# Regression Analysis: Walmart Sales Forecasting

*ECE 9063 Data Analytics Foundations*

*Yaozhuo Wang  
251009638*

*ywan3223@uwo.ca*

## 1 Problem Description

Walmart sales Forecasting Dataset provides historical sales data from more than 100 departments of Walmart stores located in different regions. As One of the leading retail stores in the US, Walmart, each stores have large number of departments. The historical sales data includes basic situation of every week from 2010 to 2013. Additionally, Walmart hosts several promotional markdown sales all year round. These reductions occur before significant vacations. There are four largest markdown which are Super Bowl (February), Labor Day (September), Thanksgiving (November), and Christmas (December). In comparison to non-holiday weeks, the evaluation of the weeks that include certain holidays is weighted five times more heavily [1].

Predicting the department-wide sales for each store is the goal of the project. Unforeseen demand is a hurdle for the company, and occasionally stock runs out because of a bad machine learning system. Modeling the effects of markdowns on these vacation periods in the absence of complete/ideal historical data is one of the challenges given by this forecasting problem. The perfect machine learning algorithm will precisely estimate demand and consider variables like the CPI, unemployment rate, and other economic situations.

Training a suitable data model to forecast department-wide sales for each store is our forecasting challenge. Data need to be explored and cleaned to understand the attributes. Determine the relevance of the attributes to weekly sales by compiling and understanding the data and make the feature selection to determine the retention or deletion of data. After developing the framework, cleaned dataset need to be evaluated by three algorithms. To measure the accuracy of the algorithms, results will be compared cross-validation using the appropriate metrics for the forecasting problem. [1]

## 2 Data for modelling

The goal of the problem is to predict the sales of the Walmart stores with the historical sales data. This historical information includes weekly sales from February 5, 2010, to July 26, 2013. The dataset is from the Kaggle dataset “Walmart Store Sales Prediction - Regression Problem” [1]. The dataset contains 6745120 data that have been categorized using 16 variables.

|  |  |
| --- | --- |
| Variable | Variable Explanation |
| Store | ID of the store |
| Dept | ID of Department in the stores |
| Date | Date of the start of accurate week |
| IsHoliday | whether the week is a special holiday week 1 – Holiday week  0 – Non-holiday week |
| Type | A to D, four types of stores |
| Size | Size of the store |
| Temperature | Temperature of the day of sale |
| Fuel Price | Average fuel price of the week |
| Unemployment | Prevailing unemployment rate |
| Markdown1 | Five different types of markdowns |
| Markdown2 |
| Markdown3 |
| Markdown4 |
| Markdown5 |
| CPI | Prevailing consumer price index |

There are four holiday events during every year. The time of the holiday also effect the weekly- sales.

|  |  |
| --- | --- |
| 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 |

Because the size of the dataset is too large, the historical sales data of the first store will be chose for the regression analysis and solve the forecasting problem. After cleaning the data, I drop some of the samples with Nan value. It includes 3657samples and 16 variables in the train dataset and 2783 samples in the test dataset. isHoliday is the binary attributes. While there is no holiday, the value will be ‘0’. Otherwise, the value will be ‘1’. And other data types are integer and float.

## 3 Background

Since the Walmart sales Forecasting problem is a regression problem, I choose three algorithm which are Linear Regression, Random Forest, and Support Vector machine to train the data.

Linear regression: To minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation, Linear Regression fits a linear model with coefficients w = (w1..., wp). When fitting a linear regression model to a set of data, it is important to pay close attention to the data's range. When attempting to forecast numbers outside of this range using a regression equation, it is frequently unsuitable and may provide astonishing results. Extrapolation is the term for this method. [2]

Random forest: A huge number of distinct decision trees work together as an ensemble in a random forest. The class with the highest votes becomes the prediction made by our model. The random forest's individual trees each spit forth a class prediction. The best results come from a large number of very uncorrelated models (trees) working together as a committee. [3]

Support Vector machine (SVM): A group of supervised learning techniques used in machine learning are called support vector machines. It works well in high-dimensional spaces and when there are more dimensions present than there are samples. Additionally, it is a flexible approach that is specified for the decision function using various Kernel functions. Both dense and sparse sample vectors are acceptable inputs for scikit-support lean’s vector machines. An SVM must fit on sparse data to be used to make predictions for such data. [4]

## 4 Methodology (Algorithms)

### Cleaning data

Before modelling, the dataset needs to be cleaned up to reduce the error. Original train dataset needed to be collected by merging feature.csv and store.csv to train.csv. All the attributes need to be in one data frame to proceed to fix NaN values and other data cleaning. Merging will also help to explore the specific data without going through the specific files. Due to the high dependencies on the date, all features will be kept in the training dataset. All the Nan values need to be filled with related numerals to keep the modelling running properly. For Temperature, Fuel price and Unemployment, Nan values will be filled by the average to keep the training dataset more reasonable than filled in with zero. Other Nan values in Markdown will be filled with zero. After that, all the zero values will be dropped to delete the extreme values to keep the clean dataset. Repeat the same operation to test.csv to get the test dataset.

### Algorithms applied

The following stage is to use algorithm with the ideal parameters to achieve the results after analysing, cleaning, and preparing the data. First, split data into 4 sets (x\_train, x\_test, y\_train, y\_test). In the process of splitting data, x\_train is the weekly sales of the train data, y\_train is other attributes. And x\_test is the data from test dataset.

Three models have the same preparation and data split process. Use three estimator (Linear regression, Random forest, Support Vector machine) in sklearn.linear\_model to train the data. The algorithm splits the samples and selects one characteristic. We can select and pass as a parameter the function that will be used to evaluate the quality of a split. In order to discover the optimal predictor and reduce overfitting, it is crucial to modify these parameters. fit( ) method is used to train and analyze the data in the code. predict() method is the model consisting of the model parameters calculated by fit(). The value of predict() obtained by making predictions on the explanatory variables.

### Evaluation procedure

Using hold-out validation to evaluate the three models. In the dataset of the Walmart historical sales data, the dataset has already split up into train and test set. The test set is used to evaluate how well the model works on data that has not yet been seen, whereas the training set is used to train the model.

## 5 Results

### Linear regression

### Support Vector machine

### Random forest

## Reference