



**Walmart Store Sales Forecasting**

Submitted to: Mr. Yogendra Singh

Submitted by: Sanjoli Jain, Jayshree Vashistha,

Diksha Kejriwal

TABLE OF CONTENT

* Certificates
* Acknowledgement
* Abstract
* Introduction
* Theory
* Problem Statement
* Data Description
* Solution Approach
* Data Pre-processing
* Machine Learning Model used
* Data Visualization
* Result Analysis
* Annexure

Certificate

Date: 29/06/19

This is to certify that **Ms. Sanjoli Jain**, student of 3rd year from Department of Computer Science, Poornima College of Engineering has undergone a Project work from June 17, 2019 to June 28, 2019 in Data Science & Machine Learning titled **Walmart Store Sales Forecasting**.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Project Incharge Seal**

**Certificate**

Date: 29/06/19

This is to certify that **Ms. Jayshree Vashistha**, student of 3rd year from Department of Information Technology, Poornima College of Engineering has undergone a Project work from June 17, 2019 to June 28, 2019 in Data Science & Machine Learning titled **Walmart Store Sales Forecasting.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Project Incharge Seal**

Certificate

Date: 29/06/19

This is to certify that **Ms. Diksha Kejriwal**, student of 3rd year from Department of Information Technology, Poornima College of Engineering has undergone a Project work from June 17, 2019 to June 28, 2019 in Data Science & Machine Learning titled **Walmart Store Sales Forecasting.**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Project Incharge Seal**

Acknowledgement

For making this project, we would like to thank our trainer Mr. Yogendra Singh, who always helped us time to time to understand our topic and the various related areas and guiding us throughout the training. We would also like to mention a word of thanks to all Training Assistants for giving us such attention and time and helping us throughout the training.

Abstract

Rapid growth of data comes with a challenge of sorting and analysing them, where raw data exists in graphical form, textual form or in images. Data science and machine learning finds its application in various fields like stock market, recommendation systems, image processing, aerial photography, military, weather forecasting etc.

This report is about our project on predicting “Walmart store weekly sales” which addresses about the data pre-processing and post processing which includes plotting and prediction of weekly sales for each department of each store of Walmart and the ability of machine learning algorithms to deal with different set of data.

In this project, we tackled a regression problem of predicting the weekly sales of each department in each store of Walmart by accessing several type of data like Unemployment rate, markdowns, CPI, store size, store type etc. We have testes and used Linear Regression, Decision Tree and Random Forest to determine results. Also, we have used some other libraries for data visualization and manipulation.

**Introduction**

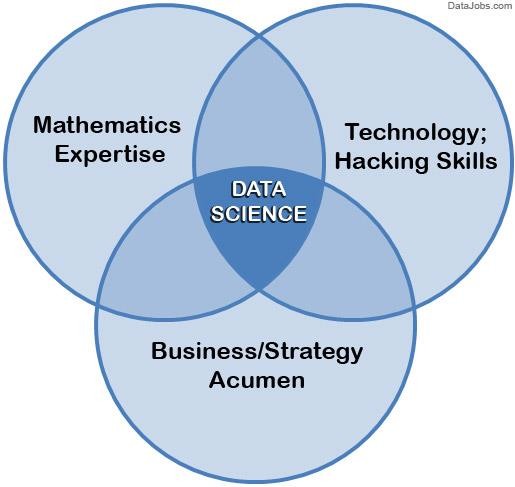
Walmart is the world’s largest company by revenue, according to the Fortune Global 500, as well as the biggest private employer in the world with 2.2 million employees. Walmart is a family owned business, as the company is controlled by the Walton family.

Walmart Stores, Inc. is an American multinational retail corporation that operates a chain of discount department stores and warehouse stores. It has over 11,000 stores in 28 countries, under a total of 65 banners. The company operates under the Walmart name in the United States and Canada. It has wholly owned operations in Argentina, Brazil and Canada. It also owns and operates the Sam’s Club retail warehouses.

With historical sales data for 45 stores located in different regions. Each store contains many departments and the aim is to project the sales for each department in each store. To add to the challenge, selected holiday markdown events are included in the dataset. These markdowns are known to affect sales.

Theory

**Data science** is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization.



**Fig. 1 Data Science**

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the data.

Supervised and unsupervised learning describe two ways in which machines algorithms can be set loose on a data set and expected to learn something useful from it.

##### Supervised: It is a learning in which we teach or train the machine using data which is well labeled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data.

Supervised learning classified into two categories of algorithms:

* **Classification**: A classification problem is when the output variable is a category, such as “Red” or “Blue” or “disease” and “no disease”.
* **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

**Unsupervised:** Unsupervised learning is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.

Unsupervised learning classified into two categories of algorithms:

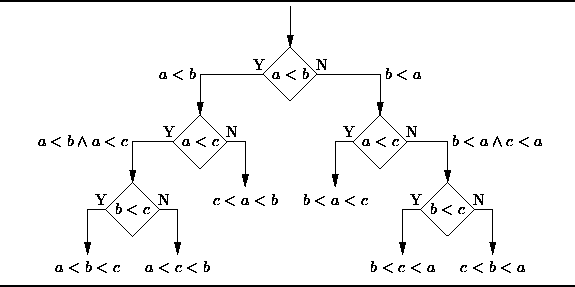
* **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.
* **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

##### Some Machine Learning Models

##### Simple Linear Regression: It is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other.

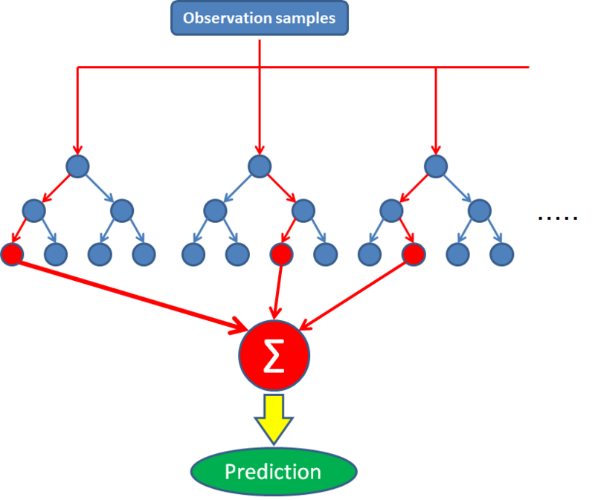
##### 

**Fig.2 Simple Linear Regression**

1. **Decision Tree:** Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any Boolean function on discrete attributes using the decision tree.

**Fig.3 Decision Tree**

1. **Random Forest:** Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results — in the case of a random forest, the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer.

****

**Fig.4 Random Forest**

Problem Statement

In the Problem, We are provided with historical sales data for 45 Walmart stores located in different regions. Each store contains a number of departments, and we are tasked with predicting the department-wise sales for each store.

In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modelling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data.

**Data Description**

To understand the methodology adopted, we first understand our dataset and variables.

**Dataset:**

Our dataset consists of 421570 total training data having weekly sales in train.csv file whereas test.csv consist of total 115064 testing data having features store type, department number, if it was holiday or not?, and date etc. The target is to train a machine-learning algorithm so that it can predict weekly sales of Walmart store based on department and store number in future.

**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 – Date of Friday of each week.

Temperature - Average temperature in the region.

Fuel Price- cost of fuel in the region

Markdown 1-5 - anonymised 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 (null or nan values).

CPI - the consumer price index

Unemployment - the unemployment rate

Is Holiday? - 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

Labour 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

**train.csv**

This is the historical training data, which covers to 2010-02-05 to 2012-11-01. Within this file you will find the following fields:

Store - the store number

Dept - the department number in a particular store

Date - Date of Friday of each week.

Weekly Sales – weekly sale of a department of a given store

Is Holiday? -whether the week is a special holiday week

**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.

**stores.csv**

This file contains anonymised information about the 45 stores, indicating the type and size of each store.

**Solution Approach**

As we have our prediction (label) as continuous values, so we applied regression algorithm to solve our problem. For the solution of the problem, At first we observed that there are some missing values in the features dataset, hence we replaced the missing values with zero. After replacing missing values, another dataframe has been created which has been merged by using "INNER JOIN" on the basis of store, date and Isholiday columns on features and store dataset. Similarly, train dataset have been merged to the previous dataset which has been made previously using "INNER JOIN". Now, we have complete set of features which is responsible for the prediction of weekly sales.

Now, in our features dataset we have categorical values as well as we will have to handle date column. To handle categorical values, we used label encoding to convert it into integer values. For date column, we have to convert date column first into date-time object and further applying toordinal() to convert it into integer format. We also have repeated weekly sales as the data is all about weekly sale of 45 stores of US in which total 99 departments are present. Hence, by applying groupby on the basis of store and date, weekly sum of sales of each store. After applying these methods, the column values have been randomized hence index of the dataset has been reset so as to arrange it in an appropriate way.

As we have total 15 columns and each column have different range values,hence we will have to scale it to same extent by using feature scaling.

Since, our training dataset is ready to process further, now we will apply various regression algorithms to check which will provide better results. We have applied

Linear Regression algorithm which yields average accuracy score approximately 83% .We also applied Decision tree regressor algorithm which yields average accuracy score approximately 91%. Finally,we applied random forest regressor algorithm and it’s average accuracy score was approximately 94%. Hence, we choose, the best result provider that is random forest regressor algorithm.

After training the model, We have saved the trained data into pickle file with all the available objects made using sklearn library and utilized those pickle files for the prediction.

**Data Pre-processing**

In the data pre-processing as we have some nan values in our data, so we replaced them with 0. We have also combined the store dataset with the features dataset. Also we have included the total features to the train and test dataset for predicting the weekly sales.

Libraries used

* numpy
* pandas
* matplotlib
* pickle
* scikitlearn
* sklearn.cross\_validation
* sklearn.preprocessing
* sklearn.linear\_model
* sklearn.ensemble
* sklearn.model\_selection

**Machine Learning model used**

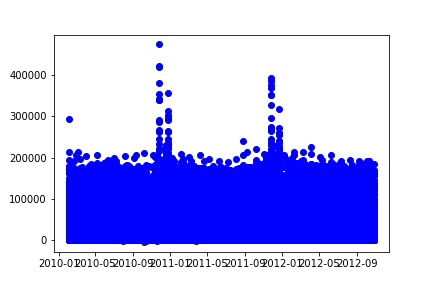
Here, In this project we have used Random forest regressor algorithm as the average score after training was about 94% which was maximum with respect to other algorithms such as Linear Regression(84%), Decision Tree (92%) approx. However as the data is very large and we had many attributes so the training part of the data took very long time. So we had trained the data single time only and had saved the trained data in the pickle file, from where we can use that train data to make the predictions.

Data Visualization

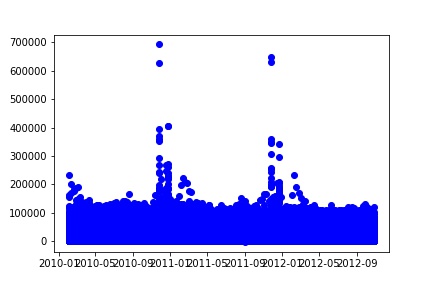
For the visualization of the data we had used matplotlib library. Through matplotlib, we were able to visualize weekly sales of the each store, weekly sales of each department on the basis of type of store. There are three type of stores in the given data i.e. Type- A, B, C. So, we are plotting the weekly sales according to the type of store. For example, we can consider that the stores are classified according to the location where store is located.

Result Analysis

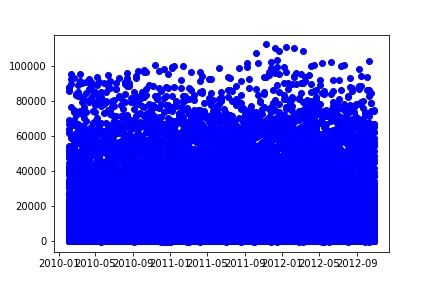
On plotting the sales of stores against the date of each Friday in the week for each store, we come to know that the sales for the Type C store is highest, sales of Type A store is medium and the sales of the Type B store is minimum. So after analysing the plotting we can predict that the type of store also plays an import role in the weekly sale of that store.



**Fig.5 Type A**



**Fig.6 Type B**

****

**Fig.7 Type C**

Conclusion

Here in this project we were able to correctly visualize the Walmart store’s weekly sales prediction, established at forty five places around the US and also their department-wise sales in those forty five places. We were able to correctly predict the weekly sales based on type of store, if it was holiday or not, date, CPI, unemployment rate, markdown prices etc. After applying different suitable machine learning models we have selected random forest regressor for the prediction of our weekly sales. On applying random forest regressor algorithm, we got an excellent average score that is 94.51%.

Annexure

Here is the link for the github repository in which you can find the above project with the source code, dataset and the framework using Flask.

Link: