Data Science Portfolio

MSc Data Science
&
MSc Management Science and Marketing Analytics
Projects

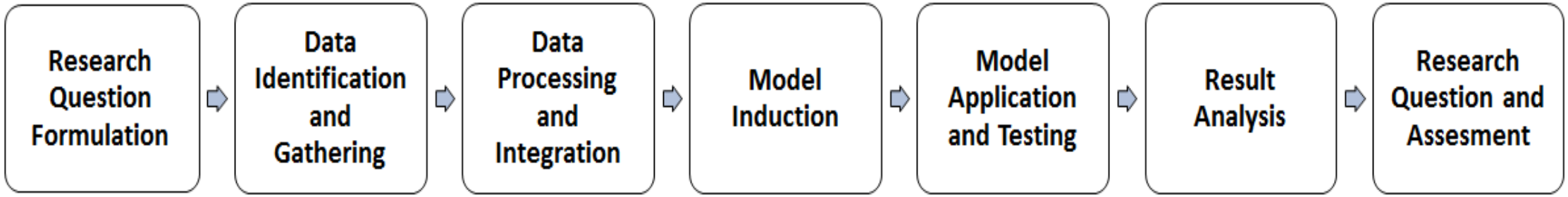
Emel Pisiren
Lancaster University
20013 -2015

Projects:

- **Project 1:** The Analytic Use of Paradata for Understanding Online User Behaviour **(R programming language)**
- **Project 2:** Modeling Financial Volatility with Change Detection Algorithms **(R programming language)**
- **Project 3:** Analysing Pet Owner's Buying Habits and Modeling Propensity for Healthy Food Products **(R programming language)**
- **Project 4:** Cluster Analysis and Classification of Mobile Phone Calls Data **(MATLAB)**
- **Project 5:** Building a Credit Scoring Model for Predicting the Probability of Default on Mortgage Loan **(SAS Enterprise Miner)**

Data Science Pipeline:

Data Science Pipeline, a scientific end-to-end process of data analysis, was implemented as a methodological framework in each project.



Project 1:

Project Name: *The Analytic Use of Paradata for Understanding Online User Behaviour*

Project Sponsor: NatCen Social Research (London)

Description: A standalone data project including data structuring, visualization, usage of natural language processing techniques and statistical methods for analysing web survey users

Software/Programming Language: R (*ggplot, plyr, stringr, reshape packages etc*.)

Project length: 3 months (summer project)

Business Objective:

This Project aimed to examine how to increase survey response rates of NatCen via web mode by understanding respondents' online behaviour through an analysis of survey Paradata.

Paradata: Additional data captured as a by-product of survey data

For example; *browser type and the operating system used by the respondent, mouse clicks, keystrokes, time spent per page*

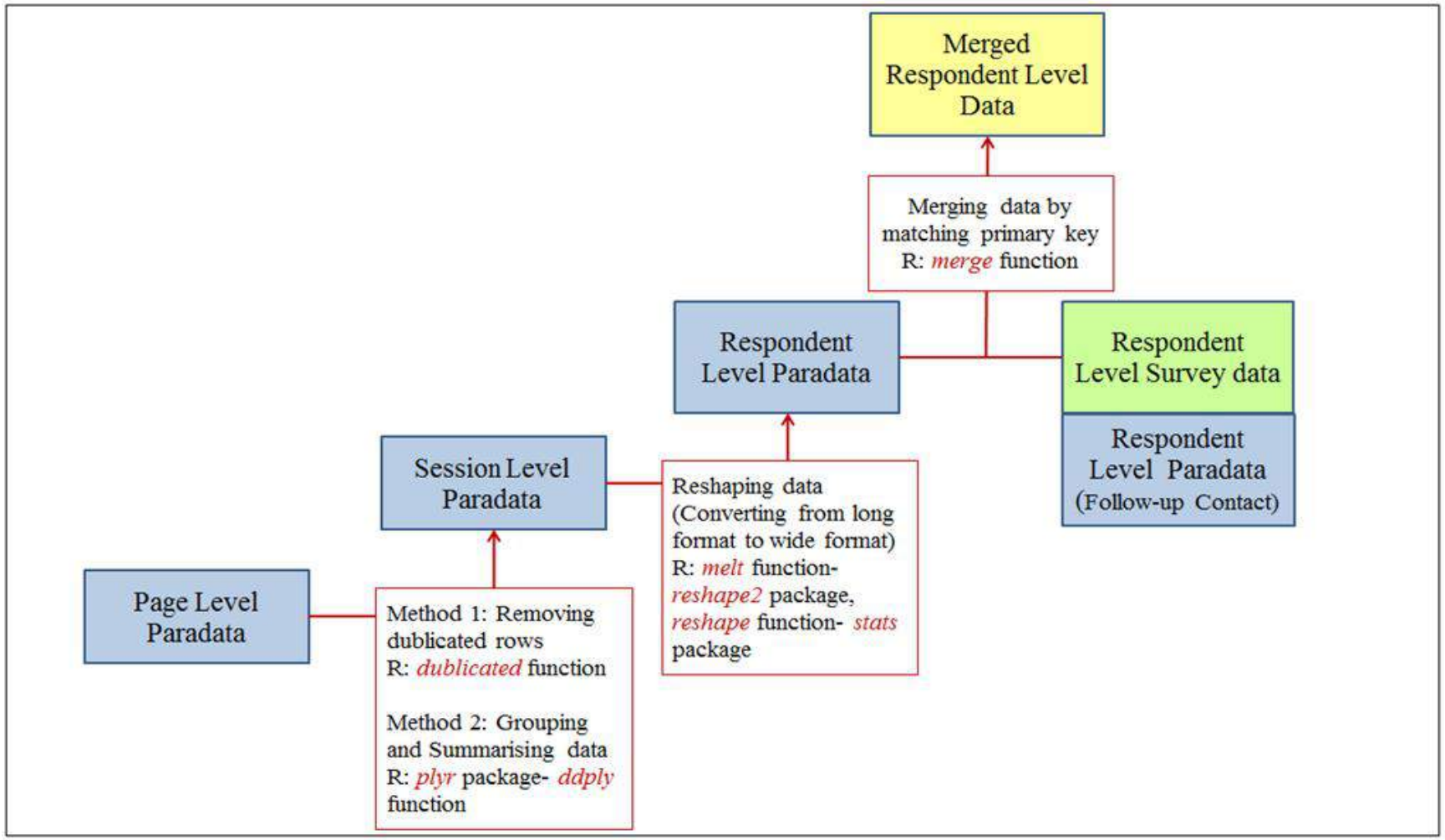
Proposed Solution:

Break-off occurs when the respondent logs into the survey, but abandons that session before finishing all questions.

Analysing user behaviour and understanding why they left the web survey would help to increase survey response rates.

Applying a detailed exploratory analysis followed by **logistic regression** for finding the relationship between *break-off* and paradata variables *(browser type, total time spent on the survey, number of questions answered etc.)* was proposed as a solution.

Data Management Methodology:



The granularity of the data files were different, so they were aggregated at the same level and then merged.

Details of the Levels

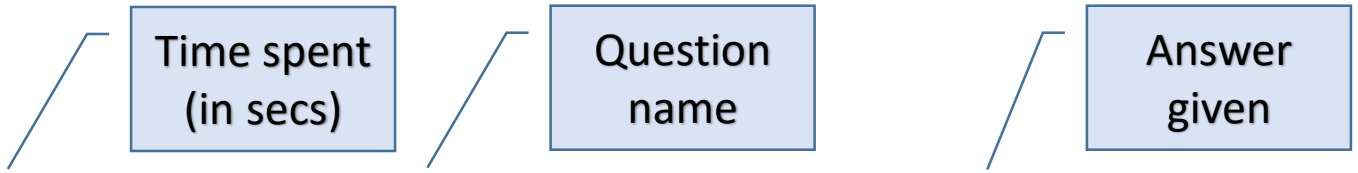
Respondent level: one case for each survey respondent

Session level: one case for each 'session' of accessing the questionnaire by a respondent

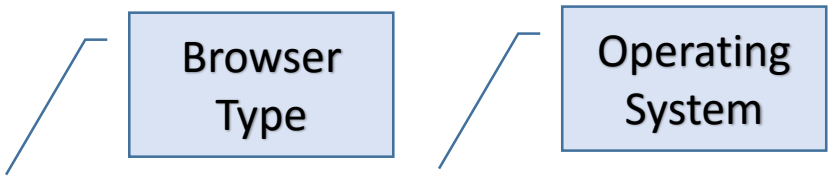
Page level: one case for each 'page' of the questionnaire accessed by a respondent

Handling Unstructured Paradata:

Text Mining Techniques was used to extract useful parts of the strings.

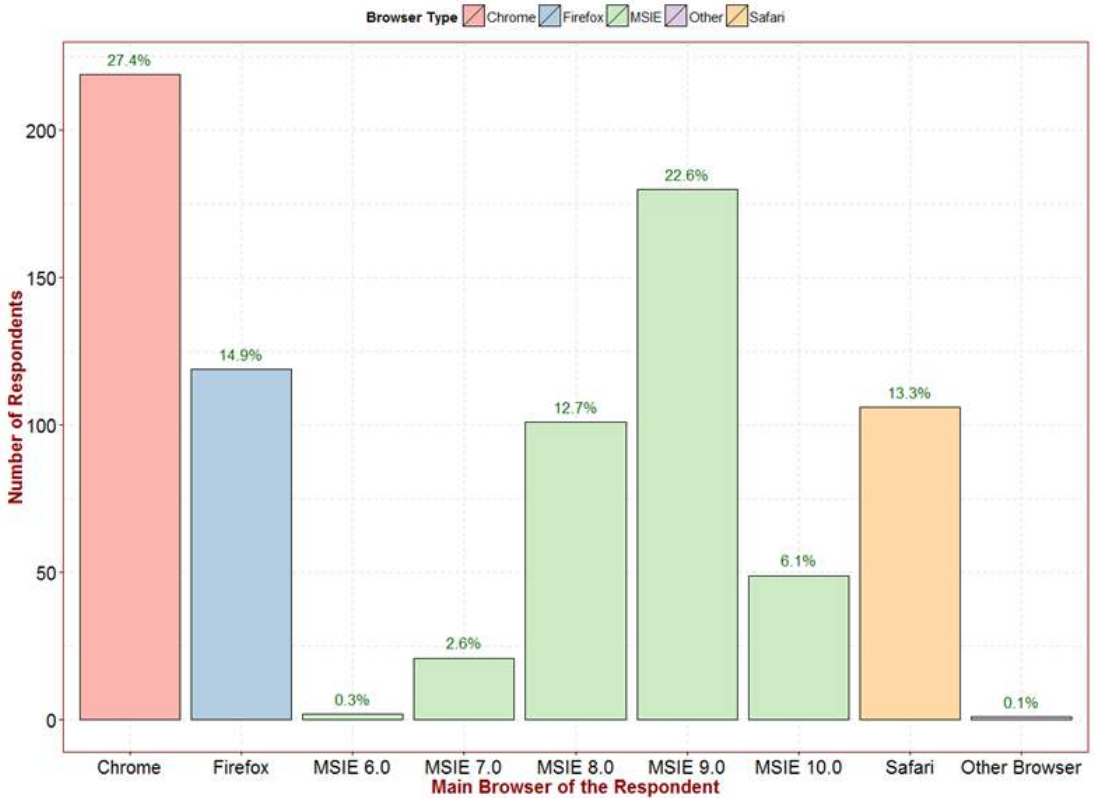


4/24/2013 8:49:04 PM:pageLoad^t=**7018**:QInd[1].**Q560Savings.save**=**2**



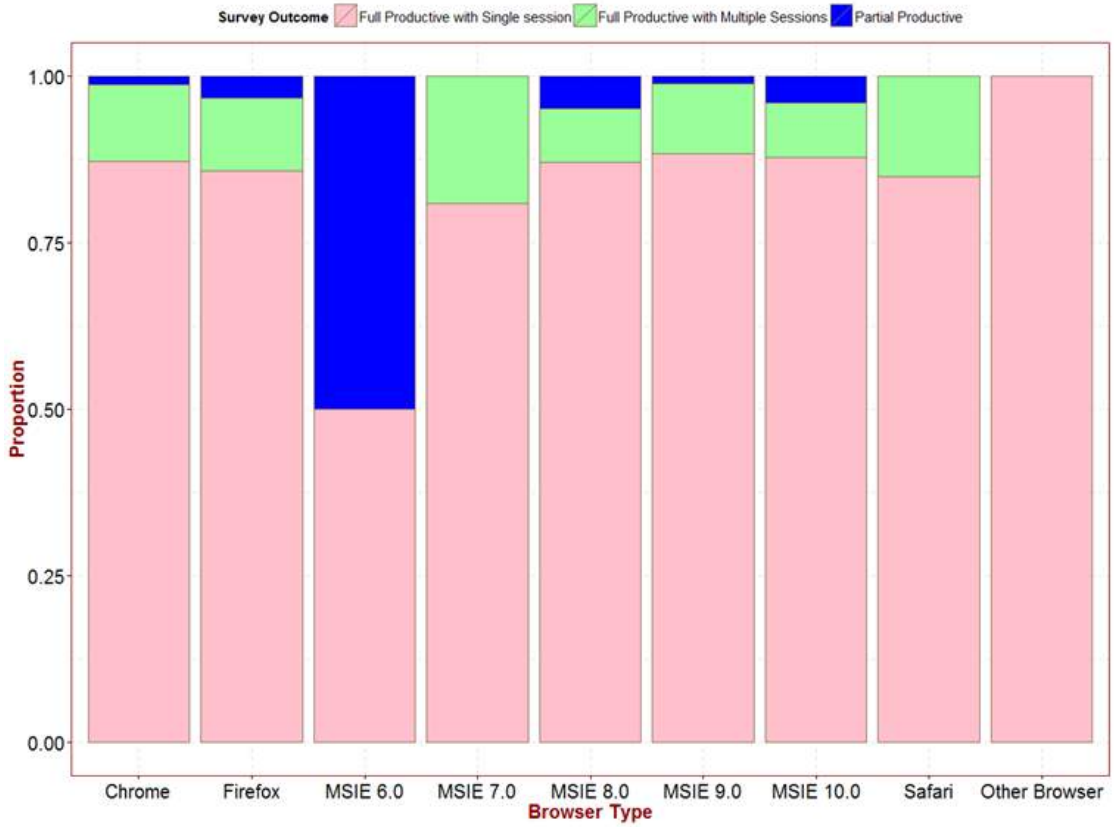
Mozilla/5.0 (compatible; **MSIE 9.0**; **Windows** NT 6.1; WOW64; Trident/5.0)

Analysing Browser Types:

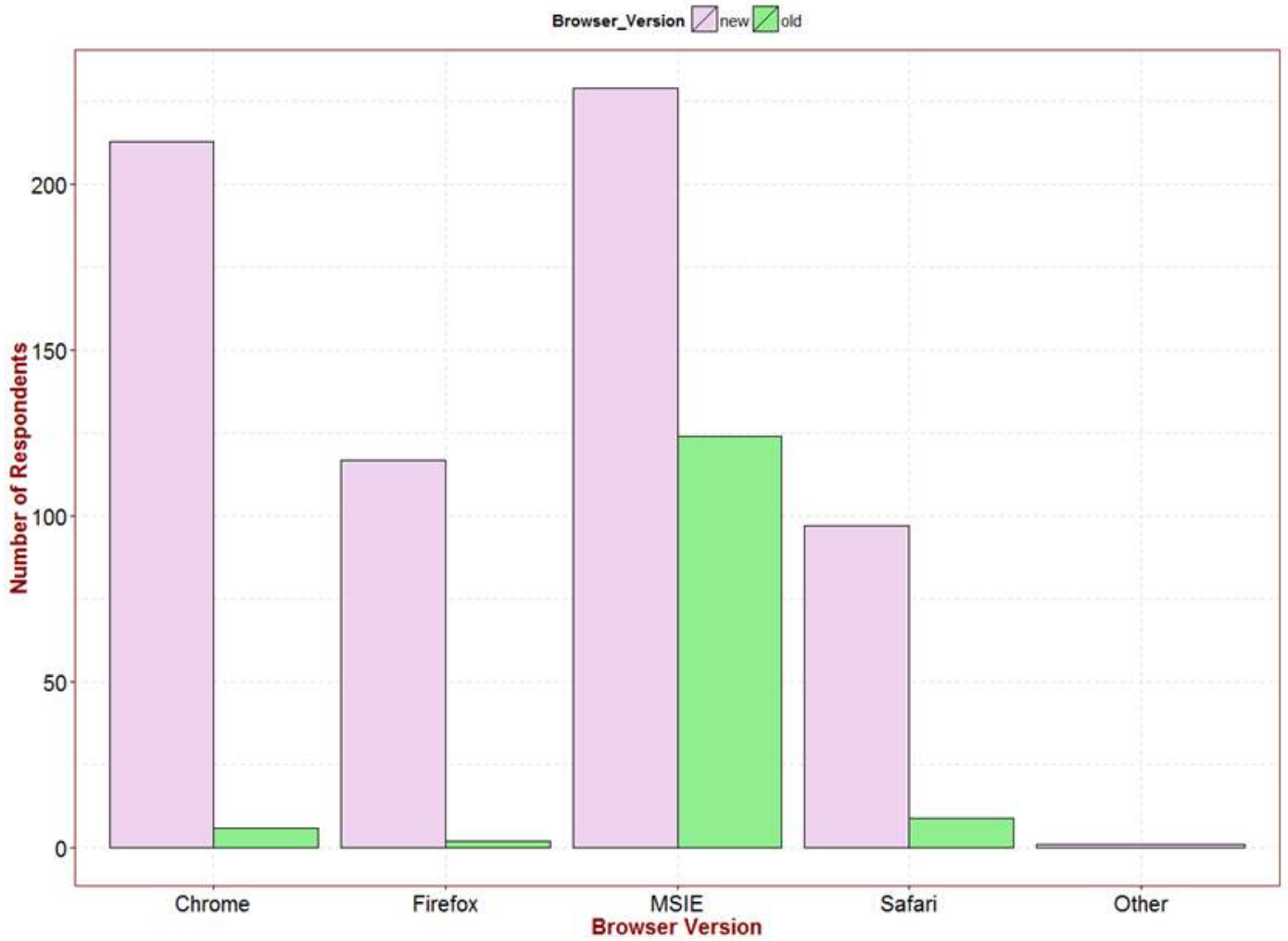


44.3% of the respondents were Internet Explorer users shown in green bars with different versions.

As the frequencies of the browser types were different, they were all stretched up to the same scale for a better comparison in terms of Survey Outcome.



Browser Types with Old/New Classification:



Browser types appeared to be not associated with break-off, so a new variable 'Browser Version' was derived. It was found that the users with old browser versions were more likely to abondon the survey.

