RESTAURANT VISITOR FORECASTING DATA ANALYTICS USING DATA MINING TECHNIQUES

Divya Samragni Nadakuditi

dnadakud@my.bridgeport.edu
Department of Computer Science and Engineering
University of Bridgeport, CT.

ABSTRACT

The amount of data that is generated is increasing exponentially day by day. Due to this, there is necessity to know information about the data. This increases the need for Data Analytics and Data mining concepts. In this paper, we perform data analytics on a chosen dataset Restaurant Visitor Forecasting to predict the number of future visitors a restaurant will receive using Data mining techniques. We further use data visualization tools to visualize the data result obtained. We also compare which model best suits the data set. The Implementation and experimental results are also mentioned in this paper.

1. INTRODUCTION

With enormous amounts of data, the need to analyze the data for various reasons like business, information, statistical analysis, reporting, multi dimensional analysis increases. We use Data Mining techniques to identify unusual situations, develop positive alternatives and also to track effectiveness. Doing so helps in improvement of business and taking required decisions for future.

In this paper, we discuss about a Data Analytics scenario and address it with data mining techniques. We analyze the project, find a suitable dataset and solve the problem.

2. RELATED WORK

Future visitor forecasting is service industry's one of the meaningful task.There are various techniques proposed by many people for predicting the future visitors.

There are various Data mining techniques that can be used for Prediction. Random search method is an algorithm that can be used for classification and regression. It can work on both linear and non linear decision trees. Random Forest builds multiple decision trees. and based on the generated random decision trees it gets the mean predictions and generate final prediction.

Neural Networks is another powerful method in which we have multiple layers of network from input to output. It is used to model non-linear relations.

Gradient Boosting is another method that without long time of computation provide high performance. It has also considered for the idea of XGBoost. It resulted in the best performance in many real world datasets.

3. PROBLEM DEFINITION

Running a well developing local restaurant is not an easy job. It involves many unexpected problems on regular basis. One of the main problem is predicting the number of visitors. This might seem to be a small measure, but many factors are dependent on this measure. The number of visitors prediction is used to effectively to purchase ingredients and in scheduling staff. The prediction would not be that simple as many factors like weather, locality, holidays affect the number of visitors.

Here we predict the number of visitors visit restaurants on some particular dates using the reservation and visitation dates along with the locality and other restaurant related details. We further predict which cuisine and areas can be visited more.

4. DATASET

The dataset we chose to perform Data Analytics was Recruit Restaurant Visitor Forecasting. We perform an extensive exploratory data analysis. The data was collected from restaurants in Japan.

The data was collected from two sites: Hot Pepper Gourmet, AirREGI/Restaurant Board(air). The first site is used by users to search for the restaurant and make

reservations online whereas the second site is used as a cash register system and a reservation control.

The data is extracted in the form of 8 files. The file air_visit_data.csv has the historic visit data of the air restaurants. The file air_reserve.csv has the data of reservations made through air system and similarly hpg_reserve.csv has the reservation data made through hpg systems. The two files air_store_info.csv, hpg_store_info.csv have the details of the restaurants like the genre and location. The file store_id_relation.csv has the connection of Air and hpg ids. The file date_info.csv flags the Japanese holidays. A specific predicted file submission format was also included in the dataset.

5. METHODOLOGY

In this project, we have used Gradient Boosting Regressor, XGBRegressor and Neural network models. The accuracy of these model is compared and the best model is chosen. Gradient Boosting is a machine learning technique which is used for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage wise fashion like other boosting methods do. The Gradient boosting builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary. The models are implemented using scikit-learn package.

The Gradient boosting regressor model is built with learning_rate of 0.2, n_estimator=200 and max_depth=10. Where as the XGBRegressor is built with learning_rate = 0.92 , n_estimators = 1000. The neural network is built with 2 hidden layers. The hidden layer 1 has 35 neutrons and hidden layer 2 has 15 neurons.

6. EXPERIMENTAL RESULT AND EVALUATION

6.1. Implementation

We implemented our model on Python 3.6 with scikit learner. We have used XGBoost package for our program. We have used the Lenovo laptop with Intel i7 CPU. We also used matplot library to plot simple plottings. We use keras model for modelling neural networks.

6.2. Regression Performance Measure

We measure the performance using Root Mean Square Error (RMSE) using the below formula where n is the number of predictions in the test set; Pi is the number i prediction value; Oi is the number i actual value of visitors.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} \left(P_i - O_i \right)^2}{n}}$$

We have used the RMSE formula and found out the Error rates of Gradient Boosting Regressor, XGB Regressor, Neural Network.

We further plotted a graph of Actual and Predicted output. To plot these graphs we got the actual values from the data of the restaurants. These actual values are represented as points. The data contains the date, month, year, mean visitors, day of week, holidays, mean reservations, mean latitude, mean longitude and many other details.

The Predicted values are the predicted number of visitors from each model.

Using the actual data of the restaurants and the predicted value of the visitors we plot the graph. The Actual data is predicted as scattered points in green color whereas the predicted visitors are represented as points in blue color.

Figure 1 shows the plot of Actual output and predicted output of Gradient Boosting Regressor.The RMSE value obtained was 0.3635.

Figure 2 shows the plot of Actual output and predicted output of XGBRegressor. The RMSE value obtained was 0.3502.

Figure 3 shows the plot of Actual output and predicted output of Neural Network. The RMSE value obtained was 0.4861.

Out of all the three, XGBRegressor has the least RMSE.

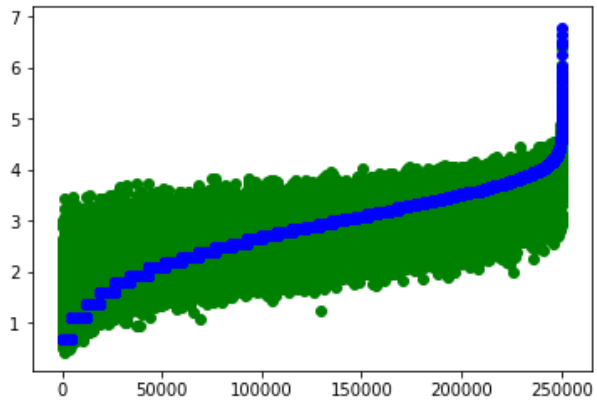

FIGURE 1. Gradient Boosting Regressor - Actual and predicted output.

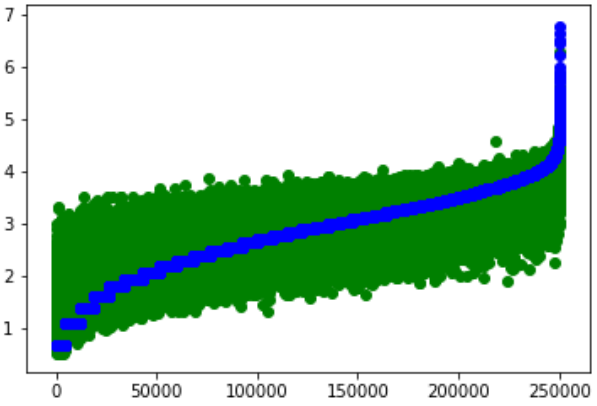

FIGURE 2. XGBRegressor - Actual and predicted output.

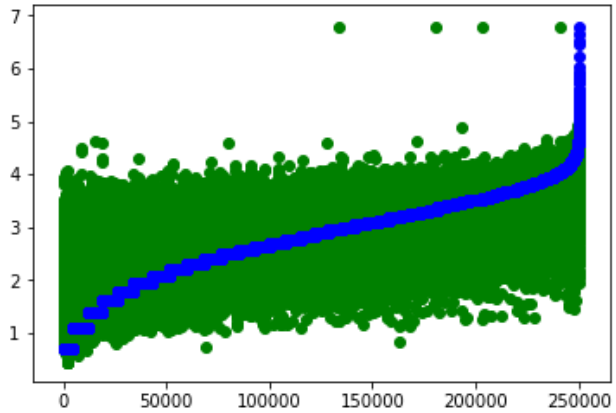

FIGURE 3. Neural Network - Actual and predicted output.

After training the model, we get the visitors values only if the store id, day of week, holiday flag column matches. We then output the values into a csv file.

6.3. Data Visualization

We then visualize the data output to represent the output data information.

Figure 4 represents the Genre name and visitors who have booked through AirREGI website. In this graph, Genere name is actual value and visitors is an actual value. The graph shows that Lzakaya will be the most visited Genre with approximately around 27 million. Italian/French will be the second most visited Genre with 18 million.

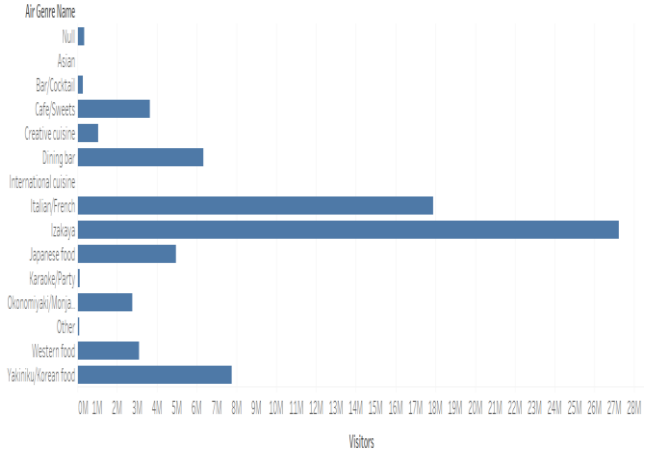

FIGURE 4 Visualization of visitors and Genre in AirREGI site

Figure 5 visualizes the areas and the predicted number of visitors in that area using AirREGI site. The areas are the actual values that are marked against the predicted future number of visitors.The area Miyagi-ken Sendai-shi Kamisugi has the maximum number of visitors with count of 5,521,175.66529829.

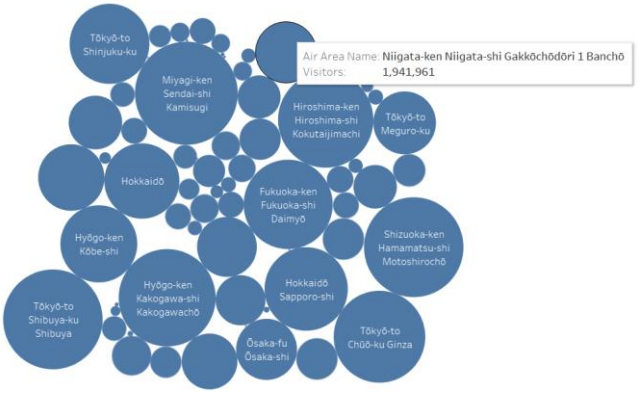

FIGURE 5 Visualization of Areas and number of visitors through AirREGI site.

Figure 6 visualizes the Areas and Genre with the number of visitors through Air REGI site. In this graph, Genere name and area are the actual value and visitors is the predicted value .

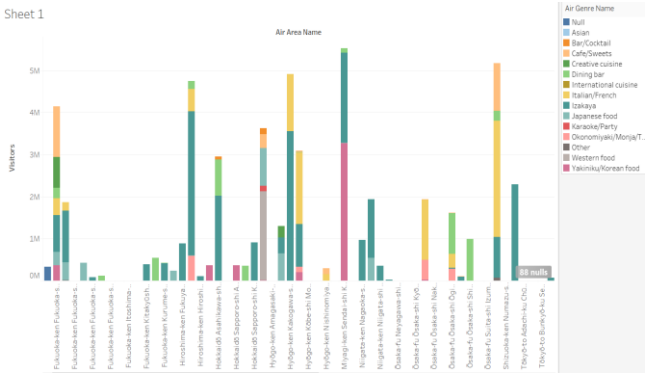

FIGURE 6 Visualization of Areas and Genre with number of visitors through AirREGI site .

Figure 7 represents the Genre name and visitors who have booked through Hot Pepper Gourmet website. Japanese Genre has the highest number of visitors. In this graph, Genere name is actual value and visitors is the predicted value.

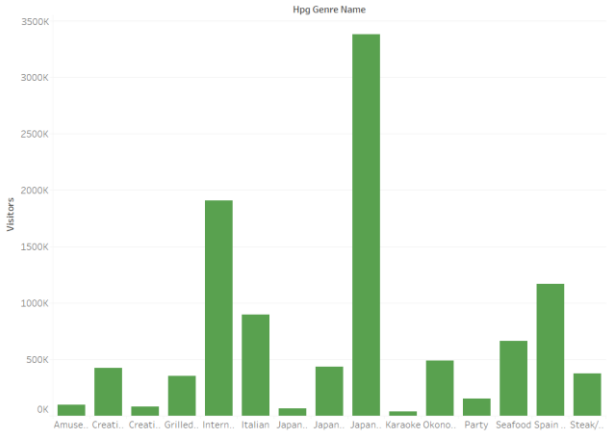

FIGURE 7 Visualization of Genre with number of visitors through Hot Pepper Gourmet site .

Figure 8 visualizes the Areas and Genre with the number of visitors through Hot Pepper Gourmet site. In this graph, Genere name and area are the actual value and visitors is predicted value.

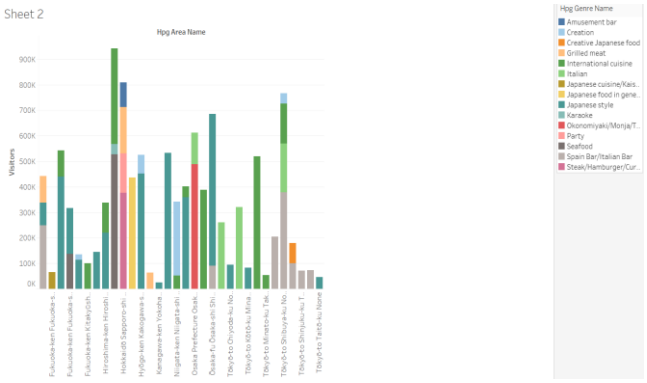

FIGURE 8 Visualization of Areas and Genre with number of visitors through Hot Pepper Gourmet site .

Figure 9 visualizes the areas and the predicted number of visitors in that area using Hot Pepper Gourmet site. In this graph, Genere name is an actual value and visitors is a predicted value.

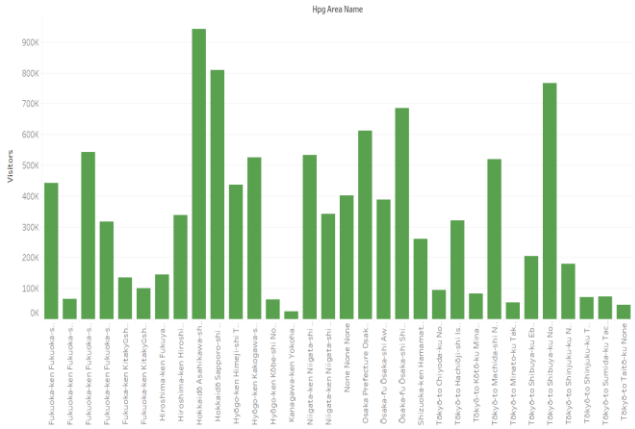

FIGURE 9 Visualization of Areas and number of visitors through Hot Pepper Gourmet site.

Figure 10 illustrates the areas where the restaurants are located in the Map.The marked spots are the only areas for which we have the data.

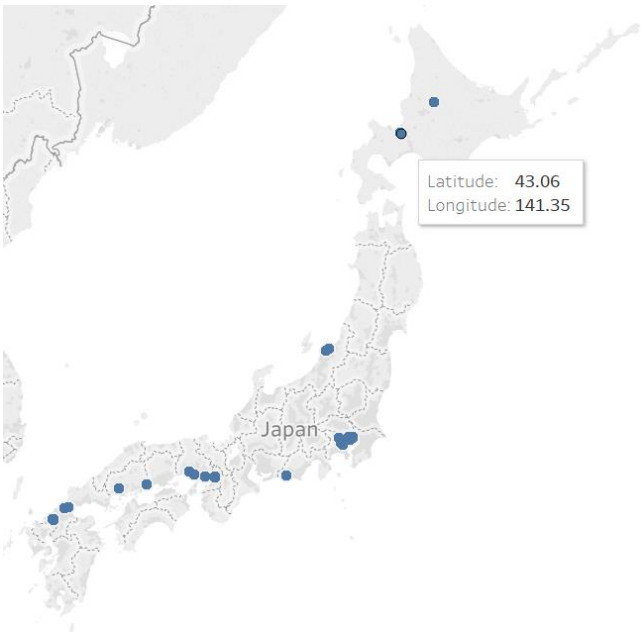

FIGURE 10 Visualization of Areas where the restaurants are located.

Figure 11 Visualizes the Areas of restaurants with the number of visitors. The areas with light color has the fewer number of visitors when compared with the Areas in Dark color. The spot color tends to become darker with the increase in the number of visitors. The latitude and longitude details used are actual values and the number of visitors is the predicted value

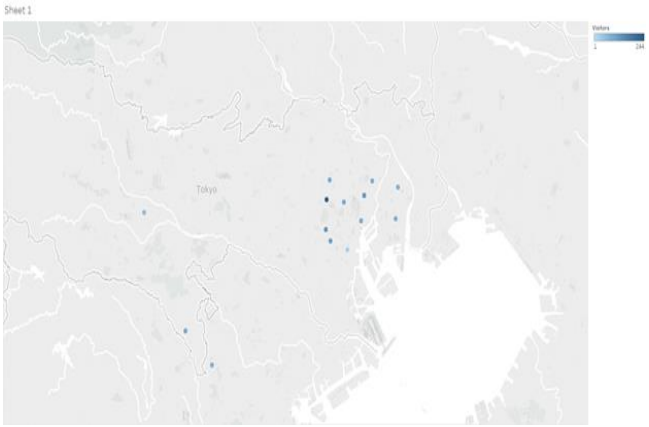

FIGURE 11 Visualization of Areas where the restaurants are located with the number of visitors.

7. PAGE NUMBERING

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

8. CONCLUSION

In this paper, we have presented our chosen approach of forecasting the future number of visitors to restaurants using restaurant information. We also used the historical visits and historical reservations to predict the future visitors. We have compare XGBRegressor, Gradient Boosting Regressor and Neural Networks to train our model and find the one with less random mean square error. This evaluation shows the effectiveness of our approaches which can be used for future work insights.

9. REFERENCES

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

[2]http://www.dbta.com/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx