Walmart sales Forecasting

Objective Forecasting the Walmart sales

Business Problem

Defining the Forecasting the weekly sales of each Store and respective department for the time period (2012-11-02 to 2013-07-26) using the historical data given.

Data Availability

- stores.csv: This file contains anonymized information about the 45 stores, indicating the type and size of store.
- train.csv: This is the historical training data, which covers to 2010-02-05 to 2012-11-01
- 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.
- features.csv: This file contains additional data related to the store, department, and regional activity for the given dates. Also contains the markdown data (promotional activity).

Approach

1 Importing the libraries and dataset

2 Preparing the combination of the dataset

The Train and Test dataset is merged with features, stores data based on the Store, dept and date respectively.

3 Part 1 - Regression Forecasting

- *Data audits*
  Check for date series, missing values and other statistics were done.

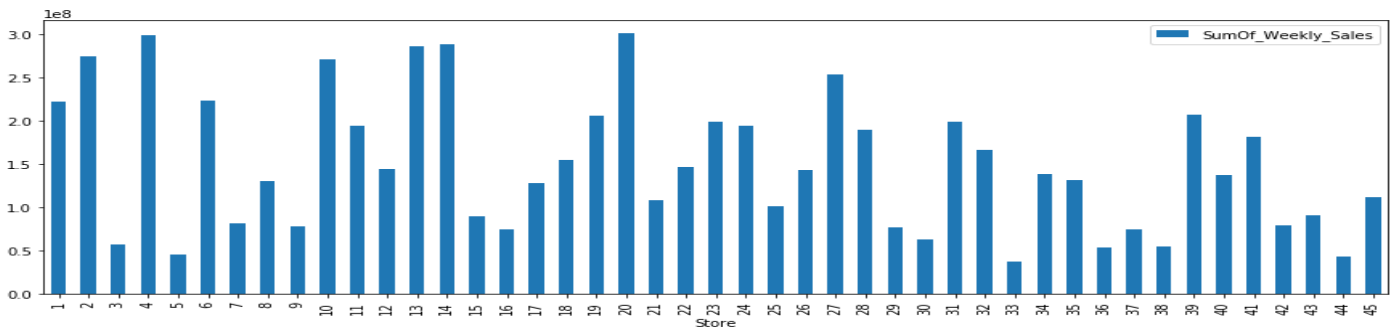
- *Data preparation*

  - Finding Numerical and Categorical variables
  - Handling Missing and Outliers
  - Handling the categorical variables

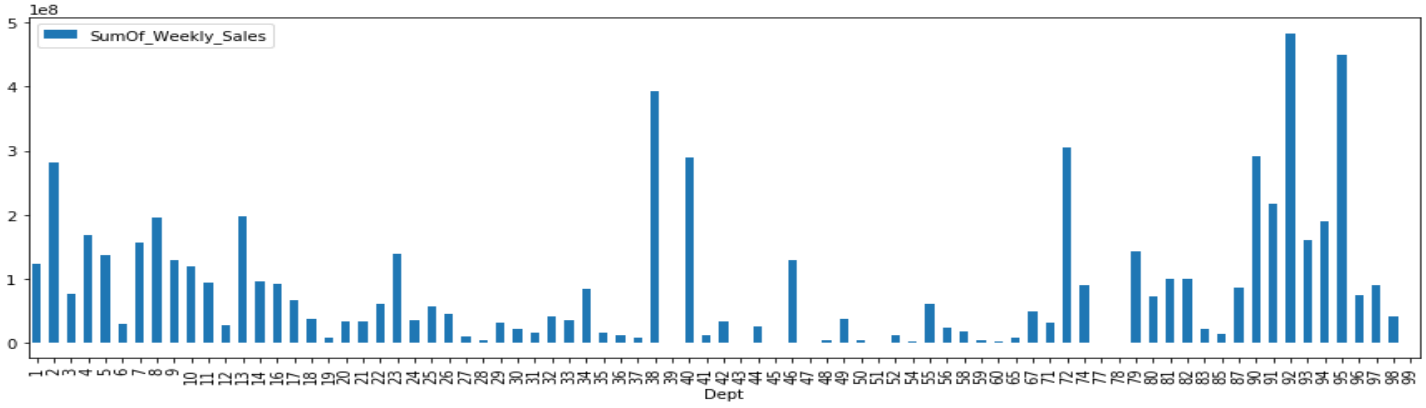
  The missing values were imputed with the median.
  Outlier capping is done at 1% and 99% percentiles.
  The categorical variables were dummified.
  The X and Y were prepared.

- *Exploratory Analysis*

  - Sales vs Stores



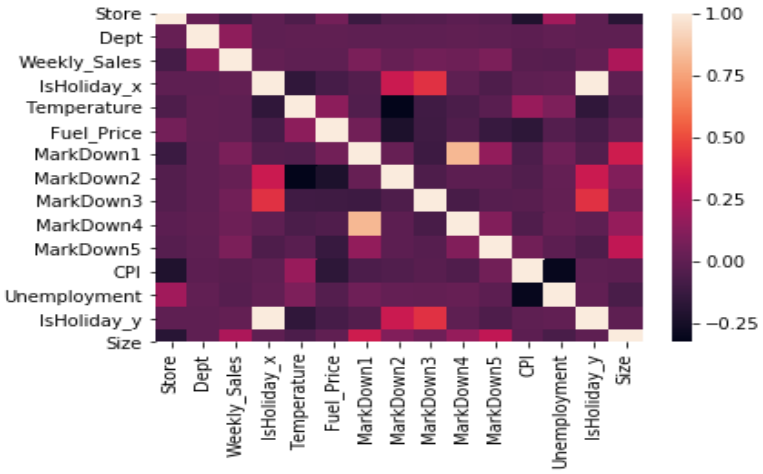
- Sales vs Dept



- Correlation of Sales with the Markdown_mean

  The sales were found less correlated with the promotional activity.

- Correlation of Xs with th Sales



- Coefficient of variance

  The coefficient of variance for the CPI, Unemployment variables were found.

- Feature Reduction

  - Using Random Forest

    The Random Forest Regressor was applied on the dataset and 20 features were found important with respect the sales.

- Train and Test split

- Model Implementation

  - Random Forest

    Best parameters: {'max_depth': 10, 'n_estimators': 100}
    Accuracy: 67 %

  - XGBoost

    Best parameters: {'max_depth': 10, 'n_estimators': 100}
    Accuracy: 76 %

Choosing the model: The MAPE and RMSE metrics were calculate manually for the train data.
It was found that XGBoost was performing better comparatively. Thus, it will be chosen for the forecasting of sales.

- Forecasting using the Regression model

  The above chosen XGBoost was implemented on the prepared test data and the predictions were done for the required time period (39 weeks).

4 Part 2 - Arima model (auto arima)

As the data was little complex, the stationarity check and tuning the p, d, q values would be tedious since these checks has to be done by separating the data into each Store and each Dept and doing it for each one separately. Thus, the auto arima model was chosen.

- Preparing Exogenous variable (Markdown)
  The Markdown (promotion) data was given in 5 separate columns which was then merged and the mean was taken row wise. Then these values were binned into 10 deciles and encoded which helps in faster processing.

- Preparing separate array for each Store
- Preparing separate array for each Dept
- Preparing Train and Test for evaluation
  The data was first separated into Store and then that data is further separated into separate arrays for each Dept.

- Model Implementation + Forescating

  - Auto ARIMA Model

    There were few challenges in implementing the model,

    As there was few Store/Dept data that has very less data – the seasonality of that dataset could not be found. So, these predictions were done assuming no seasonality.
    And the few Depts were missing completely in the dataset which were skipped.

    A user defined function was written which loops through and fits model for each Store-Dept combination data created in the data preparation and makes the predictions.
    The above the challenges were also handled here. And if at all some error occurs for a particular set it is handled by skipping it and other predictions will be proceeded. But here we were able to get predictions for every possibility.

5 Conclusion

Although ARIMA model could make better predictions considering the seasonality, trend, irregularity components present in the data, performing it for larger datasets would be very tedious task and time consuming. Thus in that case the Regression forecasting will be preferred.