The aim of the second cycle is to upon improve model reliability, robustness, performance and avoid over fitting.

In summarising the results from the models generated arrive at a conclusion, and make recommendations for future data gathering by the business. This should ensure that the future data mining results will be of benefit to the business and refine the data mining process.

# Data Exploration

A full meta data description of the data is provided in the appendix (table 1 and 2) only important features and insights that have a bearing on data mining process will be described. These insights are generated from an exploratory data analysis of the data. The analysis has been divided into two sections class variables which are qualitative and interval variables which are quantitative.

From an analysis of the business problem the target variable, the variable to be predicted has already been identified as \_\_\_\_\_. It is noted that the ratio of evidence for organic customer purchasers to non-purchasers is \_\_\_\_\_, making this a difficult data mining problem.

An indication of data quality by identifying the level of missing data is also presented, any variable with more than 50% of the data missing is normally considered unsuitable for data mining. Variables that have missing data have a bearing on regression analysis and neural network models. \_\_\_\_\_, \_\_\_\_\_ and \_\_\_\_\_ have between 5 and 11% missing data, the data quality for remaining variables is less than 5% missing. Interval variables which are heavily skewed > 3 may be transformed to normal distributions to comply with model assumptions.

Variables with outliers, extreme values are also identified. These anomalous observations may be outside the scope of models and may give indications for the predictive parameters. However the nature of outliers is that they represent only a handful of observations and insights may not be applicable to the majority of observations. The impact of outliers on modelling was found to be negligible. Model performances were practically unaffected by the removal of outliers using the filter node. The variable \_\_\_\_\_ was heavily skewed and has the majority of extreme values.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Table 1 Summarising observations, measurement levels and data mining roles.

# Data partition creation of model sets

# Data Modelling

# 

# Regression

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Run id | Missing values removed | Missing values imputed | Outliers removed | Transformation to remove skew | Scope | % True positives | % precision | AUC ROC | Estimate lift @ 25% depth |
| 1 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run Id** | **Classifier** | **%True -** | **%False -** | **%True +** | **%False +** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

Chosen Regression equation

Logit P =

# Neural Network

## Development of models

## Model performance

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Run id | Missing values removed | Missing values imputed | Outliers removed | Transformation to remove skew | Scope | % True positives | % precision | AUC ROC | Estimate lift @ 25% depth |
| 1 |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run Id** | **Classifier** | **%True -** | **%False -** | **%True +** | **%False +** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**Table 5 Neural network performance**

## Neural network architecture of best model

.

Figure 1 example neural network architecture

# Decision Tree

## Development of models

## Performance of models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Run id | Comments | Scope | % True positives | % precision | AUC ROC | Estimate lift @ 25% depth |
| 1 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run Id** | **Classifier** | **%True -** | **%False -** | **%True +** | **%False +** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

**Critical path**

**Target path of interest**

## Overfitting and limitations

# Analysis of the best model

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Run id | Missing values removed | Missing values imputed | Outliers removed | Transformation to remove skew | Scope | % True positives | % precision | AUC ROC | Estimate lift @ 25% depth |
| Logistic |  |  |  |  |  |  |  |  |  |
| ANN |  |  |  |  |  |  |  |  |  |
| DT |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Run Id** | **Classifier** | **%True -** | **%False -** | **%True +** | **%False +** |
|  | Logistic |  |  |  |  |
|  | ANN |  |  |  |  |
|  | DT |  |  |  |  |
|  |  |  |  |  |  |

**Table 8 summary results of the best performing models**

Figure 2 Cumulative lift chart of the best performing models

Figure 3 Non-cumulative lift chart of the best performing models

Table 9 summary of confusion matrix for test set

# Conclusion

# Recommendation

# References and Bibliography

# Other Resources

# Appendix

Data Mining Roles

Workflow diagram

# My Reflections on the Patchwork assignment

You may wish to support your answer with transcripts or quotes from the weekly lab journal.

You should concentrate on how your work has evolved over the term and describe your experiences of using the SAS enterprise miner software.