**Sales Estimation:**

**Project proposal**

**By:**

**Kunal Sontakke(802329466)**

**Sarthak Saxena(802325076)**

**CPSC 531-03 (20850)**

**Professor: Dr. Chun-l Philip Chen**

**Department of Computer Science**

**California State University, Fullerton**

**Spring, 2015**

**Table of Contents**

1. Introduction……………………………………………………………………………..4
2. Problem Statement………………………………………………………………………5
   1. Project goals and benefits…………………………………………………………..5
   2. Relevance and significance…………………………………………………………6
   3. Assumptions, Limitations and Delimitations……………………………………….7
3. Literature review…………………………………………………………………………8
   1. Data Mining Tasks………………………………………………………………….10
   2. KDDVM platform…………………………………………………………………..11
4. Approach…………………………………………………………………………………12
   1. Technical approach………………………………………………………………….14
   2. Proposed Methodology……………………………………………………………...17
5. References……………………………………………………………………………..…19

**Abstract**

Sales forecasts are estimates of your sales for the forecast period. The sales forecast establishes the level of activity used in all the other forecasts and budgets for the business. If your sales forecast varies wildly from your actual results, your cash flow and profitability forecasts will similarly be inaccurate. The entire sales forecasting process is intended to generate better business results from better business decisions. So what really matters is whether or not your sales forecast is set up right and broken into factors you’ll be able to track and will lead you to better management decisions.

The main objective of this project is to analyse the historical sales data available for many departments of a company by exploring different data mining techniques.

The datasets will undergo through various data mining tasks such as survey analysis, training the model using the dataset and prediction of knowledge using proper logic. Different data mining tasks, techniques and tools will be used for the sales estimation from the extracted data of the dataset.

1. **Introduction**

The most important thing in retail is to continue satisfying customer needs and their demands. As seen in today’s world competition is increasing and so is the expectations of customers. This project introduces the concept of data mining in this process. The concept behind this project is to increase the sales by analysing and keeping track of customer’s habits. This activity is known as “Knowledge Discovery in Database”. (Maimon & Rokach, n.d.) The motive of this project is to increase the sales by using data mining techniques. This includes algorithms which are used in order to provide such kind of techniques.

1. **Problem statement**

The sales of a particular store has gone bad. It depends on the quality, quantity and the prices which are available at the store. There are various discounts to be added in order to improve the sales. The store needs a proper plan about how the sales should be increased, improvement and the growth of the store. They need to find special occasions to implement it may be its holidays or special days. For implementation of this, it is essential to look for the historical data in respective departments of the company.

The holiday markdown events are included. It is very important to note down the effects of holiday markdown events and particularly how they effect it. The effect and impact of every holiday is marked and analysed. The lack of awareness and prediction techniques causes failure in the sales and it is hard to find it out.

* 1. **Project Goals and Benefits**

The benefits of using Knowledge Discovery in Database would be the increment in sales, keeping the customers happy, efficiently obtaining the goals. This includes the design of a pattern on which the activity should be working. The store is to be analysed by considering different occasions and by considering different conditions. This is beneficial in maintaining the inventory of the store and providing the customers the product they want. This is helpful in maintaining the image and the brand value of the store.

If the store needs some changes then it can even do it by following the pattern. For example, if a new product is to be introduced or may be if someone wants to try or replace some product then according to the customer need, a store can do that.

* 1. **Relevance and Significance**

As seen in today’s world statistics plays a very crucial and important role. It is very important to find the patterns on which the store is running. It helps in providing the decision criteria. The biggest significance is to give the user that confidence in taking a decision. This software or this technique will give user a confident and reliable grounds on which decisions can be made. The user can match and compare with the help of data provided.

In case of failures, data mining technique is the most efficient way to get over it. It gives the user a proper refined data on which a user can rely. For example, if a store is not giving proper sale then the owner would not be in position to take a big decision for the improvement. But after using KDD a user can think with a 360 degree approach and implement for better results. The motive of KDD is to provide a qualitative and quantitative analysis of the data.

* 1. **Assumptions, Limitations and Delimitations**

For example, if we consider animals they look for food with some kind of assumptions. Similarly a data analyst will also make assumptions for a task to be executed. (Murray & Gardiner, n.d.) This based on assumptions of sales using historical data and analysing it for mining purposes. The limitations of this process may be considered as the availability of data. The inconsistency in data availability might affect the process.

1. **Literature Review**

The knowledge data discovery process is processed in nine steps. This process is iterative and in each step going back might occur. There are different needs and possibilities at each step of this process. The entire structure of this process is shown in the Figure1.

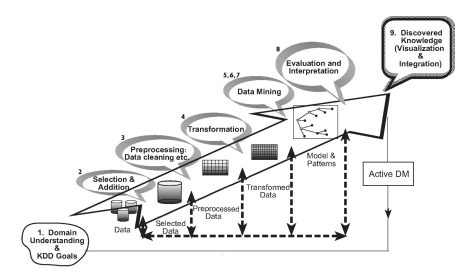


Figure1: Needs and possibilities of KDD

First step of the Knowledge Data Discovery process is to develop and understand the application domain. This is a very initial process which prepares the system for further steps. In this process the goals of the application are decided. After understanding the goals the pre-processing of data is started. Second step is to select and create the dataset on which the discovery is to be operated. In this step the availability of data is decided, which type of data is required and then all the data is integrated into one data set which includes all the important attributes of this data. In this all the attributes should be present if not then the entire process can collapse. Third step is the preprocessing and cleansing. In this particular step the data mishandling is improved. If there are certain values which are missing they are cleared and the entire data set is to be cleaned. It might be possible by using complex algorithms and patterns.

Fourth step of this process is data transformation. This step includes cropping out some irrelevant data which is not useful. This step is essential because it is providing the specifications for the entire KDD process. Fifth step is to choose the appropriate data mining task. After so much of filtering the process is ready to choose its data mining task which should be appropriate for the particular process. The objectives of this step is prediction and description. Sixth step is to choose the data mining algorithm. Before this step the strategy has been made and now the task is to make tactics. This might include finding the searching method for process or may be the technique to search.

The Seventh step is to employ the data mining algorithm. This includes the implementation of the data mining algorithm which is decided in sixth step. This step might include the iteration of that algorithm just to make sure its validity and efficiency. Eight step is to evaluate. In this step the evaluation of the pattern is done. It keeps a track of fulfilling the goals of the project and maintaining its efficiency. This step specifies the usefulness and the efficiency of the pattern used. The discovered knowledge is documented for the further use of the system. Ninth and the last step of this process is to use the discovered knowledge. This step makes it useful for the knowledge discovered for the further purpose. This step also includes the changes made into data and noticing its effects in the system. Hence, this way the KDD process is completed in its nine stages. (Maimon & Rokach, n.d.)

* 1. **Data Mining Tasks**

There are various data mining tasks which are used in the professional world. Data Pre-processing is one of the task which is used in data mining. This task is responsible for converting the present data into a suitable form for data mining. The objective of this task is to convert the information into a more useful form. Classification is another task which is used in data mining. This task assigns the different conditions and it classifies into various categories. It collects the attributes and try to match each other and then classifies it together. Regression is another task used in data mining. This is similar to classification but it works according to the most predictable attribute value. It is helpful in providing the trends of the attribute value.

Cluster analysis is another task used in the data mining process. It forms the data into clusters of information. The information is divided into particular groups and the objects in the group has same attributes as compared to the objects in another group. The cluster analysis approach is mainly used for marketing based data mining. Another task used for data mining is Sequence analysis. It predicts patterns in discrete series of information. Forecasting is one of the efficient task in data mining. It compares the data set with time and accordingly decides more values. It is basically used in predicting the information and use the most efficient data.

Deviation analysis is another classified task used in the data mining. This is used to detect the exceptional cases for the particular dataset. This technique is mostly used by Banks for fraud detection. Data Visualization is a very important task in data mining as it is makes easy to interpret ate the complex visuals into a simplified form. It is a recent boom in business applications.

* 1. **KDDVM platform**

KDDVM is service supporting Discovery, Composition and Activation. In this format users are decided on the basis of their roles which are publishers, providers and consumers.

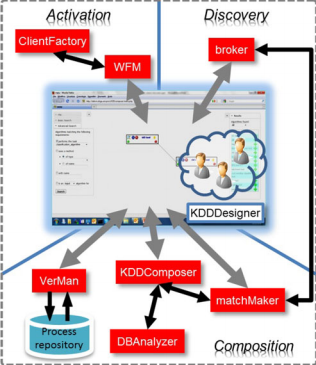


Figure2: Services by KDD

As shown in the figure 2, the services are divided into three parts which are activation, discovery and composition. The discovery service uses the browsing of basic services from the repository then retrieving the detail about every particular service. Composition service basically decides the design of the process at various levels of abstraction. On the basis of user requirements, most suitable and used service is decided. Activation service is responsible for the execution of the entire process. This includes execution of single service and the transfer of each service from one to another.

1. **Approach**

The problem solution will be written in R programming. R is a tool for statistics and data modelling. The R programming language is elegant, versatile, and has a highly expressive syntax designed around working with data. R is more than that, though — it also includes extremely powerful graphics capabilities. (code school, n.d.)

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. (The R foundation, n.d.)

The backend or the database proposed for the solution will be achieved by the use of CSV. CSV is a common file format that is widely supported by consumer, business, and scientific applications. Among its most common uses is moving tabular data between programs that natively operate on incompatible (often proprietary and/or undocumented) formats. (wikipedia, 2015)

The so-called CSV) format is the most common import and export format for spreadsheets and databases. There is no “CSV standard”, so the format is operationally defined by the many applications which read and write it. The lack of a standard means that subtle differences often exist in the data produced and consumed by different applications. (Python software foundation, n.d.)

The csv module implements classes to read and write tabular data in CSV format. It allows programmers to say, “write this data in the format preferred by Excel,” or “read data from this file which was generated by Excel,” without knowing the precise details of the CSV format used by Excel. Programmers can also describe the CSV formats understood by other applications or define their own special-purpose CSV formats. (Python software foundation, n.d.)

The dataset for this problem will be acquired from KDD. KDD (Knowledge Discovery in Database) is modelling and analysis of the large data repositories. The data mining is the basis for KDD, involving construction of the algorithm, developing the model for unknown pattern and predict the useful solution with the help of suitable logic.

* 1. **Technical Approach**

The editor for R programming of the graphical user interface to be used for R programming will be Rstudio. RStudio is a free and open source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.

RStudio Desktop, where the program is run locally as a regular desktop application; and RStudio Server, which allows accessing RStudio using a web browser while it is running on a remote Linux server. Prepackaged distributions of RStudio Desktop are available for Microsoft Windows, Mac OS X, and Linux.It is written in C++ and uses a Qt framework. (wikipedia, 2015)

CSV files can be viewed in Microsoft Excel or CSVed. CSVed 2.3.2 is the latest version of CSVed. It is an easy and powerful CSV file editor, you can manipulate any CSV file, separated with any separator. The following csv files will be included in the solution.

* + - 1. **stores.csv**

This file will contain anonymized information about the 45 stores, indicating the type and size of store.

* + - 1. **train.csv**

This will be the historical training data, which covers to 2010-02-05 to 2012-11-01. This file will contain the following fields:

* Store - the store number
* Dept - the department number
* Date - the week
* Weekly\_Sales -  sales for the given department in the given store
* IsHoliday - whether the week is a special holiday week
  + - 1. **test.csv**

This file is identical to train.csv, except we have withheld the weekly sales. You must predict the sales for each triplet of store, department, and date in this file.

* + - 1. **features.csv**

This file contains additional data related to the store, department, and regional activity for the given dates. It contains the following fields:

* Store - the store number
* Date - the week
* Temperature - average temperature in the region
* Fuel\_Price - cost of fuel in the region
* MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
* CPI - the consumer price index
* Unemployment - the unemployment rate
* IsHoliday - whether the week is a special holiday week

For convenience, the four holidays fall within the following weeks in the dataset (not all holidays are in the data):

* Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
* Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
* Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
* Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13
  1. **Proposed Methodology**

The proposed methodology for the problem solution is the quantitative approach. This will include maintaining the validity of the data sets, for this the data set will be required to be updated periodically. Updating the data sets will help us maintaining the accuracy of the predictions or the results.

The datasets will undergo through various data mining tasks such as survey analysis, training the model using the dataset and prediction of knowledge using proper logic. Different data mining tasks, techniques and tools will be used for the sales estimation from the extracted data of the dataset.

The analysis of the data set will be as discussed below.

The way holidays in the store will be dealt is to create distribution of sales for each holiday based on the date, with 3 parameters, width, skew and location (relative to the actual date). We will be making these distributions on a daily grid and then sum them up to weekly totals. Using skewed distributions will be key here for us. We will largely fit the parameters by eye to begin with.

For each store/department combination, after calculating the trend in the data, we will use a linear model with L1 regularization to fit the holidays to the detrended data. Then after subtracting the holiday fit and trend, we will take an average of value for the week over the years to find the residual weekly cycle that was not due to the holidays.

Then we will fit the trend + deseasonalized data using the Unemployment, Fuel Price and CPI, using another linear model with L1 regularization. We will calculate the missing data using a simple AR model. This fit will give a small improvement over using the pure trend.

We will use csv over the first two years / last 39 weeks split to pick whether the trend was constant or linear. For example something happened at store 14, which caused a dramatic drop in sales across all departments, so we will apply a step function to account for this.

Our conclusion is that since we only have 1 year of markdown data in the dataset it is impossible to extract anything useful from it, as we could not see the effect from one year to the next.

The dataset used here will be acquired from KDD which will provide us with a large volume of inputs to analyse and provide accurate estimates. As this is a solution to provide forecast or predictions we will require a large dataset, getting a dataset from KDD will be very useful. This data set will have two entries first the combination of item id, store id and the year, the second entry will be the total sales for that item.

1. **References**

code school. (n.d.). Retrieved from tryr.codeschool.com: http://tryr.codeschool.com/

Maimon, O., & Rokach, L. (n.d.). INTRODUCTION TO KNOWLEDGE DISCOVERY.

Murray, C., & Gardiner. (n.d.). In *Inquiring Organizations: Moving from Knowledge Management to Wisdom.*

Python software foundation. (n.d.). *CSV File Reading and Writing*. Retrieved from docs.python.org: https://docs.python.org/2/library/csv.html

The R foundation. (n.d.). *The R Project for Statistical Computing*. Retrieved from r-project.org: http://www.r-project.org/

wikipedia. (2015, April 13). *Comma-separated values*. Retrieved from wikipedia.com: http://en.wikipedia.org/wiki/Comma-separated\_values

wikipedia. (2015, March 5). *RStudio*. Retrieved from wikipedia.org: http://en.wikipedia.org/wiki/RStudio