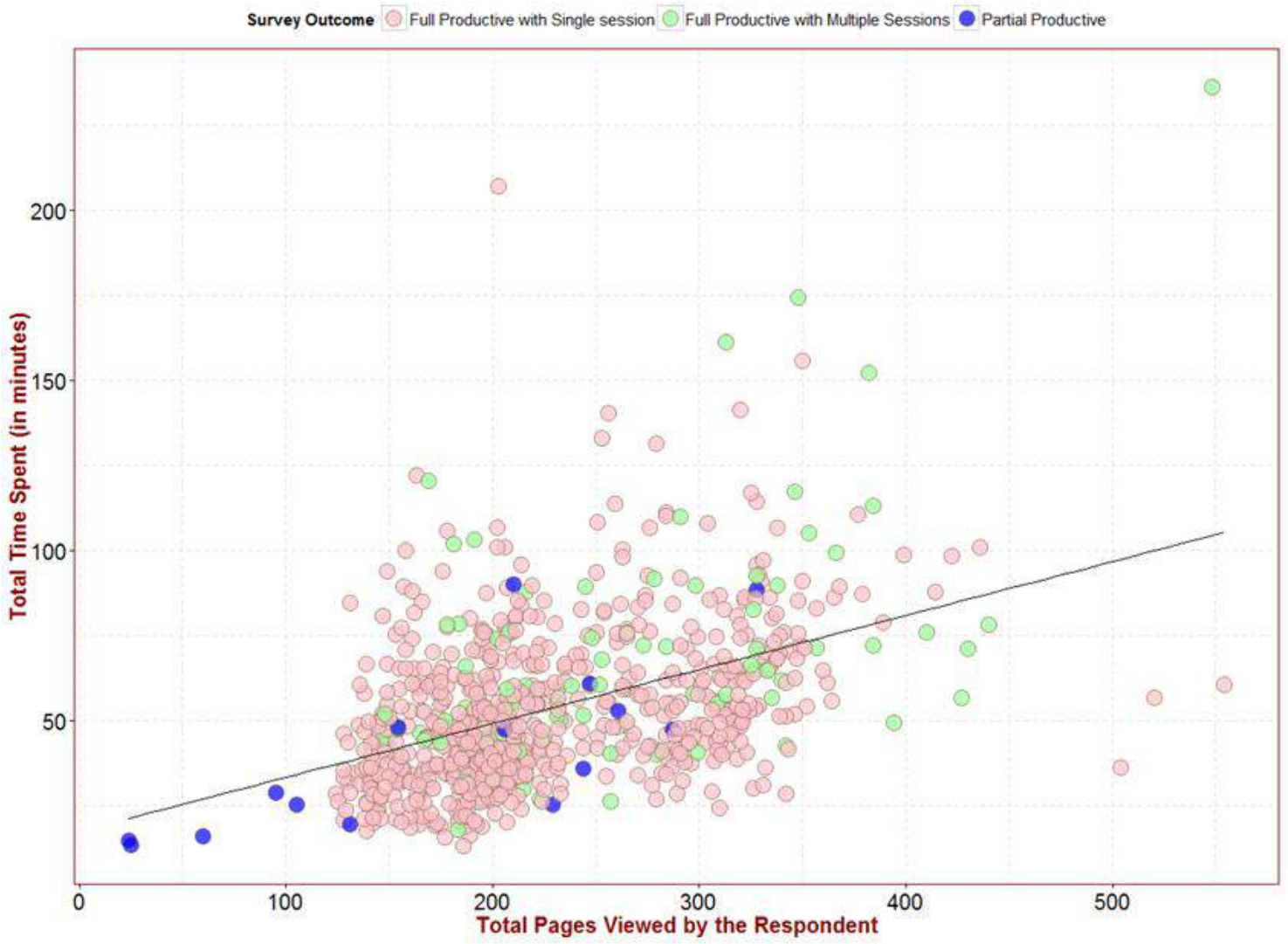
Survey Outcome	**New Browser**	**Old Browser**	**Proportion**
Full Productive with Multiple Sessions	75	14	0.157
Full Productive with Single session	572	120	0.173
Partial Productive	10	7	0.411

Total Time Spent vs. Number of Questions

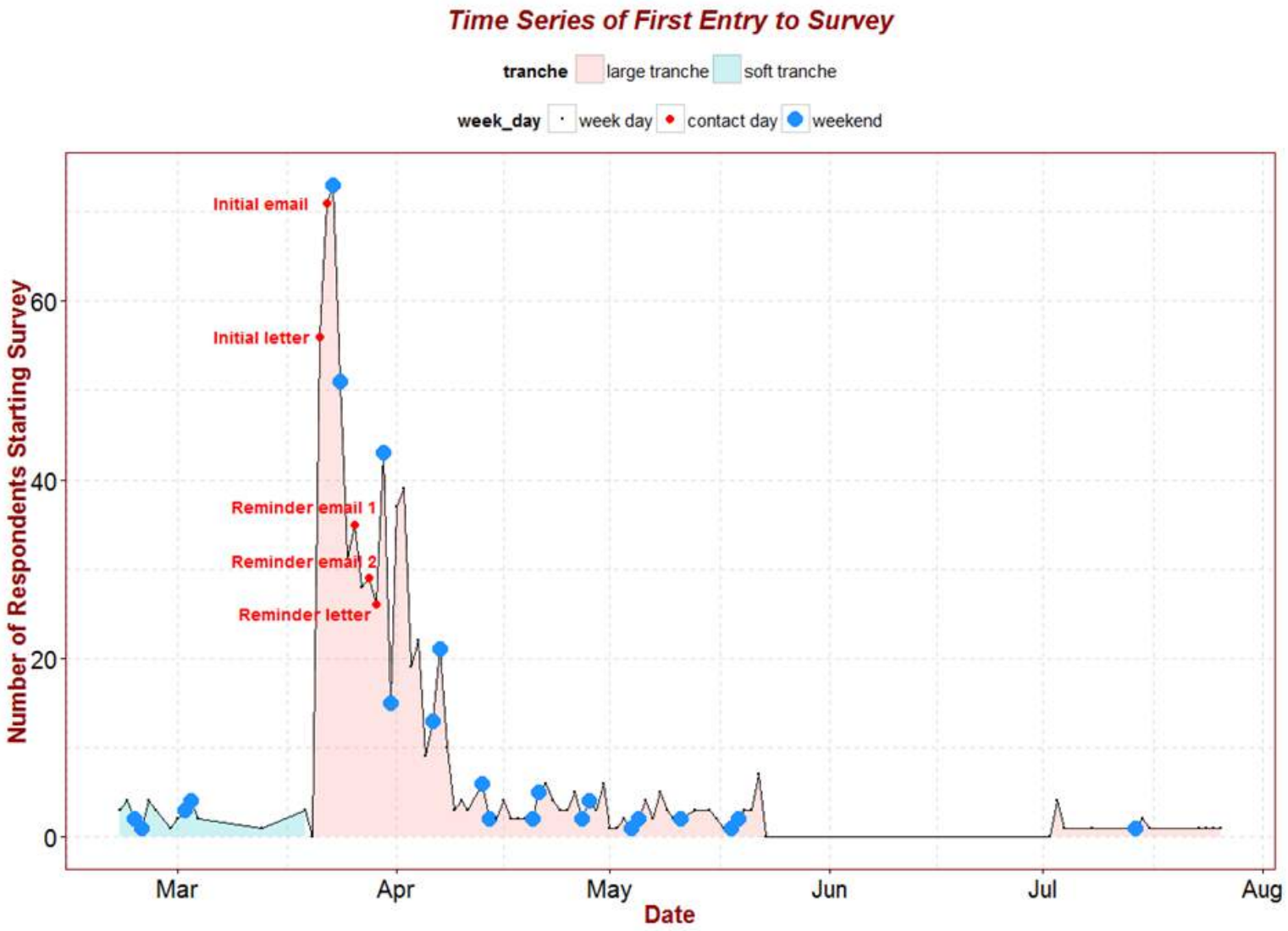
Here, each respondent are represented by a point and the points are coloured according to the survey outcome of the respondent.

There are large variations in terms of ***total number of questions*** (total pages viewed) and ***total time spent***. For instance, there are many cases spending 50 minutes and completing around 150 pages, whereas there are others spending same amount of time on more than 300 pages.

Total number of questions for completing a survey varied extensively for respondents due to routing rules which lead to different questions in the survey. Some respondents completed the survey after viewing less than 200 pages, while some of them viewed around 400 pages.



Analysing First Entry to Survey:



The time series shows the number of first entry to survey over time. The red points denote **contact days** when respondents were sent invitations or reminders. **Weekends** are shown by big blue points.

People are more likely to start the survey at the weekends, in particular Saturdays, as the peaks in the line are generally associated with weekend days.

So, Friday appears to be the best day for sending letter /e-mail invitations or reminders.

Logistic Regression Analysis:

The probability of break-off from the survey was modeled on a large set of explanatory variables containing device-type paradata, demographic information, variables derived from *average page-level timings* and *contact records.*

A binary variable for *breakoff* with two categories (*Breakoff=1, No breakoff=0)* was computed for the analysis.

Browser version, marital status, gender, education level of the respondent, *incentive offered*, *average time spent on a question* and starting the survey on *contact day* were found to be statistically significant.

Explanatory Variables	**Coefficients**
Intercept	-4.219
Browser_Version_Old	1.385
Marital_Status_Not_Single	1.172
Gender2	0.475
Education_Level_GCSE	0.379
Education_Level_A-level	0.760
Education_Level_Degree	1.111
First_session_performance	0.038
Incentive_£10 unconditional +£20	-0.323
Incentive_£30 unconditional	0.253
Regular_WebUser_Yes	-0.501
Contact_day	0.429
Browser_VersionOld:Marital_StatusNot_Single	-1.447

Suggestions to the Business:

- Regarding the last question of the survey, a critical design problem leading to confusion for users and inaccuracy in terms of survey results was detected. A new design for this page was suggested which leads to higher and accurate response rates.

- After finding out that old browser versions were causing break-offs for some users, advising users to update their browsers before starting the survey appeared to be a good practise. For instance, a notice indicating this with further guidance could be included to the invitation emails.

- The users are least likely to do a survey on Wednesdays, so instead of sending email reminders on Tuesdays, users should be contacted on Fridays just before the weekends when people are more likely to start a survey.

- NatCen was advised to apply this adaptable methodology for each test survey to get preliminary insight into the users behaviour and to detect probable issues before launching the main survey.

Project 2:

Project Name: *Modeling Financial Volatility with Change Detection Algorithms*

Project Sponsor: Lancaster University Management School

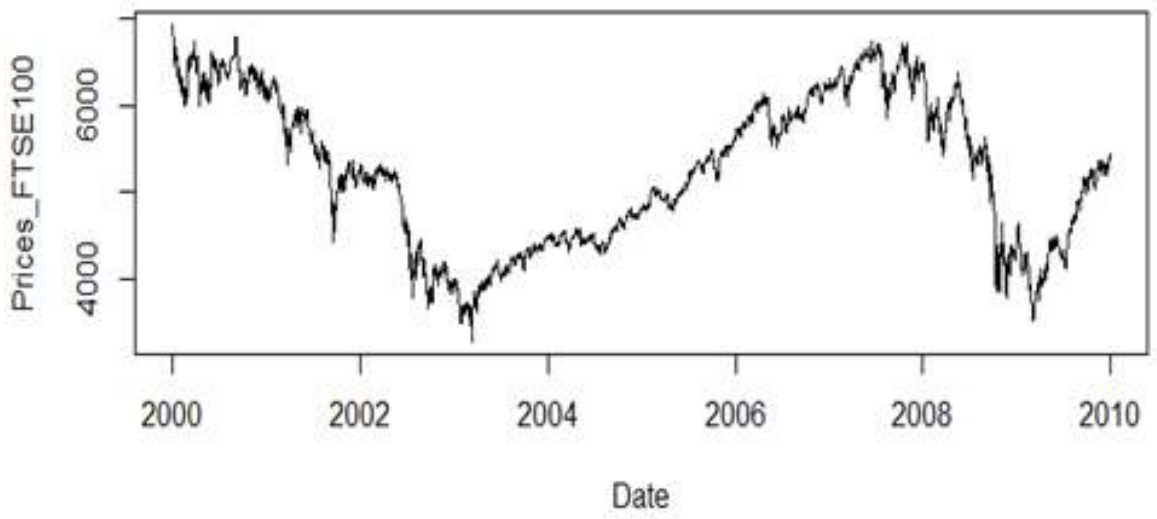
Description: A research project for detecting structural break points in financial series, volatility clustering using nonparametric statistics and applying GARCH models for capturing the behaviour of financial volatility

Software/Programming Language: R (*base R for statistics, simulation and graphics, changepoint, rugarch, cpm packages for model building)*

Project length: 3 months (summer project)

Business Objective:

Volatility refers to the variation of the price of the financial instrument over a period of time. For instance, the daily closing prices of FTSE100 index from London Stock Market show sharp up and down movements over the period from 2000 to 2010.



Volatility is core to capturing the behaviour of financial data. This project examines the change point detection models used **for identifying structural changes in the volatility** of financial time series. The objective is **to find out the contribution of change point analysis** incorporated to the financial volatility modeling.

Proposed Solution:

- In order to identify structural shifts in variance of the financial time series, first the change point algorithms were applied to simulated series to examine their performance in known settings, then they were employed to real data.

- After finding the change points which split the series into segments, dummy variables were introduced for these segments and a GARCH model was fitted on the segmented series.

- Another GARCH model was fitted to the original series as a benchmark. The results of these two models were compared to investigate the contribution of the change point analysis incorporated to the GARCH model fitting.

Change Point Analysis on Simulated Data:

To perform the change point analysis in a known setting, first of all, multiple sets of simulated data with variance shifts at particular points were generated. In this respect, various sequences which differ in their

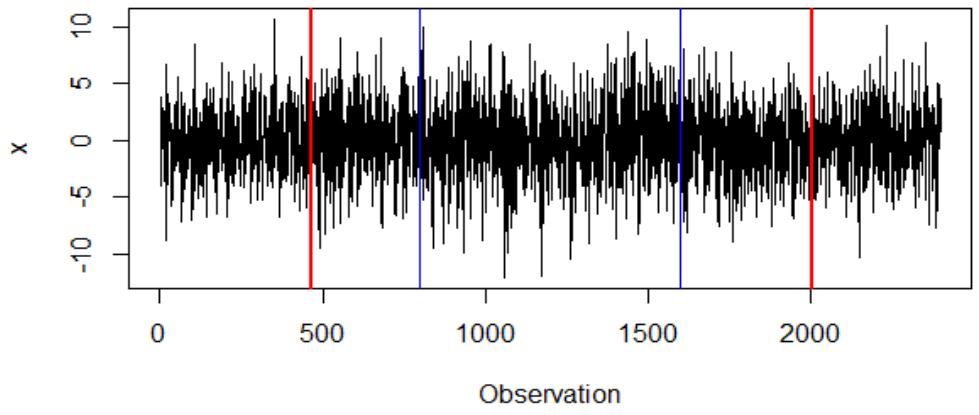
- sample size
- distributional form
- size of change
- variance level

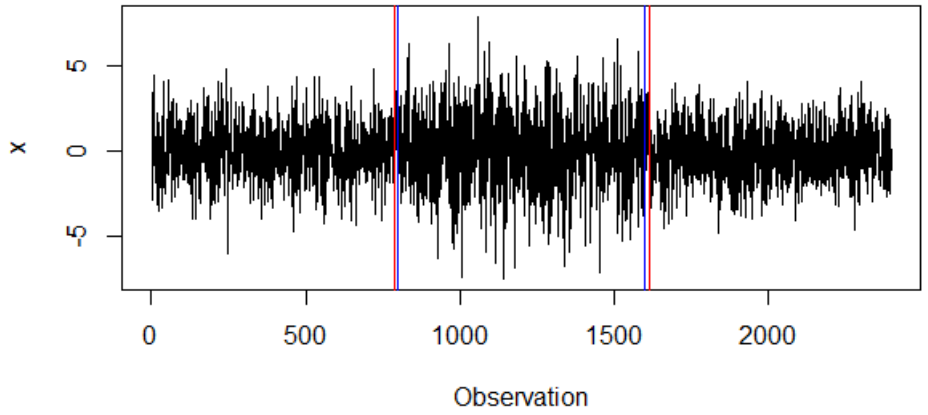
were simulated.

The simulation study compares the performance of **parametric** (*Bartlett test, ICSS algorithm*) and **non-parametric tests** (*Mood test*).

Findings from Simulation:

Below plots show two examples of simulated series. The blue lines demonstrate the locations of variance shifts predefined in the simulation and the red lines are the locations detected as the change points by the test. In some cases, the test detected false points which were quite far away from the real shift locations, as shown on the left plot. Depending on the variance of the simulated data, its distribution and the test type, the changes were captured at the right locations, as shown on the right plot.





Change Point Analysis on Real Data:

In the change point analysis on real series, as a parametric model **Iterative Cumulative Sums of Squares (ICSS) algorithm** and as a non-parametric model the **Mood test** were used, since they exhibited superior performance on simulated series. Using these tests, following financial series were investigated:

- *BIST100*, the main index for the Turkish Stock Market,
- *DAX*, one of the German Stock Index
- *FTSE100*, one of the main index on London Stock Exchange

High fluctuations occured during the year 2009 corresponding to the period with the global economic crisis were detected by each of the test.

Fewer change points were captured by the Mood test, whereas ICSS algorithm appeared to detect spurious change points far from a genuine shift in volatility.

Fitting GARCH models and Analysing the Results:

In the second stage, Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models were fitted, since it is the most widely used type and known as giving good fit in financial applications.

While running the model on the segmented series, the dummy variables were introduced into the variance equation which modify the parameters of the GARCH (1, 1) model.

The model comparison metrics used in the study were **Akaike Information Criterion (AIC)** metric which measures the goodness of fit of the model and penalizes each parameter included to the equation. The models with dummy variables provided better fit, as these models produced lower AIC values as compared to the standard GARCH model.

The models with the ICSS provided better fit for all stock indices. The superior performance of the ICSS was notable, since it flagged more change points, thus introduced more parameters to the model, yet it generated lower AIC values than the Mood test did. Overall, the results suggestted that segmented series with volatility clusters created by the change point models give better fit.

Project 3:

Project Name: *Analysing Pet Owner's Buying Habits and Modeling Propensity for Healthy Food Products*

Project Sponsor: *Acorn Pet Food Store &* Lancaster University

Description: A team project for getting hands-on experience with real world data and applying data science pipeline for solving Acorn Pet Food Store's business problem, analysing stock movement patterns of pet food, text mining and statistical modelling for assessing customers' buying habits

Software/Programming Language: R (*ggplot, dplyr, lubridate, stringr packages etc.)*

Project length: 2 months

Business Objective:

Acorn Pet Food Store was developing own label healthy pet food and wanted to supply small chain supermarkets. So, the client needed to justify the need for healthier pet food and its importance in the market.

The objective of the project was to find out the differences in the sales of **healthy** and **non-healthy** pet food and to identify patterns to shoppers' buying habits in relation to healthy and non-healthy pet food using Electronic Point of Sale (EPOS) data provided by the Client.

Proposed Solution and Data Management Strategy:

- Effectively pre-process and analyse the data in order to assess trends in buying habits of healthy and non-healthy food products

- Use text mining techniques to extract food sale records from all transaction records containing product *returns, shop use, sales of non-food items* etc. and to identify whether it is healty or non-healthy food sale

- Conduct a detailed exploratory analysis to find buying patterns of healthy food products

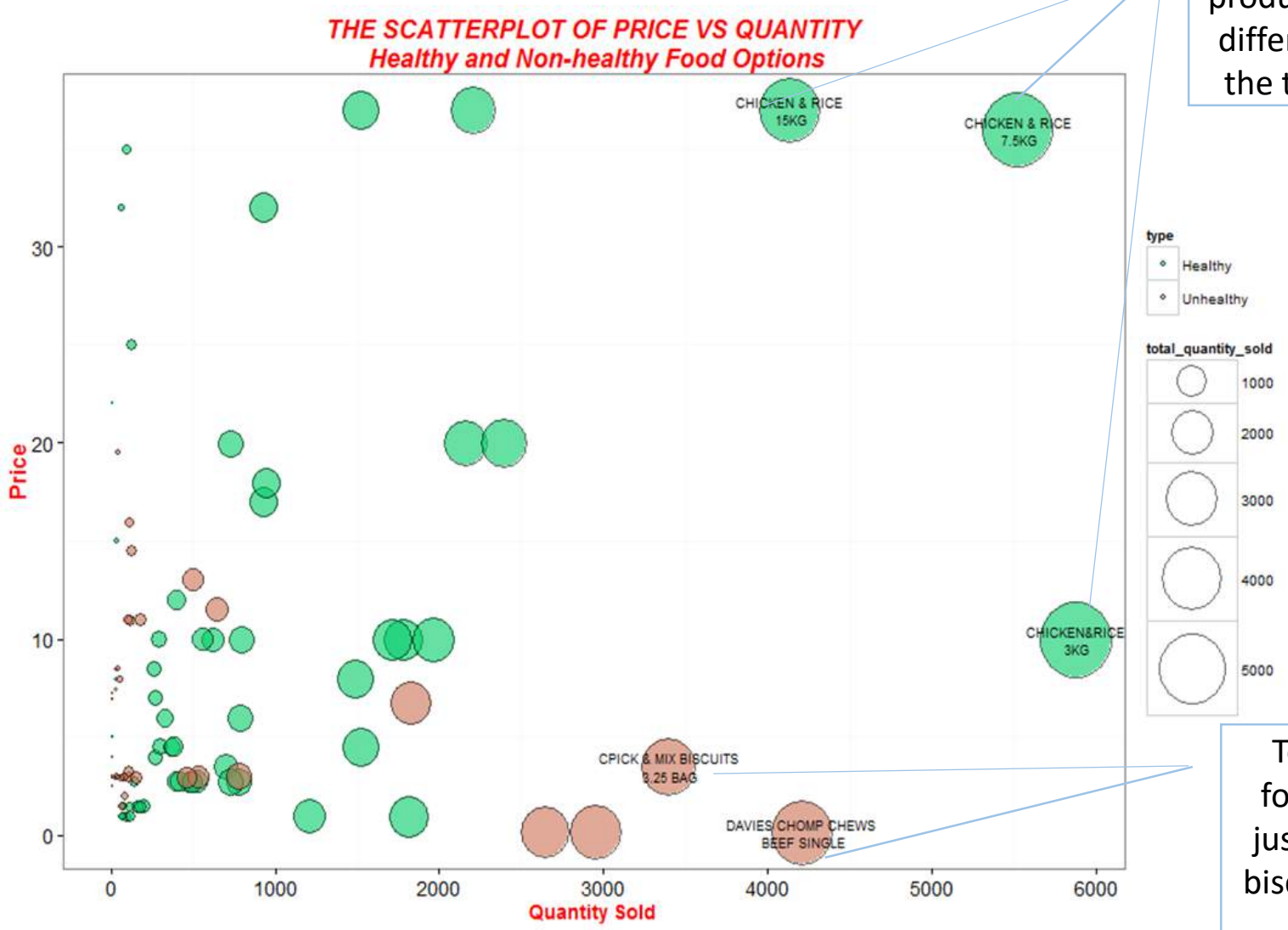
- Explore this in more detail by constructing statistical models to determine how factors such as *total money spent* and *food type* (snack/meal) affect a customer's decision to purchase healthy or non-healthy options.

Exploratory Data Analysis:

Each point represents the sum of quantities sold for each product.

Healty food products are green.
Non-healthy food products are orange.

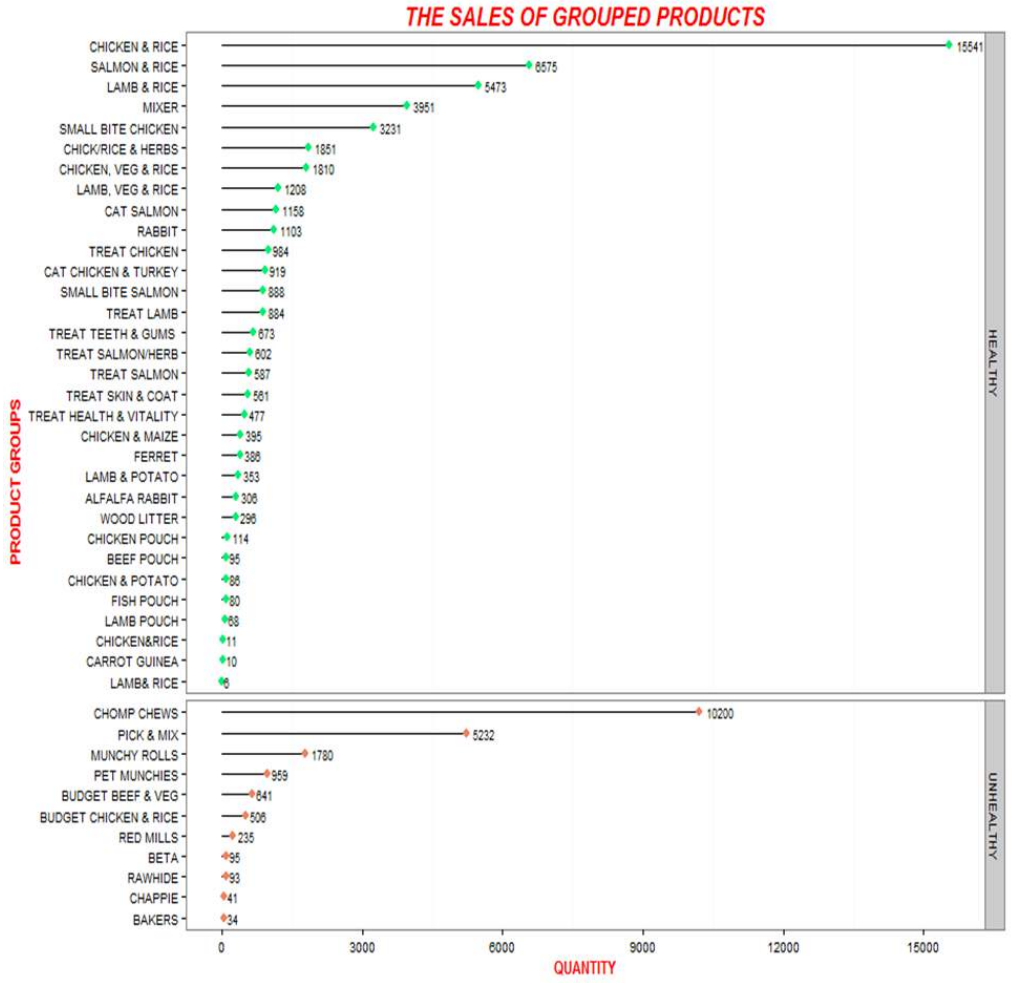
Quantity is mapped to size, so that the products with high sales are shown with larger points.

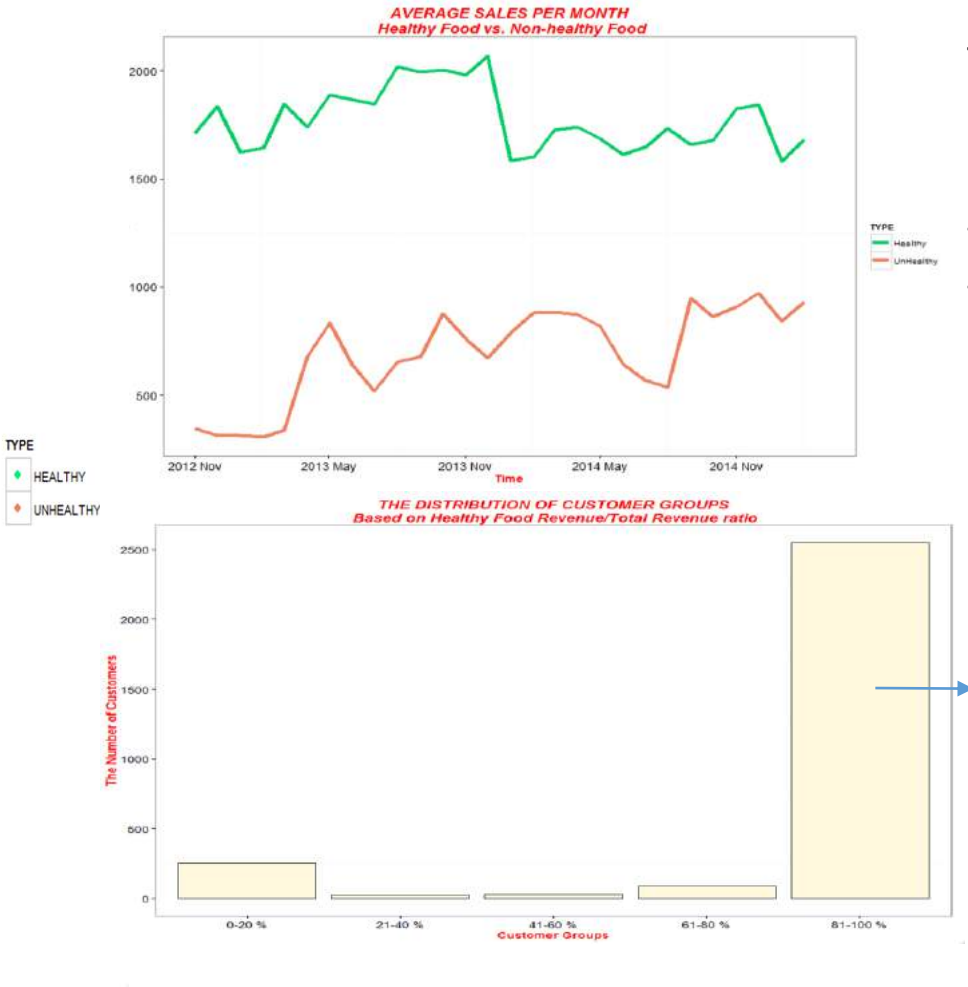


Chicken & Rice products packed in different kilos are the top products

Top non-healthy food products are just snacks (chews, biscuits), not proper meal for pets

Exploratory Data Analysis -Continued:





The sales of Healthy Food is always higher than Non-healthy Food, but there is a slight upward trend in Unhealthy Food Sales.

The largest group is composed of **customers whose healthy food revenue constitutes 81-100 % of total revenue**. Most customers in this group buy only healthy food for their pets.

Statistical Analysis:

In terms of snacks the difference of average money spent was small, but for meals the difference was higher. This suggested that customers prefer non-healthy snacks for non-price reasons (e.g. taste or larger selection of items). On the other hand non-healthy meals were noticeably cheaper. Healthy meals were much preferred, despite the availability of cheaper, unhealthy options.

Average money spent by customer on pet food

	Snack	Meal
Healthy	£2.77	£17.12
Non healthy	£3.59	£11.22

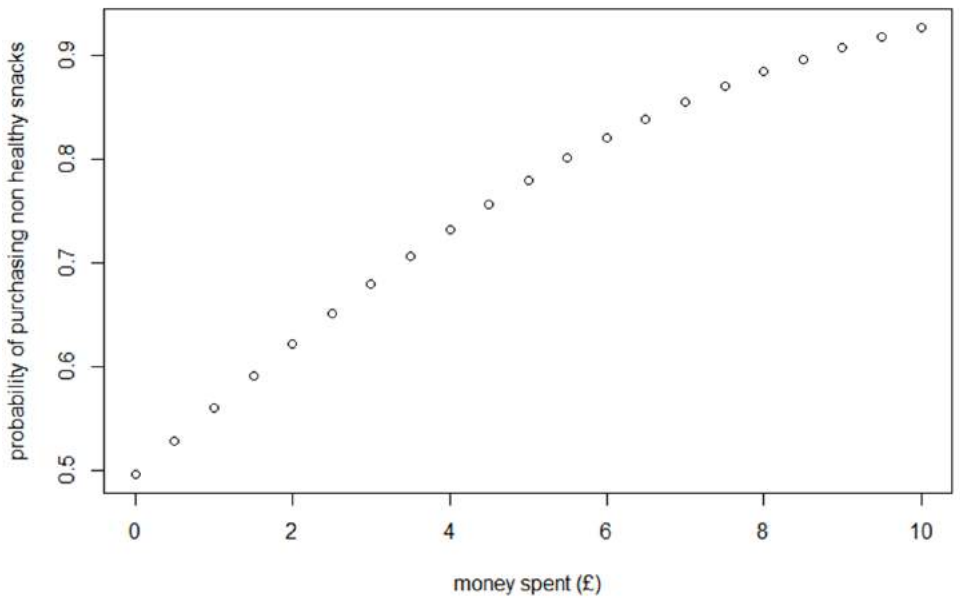
Logistic Regression Model:

A logistic regression model was employed to model the probability of a customer purchasing non-healthy pet food.

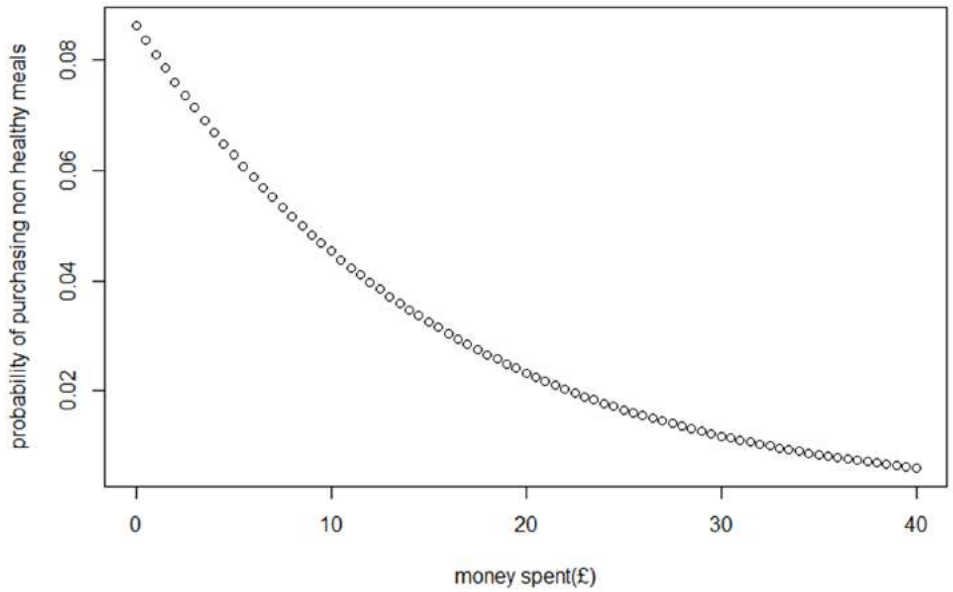
The covariates included to the model were *total cost of purchase* (which can be interpreted as the minimum amount of money a customer is willing to spend on pet food), and the *food type* (snack or meal). It was found that both covariates were statistically significant.

According to the model, at the average snack price, the probability of purchasing non-healthy snacks was **0.6975**. At the average meal price, the probability of purchasing non-healthy meal was merely **0.029**.

Plotting Probability for Non-Healty Food Purchase:



This graph shows the predicted probability of a customer purchasing **non-healthy** snacks. It can be observed that the probability of purchasing non-healthy snacks increases with the amount of money spent. So, it can be said that customers prefer non-healthy snacks.



This graph shows the predicted probability of a customer purchasing **non-healthy** meals. It is clear that customers does not prefer non-healthy options for meals.

Conclusion and Suggestions to the Client:

- Customers are currently much more likely to purchase non-healthy options for pet snacks, but healthy options for meals. As healthy meal products dominated the purchases and generated the larger part of the revenue, Acorn Pet Food Store could develop its own label healthy food product.

- It is also not uncommon for frequent customers to purchase a mix of both healthy meal and non-healthy snacks. So, multi-buy promotions could be offered to customers by matching their own labeled healthy food product with non-healthy snacks.

- Further analysis could be conducted with a more definitive list of non-healthy food products and additional data such as the types of pets customer are purchasing for and relevant demographic information.

Project 4:

Project Name: *Cluster Analysis and Classification of Mobile Phone Calls Data*

Project Sponsor: Lancaster University School of Computing

Description: A data mining project for pre-processing, feature engineering, applying k-means clustering algorithm and building a classification model by training LVQ type neural network classifier on mobile phone calls data

Software/Programming Language: MATLAB

Project length: 2 months

Business Objective:

- The objective of this project is to conduct a **cluster analysis** of mobile phone calls data for gaining insight into the groups of users with similar characteristics and to build a **classification model** for predicting the segments of unknown customers.

- By using segmentation, **groups of users with similar characteristics** can be served in line with their needs. This facilitates the development of user-centric strategies for improved customer experience and increased profit in the high competitive telecomunication business.

Proposed Solution:

- Employ data cleaning, feature extraction and normalization steps before the clustering analysis
- Explore the data and reduce the dimensionality
- Conduct k-means clustering algorithm and iterate the procedure by specifying different parameters until finding the best data partition
- Using the labels from clustering analysis, train LVQ type Neural Network classifier and asses its accuracy on validation data

Data Preprocessing and Feature Engineering:

On the dataset of mobile phone calls data, below preprocessing steps are conducted to improve the accuracy and efficiency of the machine learning algorithms:

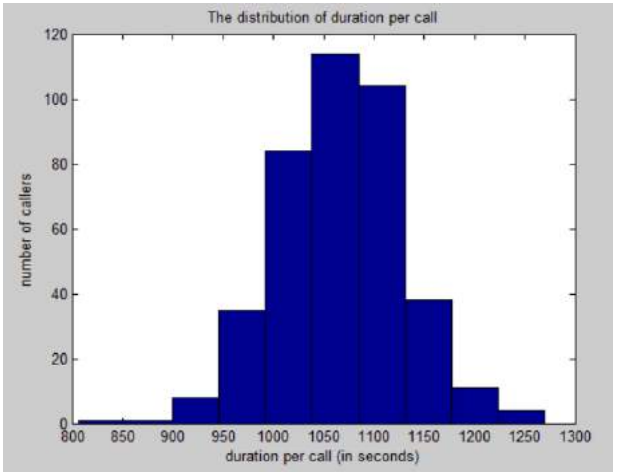
- Outlier detection by *z-score normalization*
- Missing data imputation
- Restructring data (converting to user-based dataset)
- Data Normalization

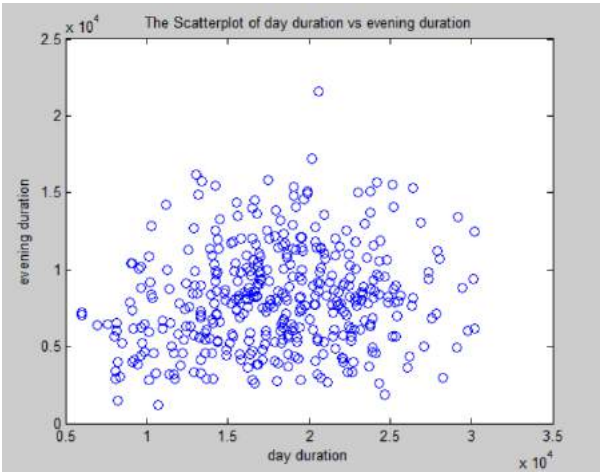
Feature Engineering steps are as follows:

- Correlation Analysis
- Features Extraction *(duration_per_call, number_evening_calls etc.)*
- Dimensionality reduction

Exploratory Analysis for Feature Selection:

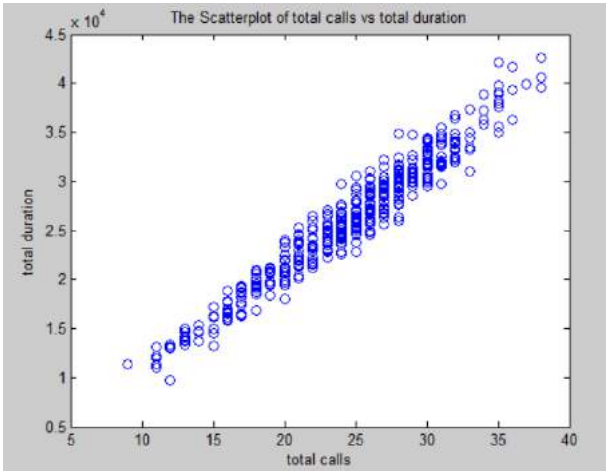
The distribution of duration per call appears to be normal with a mean of 17.7 minutes. Most users tend to spend 16-19 minutes (1000-1150 seconds) in each phone calls.

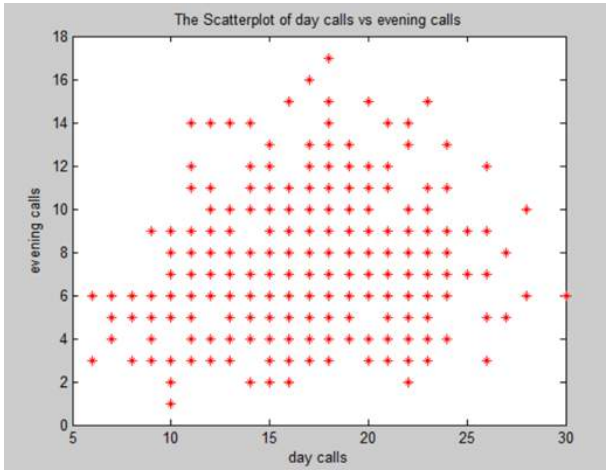




The duration of calls pattern indicates that evening calls appear to last shorter than day calls for most of the users.

It is clear that total calls and total duration are positively correlated. This means that total time spent on the phone by a user is mainly dependent on the number of calls, rather than the duration per call.

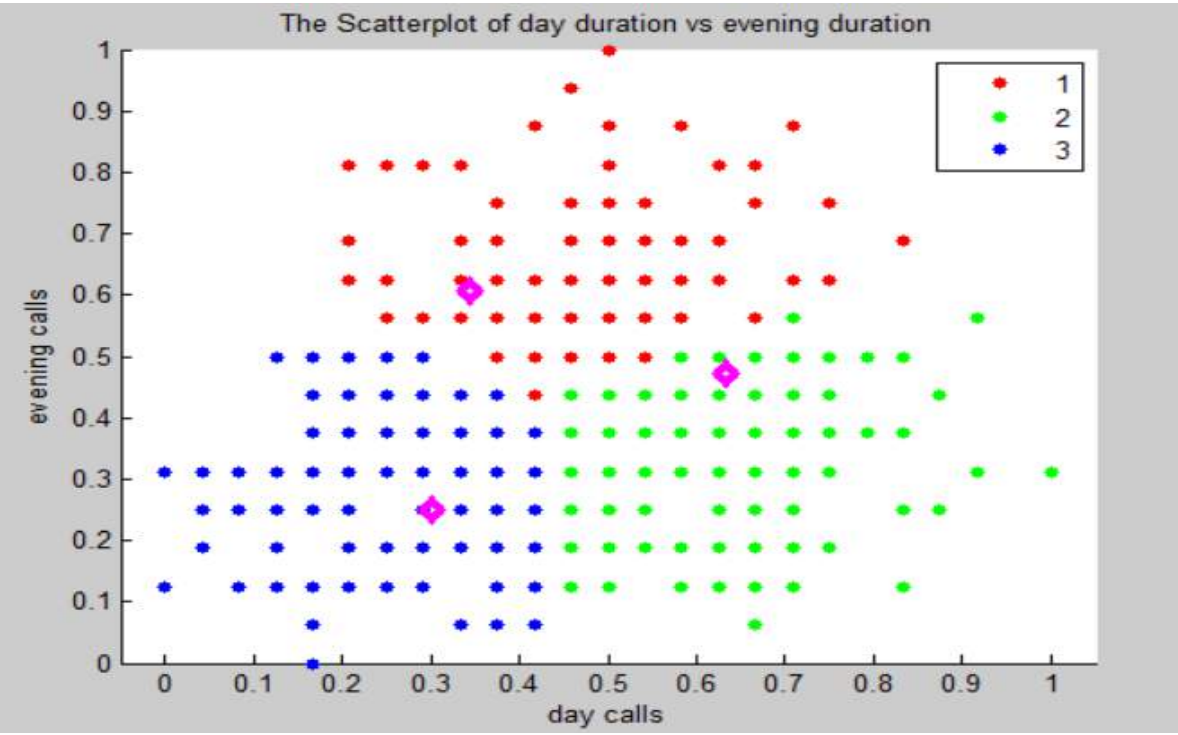




Day calls are signifincatly greater than evening calls as the means are 16.8 for day calls, 7.6 for evening calls. To understand this difference in terms of users, the scatterplot plot of number of day calls vs. number of evening calls is drawn.

Clustering Analysis with k-means Algorithm:

Two extracted features; *day_calls* and *evening_calls* were selected to be used in clustering analysis.
K-means clustering started with randomly selected cluster centers and assignment of each data objects to the closest cluster. These steps were iterated until no further improvement was observed. For finding the best data partition, different *number of clusters* and *distance measures* were used.
The final clusters with their centroids, shown as pink points, were visualized on the scatterplot.

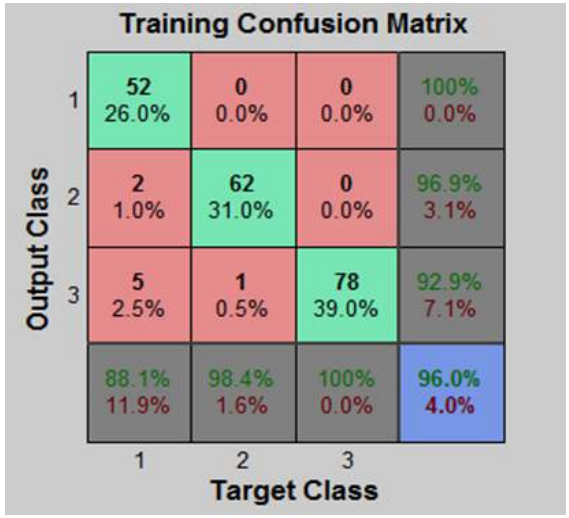


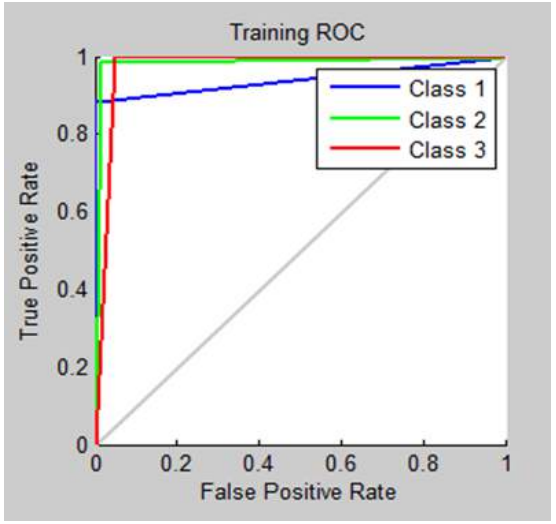
These 3 clusters were interpreted as follows:

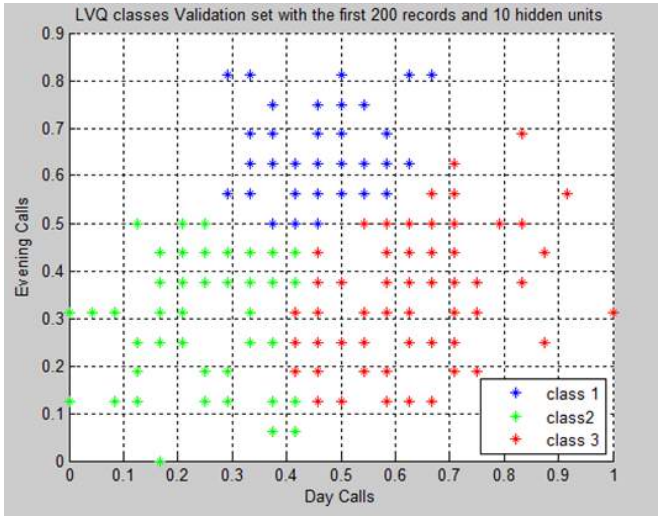
- Cluster 1 (red): Active users making both day and evening calls frequently
- Cluster 2 (green): Day time users making more day calls than evening calls
- Cluster 3 (blue): Non-frequent users making less calls both during the day and in the evening

Classification:

The clusters found by *k-means algorithm* were used for labelling classes and using these labels a classification model was conducted for mobile phone calls data by training a LVQ type Neural Network classifier.







The confusion matrix and the ROC curves having quite large area under the curves suggests high performance on the training set.

The classifier was applied to the validation set and the classes were separated fully from each other in line with the clusters of k-means algorithm.

Project 5:

Project Name: *Building a Credit Scoring Model for Predicting the Probability of Default on Mortgage Loan*

Project Sponsor: Lancaster University Management School

Description: A data mining project for building a credit scoring model using binary logistic regression and decision trees in order to predict the class label of prospect mortgage applicants

Software/Programming Language: SAS Enterprise Miner

Project length: 1 month

Business Objective:

This project aimed to identify bad customers who are likely to default on mortgage loan through building a predictive classification model based on dataset of previous customers.

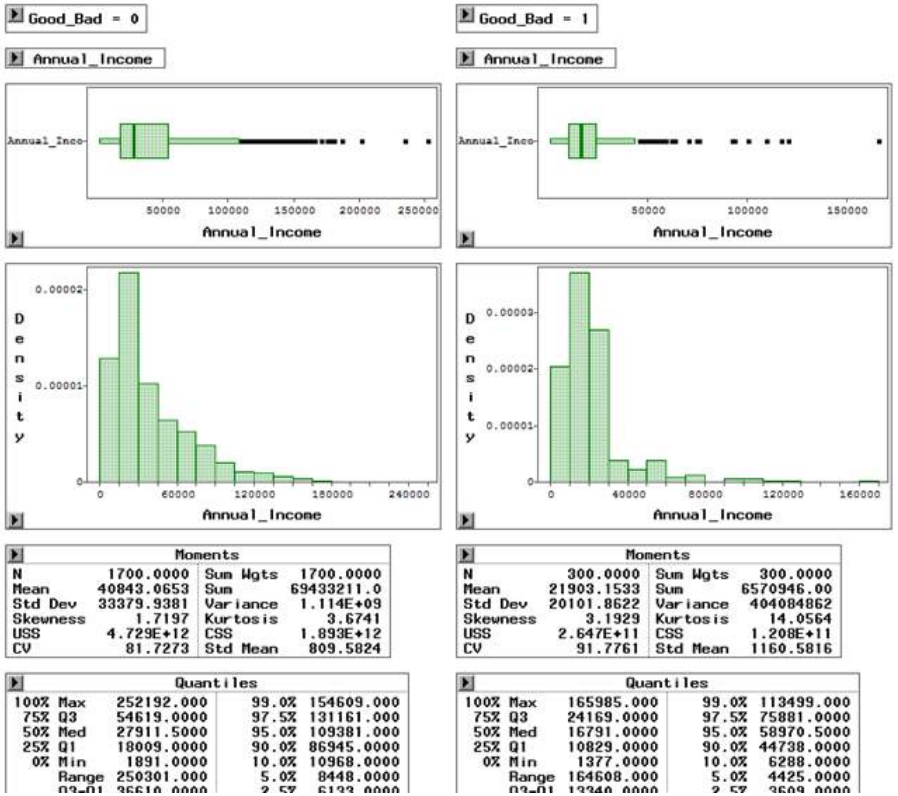
Having two class labels as Bad (Good_Bad=1) and Good (Good_Bad=0), the target variable had a binary distribution. So, in order to predict the class label of prospect mortgage applicants, classification methods such as binary logistic regression and decision tree model were applied.

Proposed Solution:

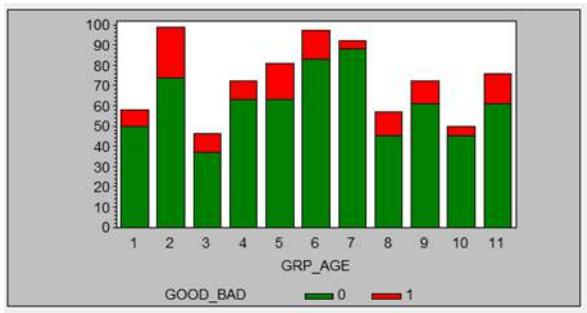
- Analyse each variable to detect their discriminatory power for the classification of good and bad customers
- Impute all missing values based on their missingness pattern
- Partition the data into training, validation and test sets according to common data partition strategy
- Apply binary logistic regression and decision tree models and compare the results

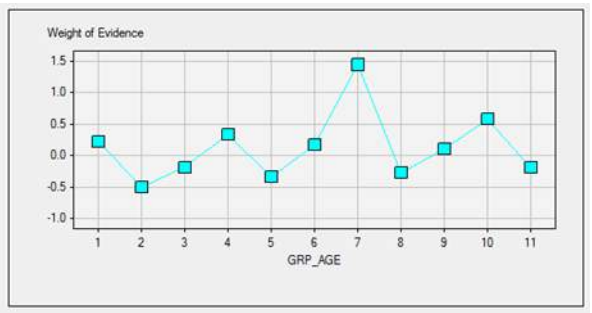
Feature Selection Techniques:

Exploratory analysis was conducted for each feature and plots with summary statistics, such as *Annual Income* plots shown below, were created for bad and good customer separately.



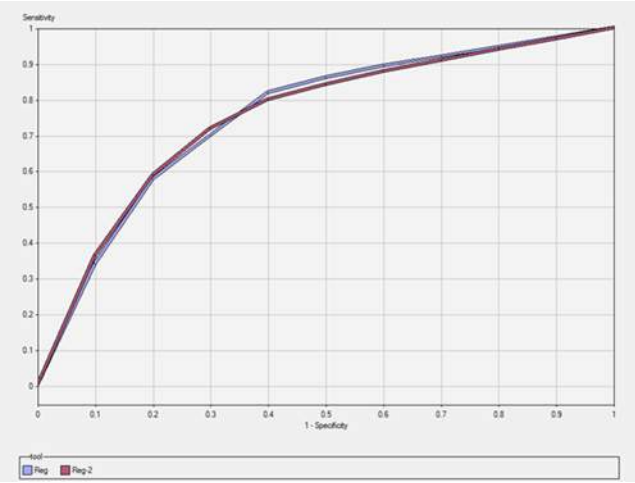
The information value of the features were analysed statistically. In order to examine the information value in more detail, weights of evidence for each bin were produced for features, such as *Age*. The groups having negative woe suggested that customers in these groups were more likely to be bad than other groups.



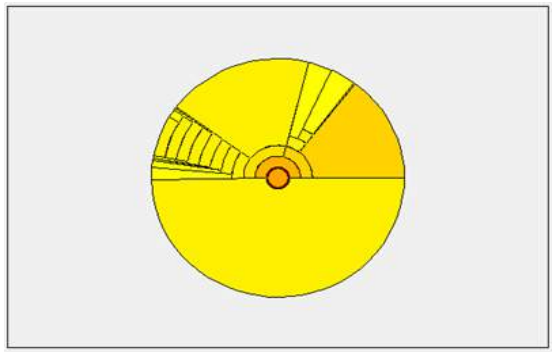


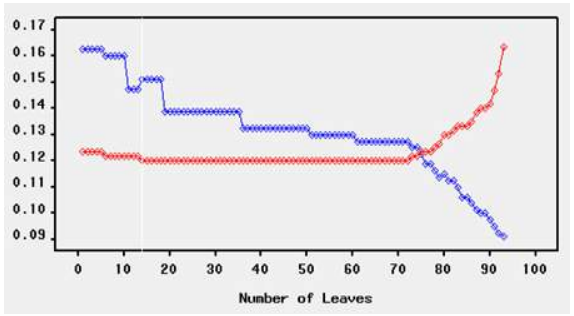
Model Building:

First, binary logistic regression model was estimated by using initial subset of variables and in order to improve model validity, new variables were added in the following trials. The performance of the model was assesed in terms of misclassification error and ROC curves.



Second, decision trees were built by specifying different combinations for the parameters; *minimum number of observations in a leaf, observations required for a split search, maximum depth of tree, maximum number of branches from a node*.





Various stopping and pruning options were also used to improve the classification. *Tree Ring Figures* were observed to assess the degree of impurity in the leaves and *Misclassification Error Table* were drawn to check over-fitting.