TIME SERIES SALES FORECASTING

Introduction to Data Management and Processing
DS 5110

Project Report

Authors:
Apoorva Surendra Malemath
Ashwin Sateesh Kumar
Barkha Saxena
Basil Varghese
Sravya Burugu

Summary

In any organization, there is an embedded desire to predict its future revenue and future sales. The basic recipe followed is "*Collect historical data related to previous sales and use it to predict expected sales".*

Following the same essence, our project entails the sales forecasting for Corporacion Favorita, an Ecuador based grocery retailer. Grocery retailers deal with large volumes of perishable goods. Hence accurate forecasts of demands and sales are essential to reduce wastage and manage operating `costs. Additionally, since grocery consumption is sensitive to holidays, festivals and seasonal changes, accurate forecasts of demands across various types of products can help the retailer capitalize on the emerging need of the consumers.

The goal of the project is to build a forecasting model that accurately predicts the unit sales of items sold at various Favorita stores. Additionally, we aim to observe and deduce trends in sales based on various factors such as stores, promotions, holidays etc. and understand how external factors affect sales.

The dataset consists of 4 years of sales data [1], at a date-store-product level, along with information on promotions run on particular days. Additionally, information about each store, holidays/events and the daily oil price for the given date range have been provided.

The various datasets were cleaned and merged to create the final dataset on which analysis was performed. EDA was done to understand how various factors such as seasons and holidays affect sales and to understand the trends in sales of a particular family of products across the years. A stationarity check was performed after which the data was made stationary as time-series models work best with stationary data.

Various modelling techniques such as Regression models (XGBoost, Random Forest) and statistical models (Holt winters exponential smoothing, ARIMA and SARIMAX) were used to predict future sales. After running and predicting using the various models, it was concluded that different models at different levels of prediction for required for various use cases.

Methods

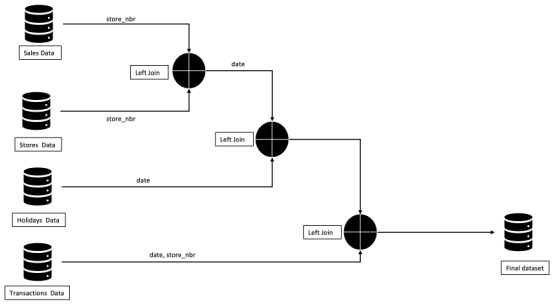
Understanding the data

1. **Sales data**(*id, date, store_nbr, family, sales, onpromotion*): The dataset has about 3 million records, which contained information about the daily sales at a day – store – product family level along with information about the number of products on promotions on a particular day at a particular store for a particular product family.
2. **Stores data**(*store_nbr, city, state, type, cluster*): The dataset contains information about the 54 Favorita stores in Ecuador, such as the store number, city of the store, state, type of store and the cluster to which a particular store belongs. Details about each type of store and each cluster is not known.
3. **Holidays data***(date, type, locale, locale_name, description, transferred)*: The dataset contains information about 350 various national, regional and local holidays that have been declared in Ecuador across the 4 years.
4. **Transactions data**(date, store_nbr, transactions): The dataset contains information about the number of transactions in a store for a particular day

Data Cleaning

- Datatype conversion: date columns were converted to datetime format
- Removal of duplicates
- Removal of missing values

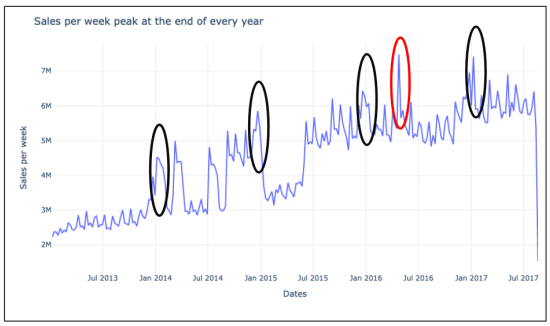
The datasets were merged as follows:



Exploratory Data Analysis (EDA)

