Paving the way for the use of prediction modelling in a health care environment

Ilse van Zyl

15324745

The final year project is presented in partial fulfilment of the requirements for the degree of Bachelors of Industrial Engineering at Stellenbosch University.

Project study leader: Liezl van Dyk

Co-project study leader: Tanya Visser

November 2011

DECLARATION

I, the undersigned, hereby declare that the work contained in this final year project is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

…….. ……

Ilse van Zyl Date

ECSA EXIT LEVEL OUTCOMES REFERENCES

The following table include references to sections in this report where ECSA exit level outcomes are addressed.

Exit level outcome	Section(s)	Page(s)
1. Problem solving	1.1 Unnecessary hospitalisation 6 Conclusions and Recommendations 8 Road ahead	1 29 30
2. Application of engineering & scientific knowledge	2 Data and Data handling	7-16
5. Engineering methods, skills & tools, incl. IT	1.7 Methodology 2.Data and Data handling 4.Technologies considered 5.Experimentation	5 7-14 22-23 23-24
6. Professional & Technical communication	Entire report ORSSA Paper ORSSA conference presentation	 Appendix C Appendix D
9. Independent learning ability	1.1 Unnecessary hospitalisation 1.2 Background 2.Data and data handling 3.Prediction techniques: CART, Neural Networks, MARS, Ensembles 4.Technologies considered	2 2-4 7-16 17-22 22-25
10. Engineering professionalism	1.Introduction 6.Conclusions and recommendations 8.Road ahead	2-6 29 30

SYNOPSIS

The high cost of hospitalisation is a challenge for many health insurance companies, governments and individuals alike. In 2006, studies concluded that well over $30 billion was spent on unnecessary hospitalisations in the United States of America, where unnecessary hospitalisations are those that could have been prevented through early patient diagnosis and treatment. Undoubtedly, there is room for improvement in this regard and it can be agreed that where lives are at stake, prevention is always better than cure; successful hospitalisation prediction may make hospitalisation prevention a realistic possibility.

The Heritage Provider Network, a health insurance and health care provider and sponsor of the Heritage Health Prize (HHP) Competition, have come to realise the potential benefits that a hospitalisation prediction model could effect (Heritage Provider Network Health Prize, 2011). The competition is aimed at producing an effective hospitalisation prediction patient admissions algorithm (PPAA) to predict the amount of days a member will be hospitalised in the next period using health insurance claims data of the current period. The goal is to ultimately prevent the unnecessary hospitalisation of identified members in their network. If successful this could have many benefits to the wider society including fewer critical medical cases, fewer claims and consequently lower expenses for all stakeholders in the affected system.

The competition serves as inspiration for this study which aims to pave the way for the research team who will be developing such a PPAA. This was accomplished by providing insights and identifying possible pitfalls in the development of a Predictive Patient Admission Algorithm (PPAA) using the Heritage Health Prize case study as a reference.

Typically available hospitalisation data that serves as input for the PPAA are briefly described, together with recommendations on methods and technologies with which to extract, transform and load (ETL) data within this context.

A list of contender techniques was assembled based on the given data, the algorithm's expected input requirements and the techniques' ability to meet these needs. The prediction modelling techniques reviewed include classification and regression trees (CART), multivariate adaptive regression splines (MARS), neural networks and ensemble methods. Techniques were compared in terms of a set of criteria needed to use the available data and give the desired outputs.

The data mining technologies considered to model with the preferred technique include Statistica data miner, SPSS Clementine, SAS Enterprise Miner, Matlab, Excel with VBA and R. These technologies were also compared on how well they can model available data with the contender techniques. The research team's compatibility with technologies was also considered.

Recommendations concerning the prediction modelling technique was using ensemble methods and the choice of technology for ETL was SQL Server and for prediction model building recommendations are Statistica, R or Matlab. Experimentation was conducted with selected CART, MARS and the Random Forests techniques in the available technologies in order to support future prediction modelling decisions of the research team. It was concluded that the included predictor variables do not have sufficient predictive power for the use of CART, MARS and Neural Networks and that Random Forests deliver more favourable results and it was recommended that this modelling should be explored further for the use of the HHP application.

OPSOMMING

Die hoë koste van hospitalisering is 'n uitdaging vir baie mediesefondse, regerings en individue. In 2006 het studies getoon dat meer as $ 30 miljard bestee is aan onnodige hospitalisering in die Verenigde State van Amerika, waar onnodige hospitalisering die gevalle is wat deur vroeë diagnose en behandeling voorkom kon word. Dit kan duidelik gesien word dat daar ruimte vir verbetering is in hierdie verband. Waar lewens op die spel, is voorkoming altyd beter as behandeling en as hospitalisering suksesvol vooruitgeskat kan word, kan hospitalisering voorkoming 'n realistiese moontlikheid word.

Die Heritage Health Provider Network, 'n gesondheid versekering verskaffer en gesondheidsdienste en die borg van die Heritage Health Prize (HHP) kompetisie, het besef wat die potensiële voordele is van hospitalisering vooruitgeskatting (Heritage Health Prize, 2011). Die kompetisie is gemik op die ontwerp van 'n effektiewe hospitalisering vooruitgeskattings algoritme wat kan voorspel wat die aantal dae gaan wees wat ' n lid gehospitaliseer gaan word in die volgende periode. So ‘n algoritme gaan opgestel word met behulp van gesondheid versekering eise en hospitalisering data. Die doel is om uiteindelik te verhoed dat die onnodige hospitalisering van geïdentifiseerde lede plaasvind. Indien dit suksesvol is kan lei tot minder kritiese mediese gevalle, minder eise en gevolglik laer kostes vir alle belanghebbendes in die betrokke stelsel.

Die kompetisie dien as inspirasie vir hierdie studie wat daarop gemik om die weg te baan vir die navorsingspan wat die algoritme gaan verder ontwikkel. Insigte en moontlike slaggate word uitgelig in die ontwikkeling van 'n vooruitgeskattings algoritme met behulp van die Heritage Health Prize gevallestudie as 'n verwysing.

In die studie word tipies beskikbare hospitalisering data, wat dien as inset vir die algoritme, kortliks beskryf, saam met aanbevelings oor die metodes en tegnologie vir die onttrek, herskep en laai (OHL) van data binne hierdie konteks.

'n Lys van die oorweegde tegnieke is saamgestel, gebaseer op die gevallestudie data, die algoritme se verwagte inset-vereistes en die tegnieke se vermoë om aan hierdie vereistes te voorsien. Die vooruitskattings tegnieke sluit in klassifikasie en regressie bome (CART), meervoudige veranderlike aanpasbare regressie latfunksies (Multivariate adaptive regression splines), neurale netwerke en kombinering metodes. Tegnieke is ook vergelyk in terme van 'n stel kriteria wat nodig is om die beskikbare data te gebruik en die verlangde uitsette te lewer. Die data-ontginning tegnologië wat oorweeg is sluit in Statistica data miner, SPSS Clementine, SAS Enterprise Miner, Matlab, Excel met VBA en R.

Hierdie tegnologië is vergelyk met verwysing tot hoe goed hulle die oorweegde vooruitskatting tegnieke kan akkommodeer. Die ondersoek span se verenigbaarheid met die tegnologië is ook in ag geneem.

Aanbevelings met betrekking tot die vooruitskatting tegnieke was om gebruik te maak van die ensemble metodes, die keuse van tegnologië vir OHL is SQL server en die bou van 'n vooruitskattings model kan gedoen word in R of Matlab en Statistica kan gebruik word vir eksplorasie doeleindes. Eksperimente is uitgevoer op CART, MARS en Random Forests (‘n kombinering metode) in beskikbare tegnologië met die doel om toekomstige besluitneming van die navorsingspan te steun met betrekking tot die modellering van die vooruitskattings algoritme. Daar was tot die gevolgtrekking gekom dat die gekose vooruitskatter veranderlikes nie effektief is met die gebruik van vooruitskattings tegnieke naamlik CART, MARS en neurale netwerke. Die eksperimente gedoen op Random Forests het meer voordelige resultate opgelewer. Dit word dus aanbeveel dat hierdie tegniek verder ondersoek word vir die gebruik in die HHP gevallestudie.

ACKNOWLEDGEMENTS

The author wishes to acknowledge the valuable time, support and guidance of the following people:

- I was blessed to have an amazing study leader Mrs Liezl van Dyk and co-study leader Mrs Tanya Visser.
- Ms Susan van Zyl for her contribution to the editing of this document
- Mr Francois van Zyl for his contribution to the structuring of this document
- Mr Francois van Zyl (jnr) and Ms Suzette van Zyl for their expert advice and contributions with data handling
- Lidia Auret for her time, help, insight and for sharing her passion for data mining with me
- Professor Martin Kidd for his help
- My class friends for motivating and encouraging me throughout this project
- Finally, I thank God for amazing opportunities and immeasurable grace

1 TABLE OF CONTENTS

Declaration i

ECSA Exit level outcomes references ii

Synopsis iii

Opsomming v

Acknowledgements vii

List of Figures xi

List of Tables xiii

List of Equations xiv

Glossary xv

1 Introduction 1

1.1 Background 1

1.1.1 Unnecessary hospitalisation 1

1.1.2 Decision support systems for hospital information system data 1

1.1.3 Heritage Health Prize competition Case study 2

1.1.4 Prediction Modeling in the Health Care Sector 3

1.1.5 Hospitalisation prevention initiatives in the Health Care Sector 3

1.2 Aim and objectives of the study 4

1.3 Composition of the research team 4

1.4 Scope of this study 4

1.5 Methodology 5

2 Data and Data handling 7

2.1 The data 7

2.2 Issues with data interpretation 8

2.3 Data handling 10
2.3.1 Extraction 10
2.3.2 Transformation 10
2.3.3 Loading 16
2.3.4 ETL Alternatives 16
3 Prediction Techniques 24
3.1 Regression modeling 25
3.2 Classification and regression trees (CART) 26
3.3 Neural Networks 27
3.4 Ensemble Methods 29
3.5 Comparing contender prediction modelling techniques 29
4 Technologies considered 31
4.1 Spss Clementine 31
4.2 SAS Enterprise miner 32
4.3 Statistica Data miner 33
4.4 Excel and VBA 33
4.5 Matlab 33
4.6 R 33
5 Experimentation 35
5.1 ETL Experimentation 35
5.2 Statistical Experimentation 35
6 Conclusions and Recommendations 48
7 Personal Experiences 51
References 52

Appendix A: The Kaggle concept and The heritage health provider network 1

Appendix B: ORSSA Paper 4

Appendix c: ORSSA Presentation..... 5

Appendix E:Full Data Dictionary..... 6

Appendix F: Excel code for summarisation 7

Appendix g: R code for Random Forests 15

Appendix H: Chi test calculations..... 18

LIST OF FIGURES

Figure 1: Roadmap to pave the way for hospitalisation prediction modelling 6

Figure 2: Roadmap to pave the way for hospitalisation prediction modelling - the data source..... 7

Figure 3: Examples of different types of variables to be found in health insurance claims and hospitalisation data. 10

Figure 4: Roadmap to pave the way for hospitalisation prediction modelling - data warehousing and ETL 11

Figure 5: Extended Entity Relationship Diagram 13

Figure 6: Extended Entity Relationship Diagram (performed in SQL Server) 13

Figure 7: Example of SQL query code 13

Figure 8: Resulting query output. 14

Figure 9: Basic ETL using Excel 18

Figure 10: Hybrid ETL using SQL Server and Excel 19

Figure 11: ETL using only 19

Figure 12: Most optimal ETL using only SQL Server 19

Figure 13: Roadmap to pave the way for hospitalisation prediction modeling - prediction modelling techniques 24

Figure 14: Roadmap to pave the way for hospitalisation prediction modeling - technologies considered 31

Figure 15:Roadmap to pave the way for hospitalisation prediction modeling – experimentation 35

Figure 16: Example of a variable's distribution 37

Figure 17: Dependent variable's distribution 37

Figure 18: Inputting variables 38

Figure 19: Stopping parameters 39

Figure 20: Importance plot for Classification tree analysis. 40

Figure 21: Importance plot for regression tree analysis where DIH >0 41

Figure 22: Observed and predicted frequencies 41

Figure 23: CART predicted vs observed 42

Figure 24: Variables read into MARS 42

Figure 25: MARS parameters 43

Figure 26: Observed and predicted frequencies 43

Figure 27: MARS predicted vs observed 44

Figure 28: Observed and predicted frequencies 45

Figure 29: Random forests predicted vs observed 46

Figure 30: Experimentation summary 47

Figure 31: Prediction error rate comparison 47

Figure 32: Roadmap to pave the way for hospitalisation prediction modeling – conclusion and recommendations 48

Figure 33: Kaggle concept 2

LIST OF TABLES

Table 1: Data dictionary for HHP data 9

Table 2: Example of converting text data into numerical data. 12

Table 3: Technology ETL decision matrix. 17

Table 4: Pairwise comparison 20

Table 5: Normalised matrix A. 21

Table 6: Criteria weights 21

Table 7: Scenario 1 - Excel and VBA 22

Table 8: Scenario 2 – Excel, VBA and SQL. 22

Table 9: Scenario 3 - Statistica 22

Table 10: Scenario 4 - SQL alone 23

Table 11: Summary of scenario weights 23

Table 12: Contender techniques decision matrix 30

Table 13: Technology decision matrix 32

Table 14: Removed variables 38

LIST OF EQUATIONS

Equation 2.1: Pairwise comparison criteria 20

Equation 2.2: Consistency index 21

Equation 5.1: Prediction error rate 36

GLOSSARY

Abbreviation	Definition
DSS	Decision support system
CART	Classification and regression trees
MARS	Multivariate adaptive regression splines
PPAA	Predictive patient admissions algorithm
ETL	Extract transform and load
VBA	Visual Basic Application
HPN	Heritage Provider Network
AHP	Analytical hierarchical Process

1 INTRODUCTION

1.1 BACKGROUND

1.1.1 UNNECESSARY HOSPITALISATION

The high cost of hospitalisation in the United States is a challenge to many health insurance companies and individuals alike. What makes matters worse is that many hospitalisations can be prevented by correctly diagnosing and treating conditions earlier. In 2006 studies have concluded that well over $30 billion was spent on unnecessary hospital (hospitalisations that could have been prevented by early diagnosis and treatment) admissions of the more than 71 million individuals that are admitted to hospitals in the United States each year (Heritage Provider Network Health Prize, 2011). The main parties affected by unnecessary hospitalisations, are health insurance companies, the patients being admitted, and the individual responsible for the hospital bill. This marks the first sign that there is a major problem to be solved in this environment.

Starfield (2000) mentions some statistic that strengthen the argument that hospitalisation should be prevented if possible, these include:

- 2000 deaths per year from unnecessary surgery
- 7 000 deaths per year from medication errors in hospitals
- 20 000 deaths per year from other errors in hospitals
- 80 000 deaths per year from infections in hospitals
- 106 000 deaths per year from non-error
- adverse effects of medications -- totalling up to 225 000 deaths per year in the US from iatrogenic causes (when a patient dies as a direct result of treatments by a physician, whether it is from misdiagnosis of the ailment, or from adverse drug reactions used to treat the illness) which ranks these deaths as the number 3 killer (drug reactions are the most common cause).

1.1.2 DECISION SUPPORT SYSTEMS FOR HOSPITAL INFORMATION SYSTEM DATA

In order to address this issue of unnecessary hospitalisation, effective decision-support systems for hospital information system data can be developed and utilised more effectively. Decision support systems are the back-bone for successful decision making in any organisation.

In large health insurance and health care companies, such as the Heritage Health Provider Network, problems are broad and complex involving high risks and uncertainty requiring such an organisation to employ a decision-making process that is structured and consistent.

According to Nykänen (2000) decision support systems in the healthcare sector have been approached from two disciplinary perspectives: Information systems science and artificial intelligence. Information systems science approach mostly supports managerial decision making (also known as managerial decision support systems) whereas an artificial intelligence-based approach (also known as clinical decision support systems) focus on the design of systems to support individual decision making in tasks that are considered to require intelligence.

Considerable research has been done in the field of decision support in health care such as the studies done by Hunt, Haynes, Hanna and Smith (1988) who found that the clinical decision support systems (CDSSs) can enhance clinical performance for drug dosing, preventive care, and other aspects of medical care, but not convincingly for diagnosis.

Now considering this, imagine a decision support system which uses hospital information systems data to identify patients that will be hospitalised in the near future, based on past patient data and then preventing such hospitalisations. This is what the Heritage Health Prize competition is attempting to do: prevent unnecessary hospitalisation with the use of prediction modelling, early diagnosis and treatment. The Heritage Health Prize competition was the inspiration for this project as was discussed in more detail in Section 1.1.3

1.1.3 HERITAGE HEALTH PRIZE COMPETITION CASE STUDY

The Heritage Health Prize competition was born out of collaboration between the internet platform Kaggle and the Heritage Provider Network (see Appendix A for more information on the Kaggle concept and The Heritage Health Provider Network). Seattle (2011) describes the aim of the competition to “jump-start a stagnating field to eke out improvements. They are meant to recruit non-conventional participants with expertise from other domains, providing fresh insights and spurring existing researchers to get on their bikes”.

Participants in the competition received a set of health insurance claims data and hospitalisation data from which participants are to construct an algorithm that can predict how many days a patient will spend in a hospital in the next year.

The winning algorithm will then be used by Heritage Provider Network to identify which patients in their network are at risk for hospitalisation and preventative measures can be applied. With the algorithm the Heritage Provider Network can determine which members are high risks for hospitalisation and act accordingly.

Ensure that these people are treated by their physicians as soon as possible, design a patient specific care plan for them and take any other necessary actions to avoid hospitalisation (Heritage Provider Network Health Prize, 2011). This rest of the project is based on the Heritage Health Prize competition's data and concept.

1.1.4 PREDICTION MODELING IN THE HEALTH CARE SECTOR

This project is not the first to consider the use of prediction modelling techniques to assist in decision making in healt care environment. For example Miyata, Hashimoto, Horiguchi, Matsuda, Motomura and Takamoto (2009) used multivariate logistic regression to predict in-house mortality in hospitalisation, using records obtained on a nation-wide administrative database in Japan. Decision tree analysis was done by Lee, Yang and Parr (1988) to predict an outbreak of dengue haemorrhagic fever (DHF) with the decision tree chi-squared automatic interaction detector (CHAID) with bi-way and multi-way splitting. The resulting trees were pruned to achieve the highest sensitivity with the shortest tree. It was concluded that this prediction technique would prevent 43.9% of mild DHF cases from hospitalisation. These predictions could help doctors decide whether to hospitalise patients or to do outpatient monitoring.

Shortcomings of these studies when compared to the HHP case project is that the response output of these models were binary in character, Miyata *et al*. (2009) predicted for mortality or non-mortality, Dwyer *et al.* (2001) used logistic regression which has binary output and finally Lee *et al.* (1988) required a prediction output of either hospitalisation or outpatient monitoring which is also binary in character.

1.1.5 HOSPITALISATION PREVENTION INITIATIVES IN THE HEALTH CARE SECTOR

Data mining have often been used in decision support systems in the health care sector. A typical initiative that used decision support system to minimize hospitalisation is the *utilization review,* used by insurance companies, which aims to ensure that the request for recommended medical treatment by a member is appropriate. Insurance companies often use statistical models to indicate which member's claim information causes an anomaly which in turn qualifies this member for a utilization review.

This procedure helps the company (and the patient) minimize costs and determine if the recommended treatment is appropriate and that the company provides adequate coverage for it. Often utilization reviews only aim to check whether or not members are exploiting the insurance company, but it is also known to have picked up cases where members would have received wrongful treatment.

1.2 AIM AND OBJECTIVES OF THE STUDY

The aim of the project is to pave the way for a research team by providing insights and identify possible pitfalls in the development of a Predictive Patient Admission Algorithm (PPAA). The following objectives are set to obtain this aim:

1. Inspect data received for the competition and determine how to extract, transform and load (ETL) data correctly
2. Compare contender techniques for development of the PPAA with given competition data.
3. Compare technologies needed for development of the PPAA, considering the appropriateness thereof, research team's resources and research team's knowledge fields
4. Experiment with chosen techniques and available technologies through demonstration in order to support prediction modelling decision of the research team.

1.3 COMPOSITION OF THE RESEARCH TEAM

The research team is a group of researchers consisting under-graduate, post-graduate students and lecturers who specialise in fields such as industrial engineering or applied mathematics. The research team has knowledge in general engineering concepts such as advanced mathematics, basic statistics, technological tools and problem solving skills. In this project the research team's resources and knowledge fields was be considered when making recommendations.

1.4 SCOPE OF THIS STUDY

This project is scoped in terms of the following aspects:

- Although it is an interesting and controversial topic in public health care in South Africa, the practical implications of the National Health Insurance Scheme was excluded from this project, because insurance information is from a private American health care system.

- This project only discussed and analysed the development of a PPAA, not the implementation thereof.
- Recommendations were made with the specified research team in mind.

