Retail Store Sales Forecasting

Name

John Otieno

https://github.com/dvhub7/Capstone

Project Abstract

One challenge of modeling retail data is the need to make decisions based on limited history. Sales forecasting is crucial for many retail operations. It is not easy for large retailers to understand the market conditions of stores in different geographical locations. Based on that prediction, resources can be allocated so business can reduce loss and generate large profits.
In this research, the goal is to forecast weekly sales of 45 Walmart stores in different geographical locations each having multiple departments. The goal is to project the sales of each department in each store using the historical data.

Introduction and Research Question

Forecasting on sales is one of the most important task of every business. I would like to analyze how internal and external factors of one of the largest retail companies can affect their Weekly Sales in the future. This project tries to achieve an approximate weekly sales prediction looking at the previous years performance per Store on a weekly basis. The number of stores are 45 with multiple departments within them and they are spread across the country.
Problem I am trying to solve:
Determine weekly sales forecast for each department in each store?
Predicting the department wide sales for each store. I will be using R for preprocessing of the data and building further models. I am also hoping to adopt ARIMA and Regression time series analysis for predictive modelling. ARIMA is a popular method of modeling time series, because of it's flexibility and generalizability.

Literature Review

There are a couple of different categories of literature that need to be reviewed to ensure best practice for the project. These are set out systematically below.

Literature About Time Series Modelling

A time series is just collection of past values of the variable being predicted. It basically working on time (years, days, hours, and minutes) Also known as naïve methods. Goal is to isolate patterns in past data, to explore hidden insights of the data and trying to understand the unpredictable nature of the market which we have been attempting to quantify.

- Literature review of modern time series forecasting methods.

http://individual.utoronto.ca/paulkara/S2012%20Literature%20Review%20-%20linear%20survey.July31.pdf

Literature About Machine Learning

This is a good first step for someone looking to learn the steps needed for exploring data, cleaning data, and training/evaluating some basic machine learning algorithms. It is also a useful resource for someone who is comfortable doing data science in other languages and wants to learn how to apply their data science skills in R.

- https://www.kaggle.com/camnugent/introduction-to-machine-learning-in-r-tutorial/notebook

Literature About Predictive Modelling

- A Novel Trigger Model for Sales Prediction with Data Mining Techniques
  Authors: Wenjie Huang, Qing Zhang, Wei Xu, Hongjiao Fu, Mingming Wang, Xun Liang

Literature About Forecasting

The predictability of an event or a quantity depends on several factors including:

1. how well we understand the factors that contribute to it;
2. how much data are available;
3. whether the forecasts can affect the thing we are trying to forecast.

- Forecasting: Principles and Practice *Rob J Hyndman*, *George Athanasopoulos (Book)*
  This textbook is intended to provide a comprehensive introduction to forecasting methods and to present enough information about each method for readers to be able to use them sensibly.

About the Dataset

In this experiment, we use Walmart's open dataset from kaggle (link:https://www.kaggle.com/c/walmartrecruiting-store-sales-forecasting/data). Multiple data sets are provided in the link above but all we use are three datasets named train.csv, store.csv, features.csv.

Stores.csv:
- Store: The store number. Range from 1-45.
- Type: Three types of stores 'A', 'B' or 'C'.
- Size: Sets the size of a Store would be calculated by the no. of products available in the particular store ranging from 34,000 to 210,000.

Train.csv: **421570 records**
- Date: The date of the week where this observation was taken .
- Weekly_Sales: The sales recorded during that Week. Weekly sales for the given department in the given store from Feb 5, 2010 to Nov 1, 2012.

- Store: The store which observation in recorded 1-45.
- Dept: One of 1-99 that shows the department.
- IsHoliday: Boolean value representing a holiday week or not. Whether the week is a special holiday week.
-The data contained 421,570 rows, with some store-specific departments missing a few to many weeks of sales.

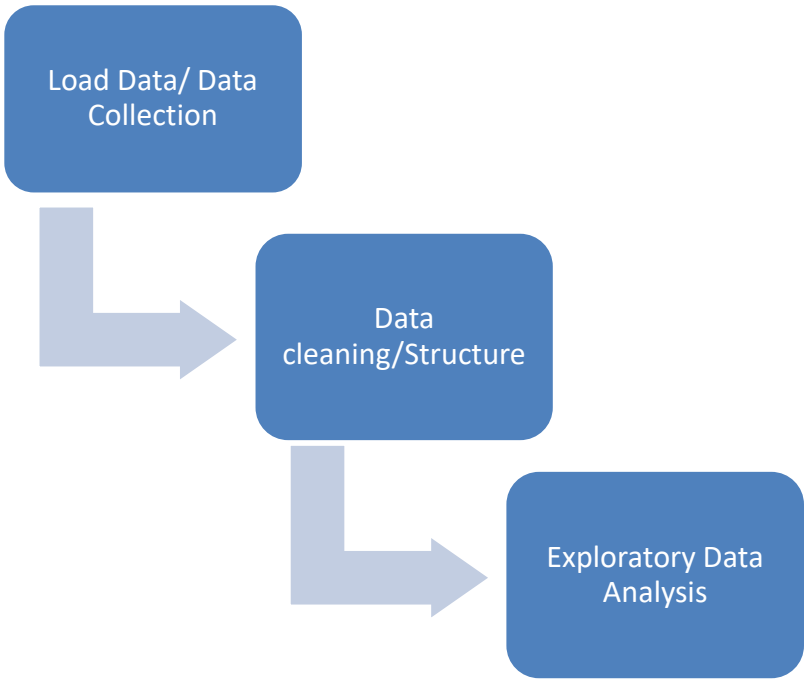
Features.csv: **8190 records**
- Temperature:Temperature of the region during that week.
- Fuel_Price: Fuel Price in that region during that week.
- MarkDown1:5 : Represents the Type of markdown and what quantity was available during that week.
- CPI: Consumer Price Index during that week.
- Unemployment: The unemployment rate during that week in the region of the store.

Test.csv: 115064 records

Data from this file has the same fields as the Train data, only the Weekly_Sales are empty.

Approach

All steps of this analysis have been brought together at this location
Github link: https://github.com/dvhub7/Capstone/blob/master/sales_forecast_v3.html
Data Exploration: Went over the provided datasets using R in detail to give an in-depth explanation of each dataset.



Step 1: Load Data/Data collection

Details:

Downloaded the data sets from the kaggle website. Stored it on the google shared drive. The data sets collected from the kaggle website include Train.csv, Test.csv, Stores.csv and Features.csv. The competition is called Walmart- recruiting store sales forecasting and can be found using the link below.
https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

```
## Load libraries and read the data files
library(tidyverse)
library(reshape2)
library(lubridate)
library(rpart)
library(rattle)
library(forecast)
library(tseries)

## Contains data with the following information
# Store Dept Date Weekly_Sales IsHoliday: 2010-02-05 ~ 2012-10-26
tr_df <- read_csv("train.csv")

## Contains data with the following information
# Store Dept Date IsHoliday: 2012-11-02 ~ 2013-07-26
tst_df <- read_csv("test.csv")

## Contains data with the following information
# Store Type Size
str_df <- read_csv("stores.csv")

## Contains data with the following nformation
# Store Date Temperature Fuel_Price Markdown1-5 CPI Unemplyment IsHoliday
ft_df <- read_csv("features.csv")
```

Figure 1: Loading of data into RStudio.

Step 2: Data cleaning/structuring

Details:

Collecting and preparing the data for analysis are often the most involved and time consuming parts of building a predictive model. While the collection was done for us, we still have to do a bit of work to prepare the data.
Change some feature types to factors for better anaylsis of the data. This was done for both the training and the test data sets.

```
# Analysis for train dataset and test dataset
# Converting the  Store, Dept and Type features to factor datatype
str_df$Store <- factor(str_df$Store)
tr_df$Store <- factor(tr_df$Store)
tst_df$Store <- factor(tst_df$Store)
ft_df$Store <- factor(ft_df$Store)
tr_df$Dept <- factor(tr_df$Dept)
tst_df$Dept <- factor(tst_df$Dept)
str_df$Type <- factor(str_df$Type)
```

Figure 2: Converting the data types of certain features.

Now in this step of the process, we need to merge the datasets to build a successive model. We first review the column from Store.csv and join it with train.csv datasets. Then with the new dataset we do another join operation with feature.csv dataset.

```
## Merge train dataset with stores dataset by store and features dataset
# training dataset from the following dates: 2010-02-05 ~ 2012-10-26
train1 <- full_join(tr_df, str_df, by= "Store")
train <- merge(x=train1, y=ft_df, by=c("Store", "Date", "IsHoliday"), all.x=TRUE, sort=FALSE)

# Export the train datatset
#Create a csv file with the merged dataset train

#write.csv(train, file = "C:/Users/endcore/Documents/Capstone/train_merged.csv",row.names=FALSE, na="")

## Merge test dataset with stores dataset by store, Date and IsHoliday features
# testing dataset from the following dates: 2012-11-02 ~ 2013-07-26
test1 <- merge(tst_df, str_df, by="Store", sort = FALSE)
test <- merge(x=test1, y=ft_df, by=c("Store", "Date", "IsHoliday"), all.x=TRUE)
test <- arrange(test, Store, Dept)

rm(tr_df,tst_df,str_df,ft_df, test1, train1)
```

Figure 3: Merging of datasets.

Next, identify all the rows without missing data and use them for the dataset.
Find the columns with the missing values for both training and test datasets.

```
     Store         Date    IsHoliday         Dept Weekly_Sales         Type         Size  Temperature   Fuel_Price
         0            0            0            0            0            0            0            0            0
 MarkDown1    MarkDown2    MarkDown3    MarkDown4    MarkDown5          CPI Unemployment
    270889       310322       284479       286603       270138            0            0
```

Figure 4: Finding columns/variables with missing values.

We can then use the complete.cases function to identify the rows without missing data:
In train dataset has 421,570 variables and 324,514 have missing values
Complete cases in train dataset are 97,056 and renamed to train_clean data set.

```
Observations: 97,056
Variables: 16
$ Store        <fctr> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ Date         <date> 2011-11-11, 2011-11-11, 2011-11-11, 2011-11-11, 2011-11-11, 2011-11-11, 2011-11-11, 2011-1...
$ IsHoliday    <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, ...
$ Dept         <fctr> 55, 91, 44, 26, 14, 22, 85, 24, 95, 18, 8, 31, 1, 13, 48, 71, 32, 42, 96, 79, 49, 35, 36, ...
$ Weekly_Sales <dbl> 23728.53, 67041.24, 5859.12, 7693.46, 14903.78, 6596.53, 2963.69, 6053.49, 115047.16, 10170...
$ Type         <fctr> A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, ...
$ Size         <int> 151315, 151315, 151315, 151315, 151315, 151315, 151315, 151315, 151315, 151315, 151315, 151...
$ Temperature  <dbl> 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, 59.11, ...
$ Fuel_Price   <dbl> 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, 3.297, ...
$ MarkDown1    <dbl> 10382.9, 10382.9, 10382.9, 10382.9, 10382.9, 10382.9, 10382.9, 10382.9, 10382.9, 10382.9, 1...
$ MarkDown2    <dbl> 6115.67, 6115.67, 6115.67, 6115.67, 6115.67, 6115.67, 6115.67, 6115.67, 6115.67, 6115.67, 6...
$ MarkDown3    <dbl> 215.07, 215.07, 215.07, 215.07, 215.07, 215.07, 215.07, 215.07, 215.07, 215.07, 215.07, 215...
$ MarkDown4    <dbl> 2406.62, 2406.62, 2406.62, 2406.62, 2406.62, 2406.62, 2406.62, 2406.62, 2406.62, 2406.62, 2...
$ MarkDown5    <dbl> 6551.42, 6551.42, 6551.42, 6551.42, 6551.42, 6551.42, 6551.42, 6551.42, 6551.42, 6551.42, 6...
$ CPI          <dbl> 217.9981, 217.9981, 217.9981, 217.9981, 217.9981, 217.9981, 217.9981, 217.9981, 217.9981, 2...
$ Unemployment <dbl> 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, 7.866, ...
```

Figure 5: Complete case for the train data set (renamed to train_clean data set).

Visualize the features of the data sets to better understand the data. I used the histogram for the numerical features

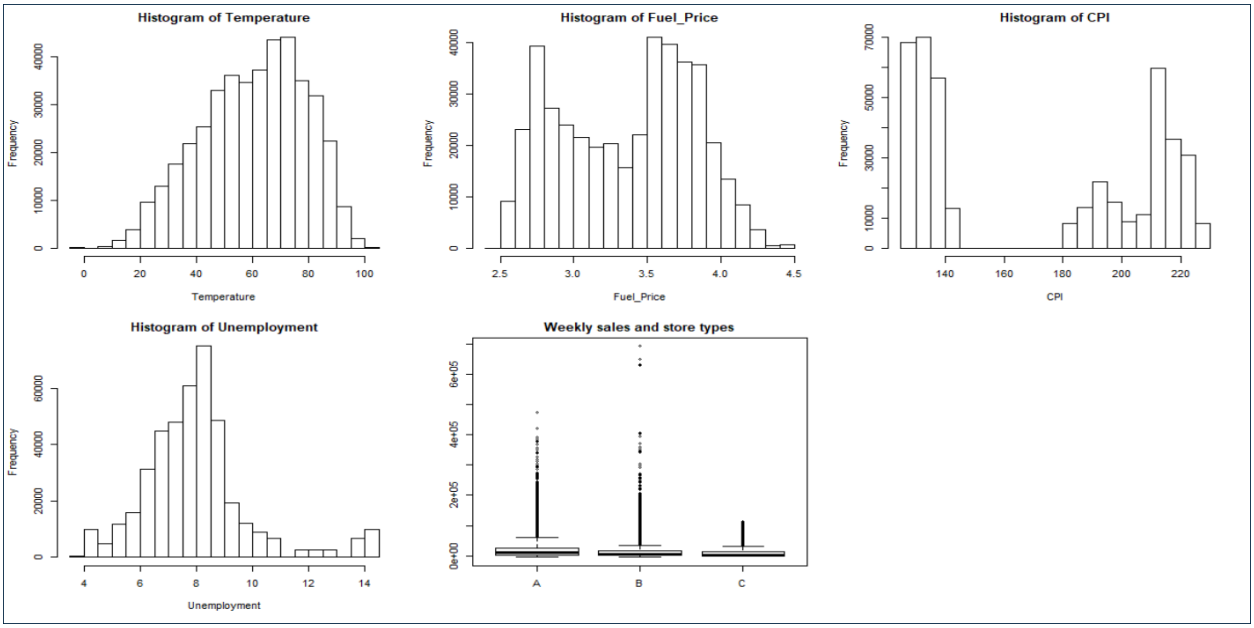


Figure 6: Histograms based on external factors.

In order for our analysis to be consistent and avoid testing on our training data, we are going to split it into a training and test data set using the caret package

The training data set has 77,648

The testing data set has 19,408

Step 3: Exploratory data analysis

Details:

- Explore relationships in the data, such as sub-questions to examine:
    - Visualizing and compare the sales of different Departments across the same store to look for similarities or differences throughout the year.

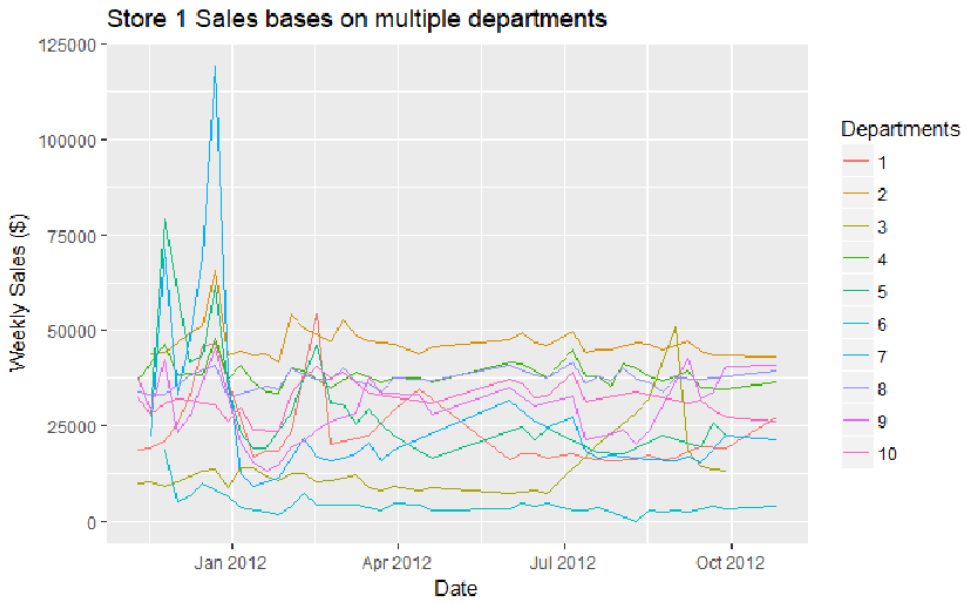
- 

Figure 7: Store sales based on multiple departments.

- Visualizing a graph that will compare same department across different stores this will determine whether all the departments perform the same in sales around the same time in the year. By plotting the sales of different departments within Store 1, the difference in departments is made evident.

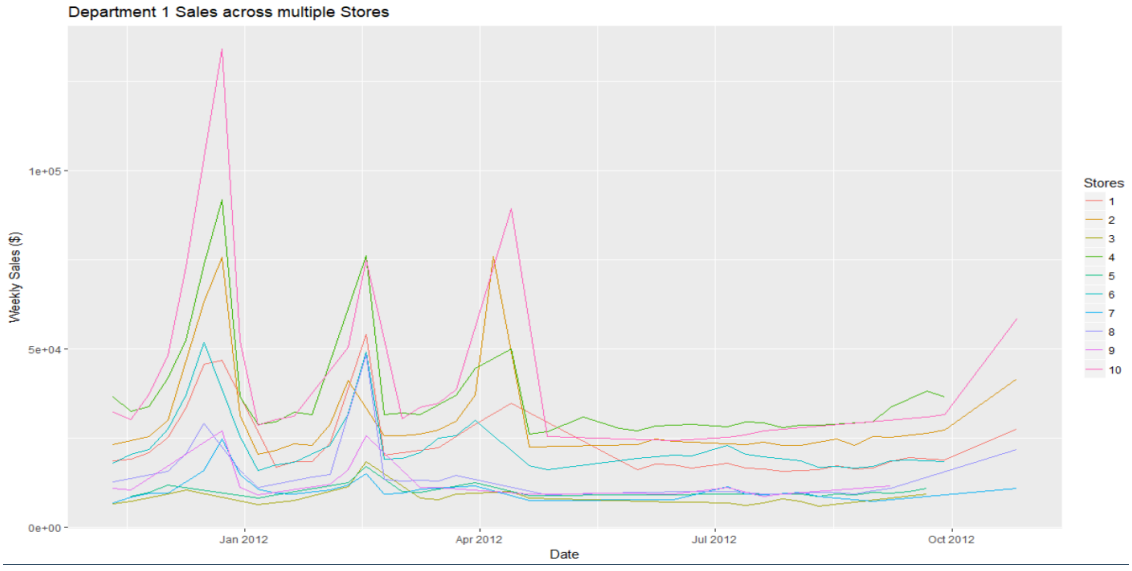
- 

Figure 8: Department across multiple stores.