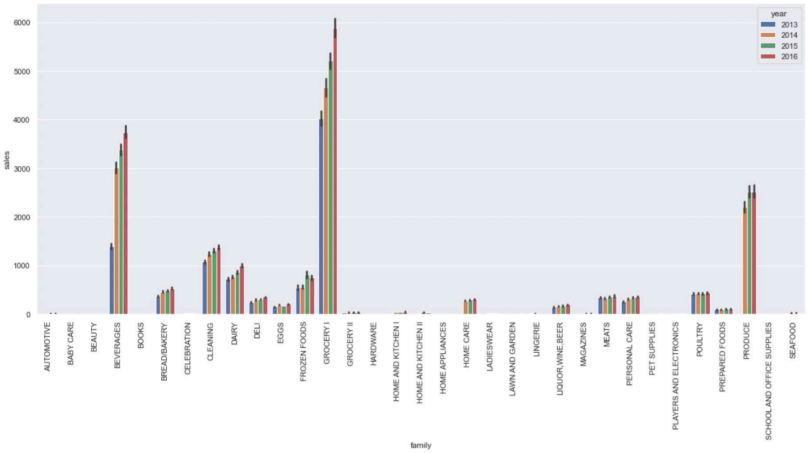
The behavior of sales based on various factors such seasonal trends, holidays and trends in sales of particular product families were visualized to identify any interesting trends.

1. Visualizing seasonal trend of weekly sales across the 4 years.



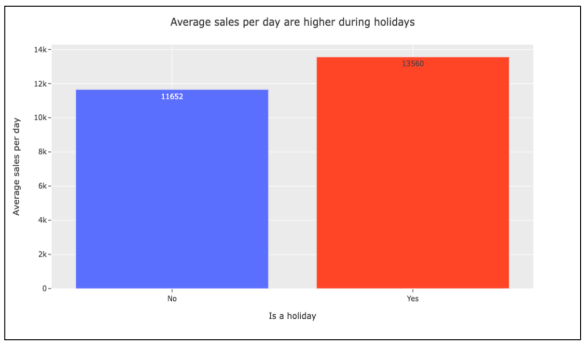
Observations:
• There is a peak in sales at the end of every year, during Dec-Jan.
• There is an increasing trend of sales from 2013 – 2017.
• An anomalous peak in sales is observed during Apr-May 2016. This is attributed to a magnitude 7.8 earthquake that struck Ecuador on April 16, 2016.

2. Visualizing the sales across product families by year.



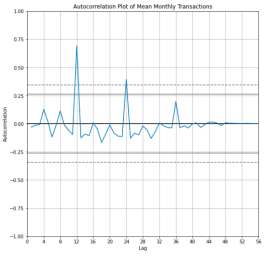
Observations:
• GROCERY I, PRODUCE and BEVERAGES are the most sold family of products across the years.

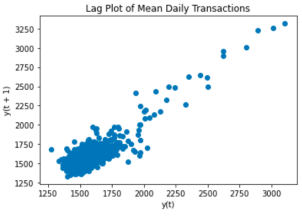
3. Visualizing the effect of holidays on the sales.



Observations:
- There is a peak in sales corresponding to most national holidays.
- Average sales per day during weeks with holidays are than weeks without holidays.
- There are certain national holidays during which sales are low.
- More promotions can be given during those holidays.

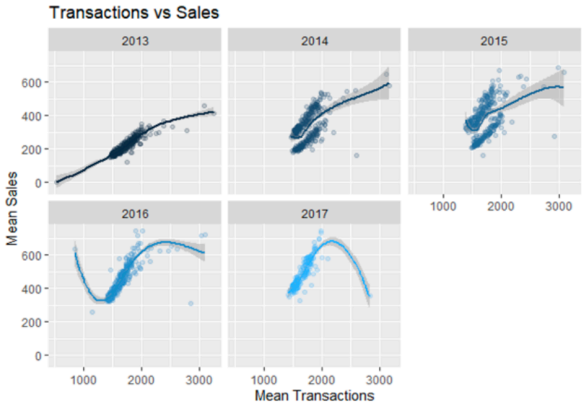
4. Visualizing the effect nature of transactions:





Observations:
•There is a peak in the transactions for every 4 months, indicating there is correlation of transactions every 4 months.
•The lag plot shows linearity between a transactions on current day and transaction on the previous day.

5. Visualizing the relationship between Sales and Transaction.



Observations:

• The sales and transactions have a positive relationship across the years.

• The gradient of the plot has increased over the years

Data pre-processing

1. Time Series Pre-processing:

- ADF(Augmented Dickey Fuller Test) test was done to test for stationarity.
- Time series models require the data to be stationary(not have an upward or downward trend). **Stationarity analysis and Seasonal decomposition** was done using ARIMA(Auto Regression Integrated Moving Average), which is a statistical technique to forecast time series data.
- Resampling to monthly averages was done for exponential smoothing.

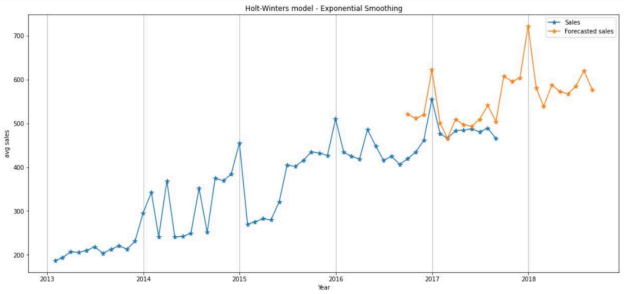
2. Machine Learning Pre-Processing:

- One hot encoding of categorical variables were done for machine learning models
- Feature engineering was done to decompose the time dimension.
- Days with zero sales values were dropped.

Modelling Techniques:

Statistical Forecasting Models

1. Holt Winters Exponential Smoothing[2] : Exponential Smoothing method uses exponentially weighted moving average(weighted with geometrically decreasing ratio) to forecast values for the future. There are 3 types of exponential smoothing techniques – Single exponential smoothing, double exponential smoothing and triple exponential smoothing. Among those we have used **Triple exponential smoothing** (Also called Holt winters Exponential smoothing) as it supports seasonality and trend in data. We have used additive trend and additive seasonality for the model parameters.



2. ARIMA[3] : Auto Regression Integrated Moving Average is a statistical modeling technique that the lags and lagged errors of a time-series data is used to predict the future data.

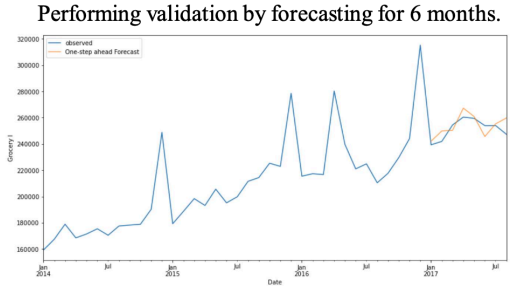
ARIMA uses 3 terms:

- p - order of the auto regressive term
- q – order of the moving average term
- d – number of differencing required to make the time series stationary

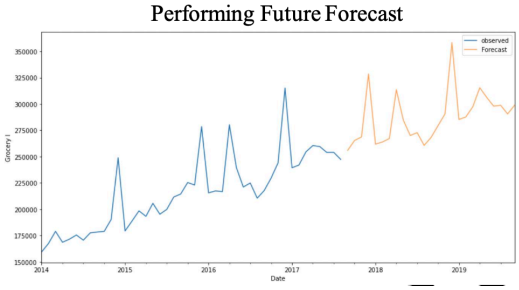
Train – test split for the analysis:-

- Training Data – 3 years of sales data from 2014 – 2016
- Validation data – 6 months of 2017
- Future prediction – 2018 - 2019

Performing validation by forecasting for 6 months.

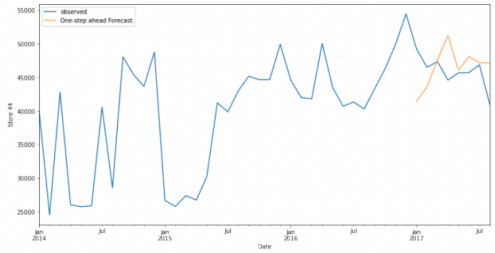


Performing Future Forecast

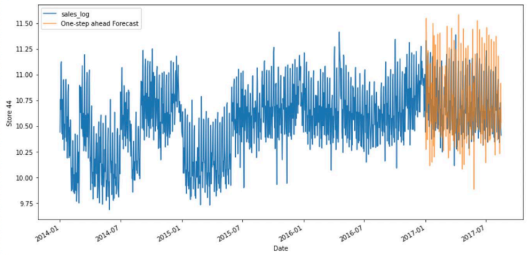


Performing forecasting for 'GROCERY I' family of products.

Forecasts using mean of sales grouped by month



Forecasting using Log Transformation on number of sales



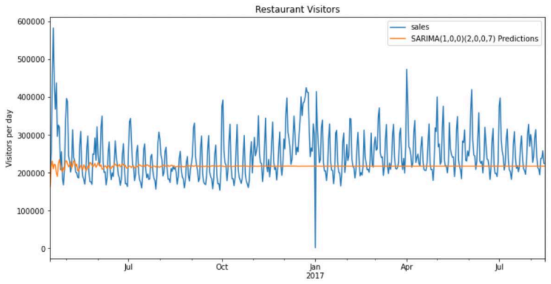
Performing forecasting for 'Store 44'.

3. SARIMAX [3]: Seasonal ARIMA model with an exogenous variable. The value chosen as an exogenous variable must have data for the future already available. Calendar data such as Holiday data can be used as an exogenous variable.
Train – test split for the analysis:-

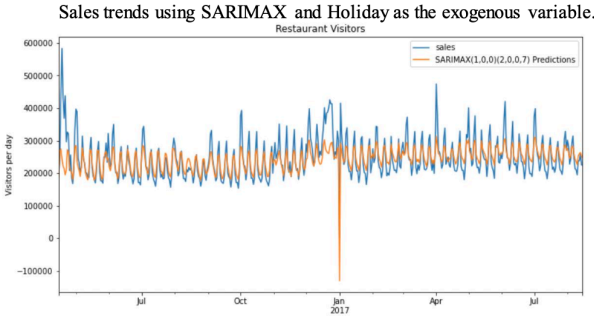
- Training Data – 3 years of sales data from 2014 – 2016
- Validation data – 6 months of 2017
- Future prediction – 2018 - 2019

Performing forecasting for 'Grocery I' using SARIMAX.

Sales trends using SARIMAX

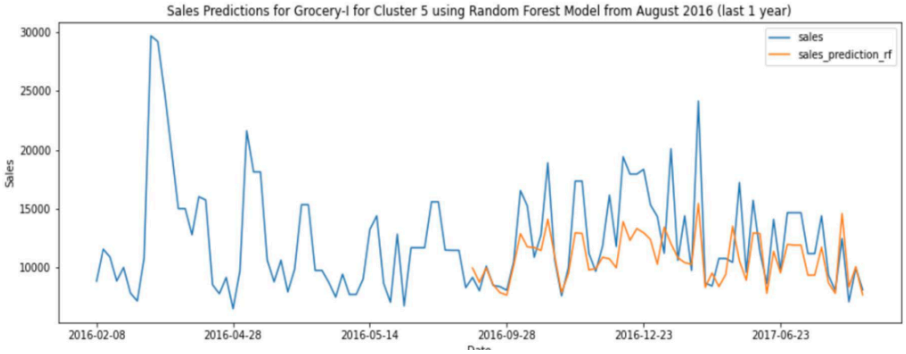


Sales trends using SARIMAX and Holiday as the exogenous variable.

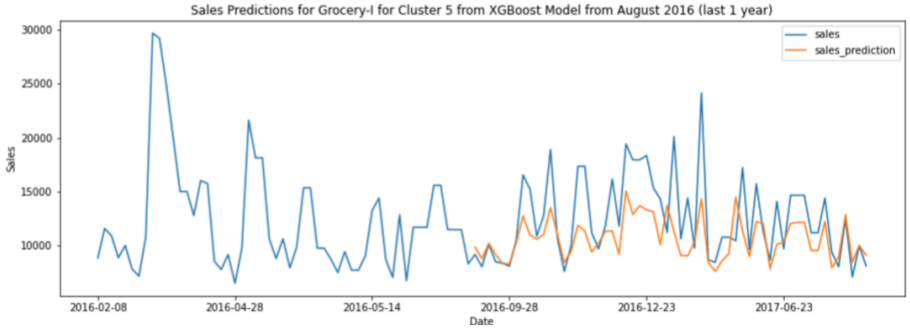


Machine Learning Models

Random Forest[4]: It is an algorithm that uses an ensemble of decision trees where each tree is created from a different bootstrap sample of the training dataset. An average of predictions across all decision trees are used, resulting in better performance than any single tree in the model.



XGBoost[5]: Extreme Gradient Boosting(XGBoost) is an ensemble algorithm similar to Random Forest, which uses multiple decision trees for prediction. Every new tree added to the model fixes the errors of the trees that are already part of the model. Addition of trees is stopped when no further improvement can be done to the model. It is a very efficient and fast algorithm that uses the stochastic gradient boosting technique.



Evaluation Metric:
Mean Absolute Percentage Error(MAPE)[5]: - MAPE is the commonly used metric to measure the accuracy of a forecast models. It is calculated as

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$ where A_t is the actual value and F_t forecast value

Results

Prediction Objective	Model	Model Type	MAPE	Training and Test Data	Usecase
Yearly per unit sales for next n months	Holt-winters Exponential Smoothing	Statistical	11.09	Univariate	Annual Budgets and Plans in Company Review
Monthly per unit sales for next n months	ARIMA	Statistical	2.25, 3.1 and 7.42	Univariate	Supply chain and demand forecasting planning
Monthly per unit sales for next n months	SARIMAX	Statistical	67.34 and 24.12	Univariate and Multivariate	Supply chain and demand forecasting planning
Daily Per unit sales per store for next 1 year	XGBoost	Machine Learning - Decision tree-based ensemble (Boosting)	26.40	Multivariate	Resource allocation and cashflow management planning
Daily Per unit sales per store for next 1 year	Random Forest	Machine Learning - Decision tree-based ensemble ML mode (Bagging)	26.80	Multivariate	Resource allocation and cashflow management planning

Discussion

Based on the results table above, we can conclude that different models can be used for different use cases. Since a retail business is dependent of multiple functions such as supply chain, budget planning, inventory management etc, different models would be required for each function in order to get the most accurate forecasts and make data driven decisions. Additionally, multivariate models are better at understanding various trends cause by external factors. Special promotions can be offered to products in the product family GROCERY I, PRODUCE and BEVERAGES as they are sold the most and inventory can be managed accordingly to make the most of the demand of these products during the holidays.

As next steps, feature engineering can be done to come up with better metrics that explain sales better than the current metrics. Additionally, more number of multivariate and hybrid models can be tried on the dataset which predict sales and transactions for all stores and products at the most granular level.

Statement of Contributors

Apoorva Surendra Malemath	EDA on Stores, Pre-Processing – Re-sampling and avergaing, Date Formatting, Log transforamtions, Forecasting using Satistical Models i.e., ARIMA and SARIMAX.
Ashwin Sateesh Kumar	EDA on Transactions, Pre-Processing – Null Imputation of Transactions, Feature Selection, Model Building and Hyperparameter Tuning.
Barkha Saxena	Data Pre-processing, Feature engineering, Time series forecasting using machine learning (XGBoost and Random Forest)
Basil Varghese	EDA on holidays and seasonal trends in sales across 4 years, Pre-Processing – Removal of duplicates, Date Formatting
Sravya Burugu	Data Preprocessing, EDA on sales and promotions, Stationarity test, Holt Winters Exponential Smoothing.

References

[1] Store Sales - Time Series Forecasting - https://www.kaggle.com/c/store-sales-time-series-forecasting/data

[2] Holt Winters Exponential Smoothing - https://machinelearningmastery.com/exponential-smoothing-for-time-series-forecasting-in-python/

[3] ARIMA and SARIMAX - https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/

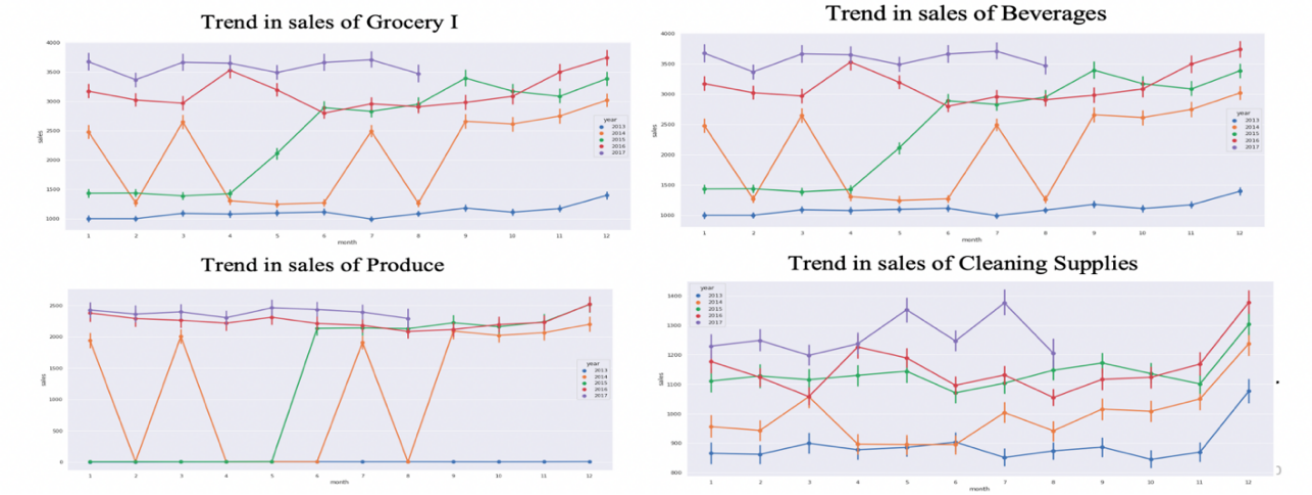
[4] XGBoost - https://machinelearningmastery.com/xgboost-for-time-series-forecasting/

[5] Random Forest - https://machinelearningmastery.com/random-forest-for-time-series-forecasting/

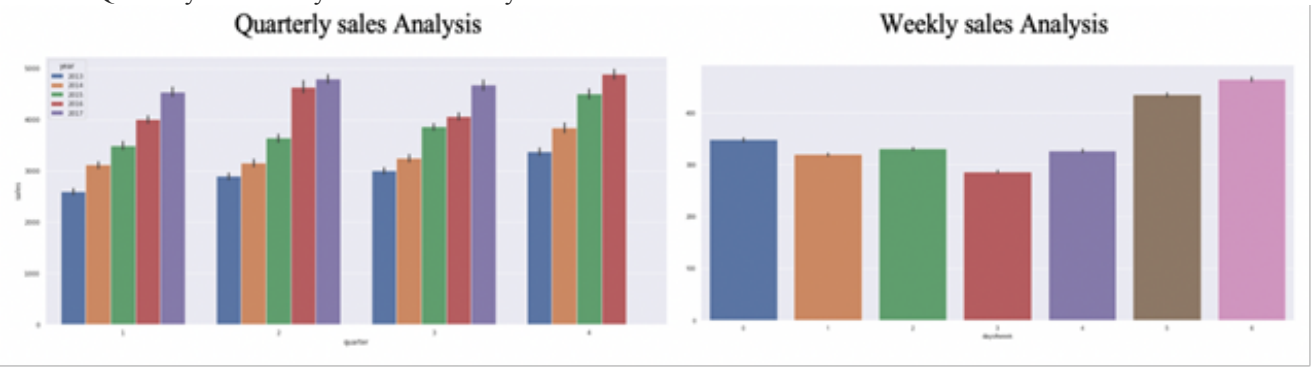
Appendix:

Github link to the codes: https://github.com/sravyaburugu01/IDMP_Project

1. Monthly sales trend of certain product families through the years



2. Quarterly and weekly sales across the years



Observations:

- Quarter over Quarter sales have been increasing over the years
- Sales are higher on weekends compared to weekdays

The parameter values given to Randomized search CV are:

```
#the rate at which the model learns
learning_rate = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]
#maximum number of levels in a tree
max_depth = [3,4,5,6,8,10,12,15]
# Minimum loss reduction required to make a further partition on a leaf node of the tree
gamma = [0.0, 0.1, 0.2, 0.3, 0.4, 0.5]
#Minimum sum of instance weight(hessian) needed in a child.
min_child_weight = [1,3,5,7,9]
#Subsample ratio of columns when constructing each tree
colsample_bytree = [0.3,0.4,0.5,0.7,0.9]
```

```
#fitting with the XGB Regressor
XGB = XGBRegressor()

#Initialize randomised search CV
random_search = RandomizedSearchCV(XGB, param_distributions = parameters, n_iter = 5,
                                   scoring ='neg_root_mean_squared_error', n_jobs = -1, cv = 10, verbose = 3)
random_search.fit(X_train,y_train)
```

The best parameters obtained are:

```
#considering best estimators from the randomised search
XGB1 = XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=0.9, gamma=0.4, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.25, max_delta_step=0, max_depth=10,
             min_child_weight=7,  monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
```

After 5 iteration and 10-fold cross validation the RMSE after hyper-parameter tuning reduced from 37.77 to 15.57 and hence the performance was boosted. The Mean Absolute Percentage Error was found to be 0.033.
The residual plot of the hyper-parameter tuned model came to be as follows: