Analysing and Forecasting Future Order Statistics

Luke Calleja
Institute of Information Communication Technology (IICT)
MCAST
Triq Kordin, Rahal Gdid
luke.calleja.a100336@mcast.edu.mt
01/06/2017

***Abstract*—With regards to businesses, it is very important to analyse current sales and trends. Analysing current data can help businesses in many different ways, such as being able to see which products are being bought the most or which type of client (ex. single, married with children, etc.) is buying a product the most. In this report, order numbers will be analysed so that future order numbers may be predicted for the next year. The analysis will be divided into quarters over two years, with the third and final year being where the forecast will take place.**

***Index Terms*—K-Means, Correlation, PLSA, Data Clustering, GA, GFS, CRISP-DM, ETL, ERD, SCD**

I. Introduction

The current dataset provided various aspects of the business. There was everything one needs to know about the business, such as product categories, the brands, as well as vast information about the client, such as marital status and number of children they may have, among other attributes. This means that a lot of information can be gathered from the dataset and analysis can be performed on various aspects. As for this report, it was decided that the forecasting of future orders will be done. The aim is to predict how orders will go in the next year based on the information provided of the previous two years. Predictions are not very easy to make and sometimes external factors come into play which may effect the margin of error, such as natural disasters which can effect sales and even company assets. Although this report may be a good indicator of future order numbers, it should be considered fact. It is just an indication of how the number of orders are expected to rise, fall or remain the same.

Throughout this research, multiple techniques of gathering data were used, such as:

- R was used to clean up the data which was not needed, such as if the client is a home owner or not
- SQL was used to organise the data
- Microsoft Excel was used to finalise neatly the report and forecasting.

The report is structured as follows:

- Literature Review: Review of current research on the provided dataset
- Research Methodology: Review of the approach taken to gather the data
- Data Gathered: Review of the data gathered
- Data Analysis: Discussion of the reports generated from the data
- Conclusion: Concluding points and recommendations

Please note that images, graphs and figures can be found in a seperate folder.

II. Literature Review

There are various studies of the current data, such as market analysis. However, one particular paper which caught the eye is one titled ”Retail Sales Prediction and Item Recommendations Using Customer Demographics at Store Level” by Michael Giering. This review paper used information from current sales to prepare customer recommendations for clients. Based on what the client buys, other similar products or products which go well together with the product bought initially by the client can be recommended to the client so that sales can be maximised.

The study was conducted by gathering data over an 18 month period in a number of different stores. The study catered for 600 different items which can be found in stores. Different customer types were also taken into consideration, depending on their age and income, among other factors. The products varied significantly and certain items were not present in all of the stores. The primary aim of this study was to develop a method to recommend products for better selection (such as which two products go well together).

The product recommender was to work similarly to Netflix and its program recommendations. Depending on what the user purchases, a similar item would be recommended depending on that item. The purpose of this review was to recommend certain items which can increase sale numbers. This can also be seen as predicting sales for items which were previously missing, as not all items were found in all the stores.

Although the dataset was very large, as with the one done on this report, there were two distinct advantages. The distributions of the data are well defined and that there is so much data that accurate statistics can easily be produced. Prediction of sales can also be made with a high enough accuracy.

As for methodology, the model was constructed by organising the stores based on their location, calculating how much each store sells and the creation of a model for store level. Organising the stores based on location was carried out in a combination of three methods: K-means clustering, Correlation clustering and PLSA clustering. This was done so that bias is minimised as much as possible. After clustering was complete, a voting system was used to assign a cluster to each store. With regards to the actual building of the model, a maximum percentage was defined so that the variance in the model can be realistic. Rank optimisation, the rank with regards to number of a particular product sold, can be described as the minimum number of singular vectors necessary to minimise error. Using this, imputation of missing data or existing data can be carried out.

Now that the model is complete, the graph can be produced. The graph was produced depending on customer type and the different products. The weights were calculated using a multi-linear regression method. However, a disadvantage of this approach is that the final result is heavily dependent on the path. On the other hand, an advantage is that results can be easily conveyed. These results were conveyed in a graph depicting the customer type, as can be seen below:

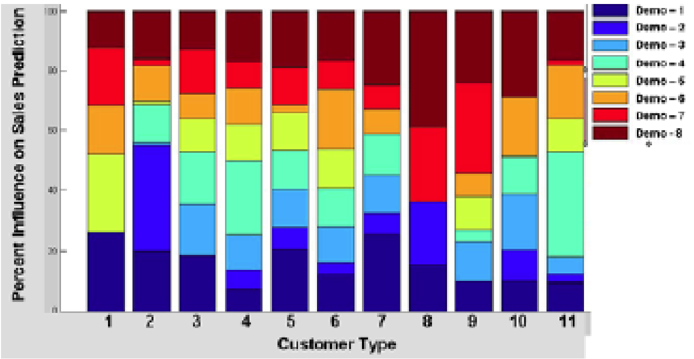


Fig. 1: Literature Review Graph

The product recommender can provide comparisons between sales of items and the expected sales of other items currently found in the system. When looking at the total sales for the store, information is mostly used for opitimising product selection. For any given store, the actual sales are constructed. This list is then sorted by total sales and the item descriptions are displayed. This can allow items to be easily compared. Items can also be colour coded so that people can see items that go well together. On a side note, values for future sales figures not currently carried out do not cater for separating of sales between items. For each store, one can obtain the distribution of items not currently being bought by certain customer types. For example, a certain class of people may not purchase a certain product.

The calculated sales values are then compared to the previously known sales values. Each non-zero value is removed and the expected value is created. Since the different segments of data were modeled at different ranks, the aggregated values are more accurate and give a better representation of the expected result. Looking at the data according to customer type, further insights can be gathered into whether outliers were being caused by store level implementation issues or shopping patters of customer sub sections.

Finally, results can now be presented and analysed. The model was run on the 6 months of out of sample data. Inputting the known sales values in each store organised by items total sales data, a linear regression of known against expected values was produced. Of the full product range, around 200 could be considered core items of high or moderate distribution. These products were the ones which benefited most from this approach. When the same test was performed across different customer types, the results varied significantly. Customer types with large shopping volumes showed larger improvement. However, this is not a surprise since they have a high level of statistical support.

This paper was able to develop a model which could predict the future sales depending on clients. This model was very useful in predicting future sales as well as being very accurate. This report was also based on a large data set over a significant period of time. The above paper was one of the reasons of choosing to forecast orders in this report.

Another interesting read is ”An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: Case study of printed circuit board” by Esmaeil Hadavandi, Hassan Shavandi and Arash Ghanbari. This paper highlights the importance of forecasting, as well as reports the performance of data clustering to be able to create a reliable and accurate sales forecasting system.

The report presented a hybrid artificial intelligence method named K-Means Genetic Fuzzy System (KGFS). This AI was used to construct an expert system for forecasting. A combination of K-means, GA and fuzzy logic approach to build up the expert system. K-means was applied to cluster the raw data. Then, the different clusters were fed into fuzzy systems. This technique is primarily used for partitioning sets of objects into a number of groups, such that each group is alike with respect to particular attributes, depending on the criteria at hand. This process includes randomly selecting an initial cluster, assigning each object to a cluster depending on how relative to that cluster it is and calculating the average for each cluster and reassigning objects. If the criteria converges, the process is stopped. If not, it starts assigning clusters again.

This report also highlights that fuzzy rule-based systems have been successfully applied in a wide range of real world scenarios and problems in different areas. However, it also highlights that several tasks need to be performed for the system to work properly and get an accurate result. One of the most important and time-consuming tasks is to collect an appropriate knowledge base regarding the scenario. The knowledge base has all the available knowledge in the form of IF-THEN rules. The difficulty presented by experts to express their knowledge in this form has made researchers develop automatic techniques which can perform this task. Therefore, a large number of methods can be automatically done.

The coding mechanism in this report is a GFS. A fuzzy decision table is able to represent a special case of crisp relation defined over the collections for fuzzy sets corresponding to the inputs and outputs. A chromosome is obtained from the decision table by analysing row by row and coding each output set as an integer number which start from 1 to the number of outputs. The fuzzy table will have two inputs (X1 and X2) as well as one output (Y). The table will also have three fuzzy sets (A1, A2, and A3) related to each input variable and for fuzzy sets (B1, B2, B3, and B4) related to the output. Next, the initial population is generated. Initial chromosomes are randomly generated. The rest have the same probability of being assigned to a gene. The next step is to calculate the values. The function is based on an application-specific measure usually employed in the design of GFSs. Next, the new chromosome is generated. A binary tournament is used for the selection procedure. In this type of selection procedure, two members of the current set are selected at random and their value is compared. The one with the best fitness value will be chosen to reproduce. Then, the best rule set in the population is newly generated to form the next population.

If the number of generations is equal to the maximum generation number, the process stops. If not, the process is repeated until the maximum generation number is reached.

After the above process, the data is tuned to adjust the shape of the membership function of the preliminary database. This approach is done as follows: each chromosome encodes a different database definition. Each triangular membership function is encoded by three values. A primary fuzzy partition is represented as an array which is made up of 3N values, with N being the number of linguistic terms for each variable. The complete database for a problem in which linguistic variables are involved is encoded into a fixed-length real encoded chromosome built by joining the representations of each one.

Finally, experimental results can be produced. In this paper, the data used was the first four years from the database. This data is used to forecast the sales over the next few years. The first stage involves records of data being inputted into the model and three different clusters are generated. Each cluster contains a part of test data. The second stage involves a GFS being built for each cluster using the related data. The final step includes sales forecasting by using each cluster's test data.

This paper, along with the previous one, show that there are many ways of forecasting data. The last paper shows how effects of messy data can be reduced by organising them into clusters. GAs are an important tool for automating certain long and repetitive processes. There can also be assurances that prime solutions can be expressed in an easy and understandable way. These two papers have also provided information that forecasting sales/orders for a business is crucial to a company. It can greatly help a business see where it is going and take decisions accordingly.

III. Research Methodology

The research methodology used was CRISP-DM, which stands for Cross-Industry Process for Data Mining. This methodology was able to provide a structured approach to this data mining report. Thanks to this methodology, data was efficiently extracted from the data set and organised neatly so that it could be used for this report. This methodology is split into six parts:

A. Business Understanding

In the initial stage of CRISP-DM, it is imperative that whatever needs to be accomplished must be understood perfectly from a business perspective. In this case, it must be understood that the desired objective is that the report produces an accurate representation of how the number of orders will perform in the next year based on previous years. The main objective is that accurate forecasting is performed. The dataset must be analysed properly and no external factors should have an effect on results, such as trying to have forecasted increased sales in each region, even if this is not true. Findings must be free from bias.

B. Data Understanding

After finalising the understanding of the business, one must now understand the data at hand. In this case, the data provided has various aspects, such as client information, including specific attributes such as how many cars they have, whether they are a home owner or not and how many children they have, as well as basic information such as name, surname and date of birth. Other data includes information about the different stores in the different cities, regions and countries, as well as information on products and orders. The possibilities for analysing data are endless. In this part of the process, the data was briefly analysed so that it could be understood properly and it was seen that it made sense and could be used for generating the report for forecasting.

C. Data Preparation

Next, after making sure that the data is correct and makes sense, it is time to prepare the data for usage. The data was first chopped down into what was needed only. That is, client information, order information and city and region information. The rest of the data was not needed for forecasting orders. The next step was to clean up the remaining data. Information such as whether the client is a home owner or not was removed as it does not effect order forecasting. The remaining data was then prepared to be used in the ETL process so that the data can be processed accordingly, depending on the order, client, region and order date. Finally, the data was integrated into the final dataset which will be used for forecasting.

D. Modelling

Now that the data is ready for use, it will now be modelled as needed. In this case, the data was set up to use information regarding the client, the location where the order was made and the date of the order. Combined, these created the order fact. In this, the order could be built up depending on the client, store location and the date of the order. These three factors were key in providing the information for the fact.

E. Evaluation

Now that the data is gathered, it can now be evaluated as needed. In this case, the data was evaluated depending on the region and the order date. Information was gathered along every quarter in every region over a two year period. For each quarter in every region, the number of orders per region were recorded and evaluated. Information was reviewed properly so that no region or quarter was overlooked in the build up of this model.

F. Deployment

Finally, the results from the evaluation were then deployed to create a chart depicting the current amount of orders and the predicted outcome for the next year. Based on orders from the previous year, a forecast for the next year was produced. Although not fully accurate, it can be a good indication of what is to come for the business and where improvements can be made.

IV. Data Gathered

The data gathered was according to date, client and store location. These three together then formed the order fact. The order fact had the order date, who made the order and where the order was made from. A diagram of the actual cube used can be found below:

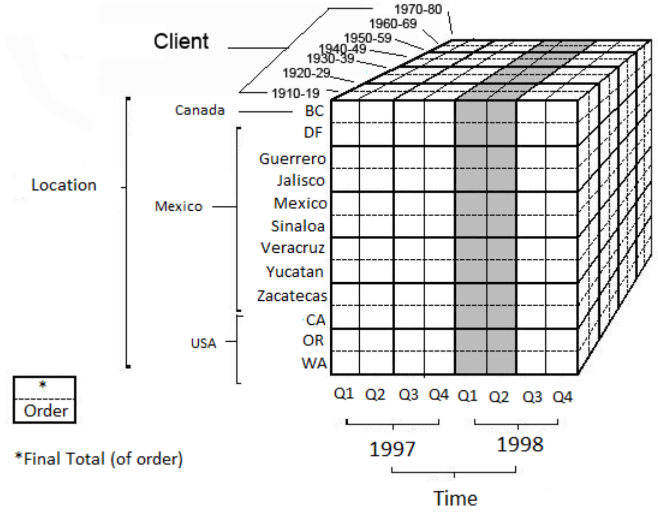


Fig. 2: OLAP Cube

As previously stated, the cube is split into three dimensions: time, location and client. Time is further split into the two years of data available which in turn is split into quarters. Locations are divided into countries which in turn are divided into regions. Finally, the clients are organised according to their age/year of birth.

From the cube, the ERD was produced. The client, location and dates together formed the order fact. The ERD can be seen in the figure below:

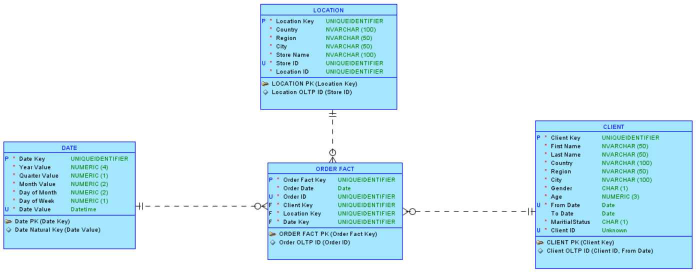
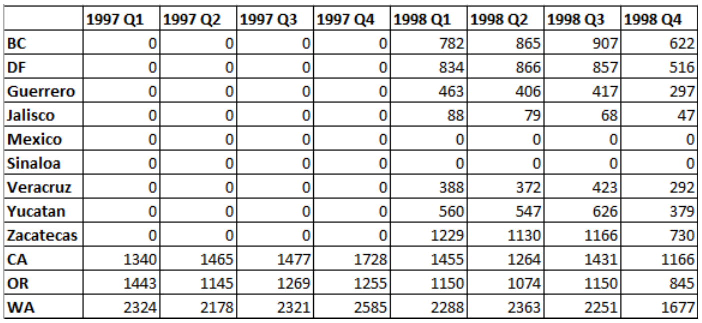

LOCATION
P * Location Key UNIQUEIDENTIFIER
* Country NVARCHAR (100)
* Region NVARCHAR (50)
* City NVARCHAR (50)
* Store Name NVARCHAR (100)
U * Store ID UNIQUEIDENTIFIER
* Location ID UNIQUEIDENTIFIER
LOCATION PK (Location Key)
Location OLTP ID (Store ID)
DATE
P * Date Key UNIQUEIDENTIFIER
* Year Value NUMERIC (4)
* Quarter Value NUMERIC (1)
* Month Value NUMERIC (2)
* Day of Month NUMERIC (2)
* Day of Week NUMERIC (1)
U * Date Value Datetime
Date PK (Date Key)
Date Natural Key (Date Value)
ORDER FACT
P * Order Fact Key UNIQUEIDENTIFIER
* Order Date Date
U * Order ID UNIQUEIDENTIFIER
F * Client Key UNIQUEIDENTIFIER
F * Location Key UNIQUEIDENTIFIER
F * Date Key UNIQUEIDENTIFIER
ORDER FACT PK (Order Fact Key)
Order OLTP ID (Order ID)
CLIENT
P * Client Key UNIQUEIDENTIFIER
* First Name NVARCHAR (50)
* Last Name NVARCHAR (50)
* Country NVARCHAR (100)
* Region NVARCHAR (50)
* City NVARCHAR (100)
* Gender CHAR (1)
* Age NUMERIC (3)
U * From Date Date
To Date Date
* MaritialStatus CHAR (1)
U * Client ID Unknown
CLIENT PK (Client Key)
Client OLTP ID (Client ID, From Date)

Fig. 3: Order Star Schema

Various attributes were kept on record, such as the clients location and age. These attributes were key to producing the final report with just the necessary data. For the client, ”From Date” and ”To Date” attributes were also included. This is because certain details about the client may change over time. This is known as a Slowly Changing Dimension (SCD). For example, a woman who gets married may change her surname. The ”from” and ”to” dates cater for this, as they represent between which dates the data is valid for.

The data gathered, as mentioned previously, is organised by region and by quarter. The number of orders per region per quarter were recorded and put in a table.



	1997 Q1	1997 Q2	1997 Q3	1997 Q4	1998 Q1	1998 Q2	1998 Q3	1998 Q4
BC	0	0	0	0	782	865	907	622
DF	0	0	0	0	834	866	857	516
Guerrero	0	0	0	0	463	406	417	297
Jalisco	0	0	0	0	88	79	68	47
Mexico	0	0	0	0	0	0	0	0
Sinaloa	0	0	0	0	0	0	0	0
Veracruz	0	0	0	0	388	372	423	292
Yucatan	0	0	0	0	560	547	626	379
Zacatecas	0	0	0	0	1229	1130	1166	730
CA	1340	1465	1477	1728	1455	1264	1431	1166
OR	1443	1145	1269	1255	1150	1074	1150	845
WA	2324	2178	2321	2585	2288	2363	2251	1677

Fig. 4: Number of Orders per region per quarter

As one can see from the above table, there were no orders in 1997 in the majority of regions. This made the forecasting procedure a bit less accurate as there was little data to compare. For Mexico and Sinaloa there were no orders in both 1997 and 1998, so forecasting could not be done on these two regions. With regards to the other regions, there were quite a good number of orders in each quarter for most of the regions, which will be further analysed in the next section of this report.

V. Data Analysis

Since the data we have is now easily readable and organised properly, it can be used to forecast future orders. All the orders were put into a line graph so that the number of orders per region per quarter can be seen better. Through this graph, forecasting was also produced.

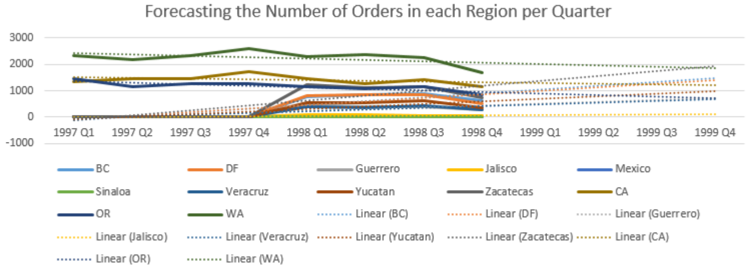


Fig. 5: Graph representation of number of orders including forecasting

The solid lines in the graph show the actual orders in the years 1997 (where applicable) and 1998. The dotted lines represent the forecast for the number of orders across the regions across the different quarters of the year.

As one can see from the graph, WA, where the most orders were made from, seem to be going on the decline. Although not a steep decline, it still might be a little worrying for the business that their most successful location is going backwards. On the other hand, Yucatan seems to be enjoying a steady rise in the number of orders, even if the final quarter of 1998 saw a much smaller number of orders when compared to the previous quarters. However, a large drop in orders was noticed across all regions in the final quarter of 1998. For example, a number of regions have seen more than half of their orders lost from the 3rd quarter to the 4th. This could be due to a number of factors, such as people not wanting to spend too much during the Christmas period.

When comparing the three regions which have orders in 1997 as well, the results were somewhat mixed. In CA, quarters 2, 3 and 4 fared much worse in 1998 than in 1997. In Q4 alone, there was a drop of around 500 orders. when compared to the same time frame the previous year. In OR, the story is worse, as all 4 quarters reported a drop in orders over the previous year, with no signs of improving. Only between the 2nd and 3rd quarter did the number of orders increase. Along the rest of the year, they kept on going down. Finally, in WA, where the most orders were made across all the regions, it was very mixed. 1997 saw the 2nd quarter drop some orders, but then there were increases in both the 3rd and 4th quarter, making it quite a positive year. However, the same heights could not be reached in 1998, with a drop in the 1st quarter, a slight rise in the 2nd, another drop in the 3rd and a significant drop in the final quarter of 1998.

VI. Conclusion

All in all, this research was a challenge but one which was fun to do. The main highlight was finally finding all the number of orders and organising them as needed. However, since in some regions there were no orders at all, it was a bit disappointing not being able to perform forecasting for these regions as well. Also, most regions did not have any orders in the year 1997, which made the predictions much less accurate than they could have been had there been orders. The more data that is available, the more accurate the forecast can be.

Although predicting the future is never easy and almost impossible, forecasting can give an indication of how a business is performing over the years. This business can see where orders are dropping and how they can be increased. As time goes by, more experience is gained and more can be learned about the market and the business can adjust accordingly. By forecasting, businesses can learn their sales curve better and act accordingly.

For future research, as previously mentioned, the more data about sales and orders provided the better. More data means that results will be more accurate. More information regarding clients or stores can also help in having an accurate forecast.