Recruit Restaurant Visitor Forecasting

by Sai Vivek Kammari (a1807677)

School of Computer Science, The University of Adelaide

Report submitted for COMP SCI 7209 Big Data Analysis and Project at the School of Computer Science, University of Adelaide towards the Master of Data Science

Abstract

For the prediction of the target variable which is the number of visitors in the given problem. I have tried to implement three regression models and compared their performance on the training dataset. The models evaluated are XGBoost, Random Forest, KNN. The best accuracy I could achieve for the experiment on the train dataset was using the XGBoost regression. Root Mean Squared Logarithmic Error (rmsle) for the XGBoost implementation was (0.489975) and the rmsle value for the random forest model was value (0.492949). I have finalized on the XGBoost algorithm and forecasted the final values for the period 23rd April-31st May 2017.

Contents

1 Introduction

Managing any local business involves many challenges and when it comes to the managing of a restraint is further challenging. In order to avoid food wastages which is of national interest, the restraint needs to anticipate the expected number of customers on a particular day before hand. This not only minimizes the food wastage, helps the restaurants to attain better profitability by managing the stock and staff efficiently.

In this project we try to predict the expected visitors for a give set of Japanese restaurants based on the historic visitors data.

2 Dataset

The datasets given for the experiment are extracts from two websites dubbed Hot Pepper Gourmet (hpg- users can browse for restaurants and also reserve online) and AirREGI / Restaurant Board (air- it's a restaurant booking and cash management system). The csv datasets 'air_reserve.csv', 'air_store.csv' and 'air_vist_data.csv' are extracted from the AirREGI website. The 'hpg_reserve.csv' and 'hpg_store_info.csv' are the extracts from the website Hot Pepper Gourmet. Additionally, 'store_id_relation.csv' and 'date_info.csv' are the two other supporting datasets provided. The dates and the stores for which the predictions are to be done are provided in the 'sample_submission.csv' file.

The historic data which is used as the training data is from 2016 until April 2017. The dates which are to be predicted fall in the range of April and May of 2017 and the problem stated also clearly mentioned that the dates for which the visitors data for the given restaurants is need to be predicted consists a 'Golden Week'. The week starting from 4/29/2017 is dubbed as 'Golden Week' because the week has 4 national holidays and the expected visitors during this week are high.

3 Implementation

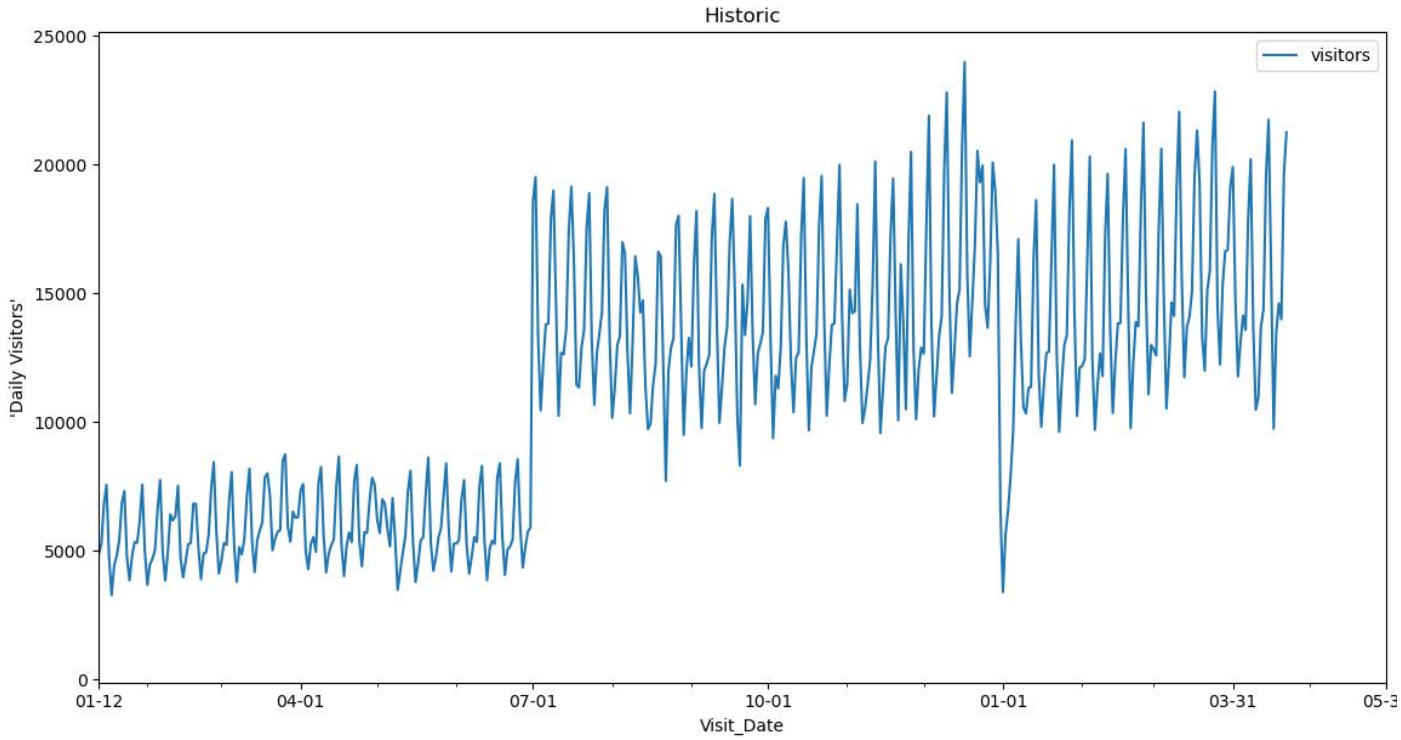
3.1 Exploratory Data Analysis- We have used Jupiter notebooks to import and analyze the data. During the preliminary examination, it has been observed that the training and test data have the following shapes (Figure 1).

Figure1

Dataset	Rows X Columns
Air stores	(829, 5)
Hpg stores	(4690, 5)
Air reserves	(92378, 4)
Air visits	(252108, 3)
Hpg reserves	(2000320, 4)
Date information	(517, 3)
Store id lookup	(150, 2)
Predictions	(32019, 2)

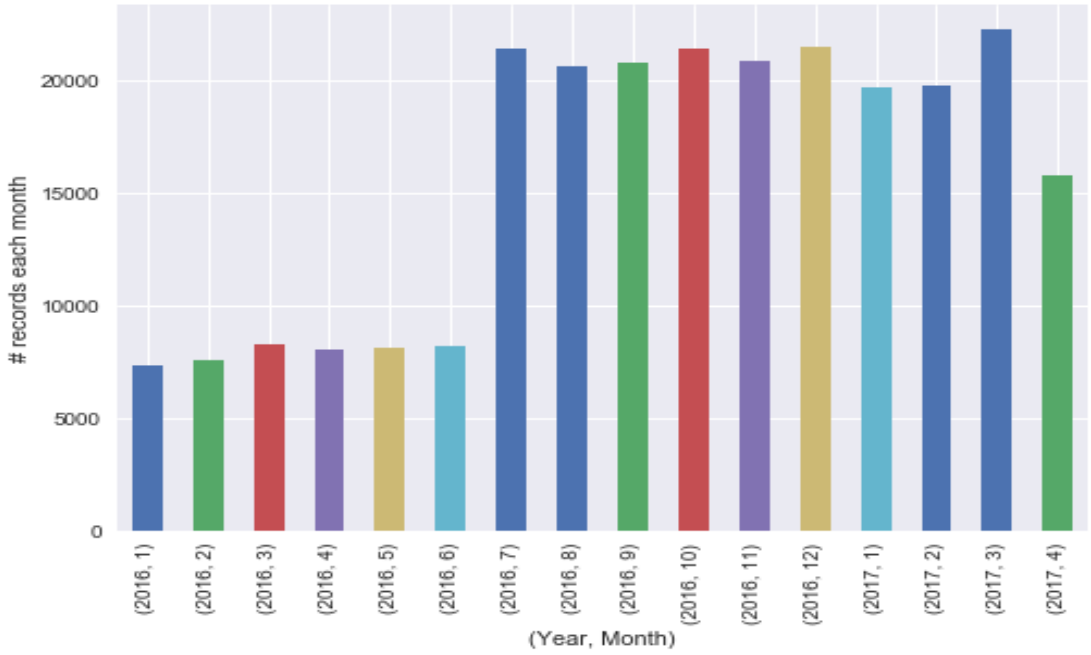
Further, deeper look into shows that the predictions are to be made for only a selected number of stores (821) and all these stores are air stores (stores registered in air website) . The other major observations are that 'The number of stores in air visits dataset is equal to number of stores in air stores dataset' & 'All the stores in air visits are present in air stores dataset.

Figure2



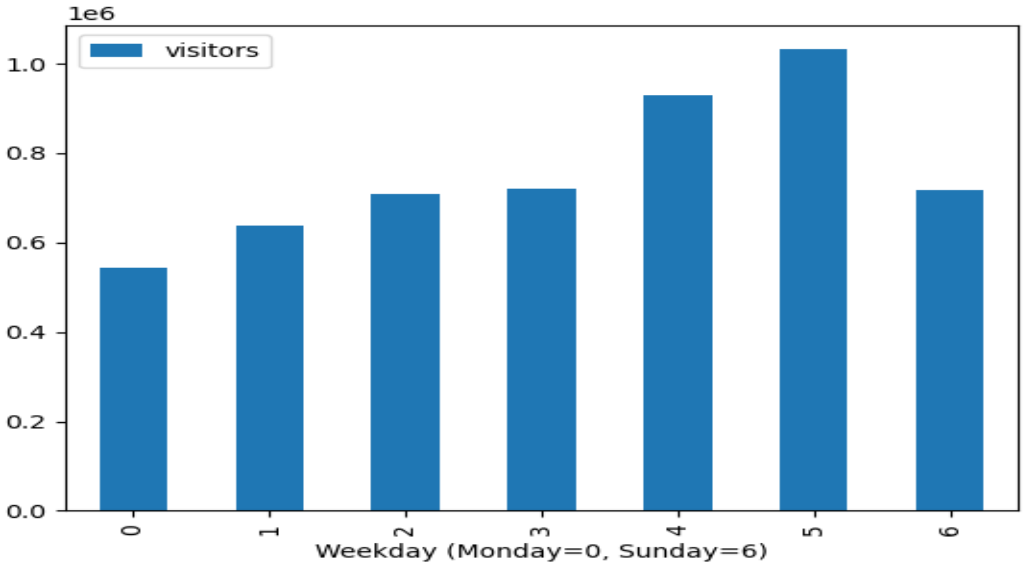
The above figure plots the total number of visitors (aggregate to all the restaurants) on each day of the historic period. The gap in the figure2 is the visitor numbers that are to be predicted (from last week of April 2017 to the end of May 2017).

Figure 3



The above figure3 summarizes the total number of visitors to all the restaurants on a monthly basis. From the above figure it is evident that the number of visitors to the restaurants has grown significantly from the month of July 2016.

Figure4



The above figure4 summarizes the total number of visitors to all the restaurants on a every day of the week. From the above figure it is evident that the number of visitors to the restaurants is highest on Saturday followed by Friday.

3.2 Feature Engineering- For predicting the numbers of visitors we will have analyze the relation between the various features given various datasets. For instance, we will have to analyze the relation between total visitors and the day of week, holidays, weekends and we will also need to analyze the relation between the location of the stores and the visitors.

For doing so, we will have to build new dataset which will combine all the datasets on various parameters. The complete restaurant dataset can be developed by merging the 'air_store.csv', 'hpg_reserve.csv' and 'store_id_relation' datasets. The total reservations dataset can be developed by combining the 'hpg_reserve.csv' and the 'air_reserve.csv' datasets.

3.3 Machine Learning Models- For the multivariate time series forecasting, I have selected to use the following regression models.

Random Forest- The primary principal involved in the random forest algorithm is that, a random sample from the data is chosen and each tree is trained. This random selection of the samples makes the algorithm very robust.

One of the major challenges in implementing the Random Forest is the hyper parameter tuning. I have used GridSearchCV technique for the tuning of hyper parameters.

GridSearchCV- the functioning of grid search is pretty simple, it will select all the combination from the given list of hyper parameters and outputs the parameters which are best and which result in minimum error. However; the grid search takes a lot of time for the computation and hence in the experiment, I have done the grid search in my local machine and for the purpose of submission, I have just used the optimized parameters for random forest.

''(n_estimators=25, random_state=3, max_depth=20, min_weight_fraction_leaf=0.0002)''

XGBoost- XGBoost is one of the most efficient algorithm which can be implemented for both classification and regression problems. The XGBoost algorithm is an implementation of gradient boosting. XGBoost can also be used to predict the time series data in multivariate scenarios.

One of the major challenges in implementing the XGBoost is the hyper parameter tuning. I have used Bayesian Optimization technique for the tuning of hyper parameters.

Bayesian Optimization- Bayesian optimization is relatively a faster method of optimization compared to grid search and random search optimization techniques. The Bayesian optimization works by defining an acquisition function and finding out the optimum values of the given range of parameter values at the point where the acquisition function maximizes. Bayesian Optimization technique learns from the previous predictions.

KNN Regression- The third model I have used is KNN Repressor, It is a more simple approach but it is a powerful algorithm for the time series predictions. The basic principle used by the KNN to predict the data is to use the most similar historical samples to the new data. I have not done any tuning for getting the optimum K value and hence, the algorithm has used the default K value.

4 Evaluation

The metric of evaluation for the models is Root Mean Squared Logarithmic Error (rmsle) which is calculated as :

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1) - \log(y_i+1))^2}$$

RMSLE parameter penalizes both under prediction and over prediction of the target variable. In this instance, the target variable is the total number of visitors. Large restraunts are relatively not so sensitive for under predictions or over predictions they can tolerate however, that is not the case for smaller restraunts.

The below table represent the rmsle for various implemented machine learning models. The below rmsle values for the XGBoost model might not be the same when the code is run but, majority of the time (around 98%) the rmsle values for XGBoost are lower.

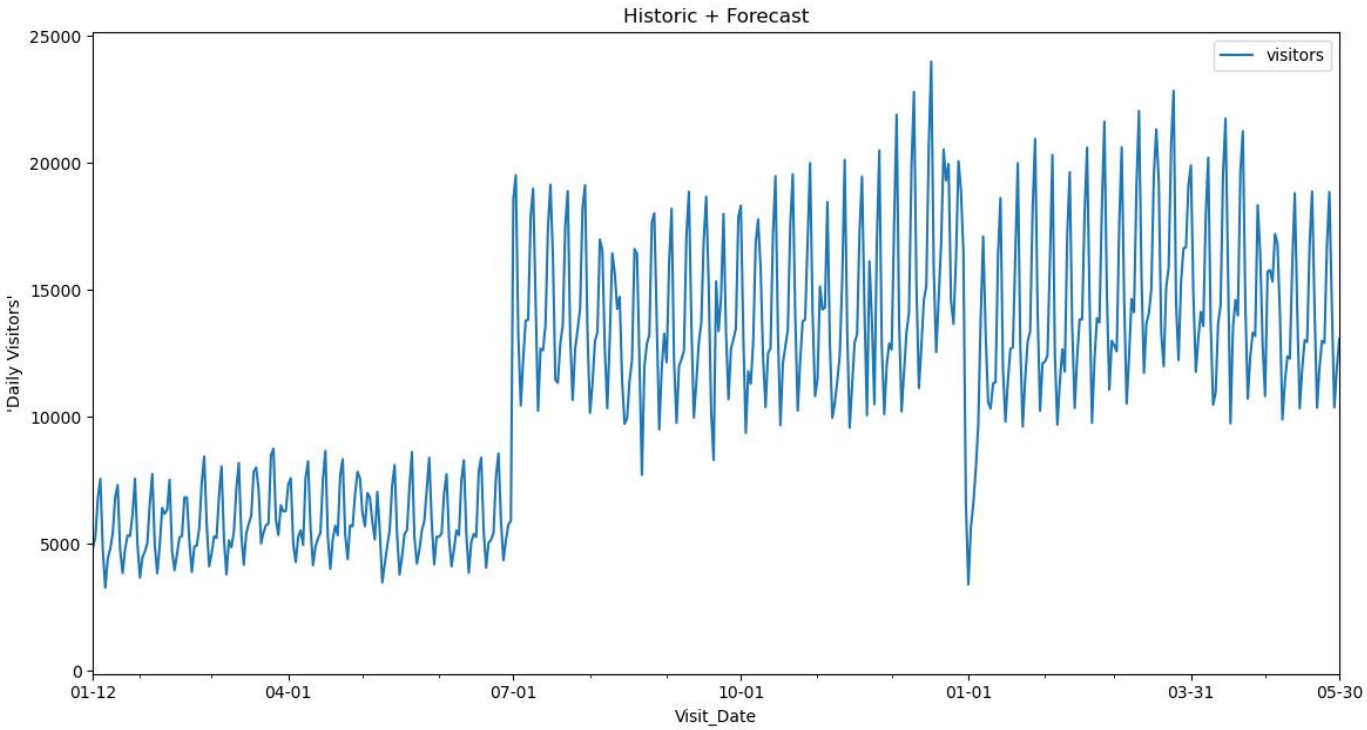
REGRESSION MODEL	RMSLE
XGBOOST	0.489975
RANDOM FOREST	0.492949
KNN	0.608833

.

5 Conclusion

As we could notice that, the rmsle value for the XGBoost model is lowest and hence I have taken XGBoost for the final implementation and predicted the value of the visitors. Below figure is a time series plot of the historic and forecasted data.

Figure5



6 Recommendations for Future Work

- Weather also plays a key role in the business of restaurants and in this whole experiment I have not analyzed the impact of weather. In future, we can use the external weather data for the entire time period and add this to the final dataset. I expect this implementation will definitely give us more precise forecast.
- We can use better tuning techniques for Random Forest, tuning the Random Forest parameters using the random search and then narrow down the grid of value. After narrowing down to the grid levels, we can then perform grid search cv to finalize the parameters. This procedure would also give better predictions from Random Forest model.
- We can implement ensemble techniques for improving the predictions. I plan to improve my model over the future after gaining significant knowledge in this direction.

7 References

https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data