- Compare the weekly sales by stores to see how the stores are performing within there locations.
- The information is based on the type of store

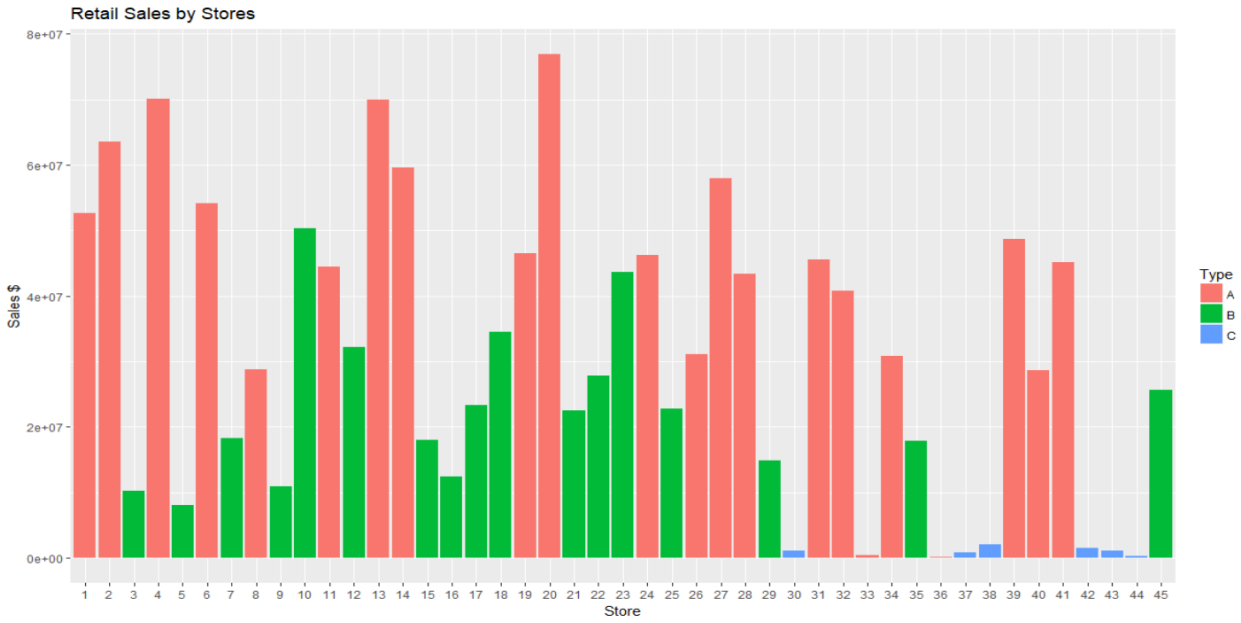


Figure 9: Weekly sales across stores.

Step 3: Modelling

Regression Analysis

Details:

- Split the data set into Test and Train data sets.
    - Also removes any NaN values
- Apply a linear regression model to the data.

Github link: https://github.com/dvhub7/Capstone

Time Series Analysis

Our goal is to aggregate these data in order to get weekly counts of the significant sales observed in this time period. To achieve this, we can use the count() function of the R package plyr to aggregate the data, and the standard ts() function to create a new time series.

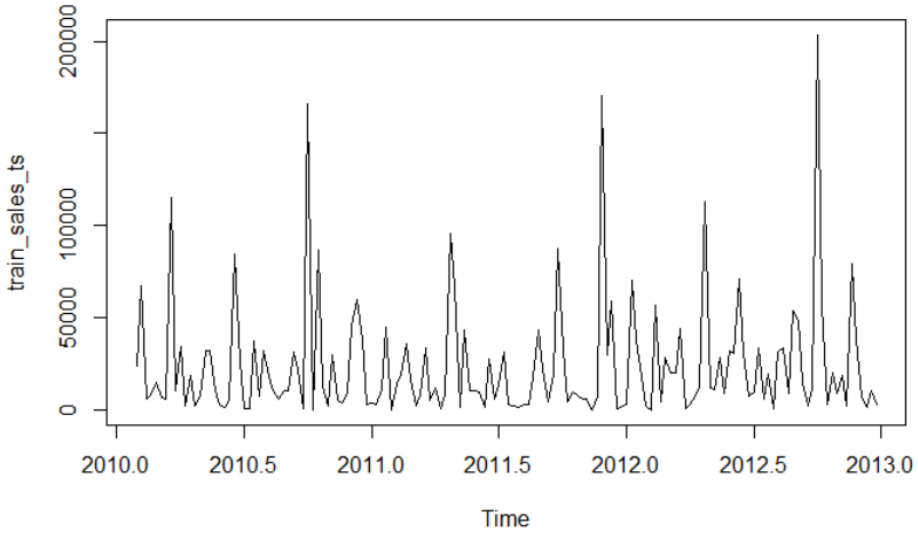
We will specify the starting and ending year and month of our time series, as well as set the freq parameter to 52 to indicate weekly readings. Finally, we will plot our data:

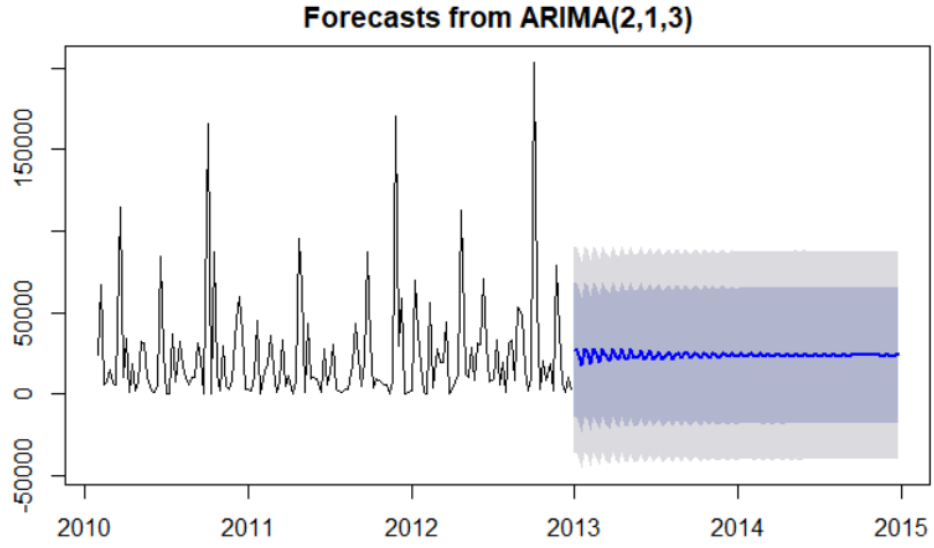
The next approach was to fit an ARIMA model since it is a popular method to model time series data. This method is popular since it has been proven to be a good way to forecast future information from the past. It checks for stationarity and adds the constant fluctuation in order to make forecasts into the

near future. The models were evaluated, and orders selected, using Akaike Information Criterion (AIC): Error handling was necessary for this process, particularly for the event of maximum likelihood failing to converge with a particular order.

RESULTS

Sales forecasts for department-average sales were overall quite accurate. A pair of plots below show a selection of forecasts overlaid with the actual sales:





CONCLULSIONS

The methods outlined in this project, are ideal for time series modeling and reduce the number of models required to make individual forecasts. Other models can be used as ell to prform the predictive modelling that may result in a better outcome. It is clear that some departments within specific stores should be modeled separately.

RECOMMENDATION

It is recommended that a further investigation on the rest of the features on the dataset should be explored further in order to generate new results and models

Bibliography

[1] A Novel Trigger Model for Sales Prediction with Data Mining Techniques
Authors: Wenjie Huang , Qing Zhang, Wei Xu, Hongjiao Fu, Mingming Wang, Xun Liang

[2] Forecasting: Principles and Practice *Rob J Hyndman*, *George Athanasopoulos (Book)*

[3] Business fluctuations: Forecasting techniques and applications.
Authors: Dale Bnails, Larry C. Peppers

[4] Souhaib Ben Taieb, James W Taylor, Rob J Hyndman (2017) Coherent Probabilistic Forecasts for Hierarchical Time Series.

[5] T.-M. Choi, Y. Yu, K.-F. Au, A hybrid SARIMA wavelet transform method for sales forecasting, Decision Support Systems, (51) (2011), pp. 130-140.

[6] Kaggle competition participants from the Kaggle Competition
www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

[7] Kaggle Walmart recruiting store sales forecasting competition discussion board.

[8] Several youtube videos and blogs about Forecasting and time series analysis

[9] Linear Regression with time series data Author Heino Bohn Nielsen
http://www.econ.ku.dk/metrics/Econometrics2_05_II/LectureNotes/regression.pdf