Inlämningsuppgift 2,5HP

IT134G Introduktion till Business Intelligence

HÖGSKOLAN
I SKÖVDE

Olof Almqvist

A15oloal

15-09-2016

Table of contents

1 Descriptive Business Intelligence

Descriptive Business Intelligence or *descriptive analytics* is the process of collecting, storing and presenting historical data in a meaningful way as to answer the question: *"What has happened?"*.

1.1 The data warehouse

A central technology to business intelligence (BI) is that of the data warehouse (DW). A DW is a specialized form of a database that is optimized for the ability to store and allow analysis of huge amounts data.

Input data into the DW can come from *e.g.* operational OLTP databases, legacy systems or web sources.

1.2 The ETL-process

DW's undergo the three step process of "Extract, Transform, Load" or "ETL". In this process the DW *extracts* data from other more specialized operational databases, loads it into a staging area where it is *transformed* into a standardized form and cleaned from errors. It is then *loaded* into the DW or a certain compartment of a DW called a data mart. Data can then be processed through reporting to a dashboard.

1.3 Characteristics of a data warehouse

Data warehouses has four important characteristics. First, they are "***subject-oriented***" which means that data is organized by a specific subject like visitors, sales or messages. Secondly, they are "***integrated"*** or data extracted from different sources are transformed into a consistent format. Thirdly, DW's must maintain ***historical data*** like for example time stamps. This permits the ability to create forecasts and see trends. The last key feature of DW's is that they are "***nonvolatile***" which means that stored data cannot be changed and updates are recorded as new implementations (Inmon 2005, pp 31).

1.4 Data cubes

One common way of carrying out descriptive analytics is by using three dimensional OLAP-cubes or simply "cubes". Cubes are small relational databases built from a star schema. These graphical tools represent a matrix of facts which refer to measured data and dimensions which represent categories of information (Figure 1).

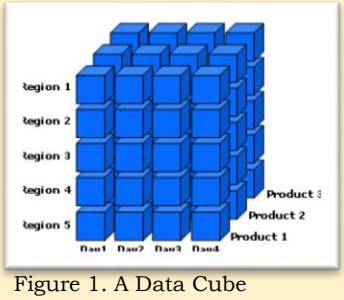

Figure 1. A Data Cube constructed in Microsoft Excel.

With this tool, it is possible to combine enormous amounts of information and get answers at very fast speed (Dubler & Wilcox 2002).

2 Predictive Business Intelligence

Predictive intelligence aims at utilizing historical data in order to create plausible scenarios for future development. This method uses several methods like statistics, machine learning and data mining.

Data mining is a computational process of discovering various patterns in a collection of data and it contains six different types of methodologies. The first is ***"anomaly detection"*** or the identification of out of the ordinary patters. The second is ***"association rule learning"*** or the finding of relationships between variables. The third is ***"clustering"***, finding groups among the data for example with Kmeans. The fourth is ***"classification"*** which is the process of storing new data into known categories. The fifth is ***"regression"*** which is a method of finding strengths of relationships between independent and dependent variables. The last type of data mining is that of ***"summarization"*** or providing reports and visualizations of data (Fayyad, Piatetsky-Shapiro & Smyth 1996).

2.1 Predictive compared with descriptive

Predictive analytics differs from descriptive analytics in that the predictive method employ inductive reasoning on historical data, thereby following trends and patters and attributing a likelihood of future development. Through this method, organizations acquire a unique way to implement novel proactive approaches like understanding when a tool will break or which customer is likely to default on a loan (Eckerson 2007).

2.2 Successful implementation by Netflix

An example of a company that have implemented a mature predictive analytics solution and acquired a competitive advantage is Netflix.

In 2006, Netflix announced a price of 1 million dollars to whoever could construct the best algorithm to predict how a viewer would rate a certain movie based on previous ratings. The winning algorithm was implemented in 2009 when only four categories of data were available to the company: customer ID, movie ID, rating and when the movie was watched. Now this and other predictive instruments have been refined with the introduction of streaming, with more types of data like the time of day when the movie was watched, the time it took to select a movie, and how often the movie was paused (Marr 2015).

Netflix analyzed the data gathered by the algorithm about information like viewing patterns and ratings and used it to produce a show. The product was the series "House of Cards" which is a drama about a senator who eventually grows to become president. The show was a tremendous success (Lazzaro 2016).

3 Prescriptive Business Intelligence

Prescriptive analytics is a method of using historical internal and external data to generate suggestions for the best possible course of action. One way to do this is to create a model, choices and data can then be inserted into the model to understand the consequences of different scenarios and which one results on the highest probability of achieving organizational objectives (appendix, figure 6). The implementation of these rules are done in Expert System (ES).

3.1 Successful examples of Expert Systems

3.1.1 Dendral

One of the first ever created ES was the hypothesis testing AI system Dendral. Dendral had the ability to help organic chemists determine the chemical composition of unknown organic molecules by analyzing their mass and previously known chemical knowledge. It was developed with limited success in regard to its science applications but had a larger impact as a stage in the development of AI systems (Lindsay et al. 1993).

3.1.2 Mycin

An interesting area of early implementation of an ES is in the medical field. A thoroughly investigated technology in this sector is that of Mycin. This system has the capacity to analyze human samples and determine the presence of bacterial infection and blood clots, and subsequently recommend the proper medication based on the patient's body weight (Shortliffe 1976). Further studies showed that Mycin could be more accurate (65%) than human experts (42.5-62.5%) at diagnosing the correct disease and subsequent treatment (Cohen et. al. 1979).

3.1.3 Possible trends

From the novel creations of Dendral and Mycin new AI tools have been developed for various sectors. Examples include ES systems for determining birth risks (Wooley & Grzymala-busse, 1994), equipment failure in the oil and gas sector (Liyanage et al. 2014), and of course IBM's Watson which is a highly advanced AI system based on self-learning with applications in many different areas (IBM u.å.)

3.2 The Industrial and Commercial Bank of China

Prescriptive analytical techniques were developed by IBM when hired by The Industrial and Commercial Bank of China (ICBC) in 2006. The aim of the collaboration was to develop a model for the maximum efficiency in where to locate bank offices to serve over 230 million individual and 3.6 million corporate customers. The quantitative model that was created incorporated parameters like customer flow, number of households and concentration of competitors. ICBC estimates that prescriptive analytics resulted in up to 1.01 billion dollars in increased deposits in the city of Suzhou alone (Zhang et. al. 2012).

4 Operational, tactical and strategic levels

One of the founders of modern military strategy Antoine-Henri Jomini describes the difference between strategy and tactics in his book "*The Art of War*" in the following manner:

"Strategy is the art of making war upon the map, and comprehends the whole theater of operations. […] Strategy decides where to act; logistics brings the troops to this point; grand tactics decides the manner of execution and the employment of the troops." (Jomini 1862).

4.1 The importance of goals

A successful organization must set future goals for what they wish to accomplish in the future. According to Doran (1981) these goals must be specific, time oriented, realistic, measurable and assignable to specific actors.

4.2 Strategy

A strategy is the overarching method of reaching those goals. A business example might be the goal to become a market leader in regard to sales within 5 years by providing Business Intelligence solutions to clients in the geographical region of Västra Götaland. A strategy to reach this objective could be "focusing on implementing BI solutions to small and middle sized companies to increase revenue in the retail sector". Strategy typically involves few people and work on long periods of time and can be supported by DSS and EIS systems.

4.3 Tactics and operations

When the goals and the strategy to reach those goals are defined, the work begins and the art and science of how to best utilize available resources to stay within the strategy and move closer to the company objectives (plan budget/production) are what is called tactics. Tactics work on time periods like weeks or months and could be supported by OAS and MIS systems.

Once plans for the deployment of resources have been finalized, operations begin. Operations describe the day to day activities whose successful implementation can be tracked by the use of the previously described DW and supported by TPS and PCS systems. This then provides a feedback mechanism for how to improve tactics and strategy in order to reach the organizations ultimate goals.

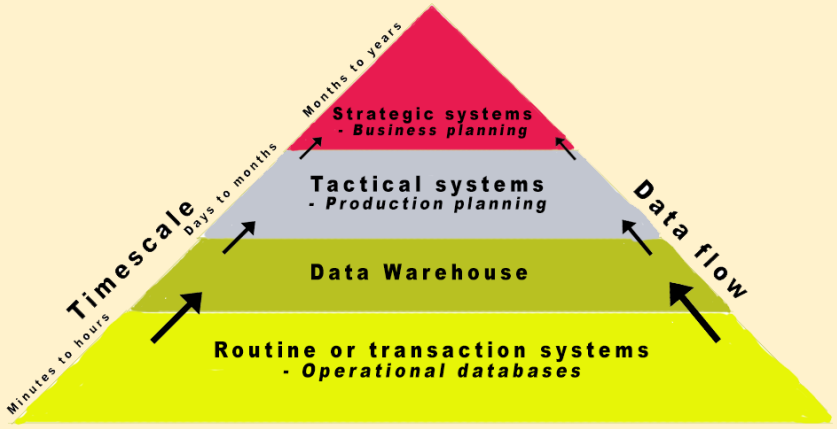

Figure 2. The flow of information within an organization supported by a data warehouse. Adapted and redrawn from Stewart (2008).

5 Decision-making

Decision making can be conducted in accordance with the approach suggested by Simon (Sharda et. al. 2014, pp 72-88) containing distinct stages for ***"intelligence"*** which translates into understanding the problem, ***"design"*** for constructing a solution in the form of a model, and ***"choice"*** where different alternatives are compared and the optimal is chosen.

External data (table 1) describe a data warehouse which has been collecting data concerning the activities of the sea transport operations of Stena Line. It is analyzed in accordance with Simons decision model and the Gorry Scott-Morton framework by implementing three BI approaches, descriptive, predictive and prescriptive.

Table 1. Categories of information related to Stena Line stored in fictional DW.

Week	Date	Delay	Delay in minutes	Direction of travel	Ship
11	*20160314*	*yes*	*20*	*Frederikshavn – Göteborg*	*Stena Jutlandica*

5.1 Descriptive BI

The implementation of a data warehouse together with complimentary BI-capacities introduce a range of new possibilities for an organization. A DW can store new data daily from daily operations and present these on a dashboard. This increases management's ability to understand whether the operational activities are in line with the organizational tactical objectives which are part of whether the organization is in line with the overarching strategy.

As an example, a dashboard could use descriptive intelligence to present the number of ships that have been delayed, how much they were delayed, and which courses that are affected and the additional cost per day (table 2).

5.2 Predictive BI

Predictive intelligence provides additional capacities than the ones provided initially by descriptive BI. Here, it is not only about reporting what has already happened but also what is likely to happen in the future based on historical information and pattern recognition (which could be done with *e.g.* the apriori algorithm, kmeans or a quantitative model). This toolbox introduces the ability to understand likely scenarios for how much the delays might cost per year and budget for these beforehand (table 2; appendix, figure 5).

It would also be possible to understand the type of ships, dates, times of day and routes with the highest risk of suffering from delays, and consequently plan ahead in order to minimize negative customer impact (table 2; appendix, figure 3, 4). Predictive BI also include statistical test which can be used to determine whether occurrences are due to chance or not (appendix, table 3, 4).

5.3 Prescriptive BI

Prescriptive intelligence utilizes the information provided by descriptive and predictive intelligence and creates a third level of analysis – that of providing suggestions for optimal outcomes or utilization of resources.

One example could be a software with prescriptive capabilities which could provide suggestions for the most profitable ways to plan company sea traffic regarding dates, time of day and routes (table 2; appendix, figure 3, 4). It could also calculate logistics using linear programming as to minimize the risk of latencies based on the information collected by the organization thus far (table 2). A software like this could be based on a quantitative model derived from an influence diagram, constructed with the available data tested using K-fold cross validation and consequently used to forecast future developments.

Table 2. A collection of suggestions for how Stena Line data could be used in Business Intelligence to improve various areas of interest.

	Operational level	Tactical level	Strategic level
Descriptive BI	What is happening right now? Key figures.	Is it in line with tactical objectives? Budget? Staff?	Is the organization developing within the bounds of its mission statement?
Predictive BI	Which vehicles are going to be late? Is there enough ships available for client needs in the near future?	Plan budget. Can the budget take estimated delays into account?	How can the mission statement and goals be adapted to projected developments?
Prescriptive BI	How to adapt current route to make up for heavy traffic or blockages along the way?	What is the optimal utilization of assets? What is optimal logistics? Should we accept an order, is it profitable? Do we have enough vehicles?	Which locations are most profitable to do business with? Which ships provide a good return in the current market climate?

