Data Mining Assignment 1: Data Mining Survey

Part 1: Survey

Introduction to Data Mining

Data Mining is a set of computational techniques that involves pattern discovery in large sets of data or specifically "the extraction of hidden predictive information from large databases" (Thearling.com, 2018). Specific techniques classed as data mining include combinations of machine learning, statistics and databases. The primary objective of data mining is to extract useful information from datasets in the form of a useful pattern or structure. Examples of these useful results are identified clusters from cluster analysis, anomalous data within a dataset, association rules that tie together elements in a system, statistical patterns in sequences and predictive analytics.

Data mining techniques coupled together in a business context allow the prediction of future trends and behaviours which allow managers to make proactive, data-driven decisions. Automation with modern software allows time-consuming analyses to be performed. More complex business questions can be asked, with questions pre-datamining being highly general using only past information and post-datamining extremely specific using predictions on time series data. For example, in the 1960's a question of "What was the total revenue for the Llanishen branch of Morrisons over 5 years?" would have been asked, whereas in 2018 it's possible to be as specific as "Within the Llanishen branch of Morrisons, what is the 6-month span estimated mean number of sales per week of beef joints and vegetable stock bought in the same basket?". Additionally, the 1960's question would require much man-hours of work to provide the result, whereas a modern computer system can supply the 2018 answer near instantly if suitable datamining techniques have been implemented. Due to the efficiency in prediction and rapid return of information, datamining is quickly becoming essential in the contemporary business environment.

Modelling Methods

Anomaly Detection

Anomaly detection involves identifying outlying or anomalous data points within a dataset which may be an interesting element that needs extra attention or an error in methodology or data collection that needs addressing. Anomalies may need to be eliminated for techniques such as K-Means clustering which need zero outliers for accurate results.

Detection can use machine learning on datasets (Allerin.com, 2018) and be implemented with supervised or unsupervised learning techniques.

Association

Association or association analysis involves simple correlation between several data items, usually of the same type or category. Correlation identifies patterns and relationships between items, with applications in the retail market. Customer buying habits can be tracked to advertise familiar products, such as a customer who typically buys milk being informed that they may have forgotten to buy milk on checkout. This association would be between a customer entity and a set of items, with milk as one of those items. The relationships between similar items is expressed as an association rule.

Classification

Classification is the establishment of descriptive attributes that can be applied to a class of customer, item or object. It's used when introducing a novel piece of information, finding closely similar information and effectively classifying said information into the correct class. As one example, a newly manufactured car may have its attributes analysed to properly categorise it's type. It is useful for vague information or situations where there are blurred lines between categories of item. Classification can also use other modelling techniques as part of its function, pre-processing data it uses or as the result of a decision tree.

Clustering

Clustering or cluster analysis is used to obtain a structure from individual pieces of data. Clusters are excellent for visual presentation of similarities in data, with different items being identified from different clusters. There are different types of clustering such as hierarchical or connectivity-based clustering, centroid based clustering, distribution-based clustering and density-based clustering.

Decision Trees

Decision trees are used in selection criteria for data. It takes the form of answering a question leading to another based on the answer until the end of the tree. This classifies data into separate, specific areas based on attributes queried.

Prediction

Prediction is used to predict future events or trends using past and/or present data. Data is collated over a given length of time and techniques are used to establish a trend from the data. This allows a business to predict sales growth or an increase/decrease in users of an application as two examples. Not only time series data is used, prediction can also be used for estimating the success/failure rates or the chance of an event occurring.

Regression

Regression or regression analysis is a very common statistical technique used to estimate relationships between data points. It is used for prediction, particularly in machine learning. Specific methods of regression include linear regression, ordinary least squares, nonparametric regression and metric regression.

Sequential Patterns

Sequential patterns are a method used for trend identification or tracking the regular frequency of related events. Much like with predictive methods, time series data is often used. A notable feature is finding missing data through observing similar patterns and noting missing values in each set of data.

Summarisation

Summarisation within data mining is an attempt to give a user an overview of the structure of the data. A simple summary can be the objective of a data mining project, for example the total sales for a given month for a business. Typically, summarisation is an early, lesser part of the overall process used to provide insight into the nature of the initially gathered data or to present the final results.

Data Mining Tools

Rapidminer (YALE)

A java-based program, Rapidminer (previously known as YALE) is a software as a service (SaaS) analytic tool that allows pre-processing, visualization of data, predictive analytics and statistical modelling. It is open source under the AGPL open source license and available free for download, although a commercial version is available for business. An online poll conducted (KDnuggets, 2013) found that Rapidminer was the most popular data analytics/big data/data mining tool in use in 2013.

A GUI interface allows the design and execution of analytical workflows called processes, consisting of multiple operators. An operator is responsible for a single task in the process and the output of an operator feeds into the next operator.

The engine can be used as an API or called directly from another program in use. Command line functionality can be used for single use functions and the program can be extended with R and python scripts.

WEKA

Another java-based tool, WEKA (Waikato Environment for Knowledge Analysis) is capable of pre-processing, clustering, classification, regression, visualisation, data analysis and predictive modelling. It notably lacks sequence modelling. It is entirely free and covered by the GNU General Public License allowing full customisation.

All the techniques in WEKA assume that data is in the form of one flat file or relation, with every data point detailed by a set number of attributes. WEKA can access SQL databases through Java to process information obtained. Using Deeplearning4j, an open source deep learning programming library written in Java, WEKA can use deep learning.

R-Programming

Project R is a GNU project primarily written in C and FORTRAN that uses considerable amounts of R language programming for its modules. Due to its ease of use and its ability to be easily extensible to accommodate a large variety of projects it has seen more use in recent years.

Orange

Python-based and open source, Orange has components for machine learning, bioinformatics and text mining. One of the major selling points of the program is that it has extensive visual programming and visual display of information. Widgets, either pre-defined or user created are used to link together workflows. Open source under the GNU General Public License, core components are written in C++ with python wrappers. Orange takes advantage of python open source libraries such as numpy, scipy and scikit-learn. The program runs within a Qt framework, an open source application framework that can be run over many software and hardware platforms.

KNIME

KNIME, written in Java, is an open source data analytics, reporting and integration tool. It uses a GUI to help the user in the assembly of nodes during pre-processing consisting of three phases; Extraction, Transformation and Loading. It does this without or with very little programming from the user. It has been used for pharmaceutical research (Tiwari and Sekhar, 2007), business intelligence and financial data analysis.

SPMF

SPMF is an open source data mining library written in Java, specialised in pattern mining. Three common algorithms used in sequential pattern mining are Generalised Sequential Pattern (GSP), SPADE and Prefix Span, also known as prefix-project sequential pattern mining (Verma and Mehta, 2014). Using SPMF, it was found that Prefix Span has better performance than GSP and SPADE, with SPADE faster than GSP. Additionally, Prefix Span was found to use less memory than GSP and SPADE which used similar amounts.

Dundas BI

Dundas BI is a browser-based business intelligence and data visualisation platform that provides a user with the ability to create their own interactive dashboards, reports, queries and analyses. The platform has an API for .Net, REST and JavaScript.

Anaconda

Anaconda is an open, data science platform programmed in Python. The open source version uses both Python and R languages and comes equipped with more than a 100 highly used Python, R and Scala modules used in data science.

CRISP-DM

The de-facto standard for developing a data mining process is a data mining process model, CRISP-DM, also known as the "Cross-industry standard process for data mining".

The development of CRISP-DM began in late 1996 (CRISP-DM 1.0, 2000) through the work of three businesses; Daimler-Benz, SPSS (Statistical Package for the Social Sciences) and NCR (National Cash Register). Each of the companies had developed data mining technologies but agreed to work together to create a standardised process model in the hopes of marketing data mining to prospective customers in an emerging market. The three companies formed a consortium called CRISP-DM, obtained funding from the European Commission and began work on drafting the process model. To gain input from practitioners of data mining, the consortium created a special interest group and hosted a day in Amsterdam to gather views and ideas on how to create a standardised process model. The response was much better than expected and over the next two and a half years, the consortium worked to develop and refine CRISP-DM. Trials were run at Mercedes-Benz and CRISP-DM was integrated into commercial data mining tools. With the end of the European Community funded part of the project in 1999, the consortium had produced a good initial draft and approximately one year later the CRISP-DM 1.0 was published.

The methodology of CRISP-DM is described as a hierarchical process model, consisting of six phases. These phases are divided into generic tasks which are in turn split into a set of specialised tasks. The specialised tasks are split into process instances which are specific instances or records of the actions, decisions and results of data mining. Ultimately this division into sub levels with tasks to be performed in a specific order is an idealised approach, it can be necessary to repeat certain tasks, which is an approach taken by the more refined ASUM-DM model which was designed in part to account for this.

CRISP-DM separates the data mining process into six distinct phases:

Business Understanding

This phase consists of four generic tasks, with this phase centred around a project management perspective that works to setup the project fully. Project managers with oversight from senior stakeholders are the primary resource involved with this phase.

The first generic task is determining the business objectives that consists of:

- Documenting the background of the project
- Determining the actual objectives of the business undertaking the project
- Defining the success criteria for those objectives having been completed

The second generic task, assessing the situation involves:

- Taking an inventory of project resources
- Recording the project requirements, assumptions made and the constraints on the project
- Conducting risk management to assess the risks and contingencies for potential issues
- Defining terminology within the project
- Detailing the costs and benefits of the project.

The third generic task, determining the data mining goals has the following outputs:

- Establishing the data mining goals for the project
- Defining the success criteria for the data mining work

The fourth generic task for business understanding, produce a project plan is composed of:

- Producing a project plan
- Making an initial assessment of tools and techniques to be used during the project

While not mentioned within the final generic task within the list of outputs in the CRISP-DM guide (CRISP-DM 1.0, 2000), it is mentioned elsewhere that a problem definition is made at this point.

Data Understanding

This phase is based on the collection of data for the project. Initial work on the data may lead to the first insights into patterns or subsets of the data that be prove useful in the future. Quick wins taken from this insight may be achieved here for the business if suitable. The primary resource in this phase should be data scientists.

This first generic task is collecting the initial data which consists of performing the initial data collection and then producing a report on the results of this initial gathering.

The second task, describing the data involves formally documenting a description of the data in an associated report.

The third task is exploring the data, taking a deep dive into the data gathered to in part, look around at the quality of the data, gain familiarity with the dataset and to find anomalous data. The results of this are recorded in a formal report.

The final generic task is verification of the data quality and then documenting the quality of the data in a data quality report. Issues with the quality of data should be resolved at this point.

Data Preparation

Data Preparation consists of five generic tasks based on getting pre-processing data for the modelling phase. Data scientists are the primary resource.

The first task, data selection consists of the data being selected based on defined rationale for the inclusion and exclusion of given data from the dataset used.

The second task, clean data involves preparing data from its raw form into a form that can be accurately read into the modelling techniques without issue. Corrupt and inaccurate records are removed at this point in addition to typographical errors. A data cleaning report is created once this task is complete.

The third task is constructing data, the putting together of data into its useable project specific form using derived attributes and generated records.

Fourth is the integration of data, where data is merged into a larger dataset. Data rationalisation, a process involving fitting together heterogeneous groups such as files with similar but different column headers occurs here.

Fifth, formatting of data is where data is formatted correctly ready for use in the modelling process. Prefixes and suffixes may be added or taken away, datatypes may be changed into another form such as integers into strings and the data is transformed from a rawer form into a form ready to be directly read into a modelling technique.

Once the tasks are complete, the dataset to be used is explicitly defined in terms of scope with a dataset description noted.

Modelling

The modelling phase is where the actual implementation of data mining takes place using the data that has been prepared in the previous phase. As with the previous two phases, data scientists should be the main resource used.

The first generic task is the selecting of the appropriate modelling techniques, documenting the techniques to be used and then stating the assumptions to be made in using said techniques.

The second task is the generation of a test design or designs for recording results later.

The third task is building the model. The model first needs to have its parameters set, modelling techniques implemented within the grander model and then descriptions of the model noted. A variety of modelling techniques are applied with their use properly calibrated. If necessary, further data preparation is used to tailor the dataset for certain techniques. This part of the process can be iterative, until the model is of high enough quality.

The final task is assessing the model. The model is given a full assessment with parameters revised as necessary.

Evaluation

Evaluation is where the data mining process is evaluated fully. The primary point of this phase is to determine if there are any business issues that have not been properly addressed. With the end of this phase, a decision on how to use the results of the data mining should be finalised. Resource wise, data scientists should present the results to project managers who then fully document the outcome of the data mining process.

The first task is to undertake a full evaluation of results with the assessment of data mining results with regards to the business success criteria and the recording of the approved models used.

The second task is, reviewing the data mining process.

The third task is to determine the next steps. A list of possible actions and decisions to take the project forward is drawn up.

Deployment

The final phase of CRISP-DM, deployment consists of the results of the model being created, being transformed into useful, organised and presented in a customer focussed manner. This phase may range from a simple report to iterating through a big data, data mining process through a full business cycle. The customer is often the one to deploy the model and so they need to be aware of the actions needed to implement it. The primary resource for this phase is project managers due to the reporting of the project's results and project closure.

The first generic task is planning the deployment with a deployment plan.

The second task, plan monitoring and maintenance, consists of developing a monitoring and maintenance plan to equip future users of the data mining model to properly maintain the model if it is to be in use into the future. Much of the time, a successful model is retained for possible future use.

The third task is the production of a final report and a final presentation which fully summarise results and end the project. The presentation should be aimed towards senior stakeholders and owners of the project.

The final task is the reviewing of the project with the experience gained from the project fully documented.

ASUM-DM

A recently developed process model for implementing a data mining or predictive analytics project, ASUM-DM (Analytics Solutions Unified Method-Data Mining) is according to Haffar:

> the "Analytical" activities and tasks of CRISP-DM but the method was augmented with missing activities and tasks as well as templates and guidelines. In other words ASUM-DM is nothing more but an extended and refined CRISP-DM. (Haffar, 2018).

The main reason for the creation of ASUM-DM is given by Haffar as CRISP-DM not covering the infrastructure and operations side of implementing a project. Project management tasks you'd expect to see in a data mining project are notably absent by and large within CRISP-DM. While the external version of ASUM-DM is available publicly, it is obvious from a cursory read through of the documentation that the process model is geared towards IBM internal processes, it is not written in the generalist form that CRISP-DM is. The "Set up Environments" stage is particularly geared towards IBM internal processes with almost the entire stage focussed on the installation of IBM software and the setting up of QA and production teams of IBM staff. Even so, if the documentation is suitably adapted to an organisation, ASUM-DM would be a better view of a data mining project plan from a project manager's or senior stakeholder's perspective. Going further, the style of the documentation is written in a manner reminiscent of Microsoft Project.

Unlike the six phases of CRISP-DM, ASUM-DM has three phases in part because the initial phase is repeatable:

Analyse, Design, Configure and Build

This phase is repeatable as data mining/predictive analytics projects are iterative in nature, according to IBM (IBM Analytics Solutions Unified Method (ASUM) - External, 2015).

There are fourteen stages within this phase:

- Prepare for Implementation
- Conduct Readiness Assessment
- Conduct Project Kick-off
- Understand Business
- Understand Data
- Design and Validate Infrastructure
- Set up Environments
- Prepare Data
- Build Model
- Evaluate Model
- Conduct Analytical Knowledge Transfer
- Define Deployment Approach
- Design Operational Testing Strategy
- Validate and Test in QA Environment

It is easy to see that compared with CRISP-DM, ASUM-DM combines the first five phases of CRISP-DM into its first phase. Everything except the deployment of the model is performed here. The

biggest differences are the addition of project-management oriented tasks which make up a large part of this phase.

Deploy

Deployment consists of six stages:

- Conduct Operational Knowledge Transfer
- Prepare for Ongoing Maintenance
- Deploy Solution
- Transit to IBM Support
- Launch
- Prepare for Project Closure

This phase is notably focussed on IBM systems, particularly in the sub tasks for each stage that I have not listed. That considered, the ASUM-DM approach is more geared towards the deployment of data mining in a business context.

Operate and Optimise

Operate and Optimise consists of five stages:

- Monitor Model
- Operate, Optimise and Improve System
- Support User Community
- Manage Infrastructure
- Govern System Lifecycle Program

As compared to CRISP-DM, this final phase has greater support for the long-term lifecycle of a project beyond the traditional project closure.

Applications and Problem Types of Data Mining

Automatic Credit Approval

Automatic credit approval has been identified as an application where data mining is used within the context of the banking sector (Chitra and Subashini, 2013). Fraud is prevented during the approval process using classification models based on decision trees, support vector machines and logistic regression techniques.

The two decision trees used within this study were C5.0 and CART, which both serve to act as classification models when selecting someone for credit approval.

C5.0 creates a decision tree from training data, splitting sets of data into subsets at each node of the tree based on the criteria of the normalised information gain measured as a difference in information entropy. An attribute with the highest information gain is used to make decisions within the context of a decision tree. While a little difficult to understand initially, this type of decision tree is used because small decision trees result from this method, meaning few questions that need to be asked.

CART (classification and regression tree) in comparison uses a binary decision tree based on splitting data into two child nodes at each level, starting with the root node that contains the whole set of the training data. A calculation used within the algorithm to generate this tree is Gini impurity, the measure of how often a randomly chosen element is correctly labelled if it is randomly labelled, relative to the distribution of labels in the subset. It is calculated by summing the probability of each item chosen times the probability of a mistake being made in categorising it.

Another technique used, a support vector machine (SVM) is used in machine learning as a polynomial kernel, a kernel function. The paper (Chitra and Subashini, 2013) is of questionable value in describing this technique as a standalone paper due to the Wikipedia page describing polynomial kernels being identical and almost confirming direct plagiarism by the authors in a simple copy and paste operation. In any case, a SVM represents the similarity of vectors in feature space over polynomials of the original variables which allows the machine to learn non-linear models. The paper is unclear on how the results are gained from these calculations.

A final technique described is logistic regression which is a type of regression analysis use for predicting the outcome of a categorical dependent variable based on one or multiple predictor variables. Logistic regression is described as easy to implement with good performance from a computational perspective.

The application of these techniques varies between different banks, ultimately the result of a decision tree will give the bank the yes or no answer when an application for credit is made.

Automatic Text Summarisation

Automatic text summarisation is the summarisation of a text document to create a shorter, more condensed version that contains the main points in the original document. There are two approaches to this summarisation, extraction and abstraction. Extraction works by picking out existing elements from the text to create the summary. Abstraction works by making a semantic representation of the document and using natural language generation to create a summary in a human-like style.

One primary example of text summarisation is key phrase extraction. According to Turney (Turney, 2000), journals ask authors to provide a list of keywords for their articles. Key-phrase extraction is supposed to serve the goals of:

- Enabling the reader to quickly determine whether the given article is in the reader's field of interest
- To be used in indexing to enable the reader to find a relevant article
- For search engines that use keywords, giving more precise results

Automatic key phrase extraction is a more general task of automatic key phrase generation where the key phrases generated do not necessarily even appear within the document used but 75% of generated key phrases do appear somewhere in said document. Key phrase extraction algorithms have been in use for decades, notably in Microsoft Word 97 and what are now defunct software products such as Metabot and Verity Search 97 in use at the time of writing. Within Word 97, keywords are automatically selected and set as the key words for the document within the document metadata.

Customer Relationship Management

Customer Relationship Management, also known as CRM involves management of the relationship between current customers, potential customer and a company. Using data mining techniques, a customer's history is used to drive retention efforts and feed sales growth.

CRM began in the 1970's with customer surveys on the front line or through annual surveys (Financesonline.com, 2018). It later evolved into using databases of customer information and statistical methods of analysis starting in 1982 with Kate and Robert Kestnbaum. By 1995, CRM as a phrase first entered the public domain with credit going to either Tom Siebel of Siebel Systems who first developed a CRM-like program or IBM. In 2003, Microsoft rolled out Dynamics CRM and in 2004 the first open source CRM was developed by SugarCRM. In the late 2000's connections were made to cloud computing and social media. In the years since CRM has used "business intelligence" to perform data driven decisions within companies, culminating in industry specific implementations of CRM.

A typical CRM is composed of data warehouse technology to provide the data used and software as a service (SaaS) CRM software. The specific role being determined by its application in operational use, analytical use, collaborative use or as a customer data platform. Operational use of CRM includes sales force automation, marketing automation and service automation. Analytical CRM systems analyse customer data and present it for business managers to make data-driven decisions. Data mining features prominently in selecting, extracting, processing and then presenting the data to power these decisions. Collaborative CRM systems are used to integrate external entities such as suppliers and distributors into a larger network to share customer information. A customer data platform is a system that collects data on individual people and builds up a large database which can be accessed from external software.

Fraud Detection

According to a BBC news article (BBC, 2016), fraud costs nearly £3000 per head of population with total fraud coming in a cost of £192 billion. The study cited by BBC (PKF, 2016), states that UK fraud by sector works out to proportions of 74% (144bn) in the private sector, 20% (37.5bn) in the public sector, 5% (10bn) by individuals and 1% (1.9bn) in the charity sector. The report goes further to say that in the private sector:

> It is estimated private sector fraud could cost the UK economy up to £143.6 billion. But further analysis suggests that may be a conservative figure, given the general sentiment among our biggest businesses against releasing commercially sensitive, or potentially

> damaging, financial fraud data. Right now no comprehensive data exists in the public domain in the UK. (PKF, 2016)

With the use of AI techniques as an emerging field, data mining is a major component of detecting where fraud occurs. Classification, clustering, segmentation and the finding of associations and patterns within the data where fraud is present is performed. Such patterns and rules found with data mining can then be used within machine learning techniques and/or neural networks to more automatically find cases of fraud in a set of data.

Healthcare

Data mining has potential to identify inefficiencies and best practices that improve care and reduce costs according to health catalyst (Health Catalyst, 2018). Some estimates give as high as 30% of overall healthcare spending being reduced consequently. Due to the complexity of healthcare and a slow rate of technological adaption, healthcare has been slow to take up data mining practices that have been used elsewhere in business and finance.

According to Koh et al. (Koh and Tan, 2005), the vast potential for data mining in healthcare can be achieved through many fields:

- Evaluation of treatment effectiveness
- Management of healthcare
- Customer Relationship Management
- Detection of fraud and abuse
- Predictive Medicine

The effectiveness of medical treatments can be evaluated through the comparing and contrasting causes, symptoms and courses of treatments. United Healthcare, the largest healthcare company in the world by revenue has used data mining on its treatment record data to explore ways in which to cut costs and deliver better medicine.

To improve healthcare management, data mining can be used to identify and track both chronic disease states and high-risk patients and reduce the total number of hospital admissions. Blue Cross Blue Shield, an American health insurance federation has used data mining to reduce spending by having better disease management using emergency department, hospitalisation claims data, pharmaceutical records and physician interviews to help identify unknown asthmatics and intervene where necessary.

Customer relationship management (CRM) as mentioned previously as an application of data mining can be applied also to healthcare. By determining the preferences, usage patterns and current and future needs of patients to improve their satisfaction, customer relationships can be improved.

A notable case of fraud and abuse detection occurred in 1998 when the Texas Medicaid Fraud and Abuse Detection System recovered $2.2 million and identified 1400 suspects in less than one year's operation. This data mining system in 2004 according to Secure Tech Alliance (Securetechalliance.org, 2006), later used data from the Smart Card Program, a program aimed at combatting fraud. It utilised data taken from verification of patients attending appointments, time stamped data signatures with verification of duration of visits and helped combat identify fraud directly by using biometric cards.

Part 2: Scenario

Introduction and Assumptions of Scenario

As specified in the coursework brief, the scenario given is to integrate a scan-as-you-shop technique with a payment application available on a mobile phone or wearable device. Such device has a potential to offer additional services beyond payment to include the contextual presentation of targeted marketing campaigns based on the consumer's behaviours and preferences.

Basic assumptions I will make about the scenario are:

- The user of a device will use a mobile phone with a store specific application in use
- Items to be bought in the store will be scanned using a barcode reader that utilises the phone's built in camera
- This additionally assumes the phone must have a camera installed to operate
- Once scanning items is complete, the user will pay at a dedicated checkout
- The scenario will be modelled on a fictional store called "ScanSave" using real world examples to draw inspiration from
- The data from a completed shop will be collected and stored after each completed shop on a central head office system

Examples of Scan-As-You-Shop Technology

Asda has working scan as you shop technology (Asda, 2018) with the website specifying a three-step process. The first step involves signing up on an instore touchscreen device, entering details and then picking up a store owned handset. After initial signing up, future trips only require a registered phone number to be entered to use a handset. The second step is the scanning and bagging stage in which items selected by the customer are scanned using the device scanning item barcodes and the items are bagged. Step three is the payment stage, in which the customer heads over to a dedicated "Scan & Go" checkout area. There they scan the QR code on the screen using their handset and then present the handset's barcode to the checkout. This transfers payment information and the customer is prompted for payment.

Tesco (Tesco, 2018) has a similar methodology to Asda with only one minor difference. The customer logs in with a Clubcard instead of a phone number.

Waitrose has another example system and according to Brown (Brown, 2015), a system that identifies items that are taken from the store but not paid for. Sainsburys (Sainsburys, 2018) is yet another system that seems to follow the same trend in use with the exception that a mobile app can also be used.

Recommendation of Data Mining Strategy

Due to data protection law, any record keeping would only apply to app users who would sign up with the record keeping outlined in the Terms and Conditions.

Association will be used to link together store items bought using correlation. Specifically, I would rank other items with a variable that acts as a level of association with other items. Tesco has approximately 90,000 items (Wood and Butler, 2015) in stock-keeping-units which are individual items with a separate listing for every size difference, flavour or other variation that makes a product not explicitly 100% identical. To get one variable for every other item, you would have 90,000 (or n-1) variables for each item which gives us a total of 8,099,910,000 variables. Items bought in the same shopping run would contribute to an increase in the correlation between items.

Such correlation would be calculated monthly, using data from all shops performed within the last month. The result of this association finding would be to provide suggestions to a customer to add a certain item based on their current basket at various points throughout their shop in non-intrusive way. For example, if a customer bought bread and eggs with a common associated item being milk, a small window in the corner of the app would suggest to the customer that they might also want to purchase milk.

Categorisation would be used to classify specific groups of items. Attributes such as the composition or use of the item would be used, for example diary products all containing a large proportion of milk. Such classifications would be used to help separate items into aisle specific areas and to be of further use in other data mining techniques such as prediction.

Prediction would be used through accessing the purchase records over the previous month, day by day to build up a picture of current trends in purchasing specific items. Such trends would be used to track the sales of goods in two ways. First, the broad large-scale trends of categories would be used as information to be fed to senior stakeholders, showing the performance of major categories of goods and their trajectory. Second, prediction would be used store by store to assess the local area's buying habits into the future, using the current trends of products to help estimate the quantity of future goods to be shipped. Effectively this would generate the trends of broad categories and the trends of individual items store by store.

Given current technology available to a large supermarket chain, accessing and using data of the size required for all the data mining techniques outlined is trivial as shown by Discord (Discord Blog, 2017), where the more than 131 billion messages contain anywhere from 1 to 2000 characters and result in a retrieve time of about 5 milliseconds.

Implementation Under CRISP-DM

In this section I have detailed the implementation of a project to be conducted in the near-term. To do this I have used the CRISP-DM framework, with an emphasis on the outputs produced.

Business Understanding

Background

ScanSave is a major supermarket chain with a 7% market share of UK supermarket sales. With many competitors embarking on scan as you shop technology, a decision has been made to develop a scan as you shop mobile application for use on mobile phones for customers to use. Associated with the use of this app, common data mining techniques will be used to enhance the performance of the company in marketing to consumers and helping predict changes in the supply chain. Successful results from a basic set of data mining techniques implemented will result in further techniques used after the initial data mining project is completed.

A set of six stores in South Glamorgan will be used for a trial run of the technology over a 2-year period, followed by a roll out of the technology to the broader area of South Wales and if successful there, a nationwide roll out.

Business Objectives

The primary objective is the integration of scan as you shop technology into stores with associated app-based data mining techniques working correctly. There are two areas this project will be focussing on; customer suggestions based on the purchasing of related goods by all customers and the increase of supply chain efficiency in reducing waste by more accurately estimating the trend of goods to be purchased in store.

Business questions to be answered:

- Does the app help point customers towards items they may not have bought without a prompt?
- Does the app help predict changes in the supply chain ahead of time?

Business Success Criteria

Success will have occurred if the following criteria is met:

- 2% of customers into stores using the technology use the app to purchase goods
- Prediction of future trends results in more than £4k in savings a month per store

Inventory of Resources

A team of ten project managers will be assigned to run the project with a team of six data scientists working to implement and use data mining techniques. Data collection will be made available via the Data Store Management team at the central office. The installation of hardware will be outsourced to a subcontractor, as well as the regular maintenance of the machines.

Access to fixed data taken from the application use will be made available monthly. Live data will not be used.

Requirements, Assumptions and Constraints

As a requirement the data mining tools would require access to relevant central office data, made available on the 1st of every month consisting of the previous month's data. An automated tool would need to be developed in order to quickly process a new dataset supplied to the system every month.

A constraint is the project needing to be run for several months before significant data can be gathered which limits measurement of the project in the first months of operation. Another constraint will be that the business objectives will have to be assessed after the deployment. Business results cannot be generated without the data mining techniques exhibiting real world tangible results.

Risks and Contingencies

The primary risk to the project is a low take up rate by customers with the result that the technology is fully in place but not utilised. A contingency will be prominent advertising used in store at the six locations and a 10% discount on the first shop performed using the app.

A secondary risk is the weakness of predictive techniques in optimising the supply chain efficiency due to reduced spoilage, this has been strongly mitigated using a low estimate of £4k per store per month. Spoilage of food items is estimated at just over 10% of produce by weight (Buzby et al., 2015) and given an average revenue of just over £514k per Tesco store per month in the UK (Tesco plc, 2018) as an example, £4k is a conservative estimate of any benefit to be had at 0.78% of store revenue.

Costs and Benefits

Initial costs are placed at approximately £180k for hardware installation, covering the entire sub-project of installing new hardware at the six stores. Continuing costs are the costs of the project staff which is given as £35k per staff member a year, equating to £46,666.67 a month, the maintenance of hardware at £1k a month and software licensing at £1k a month.

Benefits are savings made from cutting staff at the six stores with ten staff paid each £15k a year, to be let go from each store and an increase in supply chain efficiency due to improvements in predicting the flow of goods. This is given as £4k per store per month in reduced spoilage, for the most part from perishable goods. Given that data will need to be in place for a trend to be formed, the first 3 months will be unable to use prediction techniques, with a trend established after 3 months data is gathered.

A detailed two-year forecast is shown below:

Income	**Mar-18**	**Apr-18**	**May-18**	**Jun-18**	**Jul-18**	**Aug-18**	**Sep-18**	**Oct-18**	**Nov-18**	**Dec-18**	**Jan-19**	**Feb-19**
Savings From Less Staff									£75,000.00	£75,000.00	£75,000.00	£75,000.00
Increase Supply Chain Efficency									£24,000.00	£24,000.00	£24,000.00	£24,000.00
Total Income Cash Flow	£0.00	£0.00	£0.00	£0.00	£0.00	£0.00	£0.00	£0.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00
Outgoings	**Mar-18**	**Apr-18**	**May-18**	**Jun-18**	**Jul-18**	**Aug-18**	**Sep-18**	**Oct-18**	**Nov-18**	**Dec-18**	**Jan-19**	**Feb-19**
Project Staff Costs	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67
Hardware Costs	£180,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00
Software Costs	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00
Total Outgoings Cash Flow	£227,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67
Total	**Mar-18**	**Apr-18**	**May-18**	**Jun-18**	**Jul-18**	**Aug-18**	**Sep-18**	**Oct-18**	**Nov-18**	**Dec-18**	**Jan-19**	**Feb-19**
Net Monthly Cash Flow	-£227,666.67	-£48,666.67	-£48,666.67	-£48,666.67	-£48,666.67	-£48,666.67	-£48,666.67	-£48,666.67	£50,333.33	£50,333.33	£50,333.33	£50,333.33
Net Project Profit	-£227,666.67	-£276,333.33	-£325,000.00	-£373,666.67	-£422,333.33	-£471,000.00	-£519,666.67	-£568,333.33	-£518,000.00	-£467,666.67	-£417,333.33	-£367,000.00

Fig 1 – First Year Cash Flow Estimate

Income	Mar-19	Apr-19	May-19	Jun-19	Jul-19	Aug-19	Sep-19	Oct-19	Nov-19	Dec-19	Jan-20	Feb-20
Savings From Less Staff	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00	£75,000.00
Increase Supply Chain Efficency	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00	£24,000.00
Total Income Cash Flow	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00	£99,000.00
Outgoings	**Mar-19**	**Apr-19**	**May-19**	**Jun-19**	**Jul-19**	**Aug-19**	**Sep-19**	**Oct-19**	**Nov-19**	**Dec-19**	**Jan-20**	**Feb-20**
Project Staff Costs	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67	£46,666.67
Hardware Costs	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00
Software Costs	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00	£1,000.00
Total Outgoings Cash Flow	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67	£48,666.67
Total	**Mar-19**	**Apr-19**	**May-19**	**Jun-19**	**Jul-19**	**Aug-19**	**Sep-19**	**Oct-19**	**Nov-19**	**Dec-19**	**Jan-20**	**Feb-20**
Net Monthly Cash Flow	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33	£50,333.33
Net Project Profit	-£316,666.67	-£266,333.33	-£216,000.00	-£165,666.67	-£115,333.33	-£65,000.00	-£14,666.67	£35,666.67	£86,000.00	£136,333.33	£186,666.67	£237,000.00

Fig 2 – First Year Cash Flow Estimate

It is estimated that the project will turn a profit 11 months after deployment of the system, with the long term savings estimated to be £100k per store a year.

Data Mining Goals

The first primary goal of the data mining is to successfully market goods to consumers during a shop based on purchases they are making.

The second primary goal is to predict the flow of goods through stores, allowing the supply chain to be adjusted to reflect the predictions made.

Data Mining Success Criteria

The strict criteria of success are given as:

- Suggestions made to a customer result in a purchase of the suggestion 5% or more of the time
- Savings made from supply chain prediction is at least £4k a month per store or alternatively, senior stakeholders are happy with the results of trends recorded under the system

Project Plan

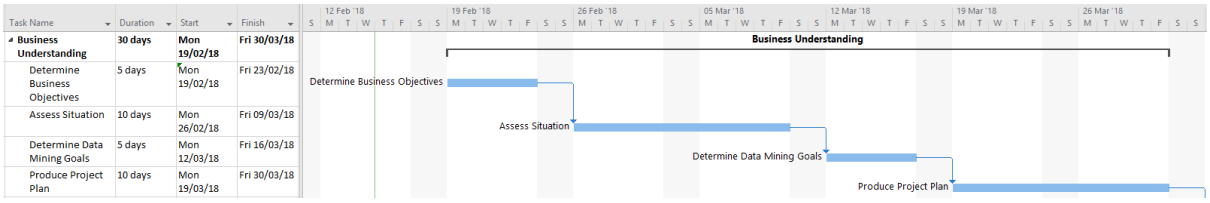
The project will go through a data mining project run through starting from the 19th of February through to the 16th of November 2018 with a broad outline of the master project's trial run of scan as you shop technology in South Glamorgan running from 19th of February 2018 to 17th February 2020. After the data mining project's initial run through, the project will be continued via the monitoring and maintenance plan until the 2-year course of the project is completed. The model can only be assessed against the project objectives once deployed for several months. If stakeholders wish to reiterate data mining methods in a different fashion or add additional components, the window between 16th of November 2018 and 17th February 2020 remains open for use on top of the existing need to measure the performance of the system post-deployment.

Hardware preparations and installation of a system with basic data gathering methods will be in place by the start of March, but this report deals primarily with the data mining aspects of said data. It is assumed that the hardware installation will be sub-contracted out. Data will be gathered and used immediately after the installation. Given the need for a full batch of at least a month's data, a start date for data understanding at a minimum of Monday 2nd of April.

Initial data collection, understanding and readying of the data mining process will occur in early April with the first data entering the system formally, giving another 5 months for the data mining development process to be underway before a go live of the system coinciding with the project's evaluation phase starting 1st of October 2018. This go live is timed with evaluation because only when months of data has been gathered can predictive inferences be made at an accurate level.

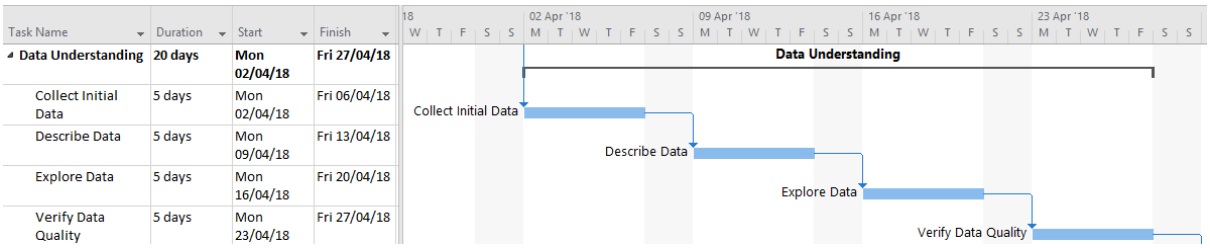
Associative data in theory could be prepared within a month however more time helps properly calibrate the association variables.

Schedule and Gantt Diagrams:



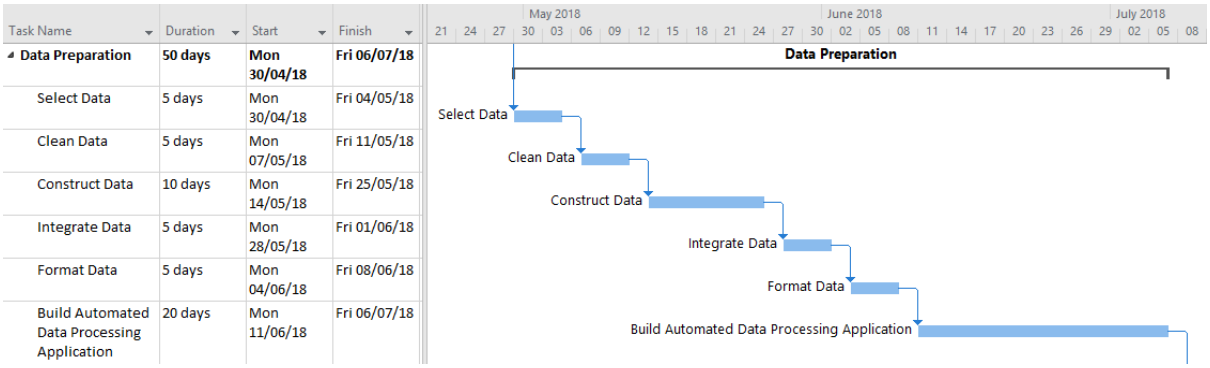
Task Name	Duration	Start	Finish
Business Understanding	**30 days**	**Mon 19/02/18**	**Fri 30/03/18**
Determine Business Objectives	5 days	Mon 19/02/18	Fri 23/02/18
Assess Situation	10 days	Mon 26/02/18	Fri 09/03/18
Determine Data Mining Goals	5 days	Mon 12/03/18	Fri 16/03/18
Produce Project Plan	10 days	Mon 19/03/18	Fri 30/03/18

Fig 3 – Business Understanding Schedule and Gantt



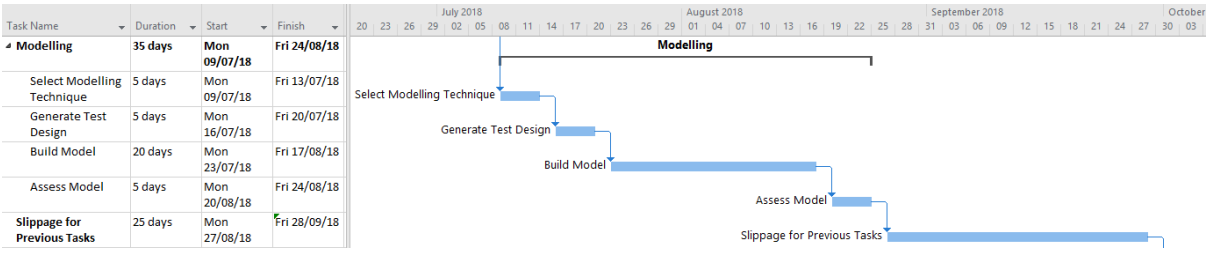
Task Name	Duration	Start	Finish
Data Understanding	**20 days**	**Mon 02/04/18**	**Fri 27/04/18**
Collect Initial Data	5 days	Mon 02/04/18	Fri 06/04/18
Describe Data	5 days	Mon 09/04/18	Fri 13/04/18
Explore Data	5 days	Mon 16/04/18	Fri 20/04/18
Verify Data Quality	5 days	Mon 23/04/18	Fri 27/04/18

Fig 4 – Data Understanding Schedule and Gantt



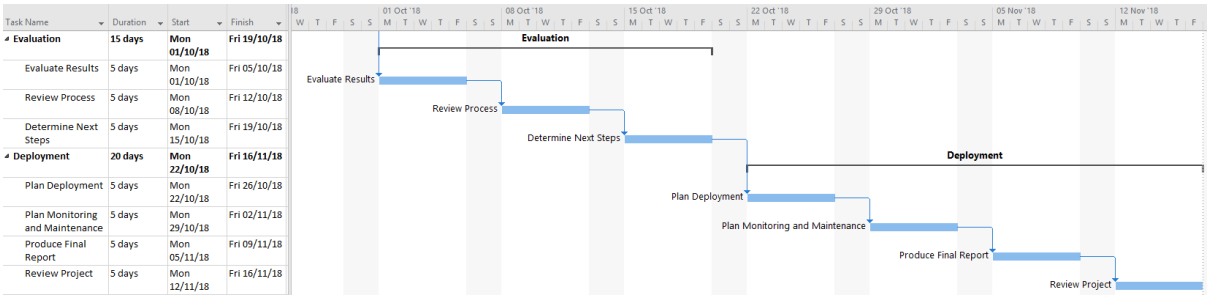
Task Name	Duration	Start	Finish
Data Preparation	**50 days**	**Mon 30/04/18**	**Fri 06/07/18**
Select Data	5 days	Mon 30/04/18	Fri 04/05/18
Clean Data	5 days	Mon 07/05/18	Fri 11/05/18
Construct Data	10 days	Mon 14/05/18	Fri 25/05/18
Integrate Data	5 days	Mon 28/05/18	Fri 01/06/18
Format Data	5 days	Mon 04/06/18	Fri 08/06/18
Build Automated Data Processing Application	20 days	Mon 11/06/18	Fri 06/07/18

Fig 5 – Data Preparation Schedule and Gantt



Task Name	Duration	Start	Finish
Modelling	**35 days**	**Mon 09/07/18**	**Fri 24/08/18**
Select Modelling Technique	5 days	Mon 09/07/18	Fri 13/07/18
Generate Test Design	5 days	Mon 16/07/18	Fri 20/07/18
Build Model	20 days	Mon 23/07/18	Fri 17/08/18
Assess Model	5 days	Mon 20/08/18	Fri 24/08/18
Slippage for Previous Tasks	25 days	Mon 27/08/18	Fri 28/09/18

Fig 6 – Modelling Schedule and Gantt



Task Name	Duration	Start	Finish
Evaluation	**15 days**	**Mon 01/10/18**	**Fri 19/10/18**
Evaluate Results	5 days	Mon 01/10/18	Fri 05/10/18
Review Process	5 days	Mon 08/10/18	Fri 12/10/18
Determine Next Steps	5 days	Mon 15/10/18	Fri 19/10/18
Deployment	**20 days**	**Mon 22/10/18**	**Fri 16/11/18**
Plan Deployment	5 days	Mon 22/10/18	Fri 26/10/18
Plan Monitoring and Maintenance	5 days	Mon 29/10/18	Fri 02/11/18
Produce Final Report	5 days	Mon 05/11/18	Fri 09/11/18
Review Project	5 days	Mon 12/11/18	Fri 16/11/18

Fig 7 – Evaluation & Deployment Schedule and Gantt

Data Understanding

Initial Data Collection Report

The initial data gathered of app using shoppers making purchases from the first month of operation will be obtained and basic details gathered into a report. The dataset for this first report will be the first month's data and the location will be held with the Data Store Management team at the central office who manage the company's data. To acquire the data, the company's intranet will be used and any issues with data access will be identified and resolved during the creation of this report.

Processes established during this first report will be used to gather data for subsequent months. An assumption made is that processes for preparing the data can be automated for future months, once an initial batch has been processed.

Data Description Report

In this report the data will be formally described, with the format, quantity, identity of datatypes and features of the data fully detailed. The form of the raw data being gathered will be effectively labelled through this stage, making a guide for future months of data gathered. The data will also be evaluated to ensure it meets the project's requirements.

Data Exploration Report

The data will be explored to find any interesting information. Given the limited nature of this first stage data, it is expected that this will be a brief report.

Data Quality Report

A quality report will be made to detail the state of the data in terms of adherence to requirements and project needs. Should the data be of high quality with few errors, in an easy to use format then the report will be brief. Any issues with the quality of the data should be documented and solutions outlined in preparation for the pre-processing of data to later come.

Data Preparation

Dataset and Dataset Description

The dataset of shopping data is to be the formal output from this phase used for the modelling phase, all activity within data preparation is to focus on getting the raw data into a prepared, pre-processed state. A formal description will also be detailed.

Rationale for Inclusion/Exclusion

Data will be selected based on its direct utility in achieving the project objectives, suggesting to shoppers' possible goods to buy and anything that can be used in predicting trends. All data to be included and excluded from the raw data will be fully documented. Given the need for attributes to be drawn from the data, I would err on the side of inclusion, so useful inferences can be made. Backups of the original data should be maintained in case culled data needs inclusion later.

Data Cleaning Report

As a successor report to the data quality report of the data understanding phase, this report will detail the actions taken to rectify quality issues outlined in the former report.

Derived Attributes

Attributes that help identify products will be gathered from the data. A primary example is the location or department a product is held in. Another attribute that could be gathered is the time and date of purchase, used in prediction.

Generated Records

Records will be generated for the classification of various goods into groups. Group categories will be created to help sum up categories of products for stakeholder evaluation.

For association, variables detailing the association of each product with other products will be made.

Merged Data

Aggregation of data from its raw form into a combined form for the purposes of classification is performed at this stage. For example, all diary products are merged into a diary category for the purposes of prediction.

Reformatted Data

All syntactic changes to the data in terms of fully readying it for direct use in a modelling system are performed here. Additional characters and unnecessary text are removed. Fields are converted from integer to string or vice versa as needed. This is entirely dependent on the format of the data as supplied by head office.

Building of an Automated Data Processing Application

To conclude the completion of the transformation of one dataset ready for the modelling phase, an automated application that transforms raw data into prepared data will be made for the project. This will allow the project to use novel data without going through the process of repeating the data steps every single month.

Modelling

With the modelling phase, approximately 4 months data will be available for use. This means initial work using prediction can be made. The data preparation phase should be quickly looped with new data using the automated application developed.

Modelling Technique

Association, Classification and Prediction are all to be used.

Modelling Assumptions

Association makes no assumptions due to its simple linkage to other products.

Classification assumes that products will already have attributes that can be derived to help classify them into the correct category or sub-category.

Prediction makes no assumptions directly but assumes that classification works properly, and data is available for at least 3 months. There should be 4 months data on hand by the modelling stage.

All models should assume that new datasets will be added to the model every month. The models should then be flexible enough to handle x-number of datasets.

Test Design

Association will be tested using selected sample data that is familiar, unfamiliar but commonly bought together and completely unrelated. For example, coffee will be compared with milk, eggs and fly spray. In theory, coffee and milk are commonly together as most people drink coffee with milk, eggs are related because people often have breakfast with coffee, but fly spray will be negligible because a food shop is quite distinct from a shop buying specific items.

Classification should require little testing due to classification using existing, tested store records, however anomaly detection should be used to find any differences.

Prediction will be tested based on checking data results manually and comparing with the output.

Parameter Settings

For association, a parameter to tweak would be one that controls the ratio of a product to another. It may make more sense for instead of a linear scale, a quadratic or logarithmic scale. This would be subject to testing of the model.

Models

There are 3 models, association of products in shopping baskets, classification of products into separate categories and prediction of trends within categories of products. The classification model feeds into the prediction model.

Model Description

Once created, the models should be fully described, and any difficulties encountered with interpretation of results.

Model Assessment and Revised Parameter Settings

The initial results of the models developed should be summarised and the accuracy of the models established. Testing will be conducted with user acceptance testing in person at each of the six stores by several of the project staff. This testing will be both in the form of using the app with suggestions prompted and in the use of predicting supply chain changes.

Any necessary changes or tweaks to the parameter settings should be noted and tuned in preparation for the next month's modelling of data.

Evaluation

With the beginning of this phase, approximately 6 months of data should be available. Automation of the data pre-processing should allow quick plugging in of new data to the models.

Assessment regarding Business Success Criteria

The results of the models cannot be evaluated against the success criteria due to the need for a length of time to elapse with the deployed models implemented in the app. Instead, the process of formally reviewing the success will be delayed until at least 3 months after deployment.

Evaluation will occur instead in the form of the results of the user acceptance testing.

Approved Models

So long as the models fulfil UAT testing, the models will be approved.

Review of Process

The project will be reviewed, and it will be assessed whether aspects have been missed or if certain parts need to be repeated. Attention will be especially paid to the design of the models as true assessment is to be conducted well after deployment.

Next Steps

At this point it will be decided to either progress to deployment, repeat an earlier phase or element of a phase or to simply end the project. This will be decided by senior stakeholders who own the project.

Deployment

Deployment Plan

A plan to deploy the technology into the stores selected for the trial will be developed. The deployment process will focus on integrating the suggestions feature via the association model for

customers. The app will experience some down time over night on a Sunday in which the application's systems will be shut down and users will be prompted for an essential update on next using the app.

In terms of supply chain optimisation, predictive data will be fed to the stores showing them a trend towards the amounts of goods to be sold in future months. Relevant staff involved with procurement will be trained to use the data to help more accurately refine the numbers of goods to be ordered.

Maintenance Plan

The monitoring and maintenance plan will be focussed on integrating new data each month into the models used and feeding back the results of the models to senior stakeholders. Of note, the business success criteria will be assessed each month through the timeframe of the wider project.

Final Report and Presentation

A report and presentation will be prepared to conclude the data mining aspect of this project, ending in November 2018. These will be used to communicate the project to members of the business joining the project and to summarise the conclusions of the project to senior management.

Experience Documentation

The overall experience of the project will be summarised into documentation, covering all the previous reports written for the project. Problems and issues come across will be addressed.

Post Data Mining Project Business Objective Evaluation

Given the need to wait until several months after deployment, at approximately the one-year mark, in February 2019 the business objectives will be formally evaluated using the recorded performance of the system.

The project will be deemed a success if:

- 2% of customers into stores using the technology use the app to purchase goods
- Prediction of future trends results in more than £4k in savings a month per store

Alternatively, if the project's objectives are not met but the project has in effect turned a profit through savings made, the project will be deemed a minor success.

If the objectives are not met and the project has not achieved sufficient savings, the project will be deemed a failure and shut down.

Future Improvements

Given the limited scope of this project, there is a large amount of scope for additional data mining techniques to be applied.

Customer Relationship Management could be integrated into the system by monitoring the type of customer that uses the mobile app to make purchases. If a customer has not made a purchase at a store within a given time frame, then they could receive an email or text offering a small discount on their next shop. Post-shop optional reviews could be recorded and assigned to each store, with trends within reviews used to show customer's opinions of a store. Positive comments could be used to help less successful stores emulate successful features and negative comments could be used to improve the store.

References

Allerin.com. (2018). *Machine learning for anomaly detection | Artificial Intelligence |*. [online] Available at: https://www.allerin.com/blog/machine-learning-for-anomaly-detection [Accessed 16 Feb. 2018].

Asda. (2018). *Scan & Go | Asda*. [online] Available at: http://www.asda.com/scan-and-go/ [Accessed 12 Feb. 2018].

BBC News. (2016). *Fraud costing UK '£193bn a year'*. [online] Available at: http://www.bbc.co.uk/news/uk-36379546 [Accessed 9 Feb. 2018].

Brown, A. (2015). *How I 'stole' my shopping from Waitrose*. [online] Telegraph.co.uk. Available at: http://www.telegraph.co.uk/comment/personal-view/11472142/How-I-stole-my-shopping-from-Waitrose.html [Accessed 12 Feb. 2018].

Buzby, J., Bentley, J., Padera, B., Ammon, C. and Campuzano, J. (2015). Estimated Fresh Produce Shrink and Food Loss in U.S. Supermarkets. *Agriculture*, 5(4), pp.626-648.

Chitra, K. and Subashini, B. (2013). Data Mining Techniques and its Applications in Banking Sector. *International Journal of Emerging Technology and Advanced Engineering*, 3(8).

CRISP-DM 1.0. (2000). 1st ed. [ebook] The CRISP-DM consortium. Available at: https://www.the-modeling-agency.com/crisp-dm.pdf [Accessed 7 Feb. 2018].

Discord Blog. (2017). *How Discord Stores Billions of Messages – Discord Blog*. [online] Available at: https://blog.discordapp.com/how-discord-stores-billions-of-messages-7fa6ec7ee4c7 [Accessed 12 Feb. 2018].

Financesonline.com. (2018). *Best CRM Software Reviews & Comparisons | 2018 List of Expert's Choices*. [online] Available at: https://crm.financesonline.com/#history-of [Accessed 9 Feb. 2018].

Health Catalyst. (2018). *What is Data Mining in Healthcare?*. [online] Available at: https://www.healthcatalyst.com/data-mining-in-healthcare [Accessed 9 Feb. 2018].

Kdnuggets.com. (2013). KDnuggets Annual Software Poll:RapidMiner and R vie for first place. [online] Available at: https://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html [Accessed 5 Feb. 2018].

Haffar, J. (2018). *Have you seen ASUM-DM? - SPSS Predictive Analytics*. [online] SPSS Predictive Analytics. Available at: https://developer.ibm.com/predictiveanalytics/2015/10/16/have-you-seen-asum-dm/ [Accessed 7 Feb. 2018].

IBM Analytics Solutions Unified Method (ASUM) - External. (2015). IBM.

Koh, H. and Tan, G. (2005). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management*, 19(2).

PKF (2016). *Annual Fraud Indicator 2016*. Annual Fraud Indicator. [online] Portsmouth: Experian. Available at: http://www.port.ac.uk/media/contacts-and-departments/icjs/ccfs/Annual-Fraud-Indicator-2016.pdf [Accessed 9 Feb. 2018].

Sainsburys. (2018). *Smartshop app | Sainsbury's*. [online] Available at: https://www.sainsburys.co.uk/shop/gb/groceries/get-ideas/our-freebies-and-competitions/our-freebies-and-competitions/smartshop-app [Accessed 12 Feb. 2018].

Securetechalliance.org. (2006). *Texas Medicaid*. [online] Available at: https://www.securetechalliance.org/resources/lib/Texas_Medicaid.pdf [Accessed 9 Feb. 2018].

Tesco. (2018). *Scan as you Shop | Tesco.com*. [online] Available at: https://www.tesco.com/scan-as-you-shop/ [Accessed 12 Feb. 2018].

Tesco plc. (2018). *Five year record*. [online] Available at: https://www.tescoplc.com/investors/reports-results-and-presentations/financial-performance/five-year-record/ [Accessed 15 Feb. 2018].

Thearling.com. (2018). *An Introduction to Data Mining*. [online] Available at: http://www.thearling.com/text/dmwhite/dmwhite.htm [Accessed 16 Feb. 2018].

Tiwari, A. and Sekhar, A. (2007). Workflow based framework for life science informatics. *Computational Biology and Chemistry*, 31(5-6), pp.305-319.

Turney, P. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval*, 2(4), pp.303-336.

Verma, M. and Mehta, D. (2014). Sequential Pattern Mining: A Comparison between GSP, SPADE and Prefix SPAN. *International Journal of Engineering Development and Research*, 2(3), pp.3016-3036.

Wood, Z. and Butler, S. (2015). *Tesco cuts range by 30% to simplify shopping*. [online] the Guardian. Available at: https://www.theguardian.com/business/2015/jan/30/tesco-cuts-range-products [Accessed 12 Feb. 2018].