Technocolabs Data Science Internship

Project Report

BigMart Sales Prediction

By Vibhu Gupta

2nd year B.tech Computer Science & Engineering

Vibhu7gupta@gmail.com

Abstract

BigMart is a big supermarket chain, with stores all around the country and it keep the track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. The aim of the project is to build a predictive model and find out the sales of each product at a particular store. In this project, we propose a predictive model using Xgboost technique for predicting the sales of a company like Big Mart and found that the model produces better performance as compared to existing model.

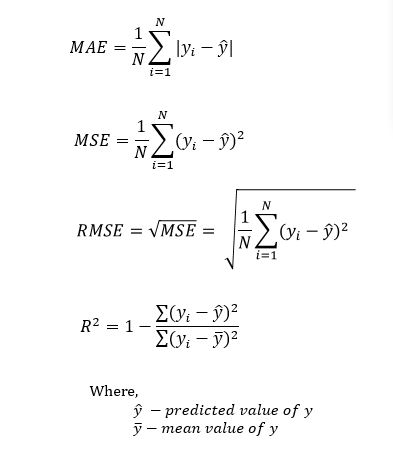
Keywords: Machine Learning, Sales Forecasting, Random Forest, Regression, Xgboost.

**Introduction:**

The Business Problem Exploring

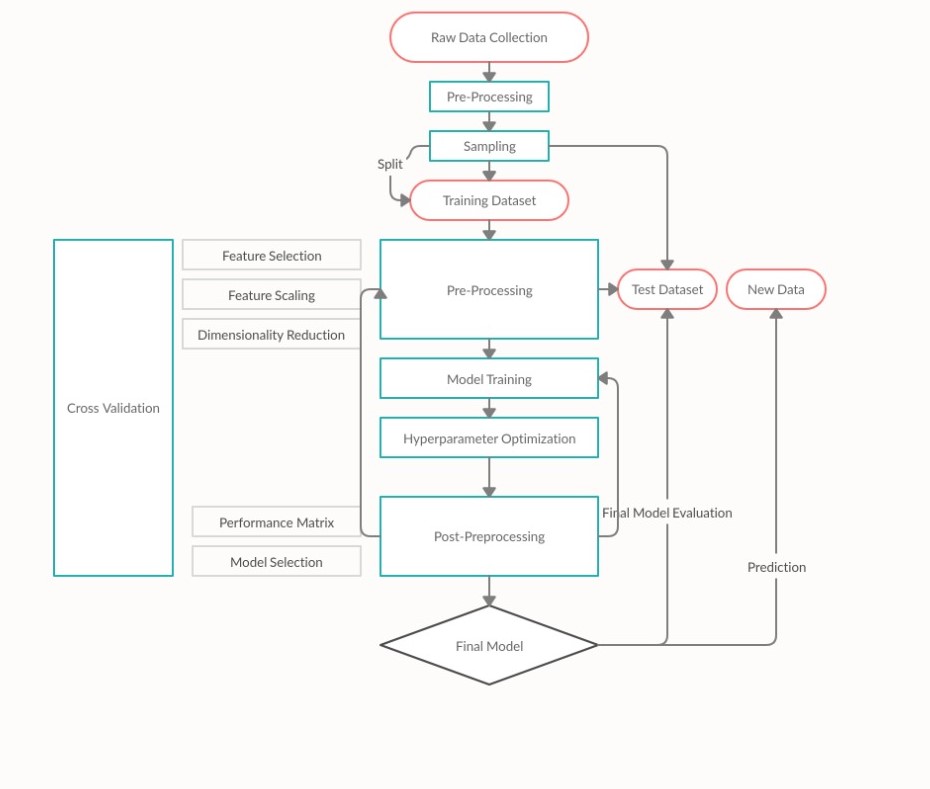
In today’s world, competition among the shopping mall and Bigmart is rapidly increasing. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service, etc.

Machine Learning Algorithms can analyze the data and helps us to find out the sales of each product at a particular store. In this project we are addressing the problem of big mart sales prediction or forecasting of an item on customer’s future demand in different big mart stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc. are used for prediction or forecasting of sales volume. To measure the performance of the models R-Square Score, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used as an evaluation metric as mentioned:



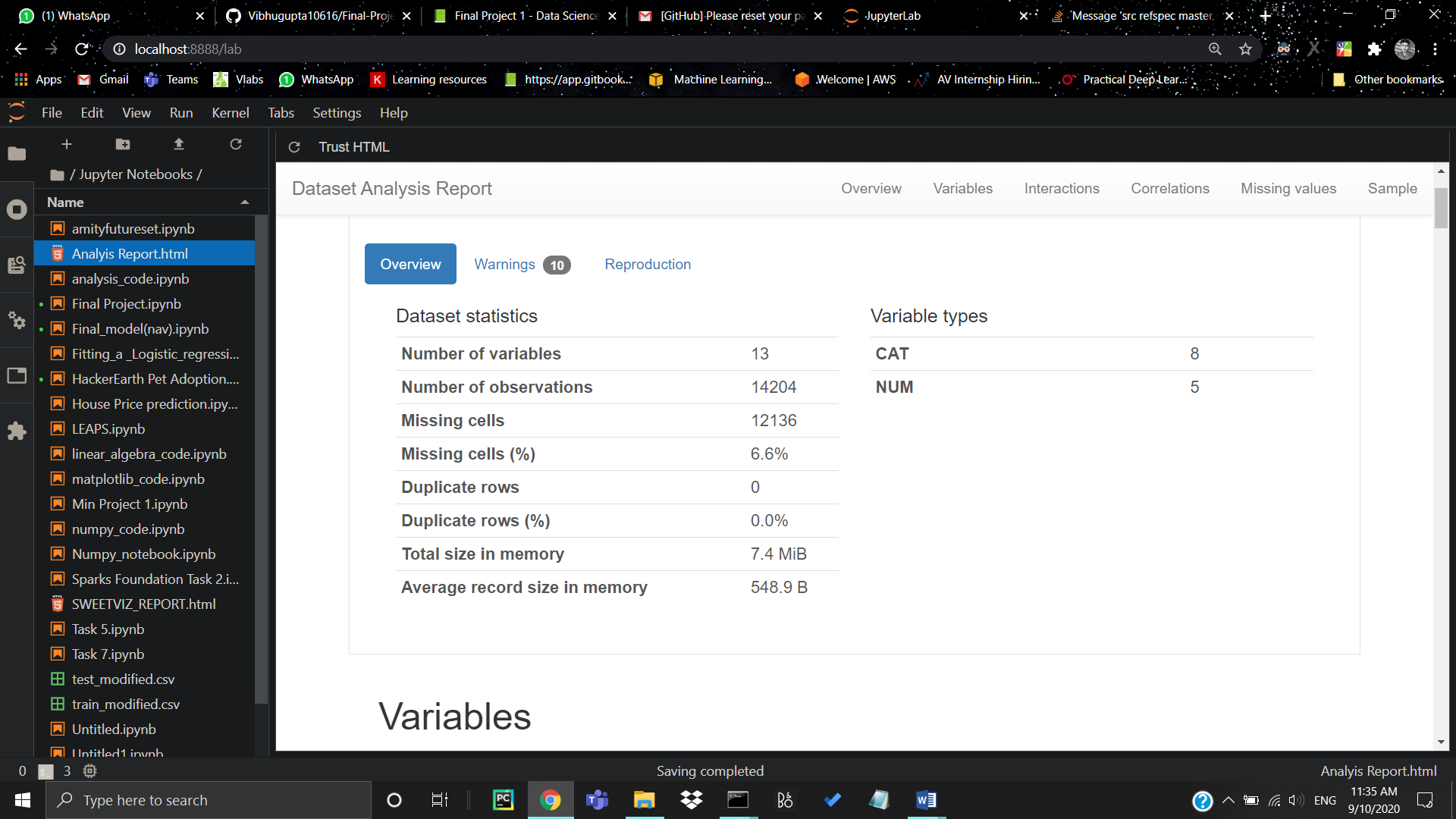
**Dataset Description**

BigMart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities. Where the dataset consists of 12 attributes like Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales. Out of these attributes response variable is the Item Outlet Sales attribute and remaining attributes are used as the predictor variables. The data-set is also based on hypotheses of store level and product level. Where store level involves attributes like: city, population density, store capacity, location, etc and the product level hypotheses involves attributes like: brand, advertisement, promotional offer, etc. After considering all, a dataset is formed and finally the data-set was divided into two parts, training set and test set in the ratio 80:20.



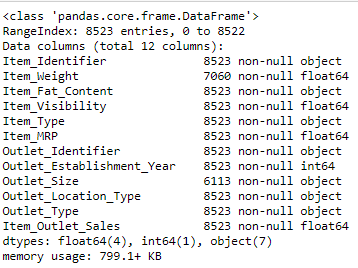
**Data Exploration**

In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible. I have used two libraries “Pandas Profiling” and “Sweetviz”. These libraries provide an easy interface for EDA and give a wider range of visualizations at one place.



**Data Cleaning**

This step typically involves imputing missing values and treating outliers. Though outlier removal is very important in regression techniques, advanced tree-based algorithms are impervious to outliers. It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing. value we replace it by the mode of that attribute and for the Item Weight missing values we replace by mean of that particular attribute.



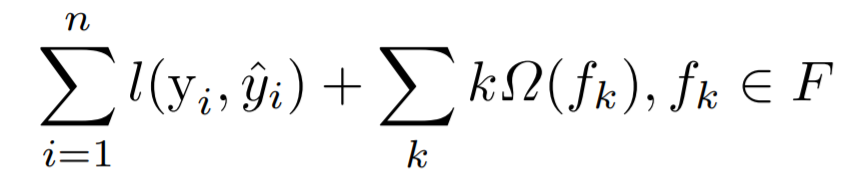
**Feature Engineering**

Some nuances were observed in the data during data exploration phase. So, this phase is used in resolving all nuances found from the data and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So, the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell. All categorical attributes discrepancies are resolved by modifying all categorical attributes into appropriate ones. In some cases, it was noticed that non-consumables and fat content property are not specified. To avoid this, we create a third category of Item fat content i.e. none. In the Item Identifier attribute, it was found that the unique ID starts with either DR or FD or NC. So, we create a new attribute Item Type New with three categories like Foods, Drinks and Non-consumables. Finally, for determining how old a particular outlet is, we add an additional attribute Year to the dataset.

**Model Building**

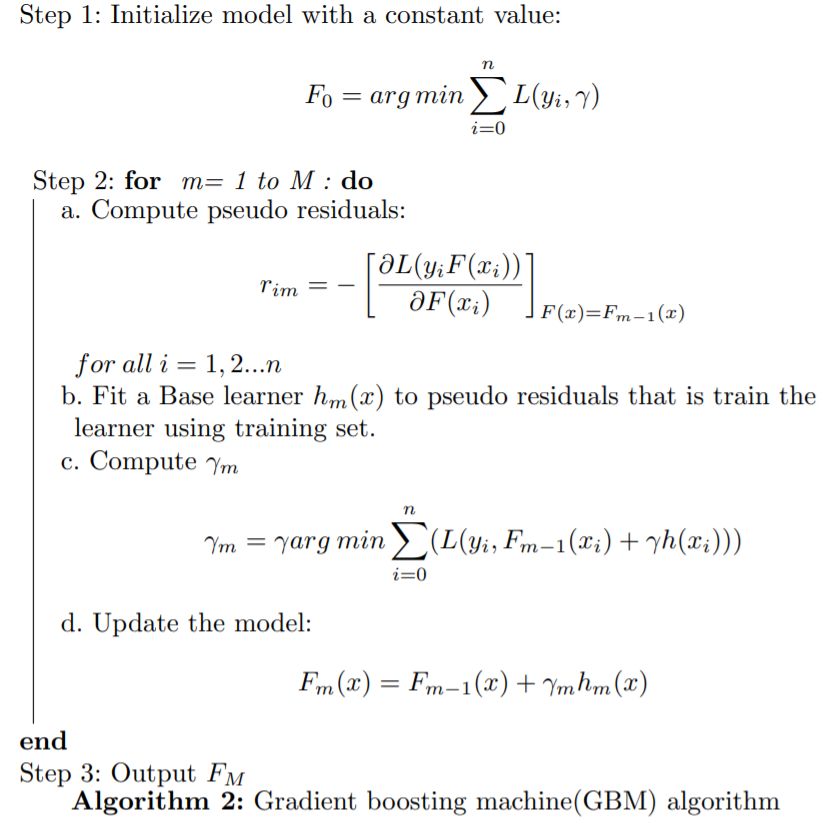
Now that we have the data ready, it’s time to make predictive models. Once the model is built it is used as predictive model to forecast sales of Big Mart. In our work, we propose a model using Xgboost algorithm and compare it with other machine learning techniques like Linear regression, Ridge regression, Decision tree etc.

XGBoost: Xgboost (Extreme Gradient Boosting) is a modified version of Gradient Boosting Machines (GBM) which improves the performance upon the GBM framework by optimizing the system using a differentiable loss function as defined:



where ˆyi : is the predicted value, yi : is the actual value and F is the set of function containing the tree, l(yi , yˆi) is the loss function. This enhances the GBM algorithm so that it can work with any differentiable loss function.

The GBM algorithm is illustrated in



The Xgboost has following exclusive features:

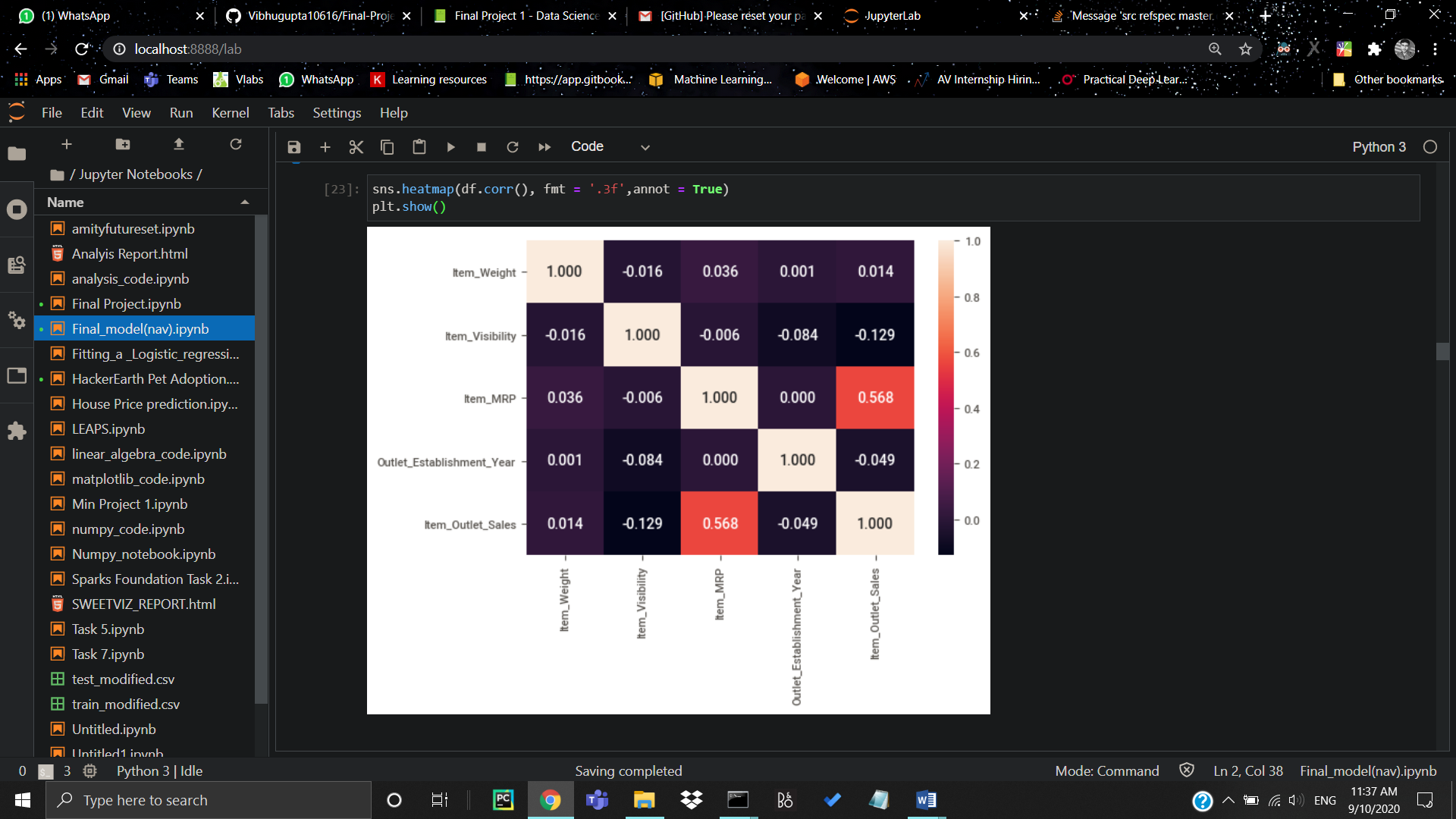
1. Sparse Aware - that is the missing data values are automatic handled.

2. Supports parallelism of tree construction.

3. Continued training - so that the fitted model can further boost with new data.

**Implementation and Results**

Correlation between the Numerical Predictors and Target Variable by using Correlation Matrix



From the current numeric variables, we can observe that the Item Visibility is the feature with the lowest correlation with our target variable. Therefore, the less visible the product is in the store the higher the price will be. This is curious since from the initial assumptions this variable was expected to have high impact in the sales increase. Nevertheless, since this is not an expected behaviour and we should investigate.

Moreover, this feature has a negative correlation with all of the other features. Furthermore, the most positive correlation belongs to Item MRP.

**Conclusions**

Day to day the companies or the malls are predicting more accurately the demand of product sales or user demands. Extensive research in this area at enterprise level is happening for accurate sales prediction. As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring more accurate prediction algorithm so that the company will not suffer any losses. In this research work, we have designed a predictive model by modifying Gradient boosting machines as Xgboost technique and experimented it on the 2013 Big Mart dataset for predicting sales of the product from a particular outlet. Experiments support that our technique produce more accurate prediction compared to than other available techniques like decision trees, ridge regression etc.

The cross-validation score along with MAE and RMSE of the proposed model and existing models is shown in Table 1 and Table 2 respectively. From the results we observe that and found that the XGBoost model is significantly improved over the other model.