5.4 Classification Tree

Data mining in the form of a classification tree was carried out in accordance with Breiman et al. (1984) on the data on ferry routes (appendix, figure 3) to determine whether some ferries, some routes, or some departure times had a significant impact on the risk of delay.

5.4.1 Analysis of the Classification Tree

Heuristic analysis was carried out to understand the cause of latencies. *Stena Danica* (S. Danica) occupied 62, 4% and *Stena Jutlandica* (S. Jutlandica) 37, 6% of all instances of a ferry arriving late (appendix, figure 4). This suggested a possibly interesting pattern. Further investigation showed that *S. Danica* made up 54% of all executed routes and *S. Jutlandica* made up 46% of all executed routes.

This suggest that *S. Danica* are suffering from 8, 4% more instances of being late when compared to *S. Jutlandica*. Company tactical efforts could thus focus on ways to bring down this number so that *S. Danica* can perform as well as similar vehicles. Ways to approach this could be technical or organizational.

Departure time and route did not seem to have an important impact on the probability of being late for the two ferries in this analysis.

5.4.2 Statistical analysis

5.4.2.1 'N-1' chi-squared test for comparison of proportions

Raw data from the generation of the classification tree is summarized in table 3 and tried in a Chi-square test – as suggested by Campbell (2007) - to understand whether the difference in probabilities of being late was statistically significant.

Table 3. Proportion of late ferry routes, total routes, and probability of being late.

Ferry	Late	Not Late	Total
S. Danica	65; P = 0.190	277	342
S. Jutlandica	39; P = 0.135	249	288
Total	104	526	630

The result of statistical analysis (table 4) shows that the difference in being late between *S. Danica* and *S. Jutlandica* is not due to chance with a probability of 93.6%.

Table 4. Summary of data from N-1 chi-squared test.

Data type	Result
Difference	5.5%
95% C.I.	-0.53 to 11.39
Chi-squared	3.43
DF	1
Significance level	P = 0.0640

Chi-square calculation was carried out by MedCalc (2016).

5.7 Kmeans and Clusplot

Lastly, an attempt was made to conduct a second type of classification analysis in the form of clustering. The two investigated variables were “number of minutes late” and “time of departure”. Statistical language R was used in accordance with (Sharda et. al. *2014,* pp 252-253) for Kmeans and Pison et al. (1999) for the Clustering plot. Functions were taken from the R clustering package.

Table 5. Cluster means acquired by Kmeans algorithm.

Cluster	Departure time	Minutes late	Fit.cluster
1	09:00	4.14	1.55
2	19:21	4.41	3.00
3	22:49	4.72	3.00
4	15:05	5.87	2.53
5	01:03	4.92	1.00

An interesting group appeared around the time 15:05 (table 5; appendix, figure 5) where there is a spike in minutes late. Speculatively because there is more sea traffic during the afternoon (figure 5).

5.8 Conclusion

The Stena Line ferry data was analyzed in accordance with Simons decision model and the Gorry and Scott-Morton framework. Number of minutes late was identified as an interesting and practical variable to investigate further. Several models were considered but clustering and classification tree was chosen as suitable. The classification tree showed that there was an interesting discrepancy between *S. Danica* (P = 0.190) and *S. Jutlandica* (P = 0.135) regarding the risk of being late (table 3; appendix, figure 4). A Chi-Square test of comparisons of proportions was carried out to confirm whether the difference was statistically significant and showed a probability of 93.6% that the two probabilities are not the same (table 4).

Furthermore, a clustering analysis was performed to investigate the relationship between departure time and number of minutes late. Five clusters were created and one interesting group appeared with a mean departure time at 15:05 o clock (table 5; appendix, figure 5).

Stena Line could use this knowledge to investigate further whether it is possible to reduce instances of being late by understanding why *S. Danica* is less likely to be on time than *S. Jutlandica* and why the afternoon departure around 15:05 increases the risk of being late for both ships. Cargos that are especially time sensitive like the transportation of vaccines could be loaded onto *S. Jutlandica* and sent around 09:00 to enjoy the smallest possible risk of being negatively affected by a scenario where it does not arrive on schedule.

6 Sources

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, CJ. (1984). Classification and Regression Trees. *Wadsworth International Group*, Belmont, CA.

Campbell, I. (2007). Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Stat Med*, 30(19), pp 3661-3675.

Cohen, S.N., Buchanan, B.G., Blum, R.L., Hannigan, J., Scott, A.C., Clancey, W.J., Wraith, S.M., Fagan, L.M. & Yu, V.L. (1979). Antimicrobial selection by a computer. A blinded evaluation by infectious disease experts. *JAMA*, 242(12), pp 1279-1282.

Doran, G.T. (1981). "There's a S.M.A.R.T. way to write management's goals and objectives". Management Review. *AMA FORUM*, 70(11), pp 35–36.

Dubler, C., Wilcox, C (2002). *Just What Are Cubes Anyway? (A Painless Introduction to OLAP Technology).* https://msdn.microsoft.com/en-us/library/aa140038(v=office.10).aspx#odc_da_whatrcubes_topic2 [08-09-2016].

Eckerson, W (2007). *Predictive analytics.* https://tdwi.org/articles/2007/05/10/predictive-analytics.aspx?sc_lang=en [07-09-2016]

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P (1996). *From Data Mining to Knowledge Discovery in Databases.* http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf [23-10-2016]

Grzymala-busse, J., Woolery, L.K. (1994). Machine learning for an expert system to predict preterm birth risk. *J am Med Inform Assoc*, 1(6), pp 439-446.

IBM (u.å.). http://www.ibm.com/watson/what-is-watson.html [10-10-2016]

Inmon, W.H. (2005). *Building the Data Warehouse*. New York: Wiley.

Jomini, B.D., Mendell, G.H. & Craighill, W.P (1862). *The Art of War*. https://archive.org/stream/artwar00mendgoog/artwar00mendgoog_djvu.txt [08-09-2016].

Lazzaro, S (2016). *Netflix Purposely Designed 'House of Cards' to Be a Major Hit—Here's How They Did It.* http://observer.com/2016/01/can-we-use-big-data-to-create-hit-tv-shows-as-addictive-as-breaking-bad/ [08-09-2016]

Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A. & Lederberg, J. (1993). DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(959), pp 209-261.

Liyanage, J.P., Raza, J. & Lokko, N.N.B.C. (2014). The Use of Expert Systems in Offshore Oil and Gas Assets: A Status Review with Respect to Emerging Needs for Innovative Solutions. *Proceedings of the 7th World Congress on Engineering Asset Management*. Switzerland: Springer International Publishing, pp 397-405.

Marr, B (2015). *The Amazing Ways Netflix Uses Big Data To Drive Success.* https://www.linkedin.com/pulse/amazing-ways-netflix-uses-big-data-drive-success-bernard-marr [08-09-2016]

MedCalc (2016). *Comparisons of proportions calculator.* https://www.medcalc.org/calc/comparison_of_proportions.php. [10-10-2016]

Pison, G., Struyf, A. & Rousseeuw, P.J. (1999). Displaying a clustering with CLUSPLOT. *Computational Statistics & Data Analysis*, 30, pp 381-392.

Sharda, R., Delen, D., Turban, E., Aronson, J.E., Liang, T-P. & King, D. (2014). *Business Intelligence and Analytics*, 10th edition. England: Pearson Education Limited.

Shortliffe, E.H. (1976). *Computer-Based Medical Consultations: MYCFIN.* New York: Elsevier.

Stewart, S (2008). *Data Warehouse Basics.* http://www.rapid-business-intelligence-success.com/data-warehouse-basics.html [15-09-2016]

Wang, X., Zhang, X., Liu, X., Guo, L., Li, T., Dong, J., Yin, W., Xie, M. & Zhang, B. (2012). Branch Reconfiguration Practice through Operations Research in Industrial and Commercial Bank of China. *Interfaces* 42(1), pp 33-44.

7 Appendix

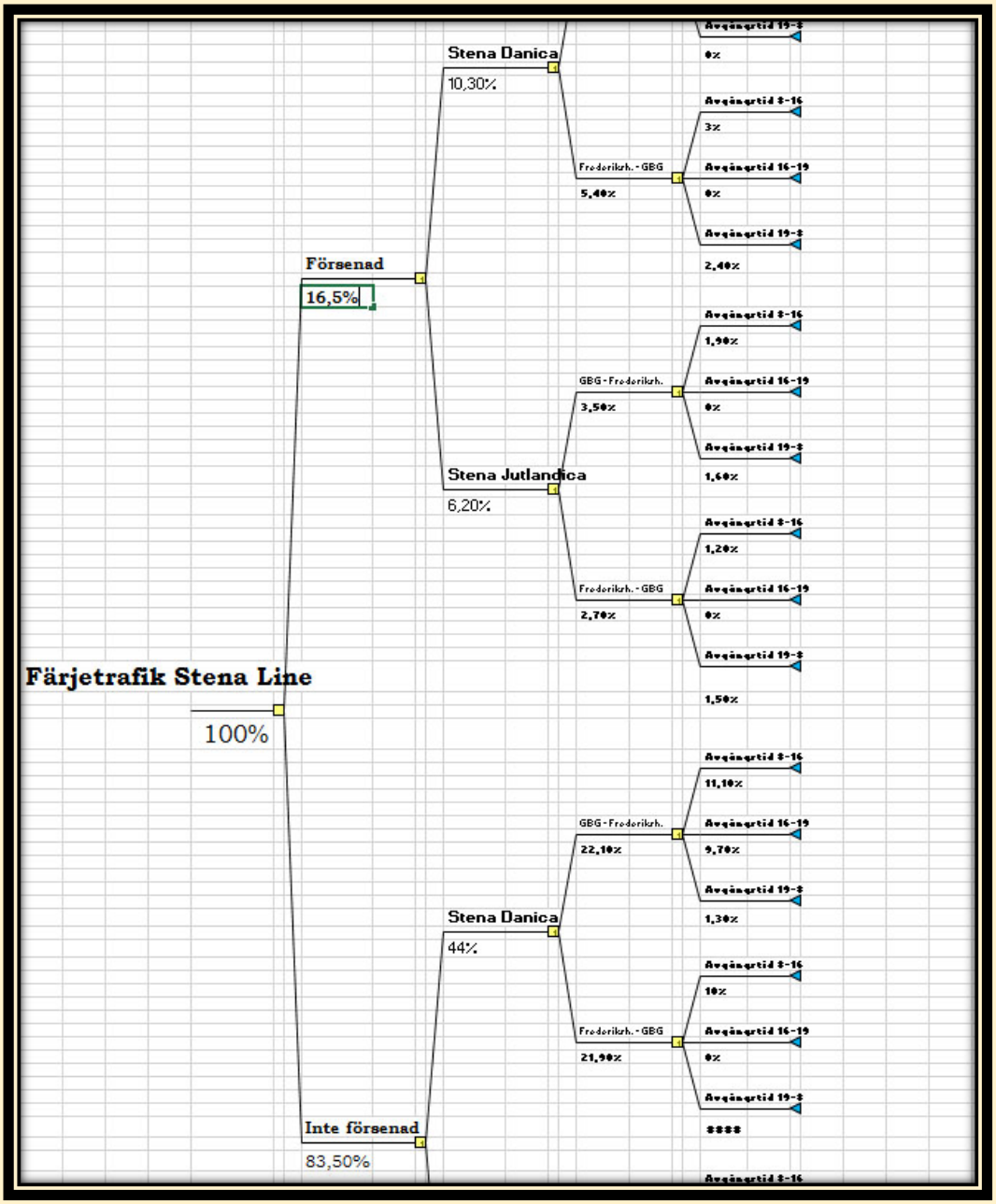

Figure 3. Classification tree for the percentage of late (16, 5%) and not late (83, 5%) travels. Data splits arbitrarily chosen and not tested for purity with the Gini Index.

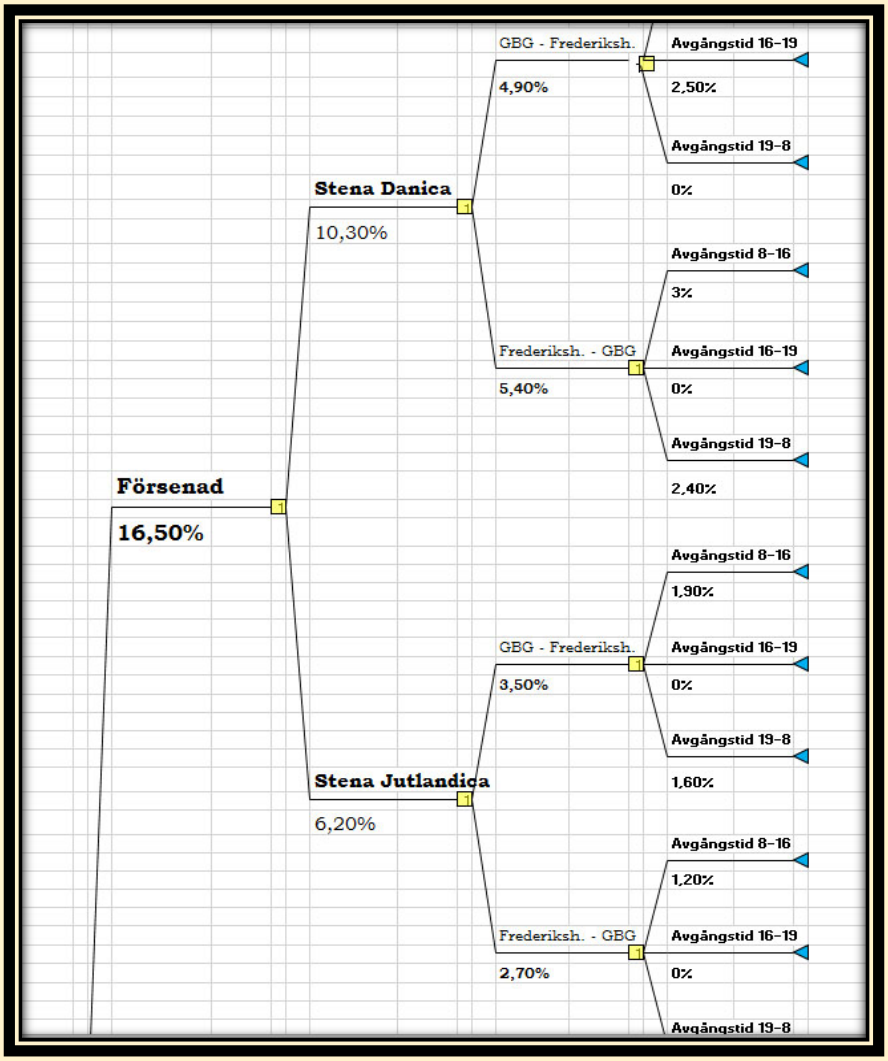

Figure 4. The proportion of late occurrences in regard to ferry, route and time of departure.

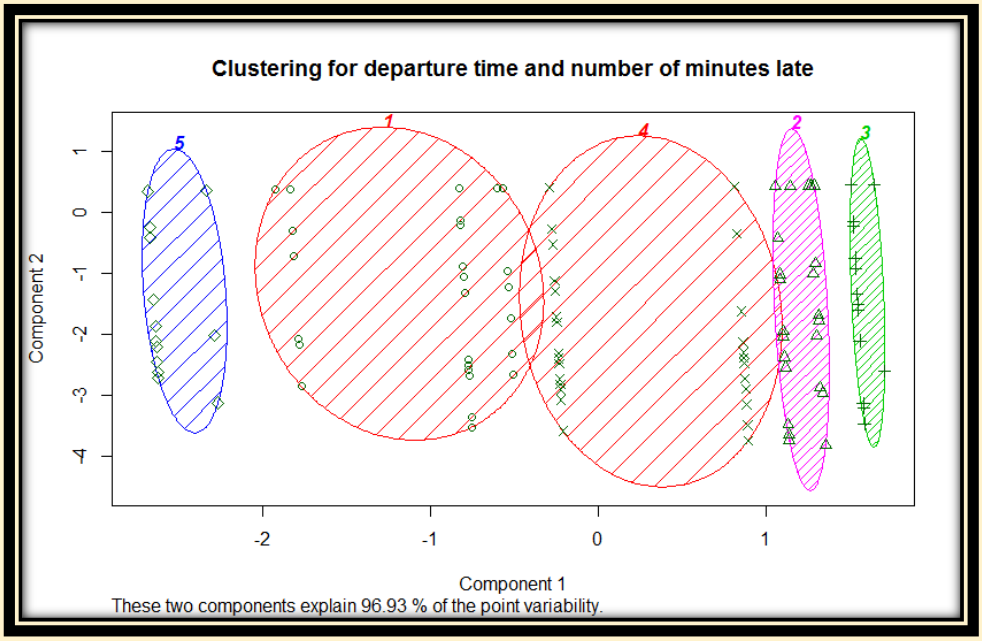

Figure 5. Clusplot for Kmeans generated clusters. Cluster number 4 describes a spike in minutes late around 15:05 in the afternoon.

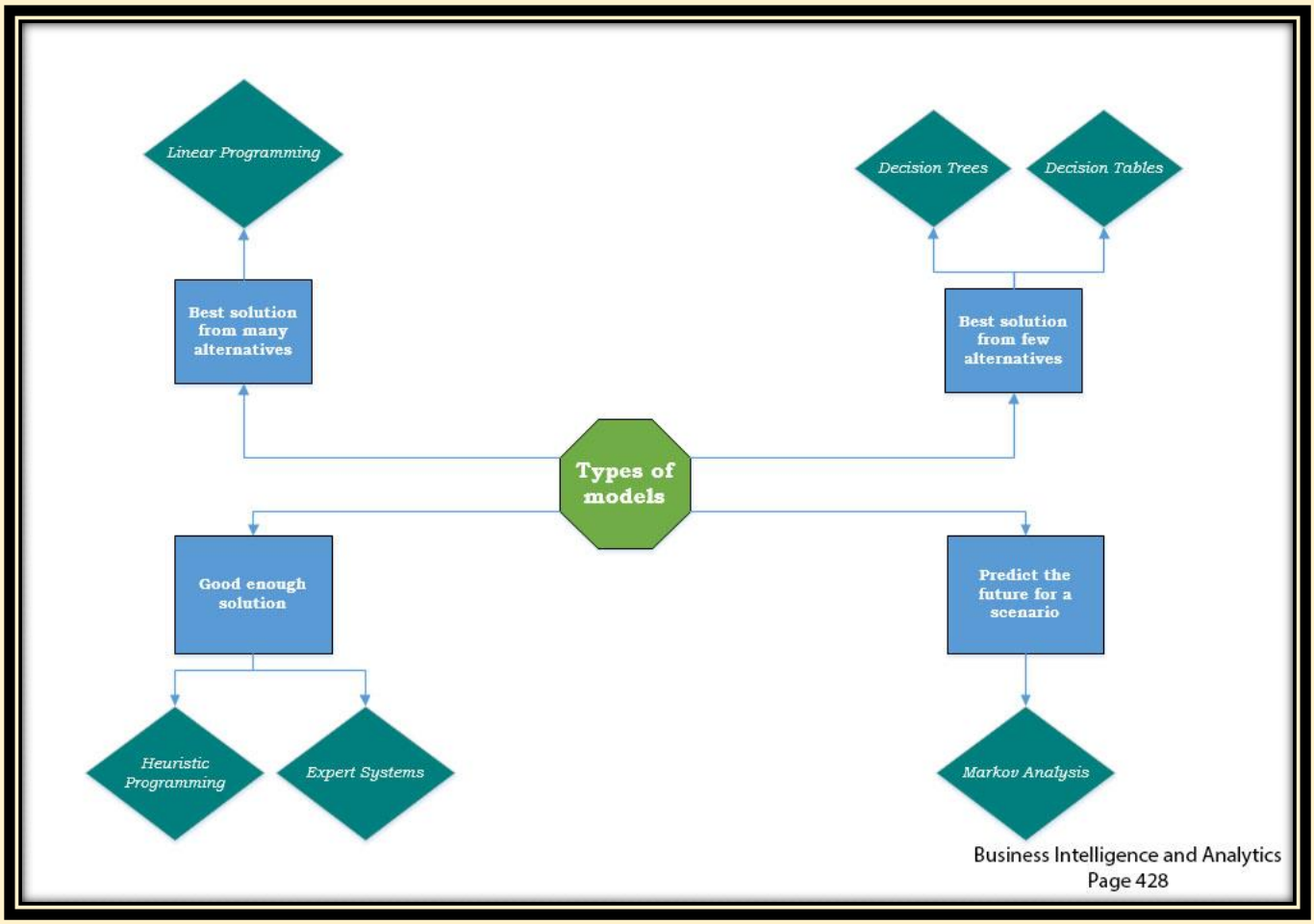

Figure 6. Different types of prescriptive models.