1.5 METHODOLOGY

Winston (2004) suggests a seven step model-building procedure when attempting to solve an organisation's problem, only the first four are applicable for this project. These steps include:

1. **Determine and formulate the problem to be solved:** The customer's problem is defined in this step. In this project the problem of unnecessary hospitalisations is defined in Section 1.1.1 and it should be noted that this is the governing problem on which this project is built, although it was beyond this project's scope to solve this governing problem. This project aims to only pave the way for a future research team (the customer for whom this project is conducted) to solving the problem of unnecessary hospitalisation. In Section 1.1 background is given on the setting of the Heritage Health Prize competition, which was the original inspiration for this project. In Section 1.1.2 decision support systems in the health care industry are discussed. Who the customer is and what the objectives and goals of this project are have been discussed in Section 1.3 and 1.2 consecutively.
2. **Observe the system:** This step includes collecting data and information for the project. Data and information collected refer to actual data received with which to do data mining as well as the information which determined the guidelines by which the project was conducted. The guidelines for this project can be deduced from Sections 1.2, 1.3 and 1.4, which describe what the "customer" expect from this project, who the customer is and what the scope of the project is.
3. **Formulate a mathematical model of the problem:** the focus of this project is not on one model alone, but rather a variety of techniques and technologies are compared and recommendations are made concerning data mining for this application. This process is described and applied in Sections 0, 3, 4 and 5.
4. **Verify and validate the model and use the model for prediction:** The approach was validated by writing a peer reviewed paper for the Operations Research Society of South Africa (ORSSA) and also presenting the paper, in September 2011, at the yearly conference of the same society. In addition the model was statistically validated by consulting the Stellenbosch University Statistics department.

Verification on the other hand was conducted by experimenting with selected data mining techniques and technologies to see whether models and technologies can handle the inputs received and produce desired outputs. This verification can be seen in Section 5.

Selecting a suitable alternative is not the main outcome required of this project, but rather paving the way to be able to build such a model successfully; these alternative models was not be compared in this project.

To accomplish the goal of this project, research was done on data mining and prediction modelling techniques and tools in context of the HHP application. Data mining and prediction modelling was performed experimentally on the competition data. Data mining refers to the process of analysing data from different perspectives and summarizing it into useful information.

Shown in Figure 1 is the roadmap followed to achieve the specified goals and objectives for this project. The focus of this project is to assist the research team to make informed decisions when choosing methods, techniques, technologies and procedures when developing an algorithm for the HHP competition. The data input for experimentation are taken from a data warehouse, which is populated by data that was extracted, transformed and loaded (ETL) from data downloaded from the Kaggle website. The knowledge input was cumulated from literature and is tested in the experimentation phase.

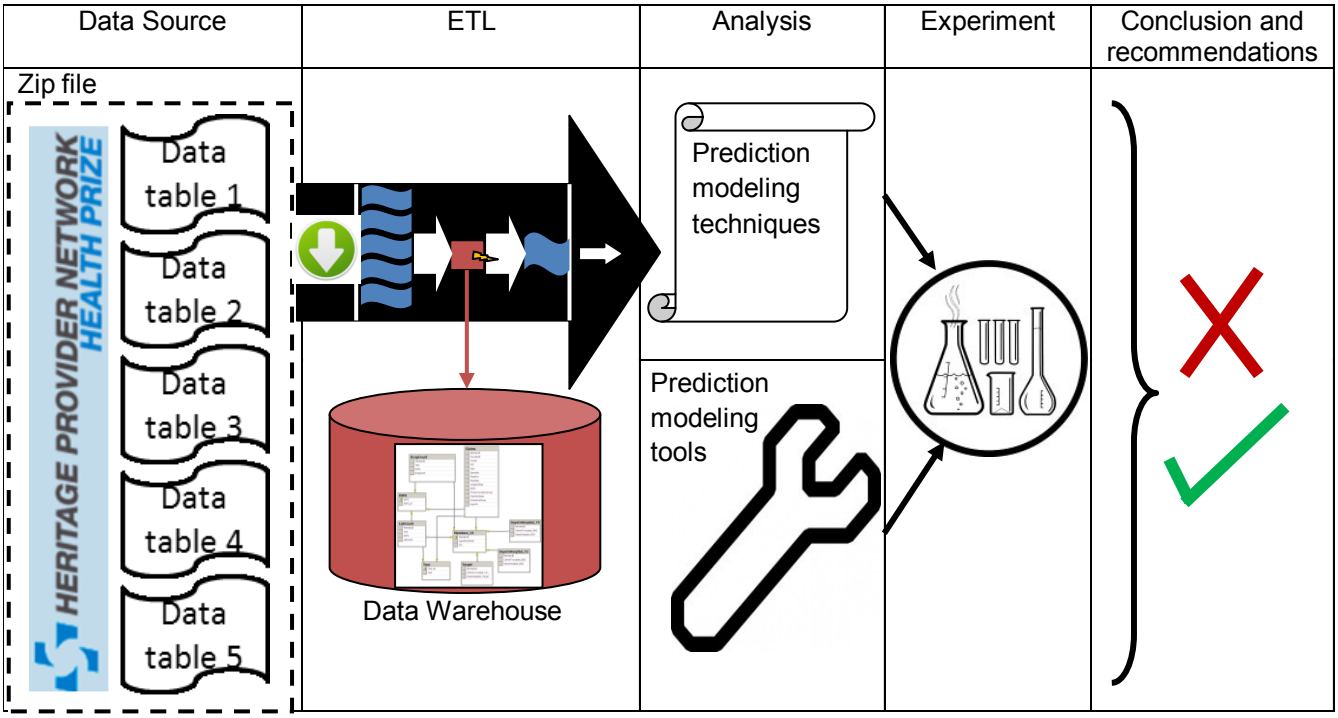

Figure 1: Roadmap to pave the way for hospitalisation prediction modelling

2 DATA AND DATA HANDLING

The problem at hand relies heavily on the handling, manipulation, analyses and interpretation of the data while still keeping data accuracy and integrity. It is therefore primarily important to first understand the available input data. Section 0 attempts to gain a good understanding of the data and to find ways to handle and manipulate data in such a way that a prediction model can be built on it while still preserving data integrity.

2.1 THE DATA

The Heritage Health Prize (HHP) data was received from the Heritage Health Provider Network (Figure 2). It is authentic data although some distortion occurred when member's identities were hidden by the competition organisers.

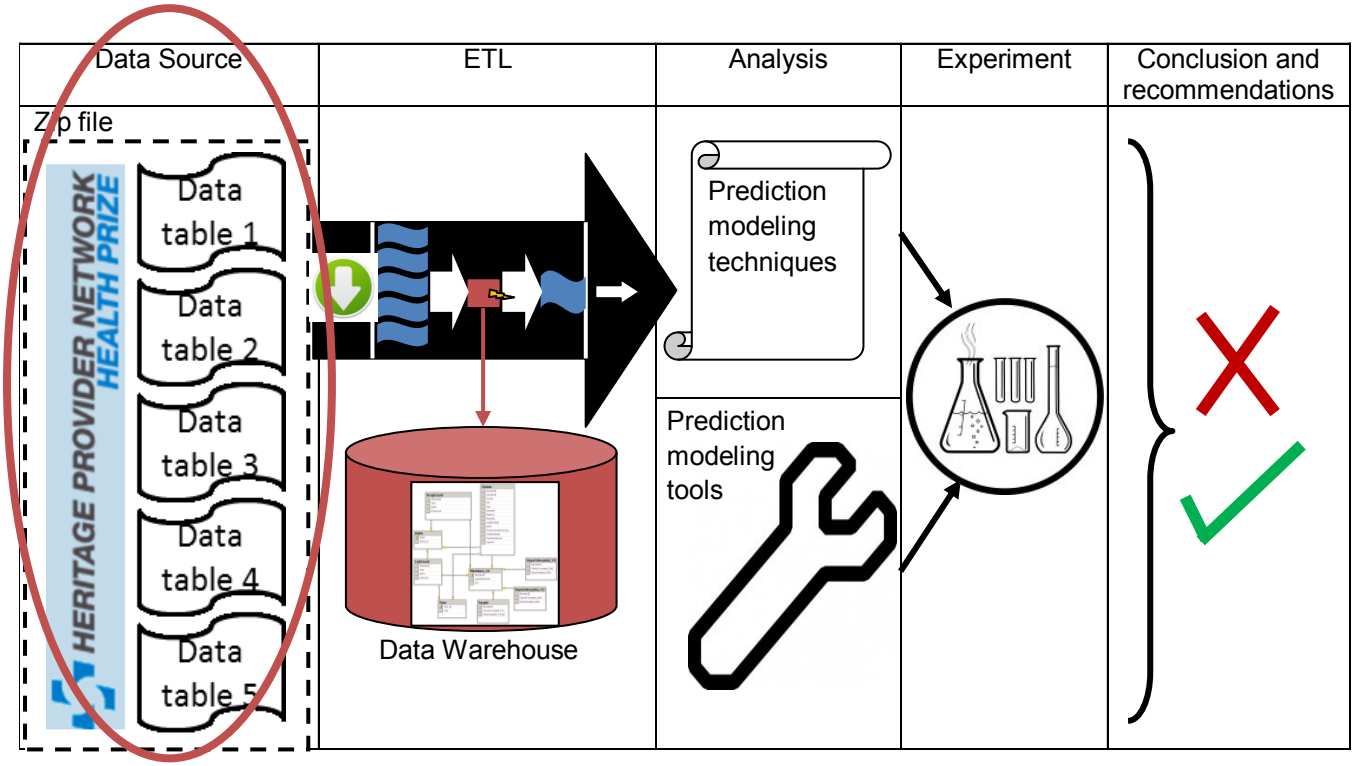

Figure 2: Roadmap to pave the way for hospitalisation prediction modelling - the data source

The data consisted of the following elements:

- General information about *members* who are part of the Heritage Health Provider Network health insurance company.
- Information about the *claims* made by members every year.
- Information about the amount of *days that members spent in hospital* every year.
- Information about *drug prescriptions* claimed by members.
- Information about *lab tests* claimed by members.
- *Metadata* to describe codes used in certain data fields.

A data dictionary is presented in Table 1, as different types of variables can be expected in health insurance claims and hospitalisation data. Firstly, the most general type is continuous numeric variables. Many data modelling techniques can only use categorical variables, and in these cases, continuous numeric variables can be converted into discrete numbers (also called discretizing). For example, if the numeric continuous variables range is 1 to 100, these variables can be discretized by dividing them into bins (sub-ranges) of four: 0-25, 26-50, 51-75, 76-100 (Nisbet, Elders, Miner, 2009: 58). Another kind of variable is categorical variables, which can be either nominal or ordinal. Examples of these different kinds of variables can be seen in Figure 3. Column DSFS_ID is an example of an ordinal categorical variable, column Procedure Group_ID is an example of a nominal categorical variable and column PayDelay is an example of a continuous numerical variable. The PPAA should be able to accommodate all these variables.

2.2 ISSUES WITH DATA INTERPRETATION

It is important to note that data sets are often riddled with ambiguities and uncertainties. Examples found in the HHP data set are listed below, and can be expected in similar data recorded in the hospital environment:

- Each member specifies a primary care physician. This could be one doctor or a group of doctors.
- A similar situation is found in MemberID, as a MemberID can represent either one person or a family. That is why, in some cases, it has been found that a male member might have a condition of pregnancy (the person who is pregnant is simply a dependent of the main member who happened to be male) (Howard, 2011).
- Where the length of hospital stay (LengthOfStay) column is blank, it is assumed that patients stayed less than a whole day (Igor, 2011).
- The amount of drugs consumed in a specified year (DrugCount) is described in the data dictionary as the "Count of unique prescription drugs filled by DSFS." This is more easily understood by means of an example: if two Paracetamol prescriptions and one Ponstan prescription are claimed in the same claim time frame (DSFS), then it will count as two unique types of drugs and will display as a 2 in the DrugCount column (Arbukcle, 2011).

Table 1: Data dictionary for HHP data (Heritage Provider Network Health Prize, 2011)

Variable	Description
MemberID, ProviderID, Vendor	Member, provider and vendor pseudonym.
AgeAtFirstClaim	Age in years at the time of the first claim's date of service computed from.
Sex	Biological sex of member: M = Male; F=Female.
PCP	Primary care physician pseudonym.
Year	Year in which the claim was made: Y1; Y2; Y3.
Speciality	Generalized specialty.
PlaceSvc	Generalized place of service
PayDelay	Number of days delay between the date of service and date of payment
LengthOf Stay	Length of stay (discharge date – admission date + 1)
DSFS	Days since first claim, computed from the first claim for that member for each year
Primary Condition Group	Broad diagnostic categories, based on the relative similarity of diseases and mortality rates, that generalize the primary diagnosis codes.
Charlson Index	A measure of the affect diseases have on overall illness, grouped by significance, that generalizes additional diagnoses.
Procedure Group	Broad categories of procedures.
SupLOS	Indicates if the NULL value for the LengthOfStay variable is due to suppression done during the de-identification process.
DrugCount	Count of unique prescription drugs filled by DSFS. No count is provided if prescriptions were filled before DSFS zero.
LabCount	Count of unique laboratory and pathology tests by DSFS.
DaysInHospital_Y2, DaysInHospital_Y3	Days in hospital Y2, Y3
ClaimedTruncated	Members with truncated claims in the year prior to the main outcome are assigned a value of 1, and 0 otherwise. If truncation is indicated (in years 2 and 3) it means that a certain member had more that 43 claims for specified year. Truncation is used as part of the suppression done during the de-identification process.

DSFS_ID
0- 1 month
1- 2 months
10-11 months
11-12 months
2- 3 months
3- 4 months
4- 5 months
5- 6 months
6- 7 months
7- 8 months
8- 9 months
9-10 months

ProcedureGroup_ID	Description
ANES	Anesthesia
EM	Evaluation and Management
MED	Medicine
PL	Pathology and Laboratory
RAD	Radiology
SAS	Surgery-Auditory System
SCS	Surgery-Cardiovascular System
SDS	Surgery-Digestive System
SEOA	Surgery-Eye and Ocular Adnexa
SGS	Surgery-Genital System
SIS	Surgery-Integumentary System
SMCD	Surgery-Maternity Care and Delivery

PayDelay
28
50
14
24
27
25
162
29
42
56
51
22

Figure 3: Examples of different types of variables to be found in health insurance claims and hospitalisation data.

2.3 DATA HANDLING

For the purposes of this project, data handling refers to the technical side of data warehousing and can be described in terms of the extract, transform and load (ETL) procedures. According to Aronson, Liang, Sharda and Turban (2005: 224-226) ETL is an integral part of any data-centric project and can often consume up to 70 % of the time in such a project . If done properly ETL will prevent a garbage-in-garbage-out situation for the life cycle of the project. Hence, this part takes up a considerable part of this project.

ETL is a three-stage process that enables integration and analysis of the data from different sources and in a variety of formats; it is visually described in

Figure 4. For typical hospitalisation and claims data, such as data used in the HHP case study, ETL steps were followed.

2.3.1 EXTRACTION

In this step data is collected from a database. Seven separate Comma Separated Values (.csv) files were downloaded from the Heritage Health Kaggle website as a zip file.

2.3.2 TRANSFORMATION

Transformation refers to the modification process the extracted data must undergo before it can be loaded into the target repository.

The transformation process includes using rules (formatting and cleaning), lookup tales (replacing of data into appropriate forms) and combination (integration) of data tables to convert the data into a desired form (Aronson, 2005: 224). Transformation can be performed in several different ways for the HHP applications, to illustrate this different alternatives were sketched and can be seen in Section 2.3.4. For the purposes of this study exploratory transformation was performed. Finding for this scenario was therefore discussed in more detail than the other recommended and alternative alternatives in Section 2.3.4.

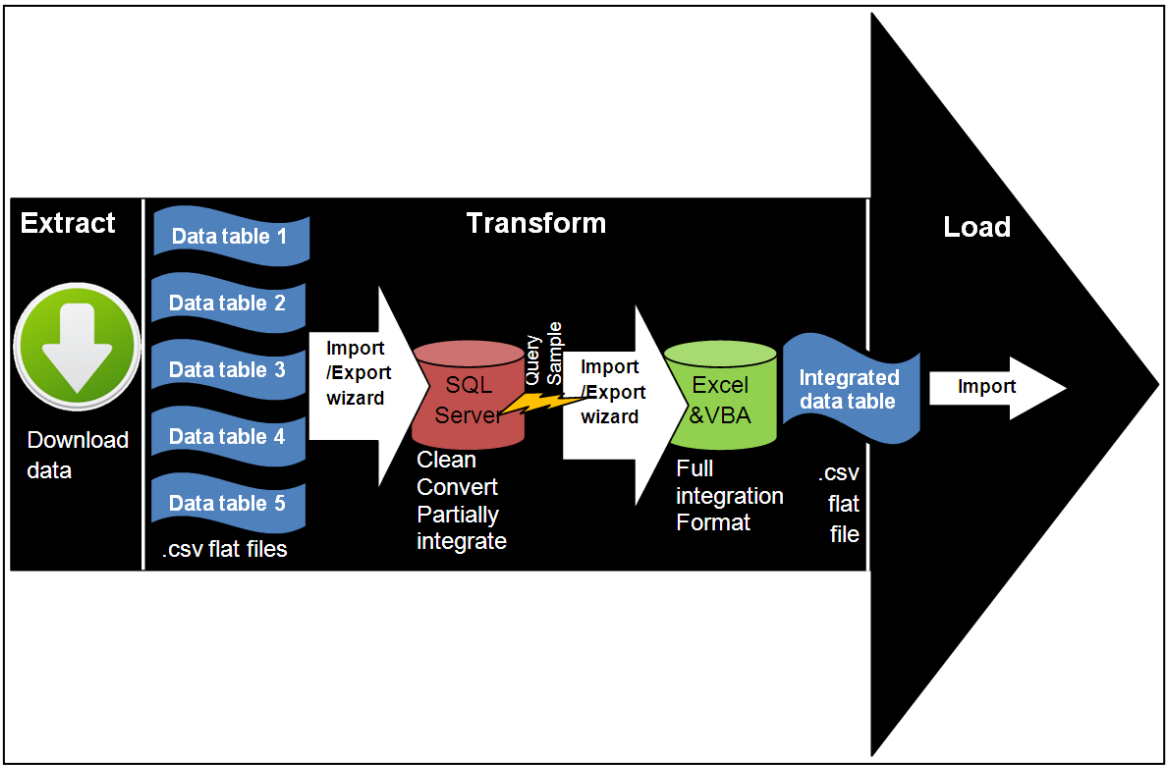

Figure 4: Roadmap to pave the way for hospitalisation prediction modelling - data warehousing and ETL

The Heritage Health Prize data required reformatting, cleaning (to remove certain duplicates and blanks) and the integration of tables.

1. Firstly, Microsoft SQL's Import/Export Wizard was used to import downloaded .csv files into Microsoft SQL Server (SQL). With the of SQL Import/Export wizard each column's data type could be specified.
2. Another important feature of the extraction process involves the parsing of extracting data. According to Caserta (2004).parsing is a procedure by which there is checked whether extracted data meets a specified structure. If not, the applicable data record is either deleted or stored in a separate table.

As little as possible parsing was done, to ensure that data was not removed that could be useful at a later stage. Necessary parsing involved the removal of records in the Claims table that had blank values on the primary keys. These blank values give errors in SQL Server, because they prevent accurate integration of tables (tables cannot be joined on blank values). Primary keys columns include: DSFS, Year and MemberID.

3. Text fields were converted to numeric values for analyses in programs like Matlab to be functional. An example of this conversion can be seen in Table 2 where "0-1 month", which is text data, is substituted with "0.5". This can be understood as the average Days from the first Claim is 0.5 (months) which is numeric. Converting data as seen in Table 2 is not always necessary when using programs with functionality such as that of Statistica, because this program has built-in functions that can manipulate text fields to either use them directly in the analysis or extract the numeric parts.
4. Next primary entity tables were added, namely: tblYear, tblDSFS. The original seven tables are: tblMembers (also a primary entity), tblDaysInHosp_Y2, tblDaysInHosp_Y3, tblClaims, tblTarget (a table of members on whom the algorithm will be applied), tblDrugCount (the count of drugs claimed by member), tblLabCount (the count of lab tests claimed by members). In Table 2 it can be seen that DSFS is the primary key and DSFS_ID is the data that was actually displayed when queries were run.

Table 2: Example of converting text data into numerical data.

	DSFS	DSFS_ID
	0- 1 month	0.5
	1- 2 months	1.5
	10-11 months	10.5
	11-12 months	11.5
	2- 3 months	2.5
	3- 4 months	3.5
	4- 5 months	4.5
	5- 6 months	5.5
	6- 7 months	6.5
	7- 8 months	7.5
	8- 9 months	8.5
	9-10 months	9.5

5. Database tables could now be integrated, and queries could be run for specified cases. A SQL Extended entity relationship diagram (EERD) of the Heritage Health Prize Data can be seen in Figure 6. One to many relationships are indicated as well as the three primary key columns: tblYear, tblDSFS and MemberID.

6. Once relationships between all tables were established, queries were constructed to extract specified combinations of tables. These queries were executed and the results were then saved as .csv files which could then be imported into Excel. An example of such a query was the combination of tables DaysInHospital_Y2 and Member_V1. An example to illustrate the SQL query code is seen in Figure 8 and an accompanying resulting query output Figure 9. It is evident from the code that, the command, DISTINCT was used to display DaysInHospital (DaysInHospital is measured in days per year) only once per MemberID, because the end goal of the combined data was to create a master sheet that contained the summarized information per MemberID. Note DISTINCT was only used for data that was already on a member level. Summing and counting had to be done for data that had to be converted from claims to member level.

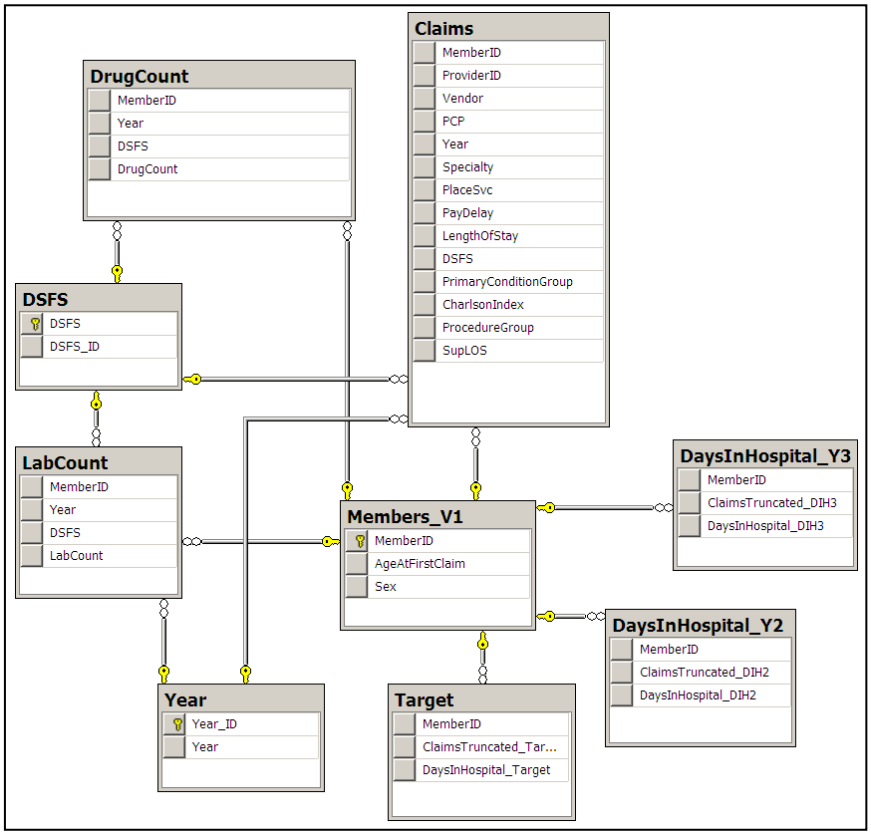

Figure 6: Extended Entity Relationship Diagram (performed in SQL Server)

```
SELECT DISTINCT
            dbo.DaysInHospital_Y2.MemberID, dbo.Members_V1.AgeAtFirstClaim, dbo.Members_V1.Sex,
dbo.DaysInHospital_Y2.ClaimsTruncated_DIH2,
            dbo.DaysInHospital_Y2.DaysInHospital_DIH2
FROM      dbo.DaysInHospital_Y2 INNER JOIN
            dbo.Members_V1 ON dbo.DaysInHospital_Y2.MemberID = dbo.Members_V1.MemberID INNER JOIN
            dbo.Claims ON dbo.Members_V1.MemberID = dbo.Claims.MemberID INNER JOIN
            dbo.DSFS ON dbo.Claims.DSFS = dbo.DSFS.DSFS INNER JOIN
            dbo.Year ON dbo.Claims.Year = dbo.Year.Year_ID
WHERE     (dbo.Year.Year_ID = N'Y2')
```

Figure 7: Example of SQL query code

	MemberID	AgeAtFirstClaim	Sex	ClaimsTruncated_DIH2	DaysInHospital_DIH2
▶	10001471	80+	F	0	0
	10002388	80+		1	0
	10004244	80+	F	0	0
	10004817	40-49	F	0	0
	10008724	50-59	M	0	0
	10009391	30-39	M	0	0

Figure 8: Resulting query output.

7. After appropriate queries's results had been saved as .csv files, these files were imported into Excel for further summarizing. The summarizing done in Excel consisted of the following: Sorting and summing the Drug and LabCount per year per member, converting categorical variables (Specialty, PlaceSvc, PrimaryConditionGroup and ProcedureGroup) to ordinal variables - in statistical terms this is called a dummy variable. Last mentioned was accomplished by counting the occurrence of each type of categorical variable per member per year, for example counting the amount of times claims were made for "Surgery" (which is a Speciality) for member 10001471 for year 2.
8. Once summarizing had been completed on one master sheet this master sheet was once again saved as a .csv file and finally uploaded into the data analysis tool where statistical analysis could be performed.

Microsoft Access was also tested for extraction, but because the import/export wizard outputted errors indicating incompatibility between the versions of the wizard and Access, this tool could not be used. The alternative would have been to copy-paste the values, but unfortunately a clipboard is limited to 65,000 records, making this approach unpractical. SQL Server could handle import and export data files containing more that 2 million records. The import/export wizard was also attempted for Access, but the wizard. For this reason it is recommended that SQL Server should be used by the research team for the Heritage Health Prize application.

An alternative approach could be to extract .csv files directly into programs like Statistica, SAS-Enterprise Miner or SPSS Clementine which also has the functionality to provide in-database access to data via low-level interfaces (Nisbet *et al.,* 2009). Last mentioned alternatives are not recommended if complex relationships are present, which is the case for the HHP application. Using alternative programs have the penalty of having to learn to use these tools whereas the research team already has basic knowledge of SQL Server.

Perl has been discussed on the Heritage Health Prize competition's online forums and is another tool to consider for this application (Howard, 2011).

Data warehouse business rules: Business rules will be described briefly in these succeeding paragraphs. For the HHP data warehouse include aspects like summarisation, standardisation and calculation rules used.

Firstly, summarising was done on a member level, because predictions have to be made per member. Excel was used to do summarising to on a more basic level, to figure out what the summarising process requires and to test different tools.

Note there in the order of 2.6 million records and an Excel sheet can only process 1.04 million records. In future it is recommended that summarising be done in SQL. SQL has the ability to handle these amounts of data, and with the pointers that was discussed next, the developer (if they are familiar with SQL coding) developing such a summarising query should find it quite straight forward. Data was firstly summarised per year and typically contained the following columns:

- Member information: MemberID (numeric), Sex (categorical variable, e.g. 0 for female, 1 for male) AgeAtFirstClaim (averaged, ordinal variable, e.g. 10-19 becomes 15).
- ClaimsTruncated (binary) and DaysInHospital (Continuous) for that specific year was added to membership information. This was done with a SQL query.
- Each type of Speciality, PlaceSvc, PrimaryConditionGroup and ProcedureGroup was listed in a separate column each and a count was done per member, e.g. how many times did member x claim for the Speciality *Emergency* in year y. This has already been mentioned that these variables were converted to counted dummy variables. By counting them it makes each of these variables a continuous variable.
- The continuous variables DrugCount and LabCount were summed per member per specified year, because there were multiple Drug and Lab claims per year per member.
- CharlsonIndex (ordinal variable) was also converted into a counted dummy variable for each occurrence: 0,1.5,3.5 and 5.5.
- Coded variables like ProdiverID, Principle Care Provider (PCP) and similar coded variables were excluded because they had an extremely high variance and would most probably be useless when considered for a predictor variable. The available predictor variables are weak, these excluded variables should once again be considered, but for scoping purposes these variables were left out.

- LengthOfStay (LOS) and Suppressed LengthOfStay (SupLOS) were also excluded, because information gathered on some of the HHP forums suggest that these two variables are derived items of DaysInHospital (Barnett, 2011).

2.3.3 LOADING

Loading was done by firstly importing a data sample into Excel (for preliminary analysis) to help understand the data better, followed by converting the bulk of the data to .csv, to be imported into statistical analysis software which was discussed in Section 4. Different technologies were tested for the ETL process and a summary of the findings for the tested technologies can be seen in Table 3.

Based on the limitations of Microsoft Access and Microsoft Excel in terms of the amount of records it can process the conclusion can be made that these two programs are probably too basic for this applications, and Statistica or SQL Server should rather be considered. The penalty to be paid for last mentioned technologies are that SQL Query language should be learnt for SQL and Visual Basic programming language and Statistica spread sheet functioning for Statistica.

2.3.4 ETL ALTERNATIVES

In the ETL process different alternatives were considered to approach a difficult-to-manage data set such as the HHP application. Alternatives are visually displayed in a series of flow charts evolving from a most basic scenario to the most optimal for the current situation.

In the first scenario, seen in Figure 9, a sample of the data was imported into Excel where integration, conversion, formatting and cleaning was done with the help of VBA. This is a useful approach for a sample of the data, but has a penalty to be paid in terms of the labour intensiveness of VBA coding required in doing this integration, conversion, formatting and cleaning as well as not being able to use the whole data set, because of capacity limitations.

In scenario 2, seen in Figure 10, SQL was employed to assist in doing the basic integration such as linking data tables that are already summarised per member and running queries to extract data per year. Scenario 2 is a useful approach, but has the limitation that only a small sample could be manipulated in Excel, because of Excel's capacity limitation.

Table 3: Technology ETL decision matrix.

Criteria for use in the HHP application		Microsoft Access	SQL Server	Microsoft Excel and VBA	Statistica
General information	**User-friendly?**	Yes	No	Moderate	Moderate
	Available?	Yes	Yes	Yes	No
	Appropriate for ETL	Basic data cleaning and integration	Advanced data cleaning and integration	Basic data cleaning	Data cleaning, basic integration
	Syntax easily used by research team	Easy, mostly menu driven, although SQL Query language can be used	SQL Query language is used which is difficult to use	VBA programming language is easy to use	Moderate, data management mostly menu driven, but VB is available.
Performance with ETL	**Extract successfully**	Works using Access's own import wizard	Works using SQL's own import wizard	Works using Excel's own import wizard	Works using Statistica's own import wizard
	Transform successfully	User friendly user interface	Not user friendly	Is not equipped to do data integration and hard coding was done in VBA	Can handle the requirements of the data
		Not as strict as SQL and often allows mistakes	Consistent and keeps data integrity	Struggles to handle the amount of data	
				Mistakes were easily made because of hard coding	
	Load successfully	Exports using clipboard which is not sufficiently large for data set size	Works using SQL's own export wizard	Works using Excel's own export wizard	Works using Statistica's own export wizard

With experimentation it was also found that Statistica, the tool used in Section 4 to do prediction modelling with, is a useful tool for integration, because it has built in functions to assist data integration and it also linked to a Visual Basic application with which more complex and/or customised integration can be achieved. Last mentioned scenario can be seen in Figure 11. SAS-Enterprise Miner or SPSS Clementine also has similar functionality according to Nisbet et al. (2009).

As mentioned these tools are not recommended for complex interactions, as is the case for the HHP data set. These tools can do integration for this application adequately, but complew summarisation Is required for which SQL Server is a better tool to use.

Finally, the recommended approach, scenario 4 seen in Figure 12, was employed using SQL alone for data integration. SQL Server is by far the most robust and fool-proof tool to use for ETL in this

application, if it is implemented correctly. It does come with a penalty of having to master the SQL Query language, which is a language the research team only has basic to no knowledge of. It is advised that the research team outsource the data integration section or acquire a team member who is a SQL expert.

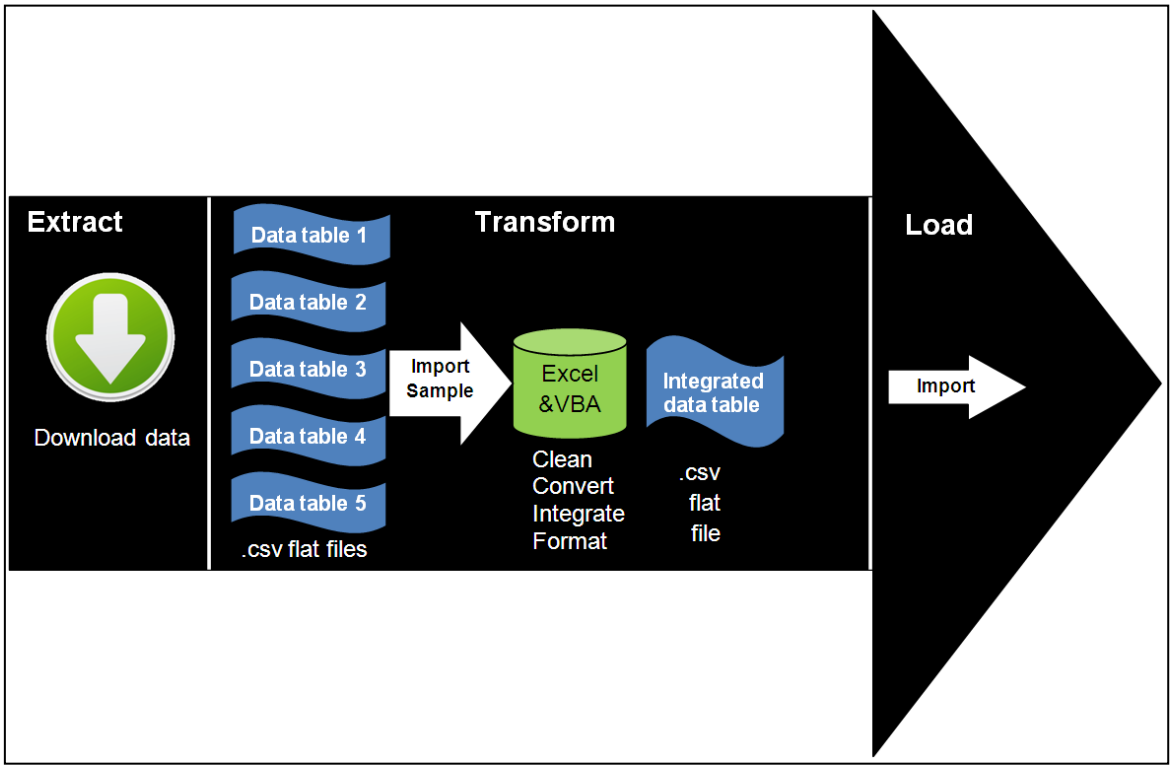

Figure 9: Basic ETL using Excel

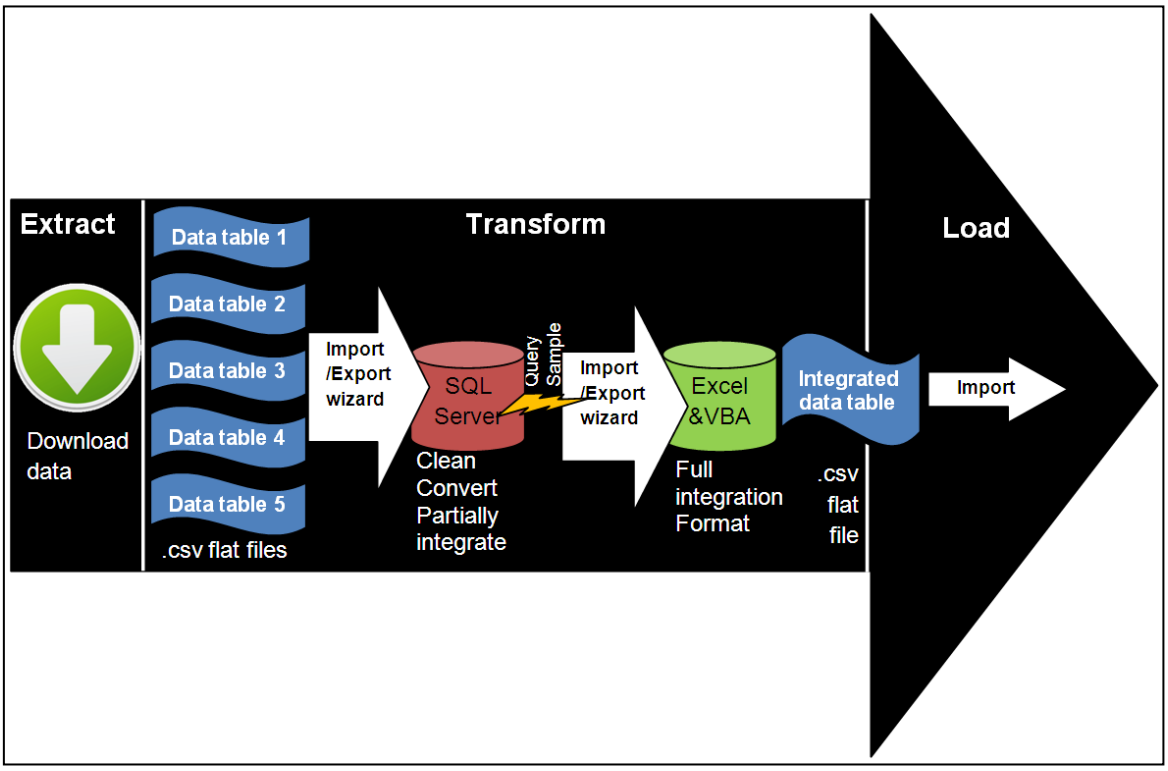

Figure 10: Hybrid ETL using SQL Server and Excel

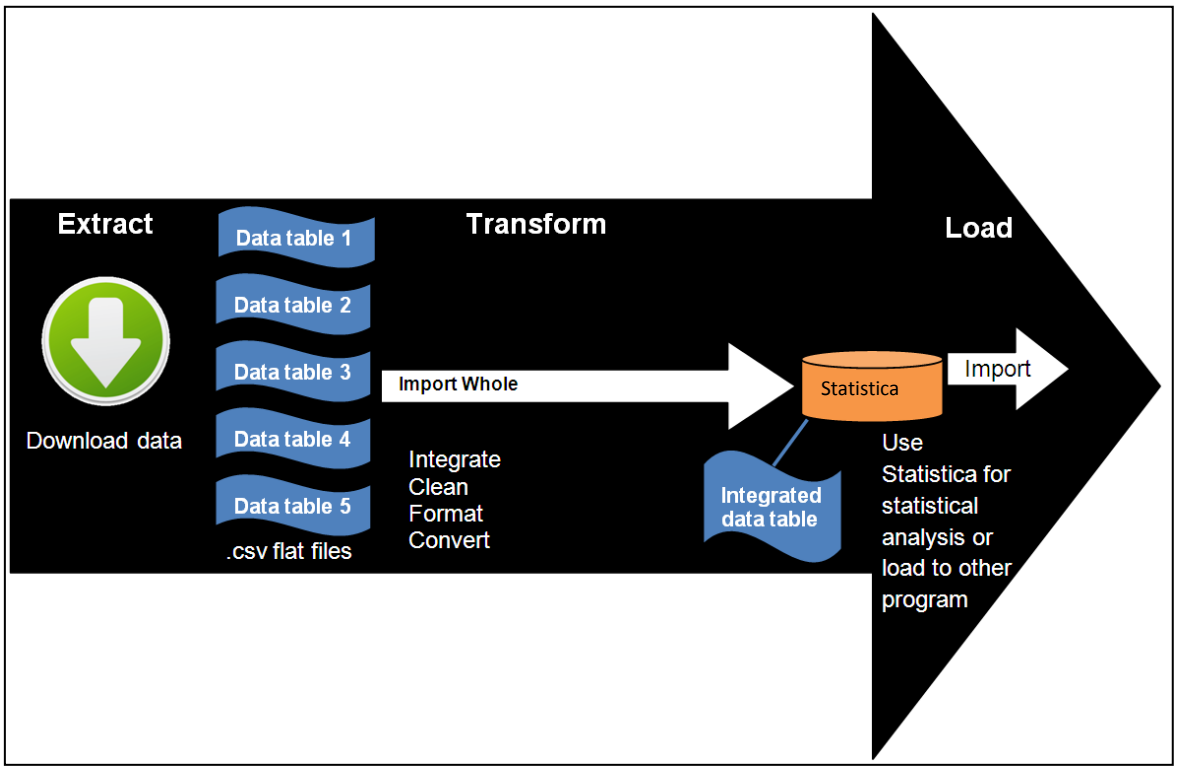

Figure 11: ETL using only

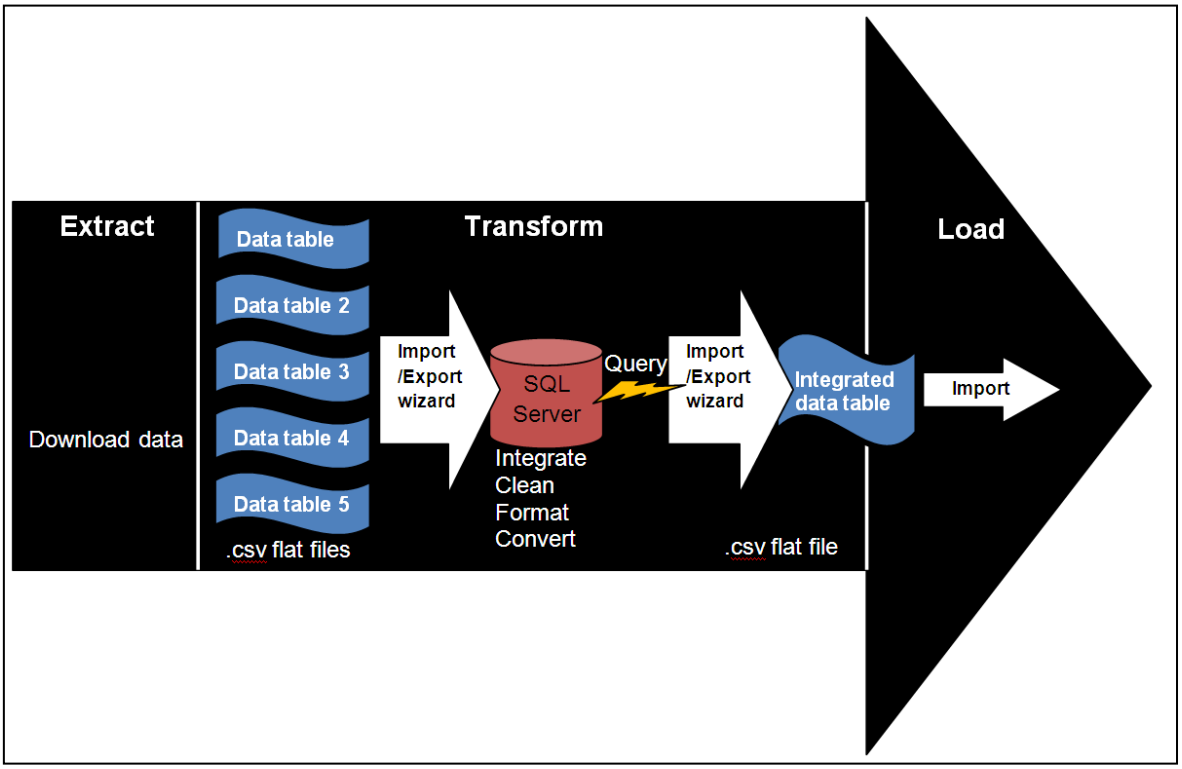

Figure 12: Most optimal ETL using only SQL Server

ANALYTICAL HIERARCHICAL PROCESS (AHP) FOR ETL ALTERNATIVES

The analytical hierarchical process was used to determine which one of the four alternatives is the most optimal. When multiple objectives are important to a decision maker, it may be difficult to choose between alternatives, as is the case for the decision to be made by the research team concerning the ETL alternatives. The process followed to perform the AHP is similar to Winston's (2004: 785-795) recommendation. The AHP was performed in terms of how well alternatives meet the following criteria:

Criteria 1	Be able to clean data easily
Criteria 2	Be able to format data easily
Criteria 3	Be able to integrate data easily
Criteria 4	Be able to convert data easily
Criteria 5	Have enough memory capacity to function properly
Criteria 6	Be able to import and export easily
Criteria 7	Be able to keep integrity
Criteria 8	The programming language must be easy to use
Criteria 9	Must be available to research team
Criteria 10	Be able t summarise data to per member per year effectively

Step 1 - Obtaining weights for each Criteria: The weight for each objective was determined by comparing criterias in a pair-wise comparison as seen in the screenshot in Table 4. Each criteria was compared to the other criterias with the constraint that

$$aij = \frac{1}{aji} \qquad \textbf{(2.1)}$$

Table 4: Pairwise comparison

Objective	Clean	Format	Integrate	Convert	Memory	Export	Integrity	Code	Availablity
Clean	1.50	0.30	0.20	0.15	0.20	0.20	0.20	0.20	0.15
Format	6.50	0.90	0.80	0.90	0.80	1.50	0.80	1.00	1.00
Integrate	8.00	1.30	0.90	1.30	1.00	1.80	1.00	1.30	1.30
Convert	8.00	1.10	0.80	1.00	0.80	1.50	0.80	1.00	1.00
Memory	8.50	1.30	1.00	1.30	1.00	1.80	1.00	1.30	0.20
Export	5.00	0.60	0.60	0.80	0.60	0.90	0.60	0.71	0.80
Integrity	9.00	1.30	1.00	1.30	1.00	1.80	1.00	1.30	1.29
Code	6.50	1.00	0.80	1.00	0.80	1.50	0.80	1.00	1.00
Availablity	7.00	1.00	0.80	0.90	0.80	1.50	0.80	1.00	0.90

Once the pair-wise comparison (call it matrix A) was done the normalised matrix was determined by dividing each entry in column i of matrix A by the sum of the entries in column i. This result can be seen in Table 5.

Table 5: Normalised matrix A.

Anorm=								
0.025	0.034	0.029	0.017	0.029	0.016	0.029	0.023	0.020
0.108	0.102	0.116	0.104	0.114	0.120	0.114	0.113	0.131
0.133	0.148	0.130	0.150	0.143	0.144	0.143	0.147	0.170
0.133	0.125	0.116	0.116	0.114	0.120	0.114	0.113	0.131
0.142	0.148	0.145	0.150	0.143	0.144	0.143	0.147	0.026
0.083	0.068	0.087	0.092	0.086	0.072	0.086	0.081	0.105
0.150	0.148	0.145	0.150	0.143	0.144	0.143	0.147	0.168
0.108	0.114	0.116	0.116	0.114	0.120	0.114	0.113	0.131
0.117	0.114	0.116	0.104	0.114	0.120	0.114	0.113	0.118

Finally, to determine the weight for each criteria, the average of row j can be calculated. The result is shown in Table 6.

Table 6: Criteria weights

W=	
0.025	
0.114	
0.145	
0.120	
0.132	
0.084	
0.149	
0.116	
0.114	

To determine how consistent weights have been allocated, a consistency index can be determined with:

$$\frac{1}{n}\sum_{i=1}^{i=n}\frac{ith\ entry\ of\ Aw}{ith\ entry\ of\ w} \qquad (2.2)$$

With: N is the number of criterias

Aw is the product of the normalised matrix and the criteria weights matrix.

W is the criteria weights

The AHP scored a consistency index of 0.05447, which is smaller than the recommended CI for 9 criterias, which is 1.45. The $\frac{CI}{RI}$ ratio is further indication that this AHP is consistent, it scores 0.038.

Step 2 - Applying AHP: Now that the weights for each criteria has been obtained, the scores for each scenario can be determined and the corresponding results for each scenario can be seen in Tables 7 to 10. Table 11 indicates a summary of all the scenario weights.

Table 7: Scenario 1 - Excel and VBA

Criteria	Importance score out of 10	Weight	Reason for score
Clean	5	0.106	Effective although cannot clean all the data, too much
Format	2	0.043	Does automatic formatting that causes problems
Integrate	3	0.064	Very labour intensive, can't do big data sets
Convert	3	0.064	Very labour intensive, can't do big data sets
Memory	1	0.021	Excel doesn't have enough memory space for the application
Export	9	0.191	Works well and easy
Integrity	5	0.106	More manual, so mistakes are made more easily
Code	9	0.191	VBA is known to research team
Available	10	0.213	Available to research team

Table 8: Scenario 2 – Excel, VBA and SQL.

Criteria	Importance score out of 10	Weight	Reason for score
Clean	9	0.150	SQL cleans data very well
Format	5	0.083	SQL can pre-format data so that Excel does not give problems
Integrate	5	0.083	A bit labour intensive, certain integration tasks can easily be done in SQL, but summarising is mostly done in VBA which is labour intensive.
Convert	9	0.150	SQL what very effective for conversion
Memory	1	0.017	Excel doesn't have enough memory space for the application
Export	8	0.133	Excel's is easy. SQL works well, but isn't always that easy, but very thorough
Integrity	6	0.100	SQL keeps integrity well, but more mistakes can be made in Excel
Code	7	0.117	VBA is known to the research team, but SQL is not
Available	10	0.167	Available to research team

Table 9: Scenario 3 - Statistica

Criteria	Importance score out of 10	Weight	Reason for score
Clean	9	0.145	Cleans well
Format	9	0.145	Formats can be specified, and there are no automatic formats that could cause problems
Integrate	7	0.113	Sufficient integration and summarisation can be done
Convert	9	0.145	Conversion is very easy and fast
Memory	7	0.113	Has sufficient memory space
Export	8	0.129	Sufficient Import/export facilities
Integrity	9	0.145	Easy to determine if integrity is violated, with graphs and plots facilities
Code	3	0.048	Statistica is a new language to the research team
Available	1	0.016	Not easily available to the research team

Table 10: Scenario 4 - SQL alone

Criteria	Importance score out of 10	Weight	Reason for score
Clean	9	0.127	SQL cleans data thoroughly
Format	9	0.127	Formats can be specified, and there are no automatic formats that could cause problems
Integrate	9	0.127	High level integration and summarisation can be done
Convert	8	0.113	Conversion is easily done
Memory	7	0.099	Has sufficient memory space
Export	7	0.099	Sufficient Import/export facilities: takes some time
Integrity	7	0.099	Can sufficiently determine if integrity is kept b using queries, not as efficient as plots
Code	5	0.070	Difficult language, even though research team has basic knowledge therein
Available	10	0.141	Available to research team

Table 11: Summary of scenario weights

Objective	Scenario 1	Scenario 2	Scenario 3	Scenario 4
Clean	0.106	0.150	0.145	0.127
Format	0.043	0.083	0.145	0.127
Integrate	0.064	0.083	0.113	0.127
Convert	0.064	0.150	0.145	0.113
Memory	0.021	0.017	0.113	0.099
Export	0.191	0.133	0.129	0.099
Integrity	0.106	0.100	0.145	0.099
Code	0.191	0.117	0.048	0.070
Available	0.213	0.167	0.016	0.141

Finally, to determine each scenario's overall score the criteria weights determined in Table 6 are multipplied with the scenario scores shown in Tables 7 to 10 using the SUMPRODUCT() function in Excel. The resluly is shown in below.

Scenario 1 overall score= 0.10584
Scenario 2 overall score= 0.10431
Scenario 3 overall score= 0.10883
Scenario 4 overall score= 0.10984

The scenario scores indicate that scenario 4 is the best alternative to follow when doing ETL.

3 PREDICTION TECHNIQUES

Now that the data is ready to be use, the next step on the road (Figure 13) is to determine which prediction modelling techniques can be used for this application.

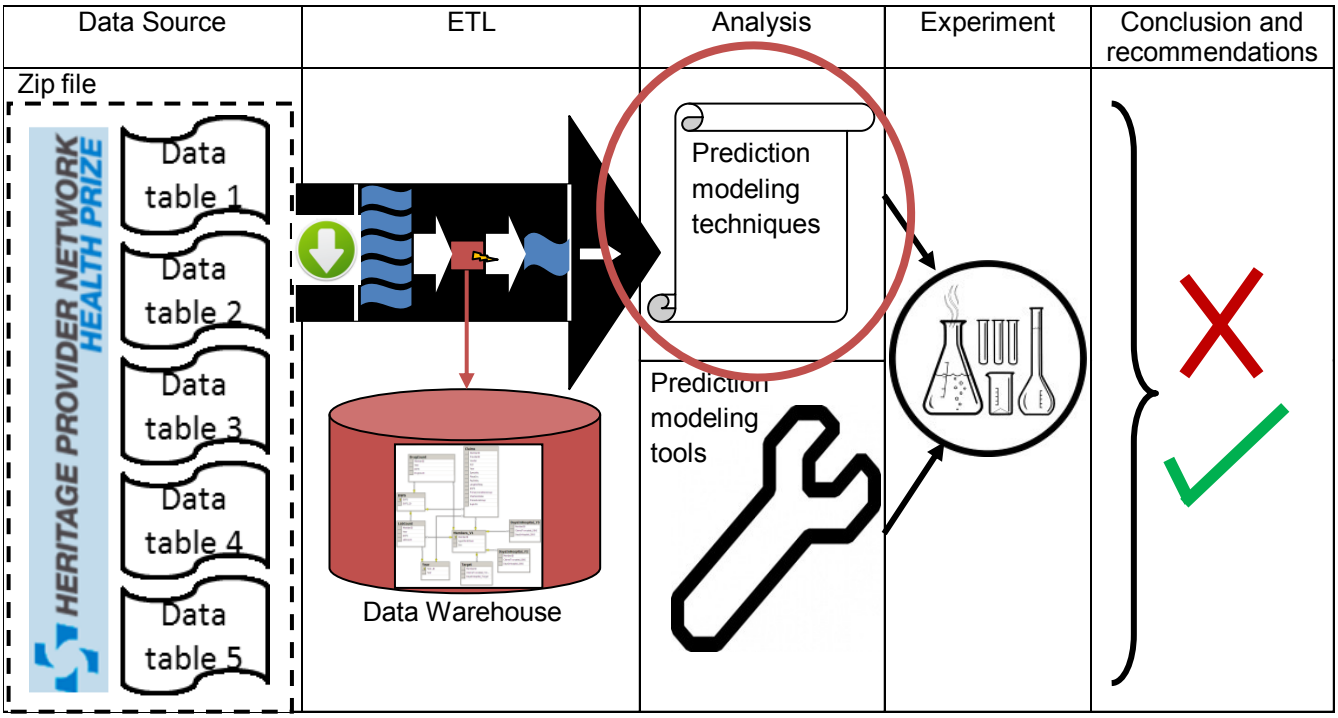

Figure 13: Roadmap to pave the way for hospitalisation prediction modeling - prediction modelling techniques

The task of the appropriate contender technique is to use the claims and member data for year x-1 and the days in hospital count for year x to build a prediction model that will be used to predict for year x+1.

There are certain characteristics that a prediction modelling contender technique needs to exhibit before being considered viable for the application in the HHP case study. These include:

1. Multivariate modelling approach: This approach encompasses the analysis of more than one predictor variable. The input data in this study consists of an n×p (n rows by p columns) rectangular array of real numbers. Claims are summarised per member and the data set then consists of a record per member, containing characteristics of such a member. Each of the n members are thus characterised with respect to p variables. The values of the p variables may be either quantitative or a numeric code for a classification scheme (Jobson, 1991). All the contender techniques were chosen on the basis that they can handle multiple predictor variables.

2. Linearity and non-linearity: Contender techniques should be able to handle linear as well as non-linear data, because variables are distributed linearly as well as non-linearly.
3. Different variable types: As described in the previous section, the dependent variables consist of continuous variables, but the predictor variables can consist of continuous, binary, ordinal or nominal categorical variables. The contender technique should therefore, be able to handle such variations in variable types.
4. Robust: This refers to the contender technique being able to model for different datasets, especially if they contain illogical data and the like (for example data sets with missing values). New data sets are made available by the competition and the technique must be able to model from these new data sets as well.
5. Resistance to over fitting: Over-fitting tends to occur when more parameters than necessary are used to fit a function to a set of data (Steig, 2009) and causes a model to generalize poorly to the new data. However, there are specific and different ways to avoid over-fitting with every technique used and these will be discussed further with each technique description.
6. Comprehensiveness of results: This refers to the ease with which the response output of the technique can be logically understood and interpreted.
7. Compatible with available technologies: It may happen that a certain technique will be able to perform prediction flawlessly in theory, but that in practice, the available technologies are limiting or too complex to use. This is an important aspect to consider in the choice of technique as well as the choice of technology. Considered technologies include: Excel and VBA, Matlab, Statistica, R, SAS and SPSS Clementine. Each tool is briefly discussed in Table 13, in terms of the: degree to which it is open source or menu driven, cost, software capabilities and the known advantages and disadvantages of each.

This study considers four multivariate prediction techniques: Multivariate Adaptive Regression Splines (MARS), Classification and regression trees (CART), Neural Networks and Ensemble Methods.

3.1 REGRESSION MODELING

Since regression is one of the simpler methods available, it is often used as the first analysis. However, basic linear regression will be insufficient as this is a complex data application and some relationships in the dataset could be linear and others non-linear. A technique used to bypass this problem is called Multivariate Adaptive Regression Splines (MARS). MARS is a nonparametric regression technique that makes no assumptions about the underlying functional relationship between the dependent and independent variables (Electronic Statistics Textbook, 2011).

Instead, it adapts a solution to the local region of the data that has similar linear responses. MARS also has a useful characteristic in that it only picks up those predictor variables that make a sizable contribution to the prediction. MARS can also handle multiple dependent variables, although this is not required for this specific application. Outputs of this model will keep only those variables associated with the bases functions that were retained for the final model solution. If no counteract measures are taken, nonparametric models may exhibit a high degree of flexibility that, in many cases, result in over-fitting. A measure to counteract over-fitting in this kind of technique, is called pruning (Electronic Statistics Textbook, 2011) and should be applied if this technique is used.

The basic MARS algorithm assumes that the predictor variables are continuous in nature, although it has been found in practice, that both continuous and categorical predictors can be used, and will often yield useful results (Electronic Statistics Textbook, 2011). If there is a continuous dependent variable, the task will be treated as a regression problem (which is the case in the HHP case study). Alternatively, if it is categorical it will be treated as a classification task (Nisbet *et al.*, 2009: 158). MARS is not robust as it is sensitive to missing values and outliers (Brookes&Kolyshkina, 2002). Missing values should not be a problem in the HHP case study as the data set is large and missing value records can simply be deleted.

3.2 CLASSIFICATION AND REGRESSION TREES (CART)

The CART methodology is technically known as binary recursive partitioning (Breiman et al, 1984). It is binary because the process of modelling involves dividing a data set into exactly two “nodes”, by asking yes/no questions (Kolyshkina,&Brookes, 2002). Typical questions for this application are, “Is the member male?”, “Is the member in the age group of 0-9?”, “Is the member suffering from cardiac problems?” and so on. Data is recursively partitioned by trees that divide data into more homogeneous sets, with respect to the response variable, than is the case in the initial data set. A tree keeps on growing until it is stopped by a criterion or if splitting is impossible.

CART is nonparametric, nonlinear and can analyse very complex interactions. Modelling variables are not selected in advance, but are picked by the algorithm. This model can use either categorical or continuous independent variables, or a combination of the two. It is also robust enough to handle missing or blank values and data sets with outliers will not negatively affect this model. CART is also said to be simple and easy to use and it can be incorporated into hybrid ensemble models with neural networks (Nisbet *et al.*, 2009: 146). They often reveal simple relationships between only a few variables that could have easily gone unnoticed using other analytic techniques (Electronic Statistics Textbook, 2011). Roman Timofeev (2004) also found CART results to be invariant to monotone transformations of its independent variables.

Some disadvantages of the decision tree models, such as CART, include:

- A small change in the value of an independent variable can sometimes lead to a large change in the predicted response.
- CART also does not capture linear structure effectively. Due to the discrete nature of the CART technique.
- A very large tree can be produced in an attempt to represent very simple linear relationships (Kolyshkina&Brookes, 2002).
- Deciding when to stop splitting trees is a well-known issue when applying CART to real life data, because real life data usually has lots of errors and random noise. An approach that can be used to address this issue is to first put a procedure in place that will stop the generation of new split nodes when improvement of the prediction is very small (Electronic Statistics Textbook, 2011).

CART trees are usually larger than is necessary and then pruned to find the optimal tree. Pruning is accomplished by testing the data set or using cross-validation or V-fold cross-validation methods. Cross-Validation can be done by comparing the tree computed from the training sample to another completely independent test sample. By doing this, one is able to see if most of the splits determined by the analysis of the training sample are essentially based on "random noise". If this is the case, the prediction for the testing sample will be poor (Electronic Statistics Textbook, 2011).

V-fold cross-validation is accomplished by repeating the analysis many times with different randomly selected samples from the data, for every tree size (starting at the root of the tree, and comparing it to the prediction of observations from randomly drawn test samples). The best tree is the one with the best average accuracy for cross-validated or predicted values (Electronic Statistics Textbook, 2011).

Most advanced statistical analytics software today have built in functions for CART, e.g. CART menu option in STATISTICA and treefit and treeprune functions in Matlab.

3.3 NEURAL NETWORKS

Neural networks were originally based on the understanding of how the brain is structured and how it functions. This model type can do both time series prediction (univariate) and causal prediction (multivariate). The latter is required for the HHP case study.

Causal prediction (multivariate) refers to an assumption that the data generating process can be explained by the interaction of causal (cause-and-effect), independent variables (Galkin&Lowell, [s.a.]).

Other features of neural network, according to Sven F. Crone (2005), are that they are non-parametric and can approximate any linear and nonlinear function to any desired degree of accuracy directly from the data. Neural networks do not assume a particular noise process, although it is considered a flexible forecasting paradigm. Input variables are flexible: binary [1;0], nominal/ordinal [0,1,2…] or metric [0.237, 7.76, ..]. This is required for the HHP case study as there are binary as well as ordinal variables. Output variables are also flexible: prediction of a single class member (binary), a multi class member (nominal) or a probability of class member (metric). Neural networks can have any number of inputs and outputs.

Neural networks are very powerful in terms of capability to model extremely complex functions, as is required in the HHP case study. Neural networks learn by example and it is therefore expected that they are quite easy to use. The user invokes training algorithms to automatically learn the structure of the representative data. The level of knowledge needed to successfully apply neural networks is somewhat lower than would be the case using other, more traditional, nonlinear statistical methods. The user needs to have some knowledge of how to select and prepare data, how to select an appropriate neural networks, and how to interpret the results (Electronic Statistics Textbook, 2011).

Neural networks are not extremely robust as they do not tend to perform well with nominal variables that have a large number of possible values. This causes a problem if data is in an unusual range or if there is missing data. As mentioned earlier, missing data are not a problem in the HHP case study. Neural networks are noise tolerant to a certain extent, but occasional outliers, far enough outside the range of normal values for a variable, may bias the training. It is best to remove outliers (Electronic Statistics Textbook, 2011).

As with most nonparametric techniques, neural networks are also prone to over-fitting. Over-fitting (over-training for Neural networks) can be prevented by validating progress against an independent test set. Validation can be done by monitoring selection error. Once the selection error starts to increase, it is an indication that the network is starting to over-fit the data, and training should be stopped. In such a situation, the network is too powerful for the problem at hand and it is recommended that the number of hidden layers should be decreased. On the other hand, if the network is not sufficiently powerful to model the underlying function, over-learning is not likely to occur, and neither training nor selection errors will drop to a satisfactory level (Electronic Statistics Textbook, 2011).

Nowadays, most advanced statistical analytics software has built-in functions for neural networks, for example, the SANN menu option in Statistica and NeuroXL add-in, in Excel.

3.4 ENSEMBLE METHODS

Ensemble methods have been called the most influential development in data mining and machine learning in the past decade. It is natural to ensemble "smooth" modelling techniques such as linear models, neural networks and MARS with decision trees in such a way that their strengths can be combined effectively (Brookes&Kolyshkina, 2002). The result of such a union is usually more accurate than the best of its components and it also improves the generalization of the model. Steps to building ensembles are firstly, to construct varied models, and secondly, to combine models' estimates. Two popular and recommended methods for creating accurate ensembles are bagging and boosting.

Bagging, also known as bootstrap aggregation, is a method of using a variety of algorithms to model a single problem and then to use the prediction of each, as a vote. The majority ruling determines the final classification for a given case and the final model is a compromise of its component models.

Boosting, on the other hand, is a method of creating variety by weighting cases, according to which models were easier or harder to model correctly (harder cases get higher weights and vice versa). Boosting works well over a wide range of different modelling approaches (Nisbet *et al.,* 2009: 306).

Criticisms of ensembles are that the more flexible an ensemble is built, the more complex it becomes to interpret its response. In addition to this criticism, is the expectancy that more complexity could also lead to over-fitting (Nisbet *et al.*, 2009: 710).

These days, most advanced statistical analytics software has built in functions for ensemble methods, for example, the ensemble menu option in STATISTICA and Treebagger functions in Matlab.

3.5 COMPARING CONTENDER PREDICTION MODELLING TECHNIQUES

Table 12 shows a comparison of the four contender techniques in terms of characteristics, needed for the HHP case study application (as discussed in the commencement of Section 2). From Table 12 it can be reasoned that CART and ensemble methods are the preferred techniques to use because of their robust nature which is necessary for the HHP case study application.

Table 12: Contender techniques decision matrix

		Contender Techniques			
		MARS	CART	Neural Networks	Ensembles
Characteristics	Categorical	X	✓	✓	✓
	Continuous	✓	✓	✓	✓
	Binary	✓	✓	✓	✓
	Robust?	No	Yes	No	Yes
	Affected by outliers	✓	X	✓	X
	Affected by missing values	X	X	X	X
	Avoiding over-fitting by applying:	Pruning	Pruning by 1)Cross-validation 2)V-fold Cross-validation	Validating progress against an independent test set	Elements contribute separately
Advantages and Disadvantages	Known advantages	-Relatively simple -Picks up only contributing variables -Not prejudice	-Flexible -Robust -Ease of use -Invariant to monotone transformations	-Powerful -Ease of use	-More accurate than the best of its components -Improves generalization
	Known Disadvantages	-Not robust Proneness to over-fit	-Many variables = very complex trees = very difficult to interpret.	-"Black box" nature -Computational burden -Proneness to over-fit	-Often difficult to interpret -Flexibility directly related to complexity

4 TECHNOLOGIES CONSIDERED

This chapter aims to introduce common data mining tools on the market today: Excel (in combination with VBA), Matlab, Statistica, SAS, SPSS Clementine and R (Figure 14). A recommendation as to which technologies are appropriate for this application was determined in this chapter. A summary of the technologies for the use in the HHP application is provided in Table 13.

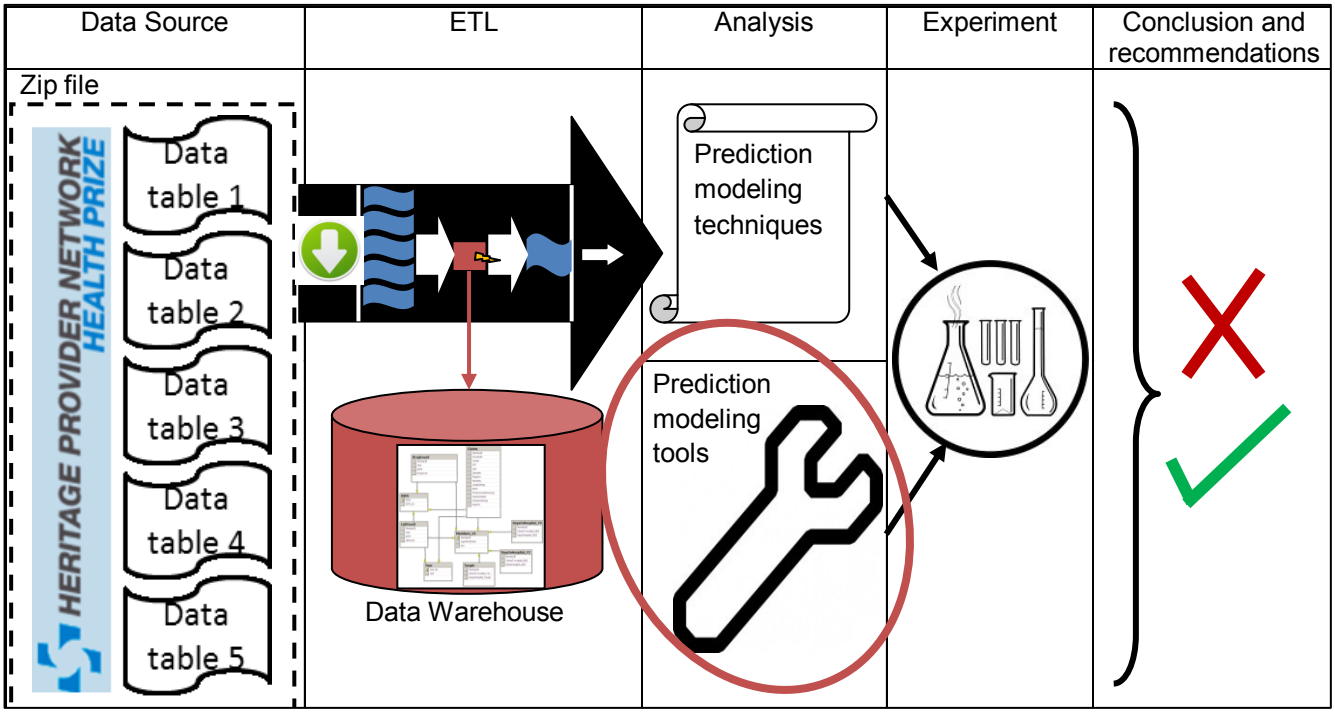

Figure 14: Roadmap to pave the way for hospitalisation prediction modeling - technologies considered

4.1 SPSS CLEMENTINE

According to Nisbet *et al*. (2009) SPSS Clementine is a mature data mining tool that has been used by many to create powerful models for business. It enables the user to develop predictive models fast and deploy them as decision support systems in business processes. It uses a graphical user-interface, making it user-friendly improving the user's ability to understand what they are doing. It also has open source facilities using 4GL command language. Considering that the prediction output for this application needs to be continuous, SPSS Clementine can only perform CART, Chi-squared automatic interaction detection (CHAID) and linear regression that could assist with this requirement.

4.2 SAS ENTERPRISE MINER

SAS Enterprise miner also uses a graphical user-interface such as SPSS Clementine. As described by the SAS website (2011) SAS was designed for use by business analysts with little statistical expertise, but who will be able to use this tool fairly easily. SAS has the advantage that one can go "behind the nodes" to customize the analytical processes using the interactive matrix language of SAS. SAS's capabilities for the HHP application include built in functions for decision trees, neural networks and two stage models (SAS, 2011).

Table 13: Technology decision matrix

Tool	***Open source/ menu driven***	***Syntax used***	***Disadvantages***	***Availablity to research team***	***Software capabilities with prediction modelling***
Excel with VBA	Both basic menu statistics and open source facilities	VB	-Not suitable for big datasets -Only very basic statistical functions	Installed on available computers	- Basic linear regression
Matlab	Open source	Matlab command language	-Not very user-friendly -Difficult to keep track of variables	Difficult to attain and expensive, especially with Statistics toolbox	-Suitable for very big data sets -Has well developed functions -Can do complex modelling for MARS, CART, Neural Nets and Ensembles.
Statistica	Advanced menu statistics and open source facilities	VB	-Build-in-functions could be limiting -Gives too much information	Difficult to attain and expensive	Very versatile, user-friendly -Spreadsheet based -Suitable for very big data sets - Has appropriate built in functions for CART, Neural Networks and Random Forests.
SAS	Advanced menu statistics and open source facilities	Interactive Matrix Language	-Implementation of a function is cumbersome -Not user-friendly	Difficult to attain and expensive	-Decision trees, neural networks and two stage models -Mostly used for business analytics
SPSS Clementine	Graphical user-interface and open source facilities	4GL command language	-Limited multivariate procedures -Slow pace of development	Difficult to attain and expensive	-CART, Chi-squared automatic interaction detection (CHAID) and linear regression
R	Open source using	*S* command language	-Memory overflow with large data sets -Steep learning curve	Free to download from the internet	-Has well developed functions -Can do complex modelling for MARS, CART, Neural Nets and Ensembles.

4.3 STATISTICA DATA MINER

Statistica Data Miner contains commands to create and maintain complex predictive models and advanced visualization properties. Statistica uses a menu-driven approach and has open source facilities available using the Visual Basic coding language. Statistica's capabilities for the HHP application include built in functions for decision trees, neural networks, MARS (has limiting cross-validation for MARS) and Random Forests.

4.4 EXCEL AND VBA

Excel and VBA is popular tools to use for small scale forecasting applications, it has limited memory capacity and it only has basic regression functions that can be of little benefit in the HHP application accept for limited exploratory analysis.

4.5 MATLAB

Matlab by Mathworks has a useful Satistical Toolox that can be added to the basic Matlab tool. This toolbox allows for very complex prediction modelling application, as in the case of the HHP application. Matlab's capabilities for the HHP application include built in functions for decision trees, neural networks, MARS, Random Forests (and other ensemble methods). On the other hand, it has the advantage that the research team is familiar with the technology and this might be an appropriate tool to use for this reason.

4.6 R

Kabacoff (2011) describes R as a tool for statistical computation and graphics consisting of an interpreted computer language allowing branching, looping as well as modular programming functions. The penalty for using R is the steep learning curve, because of the programming language the research team will have to learn as well as advanced statistical knowledge, because R does not use a plug-in-and-play approach. On the other side R is free, it contains advanced statistical routines not yet available in other packages and has advanced graphics capabilities. Software can be downloaded from one of the Comprehensive R Archive Network (CRAN) mirror sites.

Each technology has its advantages and disadvantages and often the choice of technology depends heavily on the analyst's programming ability and preference. Programs like SAS, SPSS and Excel have limited imbedded functions of contender techniques and therefore using these technologies for this application will be quite labour intensive (hard coding will have to be done to fill in any gaps that limited functions leave).

Matlab and R on the other hand have sufficient built in functions for most of the mentioned techniques and they also have appropriate visualisation resources and are powerful enough to handle the HHP case study data set size as well as the complexity thereof. Statistica is a fine tool to use for explorative purposes because of its user friendly interfaces, plug-in-and-play functionality and data management ability, but because of its plug-in-and-play approach it may be difficult to customise. Finally, Matlab and R were identified as the preferred tools for this application.

5 EXPERIMENTATION

Through experimentations with selected techniques and technologies, it can be verified whether models and technologies can handle the inputs received and produce desired outputs as required for this application. In this chapter such experiments are demonstrated (Figure 15). Future works include more extensive experiments in order to allow for cross-validation.

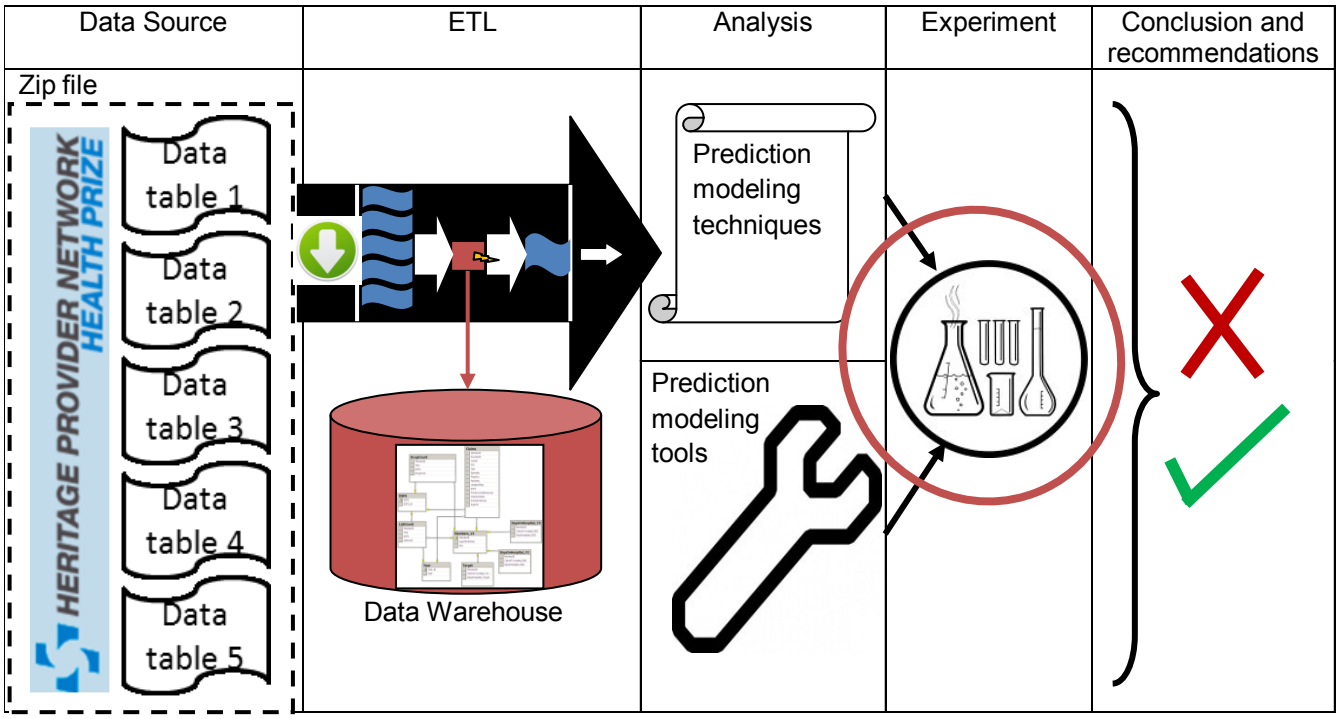

Figure 15:Roadmap to pave the way for hospitalisation prediction modeling – experimentation

5.1 ETL EXPERIMENTATION

Tests were constantly conducted in SQL to see that data's integrity is retained. This was done by running multiple queries of different forms to determine if data makes sense and if integration was done successfully. A similar validation process was repeated for integration done in Excel.

5.2 STATISTICAL EXPERIMENTATION

Exploratory prediction modelling was done by applying the contender techniques CART and MARS in Statistica and the ensemble method, Random Forests in R. Predictions were tested by looking at each of their predicted days in hospital versus observed days in hospital values. As a means to quantify the prediction accuracy, a prediction error rate was calculated for each model as suggested by the Heritage Health Prize competition website (Heritage Provider Network Health Prize, 2011).

It can also be noted that this is the same prediction error rate as that used by Heritage Health Prize judges to measure the performance of an entered PPAA.

$$Prediction\ error\ rate = \epsilon = \sqrt{\frac{1}{n}[\sum(\log(pi+1) - \log(ai+1)]^2} \qquad (5.1)$$

Where:.

i is a member;

n is the total number of members;

p is the predicted number of days spent in hospital for member i in the test period;

a is the actual number of days spent in hospital for member i in the test period.

A sample of 20 000 records was used (which is approximately 50% of the available data for that year) by selecting the first 20 000 claims for year 2. The aim of these analyses was simply to explore the data and the techniques.

Classification and regression trees in Statistica

After the sample was imported into Statistica the following steps were followed:

Step 1: Histograms of all variables were drawn with the intent to see how these variables are distributed. The percentage of each occurrence can be seen on top of each bar in the histogram in Figure 16 which is an example of a variable with almost no variance. If variables contain a value that occurs more that 95% of the time, as seen in Figure 16, these variables did not have noteworthy variance and were removed. Variables with as little variance as this are not useful in analysis and are better ignored. Variables that were removed are listed in Table 14. A histogram of the dependent variable, DaysInHospital (Figure 17), was plotted and from it it could be seen that there is also very little variance in this important variable, causing somewhat of a dilemma for the creation of a prediction model. The approach followed to try and address this dilemma, was to first do classification tree analysis dividing the dependent variable into two parts: DaysInHospital (1) less than (<) one day and DaysInHospital (2) more than (>) one day. Classification tree analysis was followed by doing regression tree analysis for number (2).

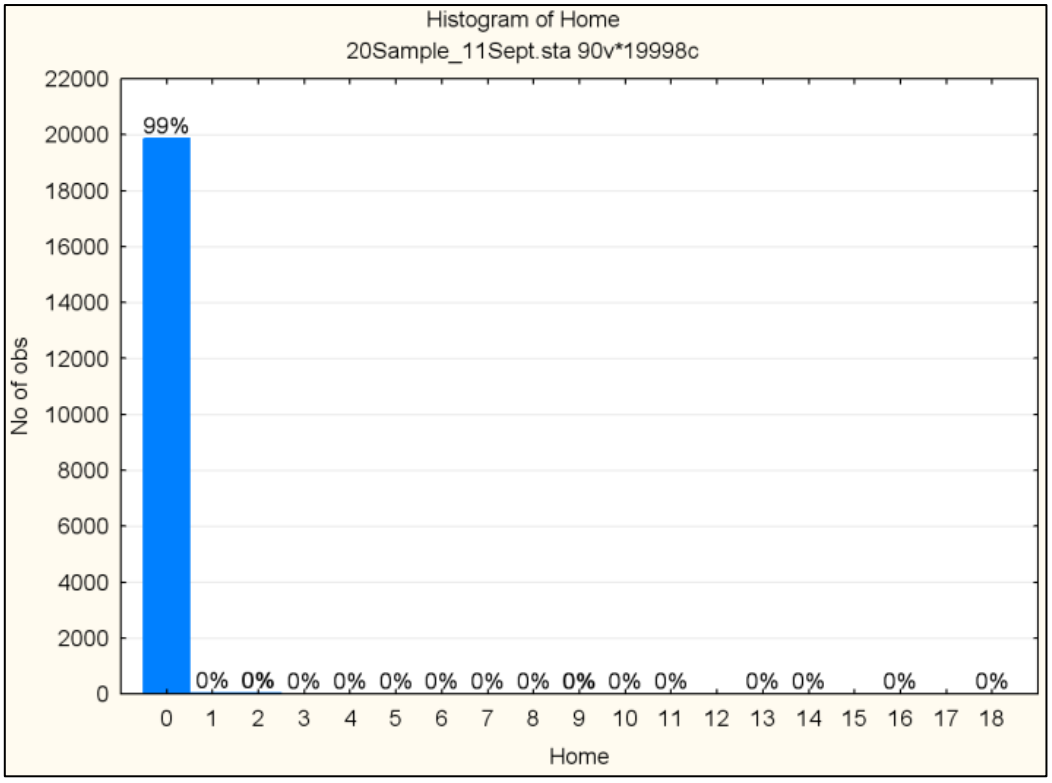

Figure 16: Example of a variable's distribution

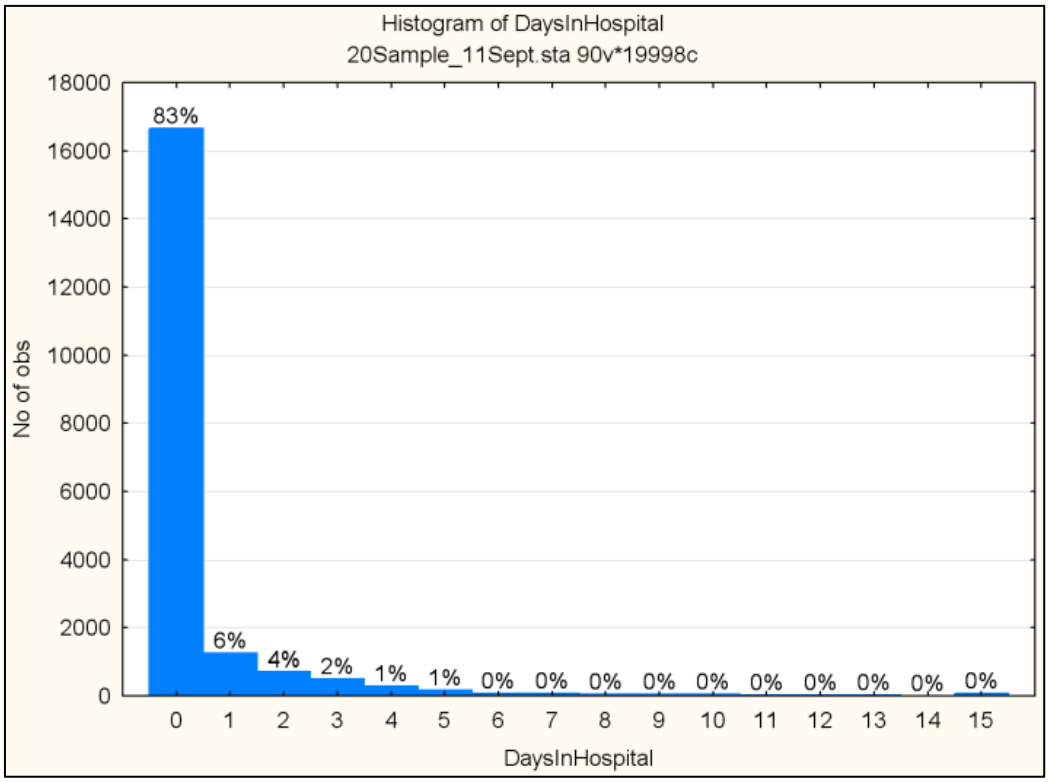

Figure 17: Dependent variable's distribution

Table 14: Removed variables

Speciality	PLaceSvc	Condition	Procedure	CharlsonIndex
Pathology	Home	APPCHOL	ANES	5.5
	Office	CANCRA	SAS	
	Other	CANCRB	SEOA	
		CANCRM	SGS	
		CATAST	SMCD	
		CHF	SNS	
		FLaELEC	SO	
		FXDISLC	SRS	
		GIOBSENT	SUS	
		GYNECA		
		HEMTOL		
		HIPFX		
		LIVERDZ		
		METAB1		
		MISCL1		
		PERINTL		
		PERVALV		
		PNCRDZ		
		PNEUM		
		PRGNCY		
		RENAL1		
		RENAL2		
		SEIZURE		
		SEPSIS		
		STROKE		

Step 2: Once the seemingly insignificant variables were removed (as described in step 1) the remaining predictor variables were bundled into categorical and continuous predictor variables with the use of the "Bundles" Statistica function. Bundling eases the monotonous task of having to select each preferred variable and only the bundle name has to be selected. Bundling was followed by inputting the variables into the CART model builder as seen in Figure 18.

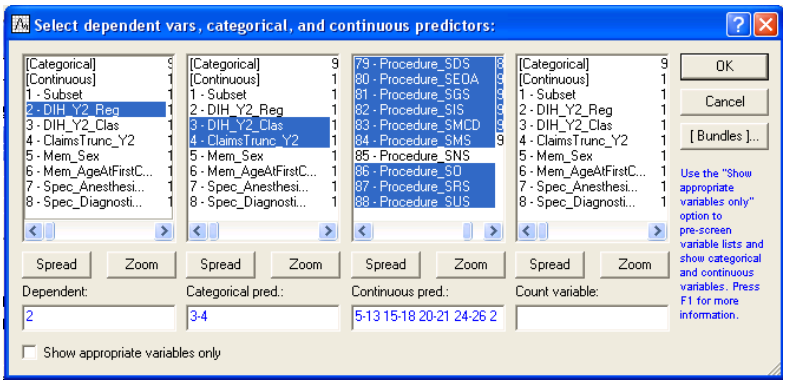

Figure 18: Inputting variables

Step 3: The final tasks, before the model was run, was to input stopping parameters and doing a validation setup. The Chi-squared measure of goodness of fit was selected and was used consistently throughout experiments. Stopping parameters are those that determine when the model will stop splitting. They consist of minimum n, minimum n in the child node, maximum n levels and maximum n nodes. These values are not crucial for the success of this model, because if a classification (or regression) tree is built to largely the tree is simply pruned back to a size that is useful to the analysis. For this reason Statistica's default values were utilised as seen in Figure 19 which was verified by statistical consultation.

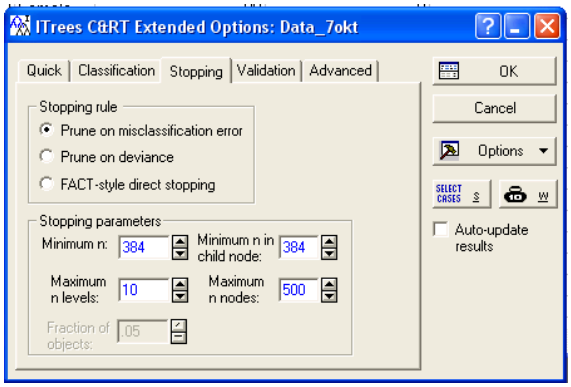

Figure 19: Stopping parameters

V-fold cross validation was used to prevent over-fitting. The test sample consisted of 30% of the observations and training set used the remaining 70% thereof. According to statistical consultation this is a sufficient proportion. The outputs that were interesting were the importance plots, indicating which variables the classification tree analysis perceived as important and the chi-squared tests for goodness of fit. As seen in Figure 20 the most important predictor variables for classification tree analysis were surgery digestive system (SDS) and Urgent care place of service.

Step 4: A similar approach was used for the regression analysis where the dependent variable, DaysInHospital, is more than (>) one day. The importance plot can be seen in Figure 21 indicating the most important variables are Charlson index 1.5, the procedure group, pathology and laboratory (PL), and inpatient hospital place of service. The predicted and observed values data was collected in Statistica and imported into Excel for goodness if fit testing.

Goodness of fit was determined by applying a chi test. Where the hypothesis was Ho: "the predicted values are following the same distribution as that of the observed values", with level of significance of 95% (alpha = 0.05) and the degrees of freedom equaling 2. The test concluded that the hypothesis Ho should be rejected because $\chi 2$ calculated > $\chi 2$critical at a significance level of 95%.

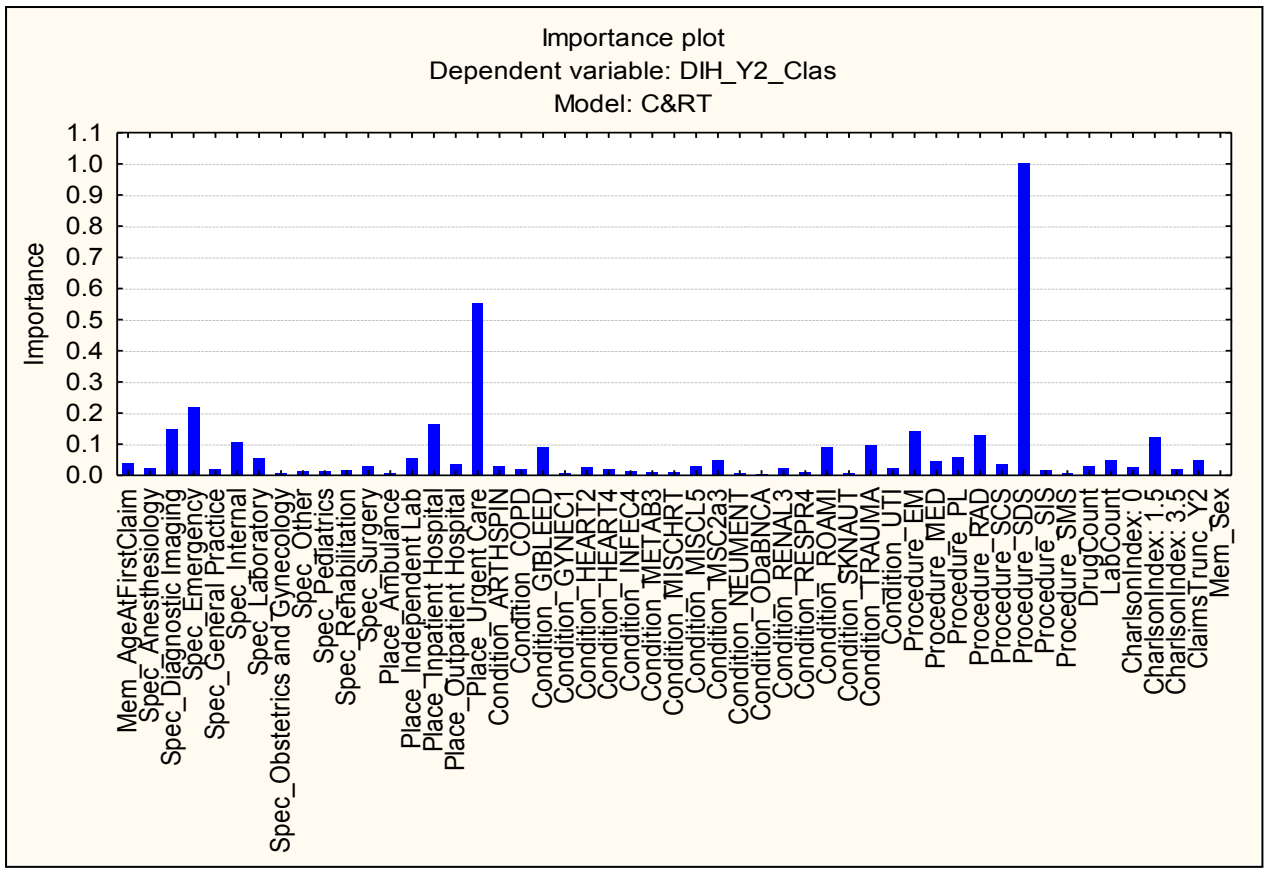

Figure 20: Importance plot for Classification tree analysis.

Calculations for this test can be seen in Appendix H. The observed and predicted frequencies, as determined in the chi test, can be seen in Figure 22. A further proof that there was unsatisfactory predictive power can be seen in Figure 23, where the observed dependent variable is plotted against the predicted values. The blue dots indicate the predictions produced by the regression tree and the red line indicates the ideal case where observed values equal predicted values. The prediction error rate for regression tree analysis scored a value of 0.353.

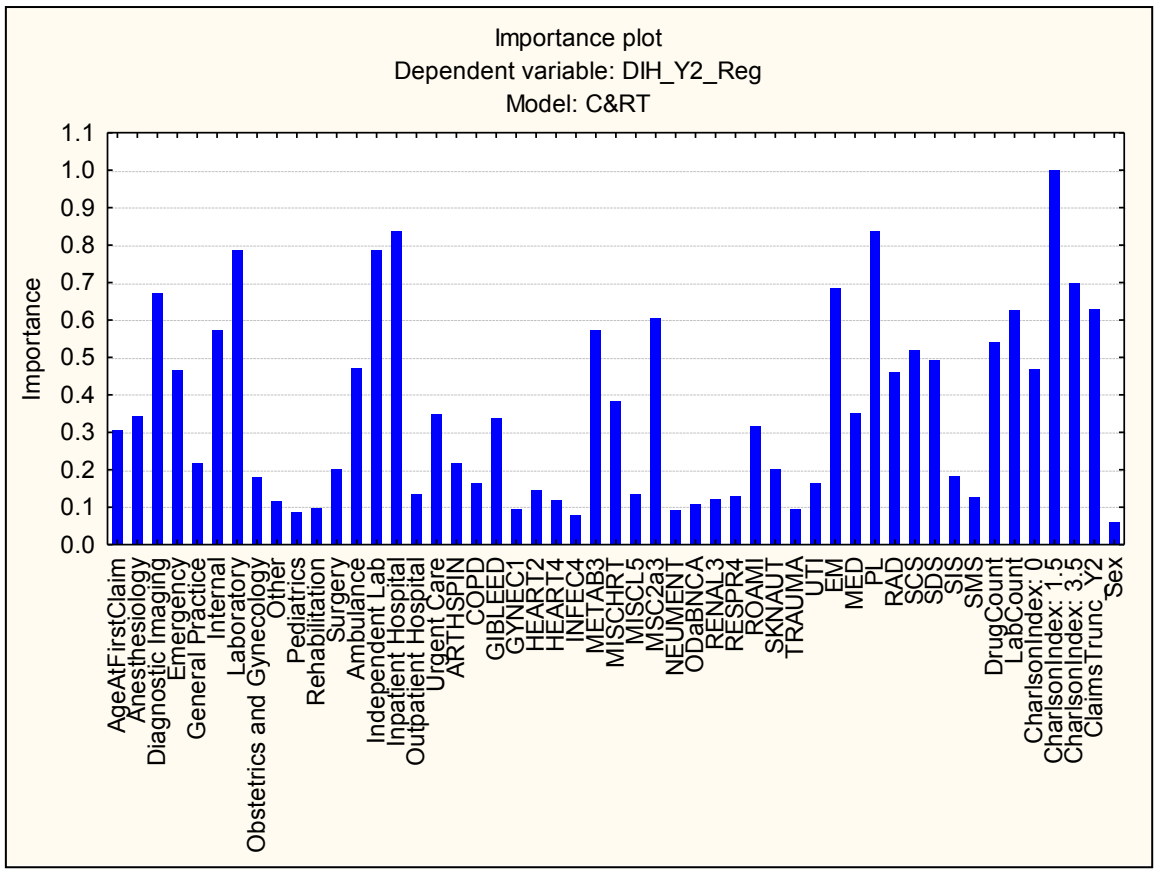

Figure 21: Importance plot for regression tree analysis where DIH >0

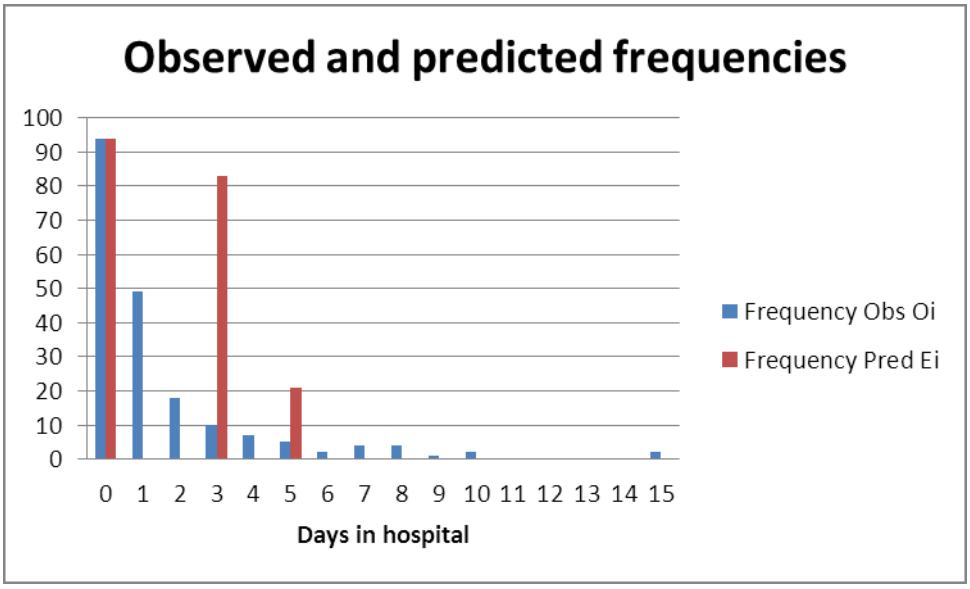

Figure 22: Observed and predicted frequencies

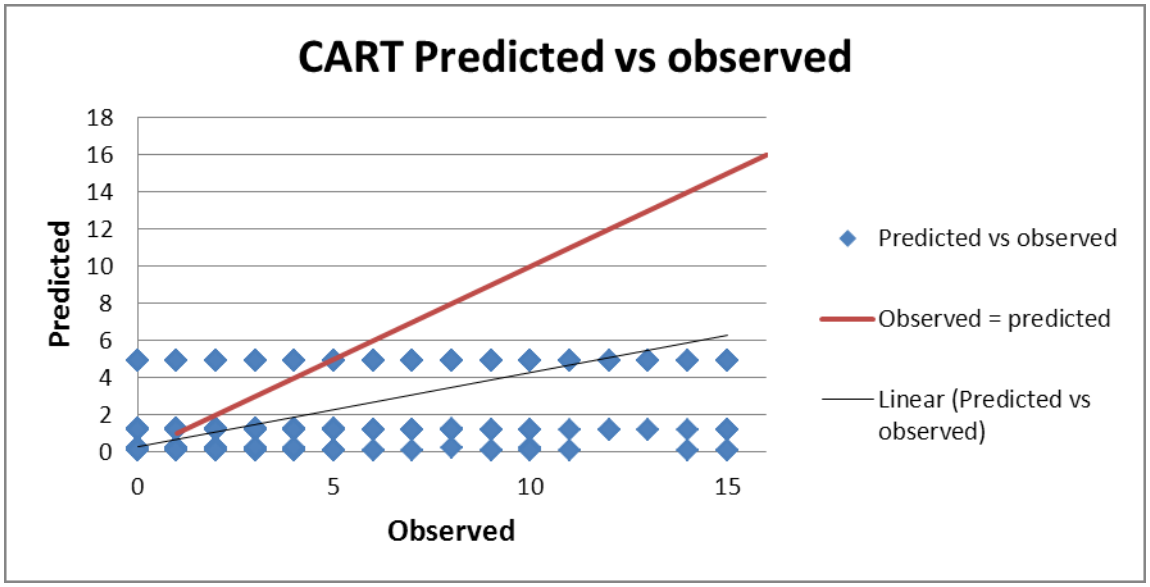

Figure 23: CART predicted vs observed

Multiple adaptive regression splines in Statistica

The same sample was used for MARS as was used for the training of the CART model. The process followed to conduct the MARS analysis in Statistica is similar to that of the CART analysis. Variables were bundled in the same manner as was the case with the CART analysis. The usable variables were then inputted into the MARS model as seen in

Figure 24. This was followed by setting the parameters of the model which is shown in the screen shot taken in Figure 25.

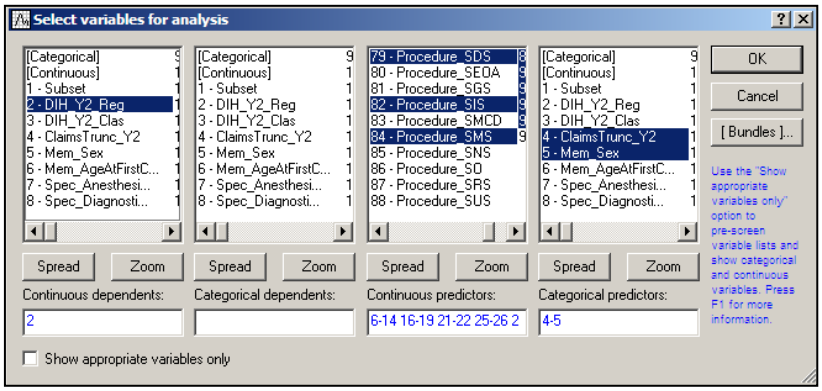

Figure 24: Variables read into MARS

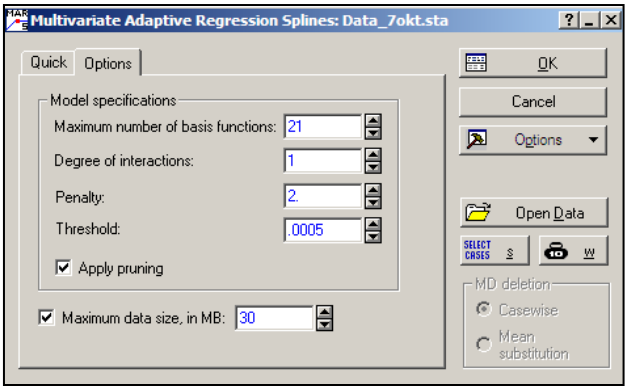

Figure 25: MARS parameters

Goodness of fit was determined by applying a chi test. Where hypothesis Ho is the same as was used with the CART analysis: "Ho: the predicted values are following the same distribution as that of the observed values", with level of significance of 95% (alpha = 0.05) and the degrees of freedom equaling 2. The test concluded that the hypothesis Ho should be rejected as $\chi 2$ calculated > $\chi 2$critical at a significance level of 95%. Calculations for this test can be seen in Appendix H. The observed and predicted frequencies, as determined in the chi test, can be seen in Figure 26, making it clear that this model does not fit the data pointedly.

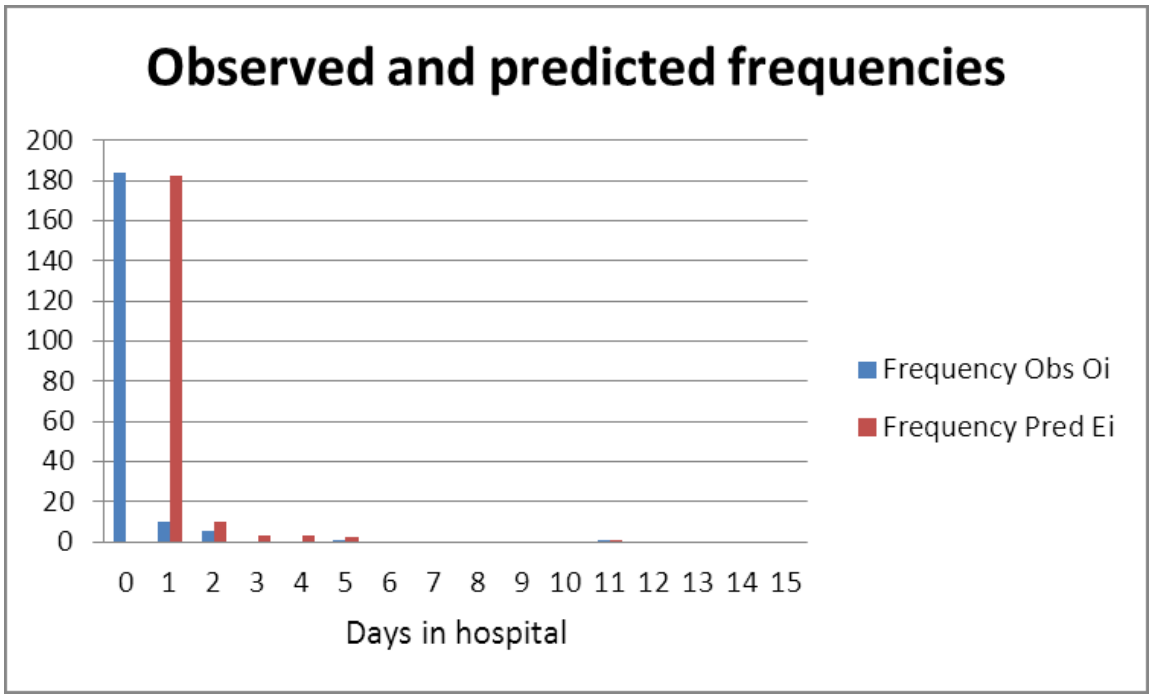

Figure 26: Observed and predicted frequencies

The observed dependent versus predicted target values where plotted in Figure 27, where the blue dots indicate the predictions produced by the regression tree and the red line indicates the ideal case where observed value equals predicted value The prediction error rate for MARS analysis scored a value of 0.358, slightly worse than CART analysis. With the help of statistical consultation it was advised that neural networks would give similarly valueless predictions and for this reason it was decided to try a completely different approach, the last of the contender techniques namely: ensemble methods.

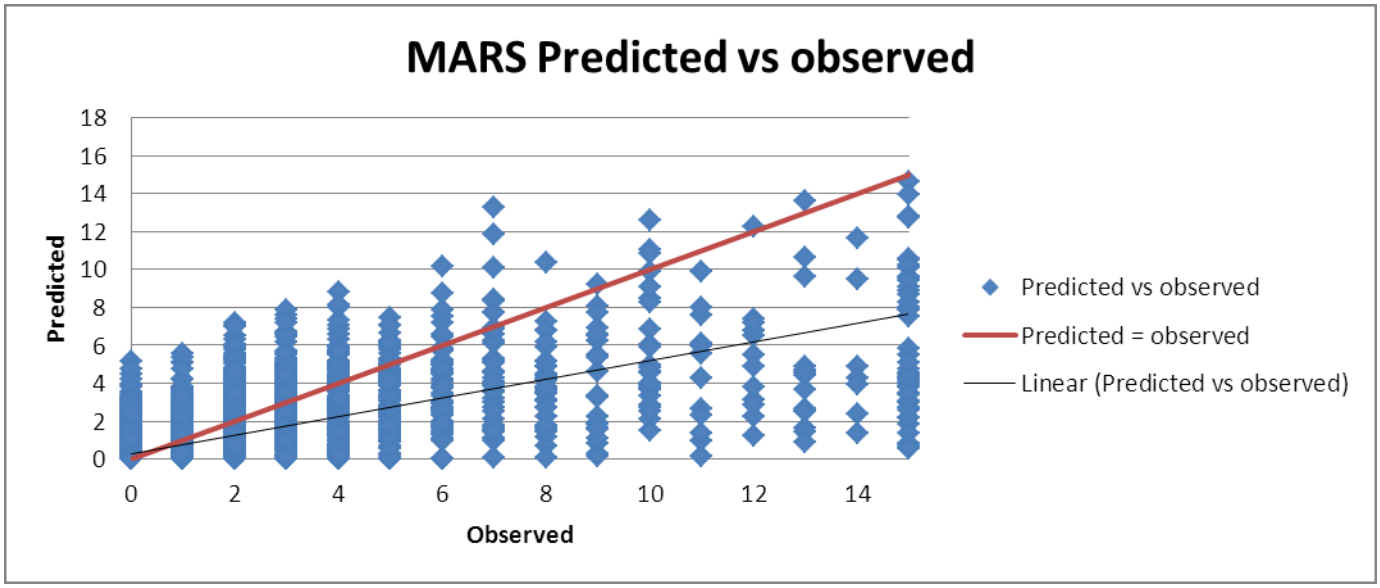

Figure 27: MARS predicted vs observed

Random Forests in R

Once it had been determined that techniques on their own did not perform well with the HHP data set it was decided to experiment with ensemble methods. There are a multitude of ensemble methods in literature, just to mention a few bayes optimal classifier, bootstrap aggregating (bagging), boosting, bayesian model averaging, bayesian model combination, bucket of models and stacking. For the HHP application the bagging method, random forests, was selected to experiment with. The reason for selecting this method was that this method had been the winner in another competition, the Netflix competition, also hosted by Kaggle, which had similar data and output requirements (Kaggle, competitions and competitive intelligence, 2010).

Unweighted regression: The number of trees to grow was set at 100, which is a sufficiently for the available data set, according to statistical consultation. After the forest was grown a chi- test was conducted to determine how well the model fits the data, using the same hypothesis as in both CART and MARS chi tests.

A significance level of 95% (alpha = 0.05) was used and the degrees of freedom equaled 12. The test concluded that the hypothesis Ho should be rejected as χ2 calculated > χ2critical at a significance level of 95%. Calculations for this test can be seen in Appendix H. The observed and predicted frequencies, as determined in the chi test, can be seen in Figure 28, where it is clear that the values of observed and predicted frequencies were similar and it seems the chi test misrepresented the goodness of fit The reason for this discrepency can be caused by either the skewly distributed dependent variable (Figure 17) or the fact that the chi test might be oversensitive for this data set size. The prediction error rate for the random forests model scored a value of 0.171, which also scores better than both CART and MARS.

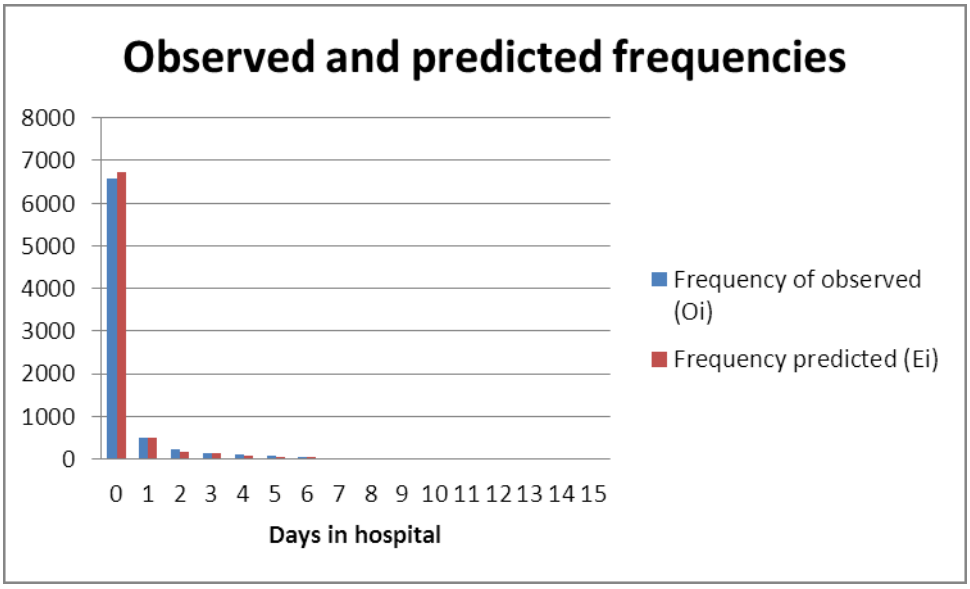

Figure 28: Observed and predicted frequencies

Finally, as seen in **Error! Reference source not found.**, the observed versus predicted values (the blue dots) are shown in a scatter plot. From this plot it can be seen even clearer that the Random Forests model performs better than both CART and MARS as the observed vs. predicted (the blue dots) more closely imitate the ideal prediction result (red line) than for the other contender techniques. Other random forests setups that can also be considered are two category classification and three category classification. Such classification analyses might result in stronger predictions because of the dependent variable's skew distribution. The code used to produce the random forests models for this application can be viewed in Appendix G.

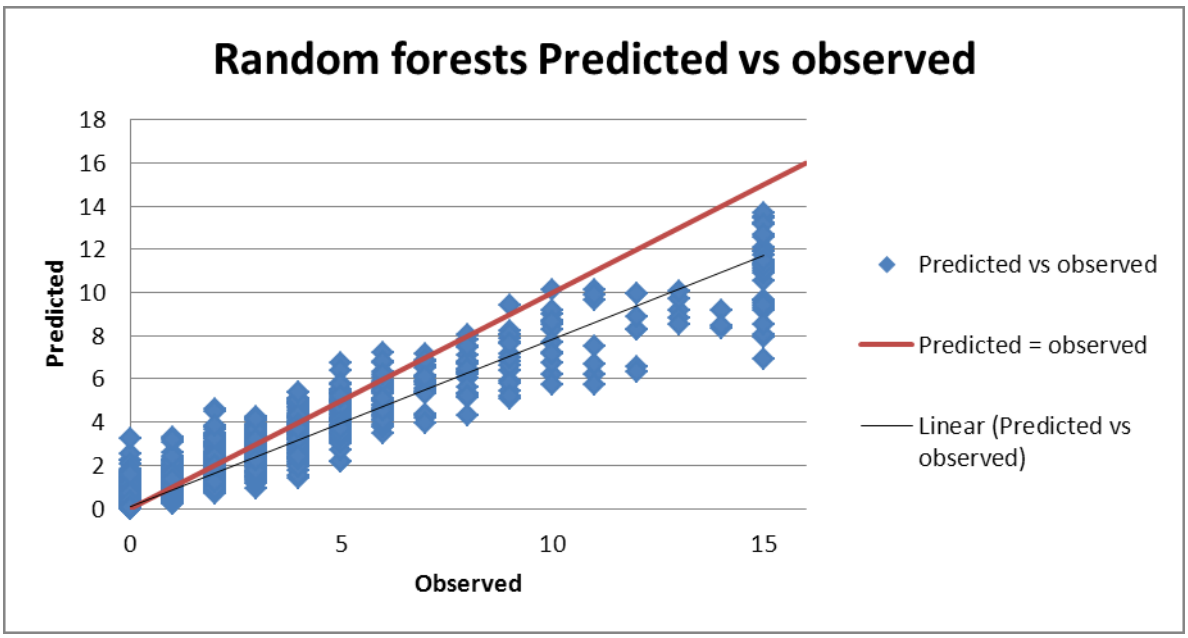

Figure 29: Random forests predicted vs observed

Results

From the exploratory tests carried out in Section 5 it can be concluded that the included predictor variables (discussed in Section 2.3.2) do not have much predictive power for a CART or MARS analysis and similar results are expected when using neural networks. It can also concluded that regression tree analysis turned up less useful information than classification tree analysis, because predictor variables were weak and did not contribute sufficiently to be able to model a realistic prediction algorithm. Even though classification tree analysis is useful for exploratory purposes, it does not output a desired continuous target value as needed for the HHP application. More simply put, the prediction model should give, as an output, a continuous number or categorisation of days as predicted value, not just a less than one day in hospital or more than one day in hospital variable.

It was further found that the bagging technique, Random Forests performed better than CART and MARS and it is recommended that further exploration should be done using Random Forests.

In Figure 30 the three models experimented with are compared, using their linear trend lines as a comparison measure. Even though these models are not necessarily linearly distributed a comparison such as this give an idea as to where most of the data is distributed and brings prediction models' performance into perspective when comparing it to the ideal observed vs predicted distribution.

The purple line is the ideal case where predicted values are equal to observed values. It can clearly be seen that Random Forests (the green line) is closest to the ideal (the purple line).

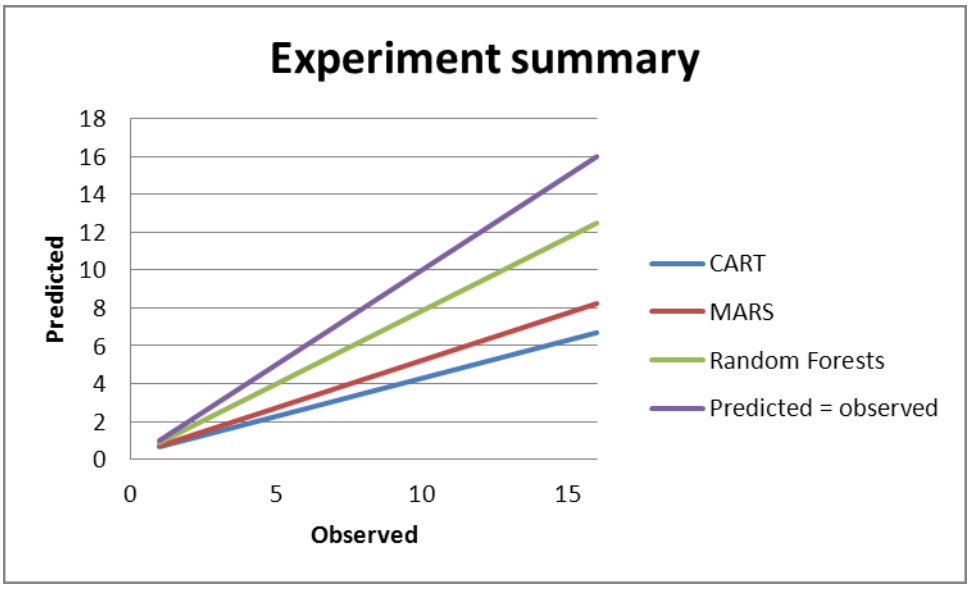

Figure 30: Experimentation summary

Further comparison can be seen in Figure 31 where each of the three models' prediction errors rates were compared, giving similar results as those seen in Figure 30.

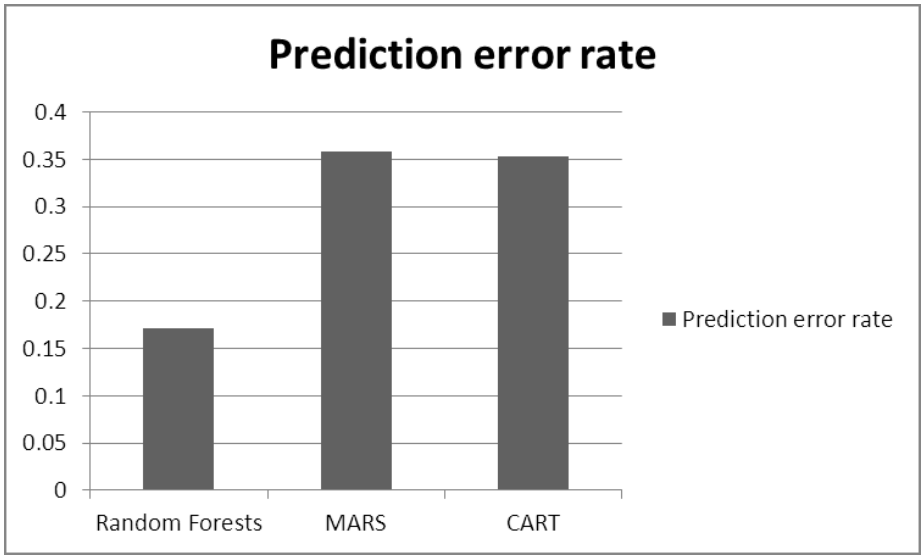

Figure 31: Prediction error rate comparison

For further in depth analyses these prediction models should be tested on more data sets to determine if similar results will effect. Also for future statistical analysis other goodness of fit tests should also be considered, because there is a possibility that the chi test may be oversensitive for this data set, possibly causing it to reject the Ho hypotheses each time.

6 CONCLUSIONS AND RECOMMENDATIONS

The aim of this project was to pave the way for a future research team by providing insights and identify possible pitfalls in the development of a Predictive Patient Admission Algorithm (PPAA) based on the Heritage Health Prize competition (see Figure 32 for the roadmap). Now that the process of evaluation, comparison and experimentation are completed conclusions and recommendations can be made on the process (Figure 31).

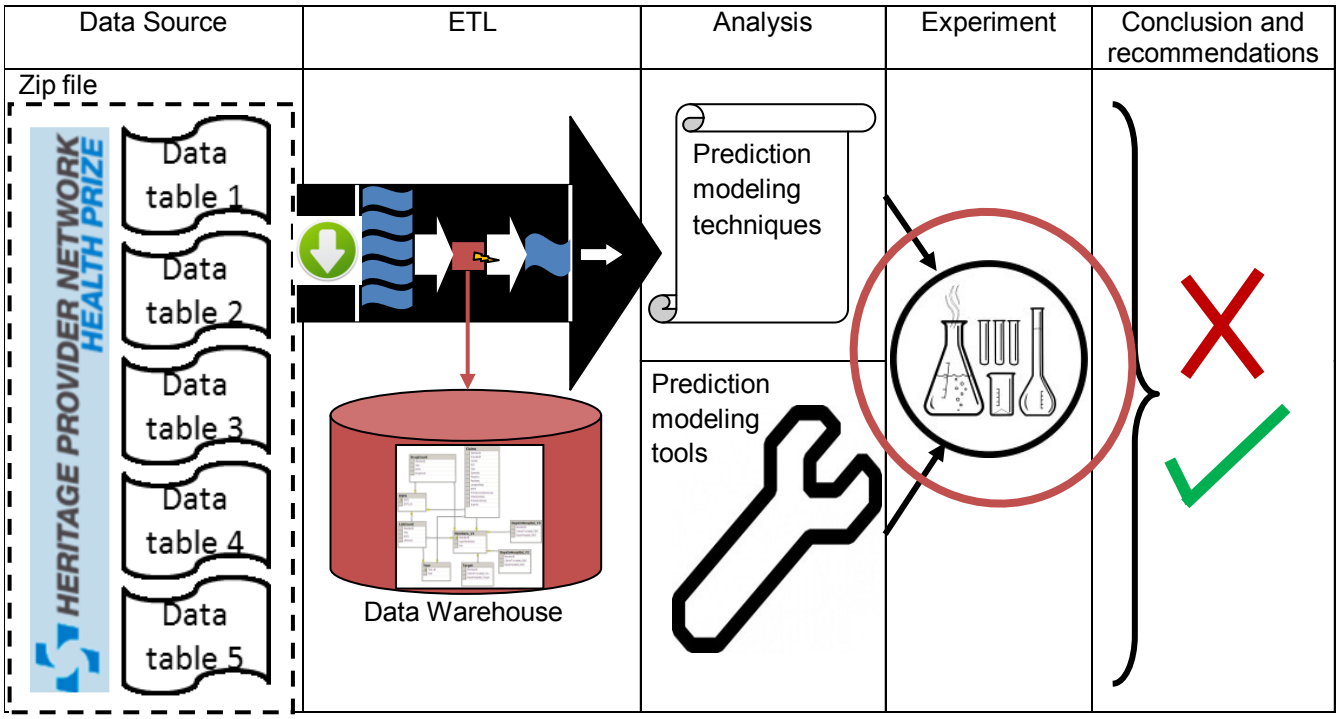

Figure 32: Roadmap to pave the way for hospitalisation prediction modeling – conclusion and recommendations

Once the original stated objectives are met it is safe to say that the aim of the project will also inherently be met. In the succeeding paragraphs a discussion exists concerning these objectives and how they were achieved. The following objectives were achieved:

Objective 1: Inspect data received for the competition and to determine how to extract, transform and load (ETL) data correctly; background was provided providing a perspective on the problem, decision support systems for similar applications, the Heritage Health Prize competition and lastly prediction modelling in the health care sector helping to understand the problem area more comprehensively. Once a better understanding of the problem area was attained data mining began by determining how the data warehouse should be constructed for use in the HHP application.

This included recommendations about procedures followed in the ETL process as well as business rules with which data warehouse should be constructed. ETL tools were compared in terms of user friendliness, availability, appropriateness, syntax used and performance with ETL processes. Finally, different ETL alternatives were considered and compared via an analytical hierarchical process which concluded that using SQL Server alone is the preferred scenario to follow for this application (within the constraints posed by the research team).

Objective 2&4: Compare and experiment with contender techniques for development of the PPAA with given competition data; contender techniques were selected based on their ability to meet criteria as required by the HHP problem and available data. Contender techniques consisted of classification and regression tree analysis (CART), multiple adaptive regression splines (MARS), neural networks and ensemble methods of which RandomForests were experimented with. Conclusions concerning the contender techniques and their ability to meet the demands of the problem and data composition were that the selected predictor variables were weak in exploratory CART and MARS analysis and could not significantly train a realistic model to predict hospitalisation (according to statistical consultation the same was suspected for neural networks). It was decided that an alternative approach should be used such as using ensemble methods. The ensemble technique Random Forests was exploratively tested on the data set which resulted in better predictive power than the CART and MARS models. It was concluded that of the considered contender techniques RandomForests yielded the most favourable results making it the preferred technique among the considered techniques for the HHP application. Other prediction modelling techniques that can also be researched for future use in this application are probabilistic Bayesian methods as well as Structural Equations Methods (SEM). These techniques were recommended at the ORSSA conference.

Objective 3&4: Compare and experiment with technologies needed for development of the PPAA, considering the suitability thereof, research team's resources and research team's knowledge fields; by researching, comparing and experimenting with data mining tools currently on the market it was determined which tools are appropriate for the HHP application within research team's constraints (experimentation was only conducted with R, Statistica and Excel). This comparison was achieved by acquiring information about each tool such as: user interface used, syntax used, availability to research team, known disadvantages of the tool and software capabilities for the application. It was concluded that SAS, SPSS and Excel have limited imbedded functions of contender techniques and therefore using these technologies for this application will be labour intensive (hard coding will have to be done to fill in any gaps that limited functions or lack of functions leave).

Matlab and R on the other hand have sufficient built in functions for most of the mentioned techniques; they also have appropriate visualisation resources and are powerful enough to handle the HHP case study data set size as well as the complexity thereof. Statistica is an appropriate tool to use for explorative purposes because of its user friendly interfaces, plug-in-and-play functionality and data management ability, but it is not as customisable as either Matlab or R. Finally, Matlab and R were selected as the preferred tools for this application.

Now that the objectives have been reached successfully it is safe to say that the road has been paved for the research team succeeding this project to model a successful hospitalisation prediction model in the health care environment. Research, results, recommendations and conclusions can be used in an effort to model the winning algorithm without having to redesigning the wheel of hospitalisation prediction modelling for this application. At the end of the day if a prediction model can successfully predict hospitalisation, there will be a major breakthrough in the industry, saving many lives and large sums of money. This conclusion not only signifies the end of this project, but also the handing over of the baton to the research team who will now take over research and hopefully achieve success.

7 PERSONAL EXPERIENCES

The interesting application of data mining facilitated by a competition such as the Heritage Health Prize competition makes for a challenging opportunity to study this interesting field. The concept of knowledge sharing was promoted by the competition through discussions on the Heritage Health Prize website, which helped with the conceptualisation of the approach used to do this final year project. The data received for this application is more complex than originally realised and a structured organise approach is required and this project was provide such a roadmap with which a structure approach can be followed.

Sufficient knowledge of the field of statistics should be acquired to successfully do prediction modelling for this application, otherwise a black-box approach could result.

As mentioned by more than one ETL expert in literature, ETL takes up to 70% of a data based project, this can be confirmed and should be considered in the scope and time of the research team.

REFERENCES

ARBUCKLE, 2011, Drugcount? Count of drugs or prescriptions? 2011, [Online], [Cited July 7th, 2011], Available: http://www.heritagehealthprize.com/c/hhp/forums

ARONSON JE, LIANG T, SHARDA R AND TURBAN E, 2005, Decision support and business intelligence systems, Pearson Prentice Hall, pp224-226.

BARNETT M, 2011, Data problems: inpatient hospital stays w/o lengthofstay & outpatient los, [Online], [Cited July 26th, 2011], Available from http://www. heritagehealthprize.com/c/ hhp/forums/t/377/data-problems-inpatient-hospital-stays-w-o-lengthofstay-outpatient-los

BREIMAN L, FRIEDMAN J, OLSHEN R & STONE C, 1984, Classification and Regression Trees, Wadsworth

BROOKES R &KOLYSHKINA I, 2002, Data mining approaches to modelling insurance risk. Paper presented at the IXth Accident Compensation Seminar.

CASERTA K, CASERTA R & CASERTA J, 2004, The Data Warehouse ETL Toolkit, Wiley.

CRONE SF, 2005, Forecasting with artificial Neural Networks, *Journal of Intelligent Systems*, 14(2-3), pp99–122.

Decision support systems from a health informatics perspective, *Citeseer,* [Online], [Cited September 26th, 2011], Available from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.9082.

Electronic Statistics Textbook, 2011, StatSoft Inc., [Online], [Cited August 23th, 2011], Available: http://www.statsoft.com/textbook/

GALKIN I. & LOWELL UM, Crash introduction to artificial neural networks, Unpublished course material, [Online], [Cited August 8th , 2011], Available: http://ulcar.uml.edu/~iag/CS/Intro-to-ANN.html

Heritage Provider Network Health Prize, 2011, [Online], [Cited August 7th, 2011] Available form http://www.heritagehealthprize.com/c/hhp

HOWARD J, 2011, Male pregnancy?, [Online], [Cited June 10th, 2011] Available from http://www.heritagehealthprize.com/c/hhp/forums

HOWARD J, 2011, [Online], [Cited May 7th, 2011], Available from http://blog. kaggle.com /2011/03/23/ getting-in-shape-for-the-sport-of-data-sciencetalk-by-jeremy-howard/

HUNT DL, HAYNES RB, HANNA SE & SMITH K, 1998, Effects of computer-based clinical decision support systems on physician performance and patient outcomes, *JAMA: the journal of the American Medical Association,* 280(15), pp1339.

IGOR , 2011, Data problems: inpatient hospital stays w/o lengthofstay & outpatient los, 2011, [Online], [Cited July 7th, 2011], Available: http://www.heritagehealthprize.com/c/hhp/forums

JOBSON JD, 1991, Applied multivariate data analysis: regression and experimental design, Springer, Faculty of Business University of Alberta Edmonton, Alberta, Canada

Kabacoff, R. 2011.

Kaggle, competitions and competitive intelligence, 2010, [Online], [Cited October 19th, 2011], Available form http://minimalstate.com/2010/04/14 /kaggle-competitions-and- competitive-intelligence/

KUKULL WA, KOEPSELL TD, CONRAD DA, IMMANUEL V, PRODZINSKI J & FRANZ C, 1986, Rapid estimation of hospitalization charges from a brief medical record review: Evaluation of a multivariate prediction model, *Medical care*, pp961-966.

LEE C, PARR RG & YANG W, 1988, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Physical Review B, 37(2), pp785.

MIYATA H, HASHIMOTO H, HORIGUCHI H, MATSUDA S, MOTOMURA N & TAKAMOTO S, 2008, Performance of in-hospital mortality prediction models for acute hospitalization: hospital standardized mortality ratio in Japan, BMC health services research, 8(1), pp229.

NISBET R, ELDER J, ELDER JF & MINER G, 2009, Handbook of statistical analysis and data mining applications, London.Nykänen, P. 2000.

SAS Introduction to SAS Enterprise Miner Software, *2011,* [Online], [Cited September 7th, 2011], Available from http://www.sas.com.

SEATTLE GF, 2011, [Online], [Cited August 23th, 2011], Available: http://www. economist. com/blogs/babbage/2011/04/incentive_ prizes

STARFIELD, B, 2000, Medical Errors - A Leading Cause of Death, *Journal American medical association (JAMA.),* 284(4).

STEIG E, 2009, On overfitting, [Online], [Cited September 17th, 2011], Available: http:// www.realclimate. org/index.php/archives/2009/06/on-overfitting/.

TIMOFEEV R & HÄRDLE W, 2004, Classification and regression trees (CART) theory and applications, Unpublished MSc thesis, Humboldt University, Berlin.

WINSTON WL, 2004, Operations research: applications and algorithms, Boston, Massachusetts: Cengage Learning.

APPENDIX A: THE KAGGLE CONCEPT AND THE HERITAGE HEALTH PROVIDER NETWORK

To understand how everything fits together it is firstly important to understand what Kaggle is and how The Heritage Health Provider Network fits in.

THE KAGGLE CONCEPT

Kaggle is a concept based on statistical and analytical outsourcing. It is the leading platform for prediction and data modelling competitions. The Kaggle concept is visually shown in Figure 33: Kaggle concept (Heritage Provider Network Health Prize, 2011).and works as follow (Heritage Provider Network Health Prize, 2011):

- Researchers, governments and other companies present a problem to Kaggle with accompanying datasets.
- Kaggle serves as a platform (via the internet) to expose these problems to the world's best data scientists who compete to find the best solution.
- Once the winning solution has been found prize money is exchanged for the intellectual property behind that winning model.

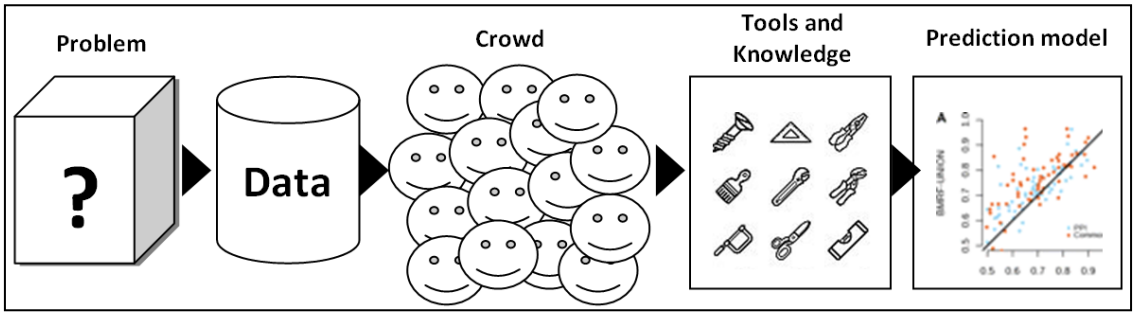

Figure 33: Kaggle concept (*Heritage Provider Network Health Prize*, 2011).

THE HERITAGE PROVIDER NETWORK (HPN)

The HPN is a privately owned organisation that provides health care and health insurance to its members. In doing so the HPN partners with certain medical groups and urgent care centres in Southern California and its affiliates with an expanding number of health care plans. The HPN puts focus on preventative health care services and is among the leaders in the field in its industry. It then makes sense that HPN is a host to this Hospitalisation Prediction Modelling competition (Heritage Provider Network Health Prize, 2011), which will effectively produce a decision support tool which will assist in their focus on preventative health care. Additionally the HPN did not host the incentivised competition for their benefit only, but they also want to achieve radical breakthroughs which they deem "necessary to begin fixing America's health care system" (Heritage Health Prize, 2011). If such a breakthrough can be achieved it will result in the increase of patient health and the decrease in the cost of care.

Health care facilities In the United States are largely owned by the private sector. Most health insurance companies are also privately owned (such as The Heritage Provider Network) with the exceptions of the government owned companies Medicare, Medicaid, TRICARE, the Children's Health Insurance, and the Veterans Health Administration. Health insurance is vital in the American health system, because patients need health insurance to cover expenses, there is not a national system that allows pro bono treatment as in certain other countries. The amount of time spent in a hospital is determined by doctors together with patients, however it is possible that individuals might decide to stay a shorter time than recommended by a doctor, or not go at all, based on the fact that their insurance (or lack there-of) will not cover medical expenses (*Heritage Provider Network Health Prize*, 2011)

APPENDIX B: ORSSA PAPER

APPENDIX C: ORSSA PRESENTATION

APPENDIX E:FULL DATA DICTIONARY

APPENDIX F: EXCEL CODE FOR SUMMARISATION

```
Sub Spec()

'' Spec Macro

'Application.ScreenUpdating = False

Dim Speciality

Dim PlaceSvc

Dim PrimaryConditionGroup

Dim ProcedureGroup

Dim CharlsonIndex

'*******************************************************************************

'Speciality

'*******************************************************************************

For a = 6 To 17

   For b = 0 To 500000

   If Cells(b, 5) <> 0 And Cells(b, 2).Value <> "" And Cells(b, 3).Value <> "" Then

      Speciality = Application.WorksheetFunction.CountIfs(Sheet2.Range("e2:e649520"),
Sheet5.Cells(2, a).Value, Sheet2.Range("a2:a649520"), Sheet5.Cells(b, 1).Value)

     Sheet5.Cells(b, a).Value = Speciality

    End If

   Next b

Next a

'*******************************************************************************

'PlaceSvc

'*******************************************************************************

 For c = 18 To 25

   For d = 0 To 500000

   If Cells(d, 5) <> 0 And Cells(d, 2).Value <> "" And Cells(d, 3).Value <> "" Then

      PlaceSvc = Application.WorksheetFunction.CountIfs(Sheet2.Range("f2:f649520"), Sheet5.Cells(2,
c).Value, Sheet2.Range("a2:a649520"), Sheet5.Cells(d, 1).Value)

      Sheet5.Cells(d, c).Value = PlaceSvc

    End If
```

```
   Next d
 Next c
'*************************************************************************************
'PrimaryConditionGroup
'*************************************************************************************
 For e = 26 To 70
   For f = 0 To 500000
   If Cells(f, 5) <> 0 And Cells(f, 2).Value <> "" And Cells(f, 3).Value <> "" Then
      PrimaryConditionGroup = Application.WorksheetFunction.CountIfs(Sheet2.Range("j2:j649520"),
Sheet5.Cells(2, e).Value, Sheet2.Range("a2:a649520"), Sheet5.Cells(f, 1).Value)
      Sheet5.Cells(f, e).Value = PrimaryConditionGroup
   End If
   Next f
 Next e
'*************************************************************************************
'ProcedureGroup
'*************************************************************************************
 For g = 71 To 87
   For h = 0 To 500000
    If Cells(h, 5) <> 0 And Cells(h, 2).Value <> "" And Cells(h, 3).Value <> "" Then
      ProcedureGroup = Application.WorksheetFunction.CountIfs(Sheet2.Range("l2:l649520"),
Sheet5.Cells(2, g).Value, Sheet2.Range("a2:a649520"), Sheet5.Cells(h, 1).Value)
      Sheet5.Cells(h, g).Value = ProcedureGroup
    End If
   Next h
Next g
'*************************************************************************************
'CharlsonIndex1
'*************************************************************************************
For k = 90 To 93
```

```
    For l = 0 To 500000

     If Cells(l, 5) <> 0 And Cells(l, 2).Value <> "" And Cells(l, 3).Value <> "" Then

        CharlsonIndex = Application.WorksheetFunction.CountIfs(Sheet2.Range("k2:k649520"),
Sheet5.Cells(2, k).Value, Sheet2.Range("a2:a649520"), Sheet5.Cells(l, 1).Value)

        Sheet5.Cells(l, k).Value = CharlsonIndex

     End If

    Next l

Next k

*******************************************************************************

  Application.ScreenUpdating = True

End Sub

'*******************************************************************************

'*******************************************************************************

Sub DrugAndLabCountAndCharl()

For i = 2 To 276100

'*******************************************************************************

'DrugCount

'*******************************************************************************

'Enigste een

'*******************************************************************************

If Sheet3.Cells(i, 1).Value <> Sheet3.Cells(i + 1, 1).Value And Sheet3.Cells(i, 1).Value <> Sheet3.Cells(i -
1, 1).Value Then ' as die enigste een

        DrugCount = Sheet3.Cells(i, 4).Value

    DrugCount = Sheet3.Cells(i, 4).Value

        On Error Resume Next 'soek waar memberID in sheets("Member")

    RyDrug = Application.WorksheetFunction.Match(Sheet3.Cells(i, 1).Value,
Sheet5.Range("a1:a552714"), 0)

    On Error GoTo 0

        Sheet5.Cells(RyDrug, 88) = DrugCount

        Sheet3.Cells(i, 5) = DrugCount
```

```
   Sheet3.Cells(i, 6) = RyDrug

      DrugCount = 0

   'Eerste een

'*************************************************************************************

ElseIf Sheet3.Cells(i, 1).Value <> Sheet3.Cells(i - 1, 1).Value And Sheet3.Cells(i, 1).Value =
Sheet3.Cells(i + 1, 1).Value Then ' die eerste een

      DrugCount = Sheet3.Cells(i, 4).Value

     'Nie eerste of laaste een nie

'*************************************************************************************

ElseIf Sheet3.Cells(i, 1).Value = Sheet3.Cells(i + 1, 1).Value And Sheet3.Cells(i, 1).Value =
Sheet3.Cells(i - 1, 1).Value Then ' nie die eerste een

      DrugCount = DrugCount + Sheet3.Cells(i, 4).Value

   'Laaste een

'*************************************************************************************

ElseIf Sheet3.Cells(i, 1).Value <> Sheet3.Cells(i + 1, 1).Value And Sheet3.Cells(i, 1).Value =
Sheet3.Cells(i - 1, 1).Value Then ' die laaste een

      DrugCount = DrugCount + Sheet3.Cells(i, 4).Value

      On Error Resume Next 'soek waar memberID in sheets("Member")

   RyDrug = Application.WorksheetFunction.Match(Sheet3.Cells(i, 1).Value,
Sheet5.Range("a1:a552714"), 0)

   'On Error GoTo 0

      Sheet5.Cells(RyDrug, 88) = DrugCount

      Sheet3.Cells(i, 5) = DrugCount

   Sheet3.Cells(i, 6) = RyDrug

      DrugCount = 0

   End If

'*************************************************************************************

'LabCount

'*************************************************************************************

'Enigste een
```

```
'*********************************************************************************

If Sheet4.Cells(i, 1).Value <> Sheet4.Cells(i - 1, 1).Value And Sheet4.Cells(i, 1).Value <> Sheet4.Cells(i + 1, 1).Value Then ' die eerste een

      LabCount = Sheet4.Cells(i, 4).Value

   LabCount = Sheet4.Cells(i, 4).Value

      On Error Resume Next 'soek waar memberID in sheets("Member")

   RyLab = Application.WorksheetFunction.Match(Sheet4.Cells(i, 1).Value, Sheet5.Range("a1:a552714"), 0)

   On Error GoTo 0

      Sheet5.Cells(RyLab, 89) = LabCount

   Sheet4.Cells(i, 5) = LabCount

   Sheet4.Cells(i, 6) = RyLab

      LabCount = 0

'Eerste een

''*********************************************************************************

ElseIf Sheet4.Cells(i, 1).Value <> Sheet4.Cells(i - 1, 1).Value And Sheet4.Cells(i, 1).Value = Sheet4.Cells(i + 1, 1).Value Then ' die eerste een

      LabCount = Sheet4.Cells(i, 4).Value

     'Nie eerste of laaste een nie

''*********************************************************************************
ElseIf Sheet4.Cells(i, 1).Value = Sheet4.Cells(i + 1, 1).Value And Sheet4.Cells(i, 1).Value = Sheet4.Cells(i - 1, 1).Value Then ' nie die eerste een

      LabCount = LabCount + Sheet4.Cells(i, 4).Value

   'Laaste een

'*********************************************************************************

ElseIf Sheet4.Cells(i, 1).Value <> Sheet4.Cells(i + 1, 1).Value And Sheet4.Cells(i, 1).Value = Sheet4.Cells(i - 1, 1).Value Then ' die laaste een

      LabCount = LabCount + Sheet4.Cells(i, 4).Value

      On Error Resume Next 'soek waar memberID in sheets("Member")

   RyLab = Application.WorksheetFunction.Match(Sheet4.Cells(i, 1).Value, Sheet5.Range("a1:a552714"), 0)

   On Error GoTo 0
```

```
      Sheet5.Cells(RyLab, 89) = LabCount
   Sheet4.Cells(i, 5) = LabCount
   Sheet4.Cells(i, 6) = RyLab
      LabCount = 0
   End If
Next i
For c = 276101 To 649520
''replace DrugCount blanks met 0
''*****************************************************************************
   If Cells(c, 88).Value = "" Then
   Cells(c, 88).Value = 0
   End If
''replace LabCount blanks met 0
'*****************************************************************************
    If Cells(c, 89).Value = "" Then
   Cells(c, 89).Value = 0
   End If
   'CharlsonIndex
'*****************************************************************************
x = Sheet2.Cells(c, 11).Value
Select Case x
Case Is = "1-2"
Sheet2.Cells(c, 11).Value = 1.5
Case Is = "2-3"
Sheet2.Cells(c, 11).Value = 2.5
Case Is = "3-4"
Sheet2.Cells(c, 11).Value = 3.5
Case Is = "5+"
Sheet2.Cells(c, 11).Value = 5.5
```

```
End Select
    Next c
End Sub
```

APPENDIX G: R CODE FOR RANDOM FORESTS

```
library(randomForest)

temp = read.table("E:/SKRIPSIEFlash_OKT/DATA/Data_7Okt.csv",sep=",",header=TRUE)

X = temp[,3:ncol(temp)]

y = temp[,1]

#####################################################################Unweighted regression

rfr = randomForest(X,y,ntree=100,keep.forest=TRUE,proximity=FALSE)

ypred1 = predict(rfr,X,type="response")

plot(y,ypred1,col="blue",xlim=c(min(c(y,ypred1)),max(c(y,ypred1))),ylim=c(min(c(y,ypred1)),max(c(y,
ypred1))))

lines(c(min(c(y,ypred1)),max(c(y,ypred1))),c(min(c(y,ypred1)),max(c(y,ypred1))),col="red")

#############################################################Classification (three categories)

y1rawind = which(y==0)

ind1 = sample(y1rawind,1000)

N1 = length(ind1)

ind2 = which((y>0)&(y<5))

N2 = length(ind2)

ind3 = which(y>4)

N3 = length(ind3)

ynew = array(0,c(N1+N2+N3,1))

ynew[1:N1] = 1

ynew[(N1+1):(N1+N2)] = 2

ynew[(N2+1):(N1+N2+N3)] = 3

Xnew1 = X[ind1,]

Xnew2 = X[ind2,]

Xnew3 = X[ind3,]
```

```
Xnew = rbind(Xnew1,Xnew2,Xnew3)

 rfc3 = randomForest(Xnew,as.factor(ynew),ntree=100,keep.forest=TRUE,proximity=FALSE)

ypred3 = predict(rfc3,X,type="response")

plot(y,ypred3,col="blue",xlim=c(min(c(y,ypred3)),max(c(y,ypred3))),ylim=c(min(c(y,ypred3)),max(c(y,
ypred3))))

lines(c(min(c(y,ypred3)),max(c(y,ypred3))),c(min(c(y,ypred3)),max(c(y,ypred3))),col="red")

################################################## Classification (two categories)

y1rawind = which(y==0)

ind1 = sample(y1rawind,1000)

N1 = length(ind1)

ind2 = which((y>0))

N2 = length(ind2)

ynew = array(0,c(N1+N2,1))

ynew[1:N1] = 1

ynew[(N1+1):(N1+N2)] = 2

Xnew1 = X[ind1,]

Xnew2 = X[ind2,]

Xnew = rbind(Xnew1,Xnew2)

 rfc2 = randomForest(Xnew,as.factor(ynew),ntree=100,keep.forest=TRUE,proximity=FALSE)

ypred2 = predict(rfc2,X,type="response")

plot(y,ypred2,col="blue",xlim=c(min(c(y,ypred2)),max(c(y,ypred2))),ylim=c(min(c(y,ypred2)),max(c(y,
ypred2))))

lines(c(min(c(y,ypred2)),max(c(y,ypred2))),c(min(c(y,ypred2)),max(c(y,ypred2))),col="red")
```

APPENDIX H: CHI TEST CALCULATIONS

CART regression – chi test calculations

Bin	Frequency Obs Oi	Frequenc y Pred Ei	Ei'	Oi'	(Oi'-Ei')^2/Ei'
0	94	94	94	161	47.7553191
1	49	0			
2	18	0			
3	10	83	83	17	52.4819277
4	7	0			
5	5	21	21	20	0.04761905
6	2	0			
7	4	0			
8	4	0			
9	1	0			
10	2	0			
11	0	0			
12	0	0			
13	0	0			
14	0	0			
15	2	0			
	198	198			

Xcalc= 100.284866
v= k-m-1 2
Xcrit= 5.99

Ho: predicted values are following the same distribtion as that of t
a= 0.05
Xcalc<Xcrit then Ho is rejected

MARS– chi test calculations

Bin	Frequency Obs Oi	Frequency Pred Ei	Ei'	Oi'	(Oi'-Ei')^2/Ei'
0	184	0	182	194	0.791208791
1	10	182			
2	5	10	10	5	2.5
3	0	3	9	2	5.444444444
4	0	3			
5	1	2			
6	0	0			
7	0	0			
8	0	0			
9	0	0			
10	0	0			
11	1	1			
12	0	0			
13	0	0			
14	0	0			
15	0	0			
	201	201			

Xcalc= 8.735653236
v= k-m 2
Xcrit= 5.99

Ho: predicted values are following the same distribtion as that
a= 0.05
Xcalc>Xcrit then Ho is rejected

Random Forests regression – chi test calculations

Bin	Frequency of observed (Oi)	Frequency predicted (Ei)	Ei'	Oi'	(Oi'-Ei')^2/Ei'
0	6586	6728	6728	6586	2.997027348
1	487	507	507	487	0.788954635
2	240	176	176	240	23.27272727
3	153	144	144	153	0.5625
4	97	75	75	97	6.453333333
5	66	58	58	66	1.103448276
6	42	36	36	42	1
7	19	16	16	19	0.5625
8	23	20	20	23	0.45
9	17	16	16	17	0.0625
10	13	8	8	13	3.125
11	8	8	8	8	0
12	7	3	9	49	177.7777778
13	6	6			
14	3	0			
15	33	0			

Xcalc=	218.1557686
v= k-m-1	12
Xcrit=	21.03

Ho: predicted values are following the same distribtion as that of the

a= 0.05

Xcalc>Xcrit then Ho is rejected