



**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
* Methodology
* Data Pre-processing
* Machine Learning Model used
* Data Visualization
* Result Analysis
* Annexure

Certificate

Date: 29/06/19

Certificate

Date: 29/06/19

Certificate

Date: 29/06/19

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 have 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 ratio, markdowns, CPI, store size, store type etc. We have tested and used Linear Regression, Decision Tree and Random Forest to determine the 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.[3] 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.

**Data science – discovery of data insight**

This aspect of data science is all about uncovering findings from data. Diving in at a granular level to mine and understand complex behaviours, trends, and inferences. It's about surfacing hidden insight that can help enable companies to make smarter business decisions.

For example:

Netflix data mines movie-viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.

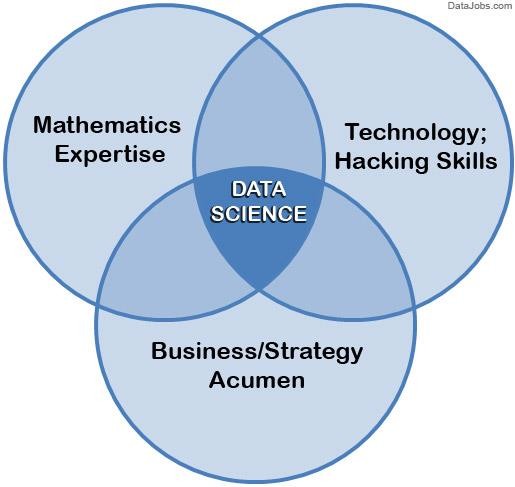
**Data science – development of data product**

A "data product" is a technical asset that: (1) utilizes data as input, and (2) processes that data to return algorithmically generated results. The classic example of a data product is a recommendation engine, which ingests user data, and makes personalized recommendations based on that data.

**For example:**

Amazon's recommendation engines suggest items for you to buy, determined by their algorithms. Netflix recommends movies to you. Spotify recommends music to you.

Machine learning and statistics are part of data science. The word learning in machine learning means that the algorithms depend on some data, used as a training set, to fine-tune some model or algorithm parameters. This encompasses many techniques such as regression, naive Bayes or supervised clustering.



**Fig. 1 Data Science**

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:

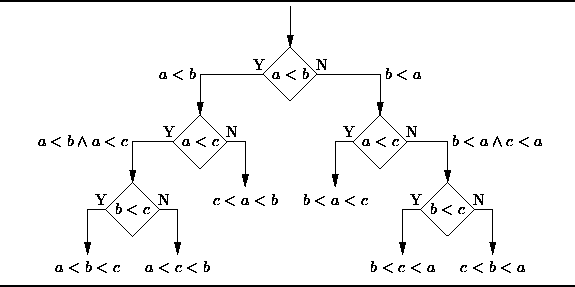
If we are training our machine-learning task for every input with corresponding target, it is called [supervised](https://en.wikipedia.org/wiki/Supervised_learning) [learning,](https://en.wikipedia.org/wiki/Supervised_learning) which will be able to provide target for any new input after sufficient training. Our learning algorithm seeks a function from inputs to the respective targets. If the targets are expressed in some classes, it is called *classification* problem. Alternatively, if the target space is continuous, it is called *regression* problem.

* **Regression** analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables.
* **Classification** model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. Outcomes are labels that can be applied to a datasets.

**Unsupervised:**

If we are training our machine-learning task only with a set of inputs, it is called unsupervised learning, which will be able to find the structure or relationships between different inputs. Most important unsupervised learning is *clustering*, which will create different cluster of inputs and will be able to put any new input in appropriate cluster.

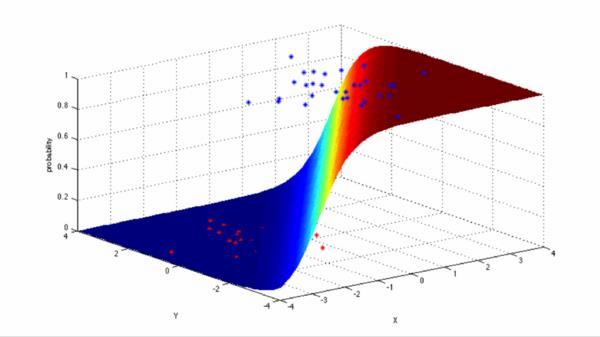
* **Cluster** analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory [data mining,](https://en.wikipedia.org/wiki/Data_mining) and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [machine learning](https://en.wikipedia.org/wiki/Machine_learning), [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis,](https://en.wikipedia.org/wiki/Image_analysis) [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [bioinformatics,](https://en.wikipedia.org/wiki/Bioinformatics) [data compression,](https://en.wikipedia.org/wiki/Data_compression) and [computer graphics.](https://en.wikipedia.org/wiki/Computer_graphics)

1. **Decision Trees:** A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility.

**Fig. 2 Decision Tree**

From a business decision point of view, a decision tree is the minimum number of yes/no questions that one has to ask, to assess the probability of making a correct decision, most of the time. As a method, it allows you to approach the problem in a structured and systematic way to arrive at a logical conclusion.

1. **Logistic Regression:**  It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.



##### Fig. 3 Logistic Regression

1. **Linear Regression:** It is a linear approach to modeling the relationship between a scalar response ( or dependent variable) and one or more explanatory variables (or independent variables).