Restaurant Visitor Forecasting

Team Name: Error 404

Riya Ghosh
MT2020125
riya.ghosh@iiitb.org

Parijat Moulik
MT2020059
parijat.moulik@iiitb.org

Anjishnu Chakrabarti
MT2020116
anjishnu.chakrabarti@iiitb.org

***Abstract*—This is a detailed report on our work on predicting the total number of visitors to a restaurant for future dates using reservation and visitation dates.**[1]

***Index Terms*—Feature Engineering, Label Encoding, RMSLE, Linear Regression, Decision Tree Regressor, Gradient Boosting Regressor, XGBoost, LightGBM, ensemble model**

Problem Statement

Running a thriving local restaurant isn't always as charming as first impressions appear. There are often all sorts of unexpected troubles popping up that could hurt business.

One common predicament is that restaurants need to know how many customers to expect each day to effectively purchase ingredients and schedule staff members. This forecast isn't easy to make because many unpredictable factors affect restaurant attendance, like weather and local competition. It's even harder for newer restaurants with little historical data.

Recruit Holdings has unique access to key datasets that could make automated future customer predictions possible. Specifically, Recruit Holdings owns Hot Pepper Gourmet (a restaurant review service), AirREGI (a restaurant point of sales service), and Restaurant Board (reservation log management software).

Dataset

A time-series forecasting problem centered around restaurant visitors is provided. The data comes from two separate sites:

- Hot Pepper Gourmet (hpg) : similar to Yelp, here users can search restaurants and also make a reservation online
- AirREGI / Restaurant Board (air) : similar to Square, a reservation control and cash register system

The reservations, visits, and other information from these sites is to be used to forecast future restaurant visitor totals on a given date. The training data covers the dates from 2016 until early (first week) April 2017. The test set covers the mid weeks (second and third weeks) of April 2017. The training and testing set both omit days where the restaurants were closed.

This is a relational data set from two systems. Each file is prefaced with the source (either air_ or hpg_) to indicate its origin. Each restaurant has a unique air_store_id and hpg_store_id. Note that not all restaurants are covered by both systems, and that you have been provided data beyond the restaurants for which you must forecast. Latitudes and Longitudes are not exact to discourage de-identification of restaurants.

1) Air_reserve.csv: This file contains reservations made in the air system. Note that the reserve_datetime indicates the time when the reservation was created, whereas the visit_datetime is the time in the future where the visit will occur.
2) Hpg_reserve.csv: This file is similar to air_reserve.csv and contains reservations made in the hpg system.
3) Air_store_info.csv: This file contains information about select air restaurants like genre, area, longitude and latitude. [Note: latitude and longitude are the latitude and longitude of the area to which the store belongs]
4) Hpg_store_info.csv: This file is similar to hpg_store_info.csv and contains information about select hpg restaurant.
5) Store_id_relation.csv: This file allows you to join select restaurants that have both the air and hpg system.
6) Date_info.csv: This file gives basic information about the calendar dates and holidays in the dataset.
7) Train.csv: This file contains historical visit data for the air restaurants.

I. Introduction

Maintaining restaurants is an ambitious undertaking. A lot of effort and resources is poured into such ventures. So for them to be successful, knowing the target market's preferences and habits is very important. Restaurants could be managed proactively by optimizing their resources and cutting off unwanted expenditures.

Few trends in the visitor's data observed over an year can be useful for predicting the number of future visitors. Along with that, the reservation data from different reservation services in Japan will provide more information to understand

[1]The problem statement is a contest hosted on kaggle.

the customer's motivation. Observing the inclination of the target market to a particular area and finding out the popular genres gives a better idea to the restaurant market about exploring new strategies.

The final model we'll generate gives a prediction on the total number of visitors on a future date for a given restaurant. Knowing whether a day is going to be busy can help the restaurants to better prepare their food and schedule their staff members on such days. This can affect the customer satisfaction directly helping the restaurants to maintain consistency in the market. In a similar manner, being aware of whether a day would go by slowly in future, could help in stocking up less ingredients and prioritizing staff retention.

II. Data Pre-Processing

The train dataset contains information about the number of visitors for a particular air_store on a given day. So, we need more information about the air_stores. The hpg_reserve dataset contains all the reservation data entries for hpg restaurants. However, not all hpg restaurants are mapped with air store id. Therefore hpg_reserve dataset is inner joined with the store_relation_id dataset to get the air_store_id for available hpg_store_id.

The reserve datasets contain only the information about reservation date and number of reserved visitors for an air_store_id on that day. So these datasets are joined with their corresponding store_info datasets to get the information about genre, area and location (latitude and longitude) of the restaurants. All these reservation data and restaurant information is joined with the train dataset that contains the number of visitors.

Date_info gives the data whether a particular date was a holiday in Japan or not. This is important information to predict the number of visitors and hence this dataset is joined with the train dataset. From the visit date column present in the train dataset, we can avail more information about the day of week, month, year to check for trends in the number of visitors corresponding to these features.

None of the datasets had any null values. So there is no need of performing any kind of missing data imputation.

III. Data Visualization

A. Basic Overview

The train dataset contains the number of visitors for different air stores from January 1, 2016 to April 7, 2016. We have to predict the number of visitors for dates lying between September 22, 2016 to April 22, 2016. The distribution of number of visitors for each visit date is skewed with a large number of restaurants having mean number of visitors less than 20.

The air reserve data is compared with the hpg reserve data. There are a lot more hpg restaurants (13335) when compared to the number of air restaurants(314). However, only 150 hpg restaurants are mapped with air store id and only those hpg restaurant data is considered because the train dataset contains data for restaurants with valid air store id. The air reservation data contains restaurants that serve 14 different genres of food located in 103 different areas in Japan. There are 34 unique genres of hpg restaurants located at 119 locations across Japan.

B. Train data visualization

On plotting the number of unique restaurants against the train data,we observe a 150 percent hike in the number of restaurants during mid of 2016 (shown in Fig. 1).

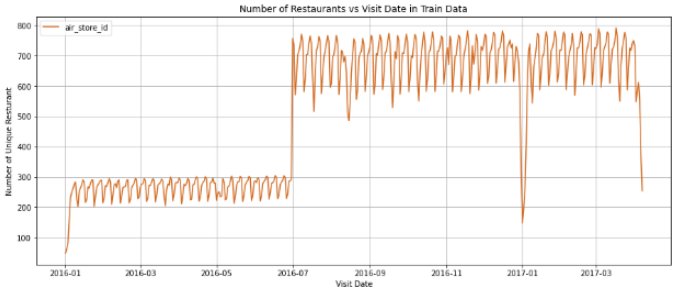

Fig. 1.

The reason behind the hike is that there is an addition of 500(approx) new restaurants to the air database in mid 2016. This caused a hike in the number of visitors after mid 2016. The comparison between the number of visitors with that of air reservation and hpg reservation is shown in Fig. 2. It is clear that the majority of people do not make a reservation before visiting a restaurant. There is a sharp decline in visitors and reservations at new years eve as most of the restaurants remain closed on new year's eve.

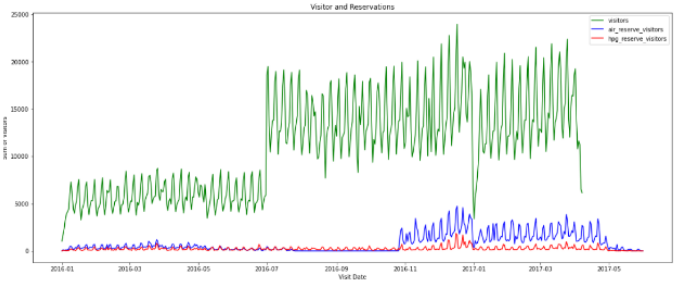

Fig. 2.

C. Visualization based on Area name

Fig. 3 shows the distribution of restaurants in Japan as given in the train dataset. The restaurants are located in 7 major areas. Zooming on each of these 7 locations shows the distribution of restaurants in each of the locations. Tokyo has the largest number of restaurants. Hokkaidō Asahikawa-shi 3 Jōdōri is that area that has the most number of hpg reservations, whereas, Tōkyō-to Shibuya-ku Shibuya has the largest number of air restaurants.

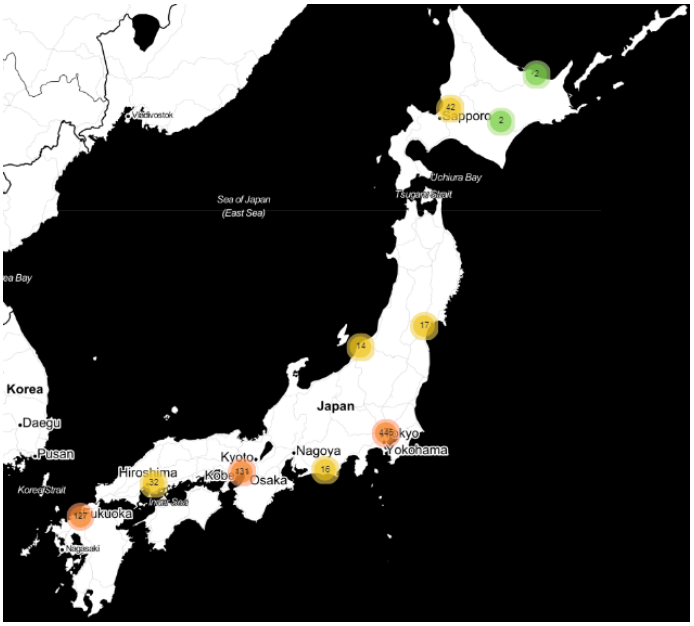

Fig. 3.

D. Visualization based on Genre name

Individual data visualization on number of visitors against genre name in the air and hpg reservation dataset shows that Japanese type is the most popular genre of food for hpg users which is quite understandable and, Party, japanese cuisine/Kaiseki. and Karaoke are not so common among hpg users who make reservations. However, findings from the air reserve dataset shows that Izakaya is the most popular genre and the second most popular genre in Japan is Italian/French. International cuisine,Asian and Karaoke/Party are the least preferred genre for air users. The choice of air users dominates and hence barplot on train dataset shows similar findings as air reserve dataset. This is shown in Fig. 4.

IV. Feature Engineering

From both the air and hpg reservation information we observed that most people visit the restaurant without any prior reservation. A 24-hour periodic trend (shown in Fig. 5) is noticed in the time difference between the visit time and the reserve time. More people make reservations at a gap of

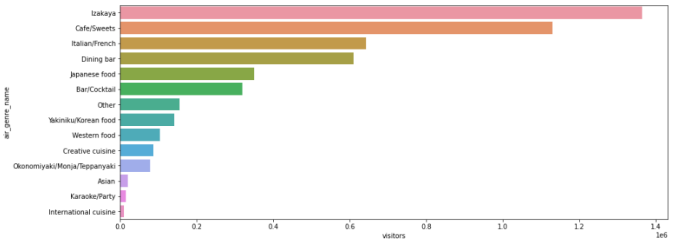

Fig. 4.

24, 48, 72... hours, whereas very few visitors reserve 12, 36, 60... hours prior to visitation.

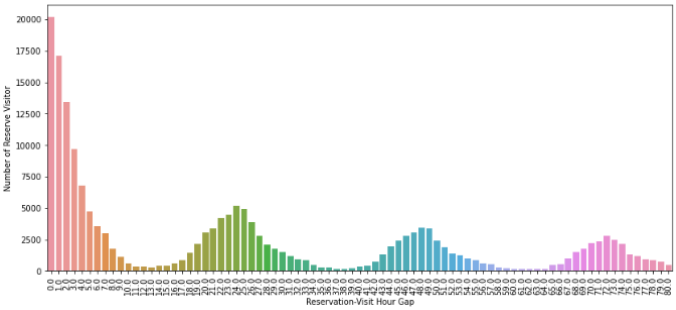

Fig. 5.

We are trying to gather more information from this feature by finding out the mean reserve_datetime_difference for each air_store_id per day. Along with that, we are finding the total reserve visitors and mean reserve visitors for each air_store_id per day.

The restaurants observe higher numbers of visitors on Saturdays and Sundays. On Monday and Tuesday there are the least number of visitors. In case of the month data, we can observe visitors hike in the month of December, as it is a festive month,The month of August and November has least average visitors. Therefore, we extract the day of week, month and year information from the given visit_date column. Since the day of the week and month name are categorical values and hence are label encoded. One more operation that is needed to be done is, creating a new column 'id' which is same as air_store_id merged with visit_date as given in sample submission.

As we have seen previously in data visualization, the difference between the mean number of visitors of the holiday and non-holiday is not high since weekends are not considered as holidays. However the mean number of visitors on a holiday is comparable to that on the weekend.This is why we are adding a new feature non_working that takes into account all the holidays as well as the weekends. The new feature non_working store 1 for every holiday and weekend. The rest of the days are marked as 0.

Three derived columns are added from the latitude and longitude data. Two of them are the relative distance from the maximum latitude and longitude and the other column is summation of latitude and longitude. This helps to give proper mapping for each of the restaurant with respect to location.

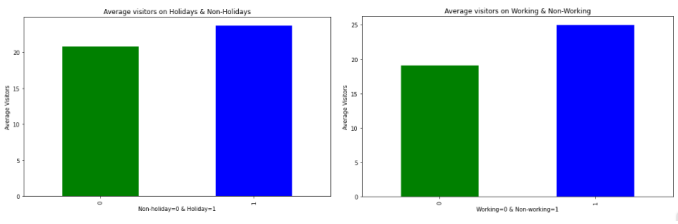

Fig. 6.

Now we can see that the mean visitors in a working day is quite low as compared to that of a non-working day. This makes it easier to predict future visitors for a given visit_date.

The genre name and area name of a restaurant are categorical values, therefore they are label encoded before feeding into the training model. The air_store_id information is also categorical data, it is required to be label encoded in the train and the test dataset.

V. Traning and Results

Regression analysis is used when we want to predict a continuous dependent variable from a number of independent variables. In this problem, we have to predict the number of visitors corresponding to a given restaurant on a particular day. Since, the number of visitors is a continuous variable, we are treating this problem as a regression problem.

To find out the proper hyperparameters required to train the model using tree-based models, we are using GridSearchCV that determines the optimal values for a given model.

Root Mean Squared Logarithmic Error (RMSLE) is the evaluation metric used here. It is a common metric for regression problems when predictions have large deviations.

TABLE I
RMSLE on various algorithms

S. no.	Algorithm	RMSLE
1	Linear Regressor	0.77995
2	Decision Tree Regressor	0.76676
3	Gradient Boosting Regressor	0.55771
4	XGBoost Regressor	0.50872
5	Ensemble Model of Gradient Boosting and XGBoost	0.52983

VI. Improvement of Previous Model

A. Feature Engineering on Latitude and Longitude

Since the latitude and longitude had not been considered in the feature selection previously, three derived columns are added from the latitude and longitude data. Two of them are the relative distance from the maximum latitude and longitude and the other column is summation of latitude and longitude. This helps to give proper mapping for each of the restaurant with respect to location.

B. More Training Models

Building a model using a decision tree can tend to overfit. Hence to reduce the variance of the model we are using bagging technique like random forest. Random Forest which is an extension over bagging uses random row sampling and random feature sampling which improves the performance of individual weak learners(decision tree). The accuracy we got using Random forest is 0.72445.

Since our training set has more than 2,39,000 entries, XGBoost takes a long time to train the model. Therefore, we selected LightGBM which is much faster than XGBoost regressor.

Previously, we were using equally weighted ensemble model. To find the proper weight in which two different models should be aggregated, we are checking for all possible combinations to find out the ensemble model to give minimum RMSLE value.

TABLE II
RMSLE on more algorithms

S. no.	Algorithm	RMSLE
1	Random Forest Regressor	0.72445
2	LightGBM Regressor	0.50286
3	Ensemble Model of XGBoost and LightGBM	0.50206
4	31.7% XGBoost and 68.3% LGBM	0.50148

VII. Conclusion

With an RSMLE value of 0.50148 the ensemble model with 31.7% XGBoost Regressor and 68.3% LightGBM Regressor is quite able to predict number of visitors on future dates. The ensemble model helps to avoid over-fitting and generalize our model for better prediction.

Acknowledgment

We would like to thank our teaching assistant Arjun Verma for keeping discussion sessions which helped us understand the problem in the initial stages and, on numerous other

occasions. Special thanks to Raghavan sir and Neelam ma'am for encouraging us throughout the course curriculum. Their highly detailed lectures on various topics helped us understand what we were actually doing in every step of this project.

REFERENCES

[1] Takashi Tanizaki, Tomohiro Hoshino, Takeshi Shimmura, Takeshi Takenaka, Demand forecasting in restaurants using machine learning and statistical analysis, Procedia CIRP, Volume 79, 2019, Pages 679-683, ISSN 2212-8271, https://doi.org/10.1016/j.procir.2019.02.042.

[2] RONCHETTI, Elvezio. "Regression and Time Series Model Selection." Journal of the American Statistical Association, vol. 95, no. 451, 2000, p. 1008. Gale Academic OneFile, . Accessed 29 Nov. 2020.

[3] Paper D. (2020) Scikit-Learn Classifier Tuning from Simple Training Sets. In: Hands-on Scikit-Learn for Machine Learning Applications. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-5373-1_5

[4] Duffy, N., Helmbold, D. Boosting Methods for Regression. Machine Learning 47, 153–200 (2002). https://doi.org/10.1023/A:1013685603443

[5] B. M. Pavlyshenko, "Linear, machine learning and probabilistic approaches for time series analysis," 2016 IEEE First International Conference on Data Stream Mining Processing (DSMP), Lviv, 2016, pp. 377-381, doi: 10.1109/DSMP.2016.7583582.

[6] Mariana Oliveira, Luis Torgo ; Proceedings of the Sixth Asian Conference on Machine Learning, PMLR 39:360-370, 2015.

VIII. PROJECT FILE LINK

https://drive.google.com/drive/folders/1mt7o-GZMWw0Vq5H349ET1oN8rCBz3cvz?usp=sharing

The link contains the well-commented jupyter notebook with EDA, Feature Engineering and Selection and Model Building along with the submission.csv file.