Restaurant Visitor Forecasting: Project Report

Neetha Reddy
IMT2018050
neetha.reddy@iiitb.org

Nikitha Adivi
IMT2018051
nikitha.adivi@iiitb.org

Rutvi Padhy
IMT2018519
rutvi.padhy@iiitb.org

***Abstract*—This is a detailed report on our work on predicting the number of visitors visiting certain restaurants in Japan using big data and supervised learning. The big data we are provided with contains restaurant information, historical visits and historical reservations. Using the features given in the model, we build a model which accurately predicts the number of visitors a restaurant receives on a particular day. We also attempt to evaluate the effectiveness of our models.**[1]

***Index Terms*—Feature Engineering, Machine learning, XGBoost, Light GBM, Grid Search, Walk Forward Cross Validation, Time Series Split**

DATASET

We are given data collected from two restaurant booking websites: Hot Pepper Gourmet(hpg) and AirREGI(air). The training data covers the dates from 2016 until early (first week) April 2017. We need to predict the number of visitors per given restaurant for mid weeks (second and third weeks) of April 2017.

We are given the historical data regarding air reservations, hpg reservation, air visits, mappings for some restaurants having both hpg and air IDs, and holidays. Data is skewed so appropriate normalisations have been made wherever required.

Dataset Description:

- **train.csv**: air_store_id, visit_date, visitors
- **air_reserve.csv**: air_store_id,visit_datetime, reserve_datetime, reserve_visitors
- **hpg_reserve.csv**: hpg_store_id,visit_datetime, reserve_datetime, reserve_visitors
- **air_store_info.csv**:air_store_id,air_genre_name, air_area_name, latitude, longitude.
- **date_info.csv**: calendar_date, day_of_week, holiday_flg
- **hpg_store_info.csv**:hpg_store_id,hpg_genre_name, hpg_area_name, latitude, longitude.
- **store_id_relation.csv**: air_store_id, hpg_store_id
- **test.csv**: air_store_id, visitors

I. INTRODUCTION

There has been a tremendous increase in the number of people eating out at restaurants. To keep up with this demand, more and more new restaurants are being opened every year. The steady growth in the number of new restaurants and fast-food chains meant that local restaurants are struggling to keep up. To maximise their profits, the restaurants need to know how many customers to expect on a particular day to effectively purchase ingredients and schedule staff members. Based on demand, they can also choose to hire/fire seasonal workers. But these predictions are not easy to make.

These predictions depend on many unpredictable factors such as weather conditions and local competition. They are even harder to make when there is little or no historical data for a restaurant (this is usually the case for new restaurants) [1]. Even within local restaurants, there are several factors (popularity of the food cuisine, number of restaurants serving a particular cuisine, their locations, etc to name a few) which might influence the predictions. It is safe to assume that holidays also might affect our predictions. Apart from these factors, there might be several other factors which might have a direct or indirect impact on our predictions.

Taking all these points (and more) into account, we visualise which of these direct or indirect features impact our predictions and make note of them. We then try to train a model using these select features to accurately predicts the number of visitors per restaurant per day.

II. DATA VISUALISATION

Exploratory Data Analysis (EDA) is an approach for data analysis that employs a variety of graphical techniques to meet the following goals :

- Maximize insight into a data set
- Extract important variables
- Detect outliers and anomalies. Also, check if there is missing data.

EDA helps us develop simple models with great explanatory predictive power. The EDA about the given Restaurant Visitor Forecasting problem was done in steps, first, a thorough analysis was done for within a given data set followed by analysis across data sets.

From Fig. 1., we see that Cafe/Sweets and Japanese food have the highest median number of visitors.

[1] This is an InClass Competition for AI511 course 2020 at IIIT Bangalore. Link to competition- https://www.kaggle.com/c/restaurant-visitor-forecasting/notebooks?competitionId=22533sortBy=dateRuntab=collaboration

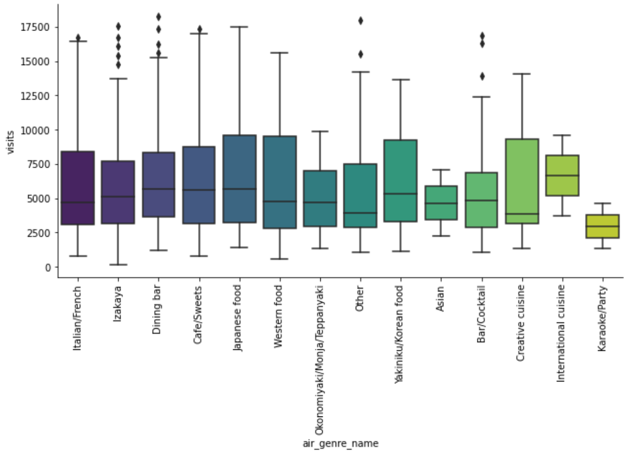

Fig. 1. Number of air visits per air genre

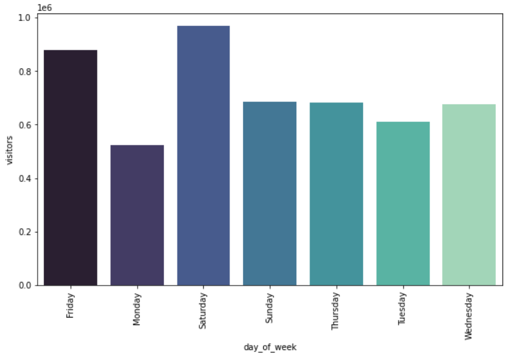

Fig. 2. Number of air visits per day of week

From Fig. 2., we can infer that on fridays and saturdays, restaurants have the most number of visitors. Sunday being a weekend doesn't have more number of visitors than most weekdays. Monday has least number of visitors. Other weekdays perform more or less the same.

From Fig. 3., we can infer that people generally visit restaurants more on holidays. Weekend being a holiday does not impact the number of visitors much. Immediately after a weekend, if there is a holiday on a weekday, people tend to visit the restaurant more.

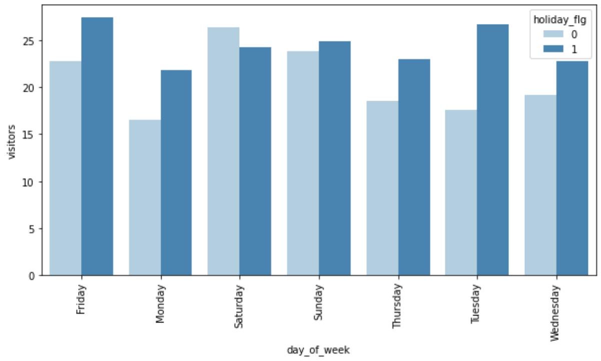

Fig. 3. Number of air visits based day of week and holiday flag

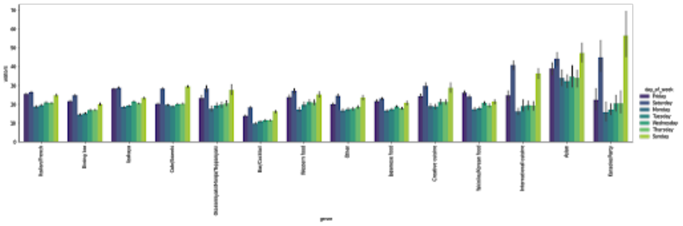

Fig. 4. Number of air visits per genre per day of the week

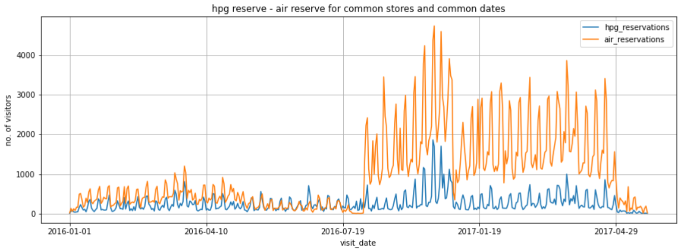

Fig. 5. Hpg reservations-air reservations for common stores

From Fig. 4., we can infer that for any given day of the week, Bar/Cocktail is the least preferred cuisine while Asian is the most preferred cuisine. Friday, Saturday and Sunday are the most preferred days for all the cuisines. All cuisines have similar number of visitors for the weekdays. Specific inferences can also be made. For example, Karaoke/Party is most preferred on weekends and Asian on weekdays. There is a lot of weekday-weekend variance for cuisines like International Cuisine and Karaoke/Party.

From Fig. 5., we can infer that for stores that are present in both hpg and air reserve data, Jan 2016 to July 2016, both the data almost overlaps. But from Aug 2016, hpg reservations are actually more. So taking hpg reservations will give us more data for the same. On further visualization, we see that all these 150 common stores serve Japanese cuisine.

III. Data Pre-Processing

A. Feature Selection

`Initial Phase:`

After a thorough Exploratory Data Analysis, the following conclusions regarding feature selection were drawn:

The features *latitude* and *longitude* were discarded due to their high correlation with the air_area_name. Features from $hpg\_store\_info.csv$ were discarded because the stores that were common to both air and hpg were found to be of Japanese genre only. The features selected are:

- holiday_flg
- visit_date
- air_area_name
- air_genre_name
- day_of_week

`Final Phase:`

Apart from the above mentioned features, we also:

- ***Deleted part of 2016 data***: From EDA, we can observe that no. of visitors from Jan 1 2016 to July 1 2016 didn't correspond to the no. of visitors in 2017 for the same date. Deleting this data improved our accuracy to 0.51985 using a basic XGB model.
- ***Included reservation data***: Initially, we neglected air reservation data because there was information corresponding to only 330 stores while there were 829 unique stores in the train data. We neglected hpg reservation data since there was no corresponding visitors data provided for the hpg stores. Later on, we mapped the hpg stores given in hpg_reserve.csv to their corresponding air_store_id using store_relation_info.csv for those 150 common restaurants. We then pre-processed the visit_datetime and reserve_date present in both air_reserve and hpg_reserve. We added an additional feature called time_diff which is the number of days between the reservation date and the actual date of the visit. We then merged both air_reserve data and hpg_reserve data with the other selected features on air_store_id and visit_date. Adding reserve_visitors and time_diff to our model decreased the error by a significant amount.

B. Additional Features

`Initial and Final Phases:`

The features added for better performance include the following :

- ***Next day holiday*** :It was observed in the EDA that days preceding a holiday have higher number of visitors.
- ***Size*** : As it captures information pertaining to number of stores per genre per area. This takes into account local competition, if any.
- ***Month, year***: $visit\_date$ was split into month and year. $date$ was ignored as it didn't matter.
- ***Minimum visitors, mean visitors, median visitors and maximum visitors, count***: This drastically improved the accuracy as this provided a range for the no. of visitors. $count$ helps average the sum of the no. of visitors [4].

C. Feature Encoding

- month : label-encoded on the basis of number of visitors per month
- air_area_name : label-encoded on the basis of number of visitors per area because there was a visible order and it was very dependent on the no. of visitors.
- air_genre_name : label-encoded on the basis of number of visitors per genre.
- Day of the week : one-hot encoded to get the number of visits per genre per day of week.

We mostly chose label-encoding because, from the EDA, we could see an inherent order.

IV. Model selection and Hyperparameter Tuning

KNN

- On tuning(see Fig. 2 for code), we got n_estimators= 170 for this data. For this model, we did not consider the features from reservation data (neither air_reserve data nor hpg_reserve data). It did not perform as well as expected.
- RMSLE score: 0.53015

XGBRgressor

- This tree based boosting ensemble, though slow, proved to be very helpful in drastically making the RMSLE better.
- The following are the tuned set of hyperparameters(see Fig. 4 for code): max_depth=8, gamma=0, min_child_weight=1, subsample=1, colsample_bytree=1
- n_estimators = 500, learning_rate = 0.01 were set according to intuition.
- This model with the above tuned hyperparameters gave a RMSLE score of 0.50740.

LGBMRegressor

- Light GBM was the best -performing model. Hyper parameters were tuned using hyperopt library.
- The following are the values for the tuned set of hyperparameters(see Fig. 6 for code): 'sub_feature': 0.5417895574997428, 'num_leaves': 105, 'min_data': 42
- Other hyperparameters used: 'metric': 'rmse', 'min_hessian': 1, 'n_estimators': 500, 'learning_rate': 0.02, 'boosting_type': 'gbdt', 'objective': 'regression'
- RMSLE score for this model with the above hyperparameters is : 0.50231

Ensemble of XGBRegressor and LGBMRegressor

- As both XGBRegressor and LGBMRegressor individually performed very well on the dataset, a weighted ensemble of both these models was tried to get a better RMSLE score.
- The weights were manually tuned and the best score was possible with 0.84 as the weight for LGBM and 0.16 for XGB.
- RMSLE score obtained was: 0.50286. This didn't perform as better as the LGBMRegressor alone.

V. Training and Results

From TABLE I, we observe that LGBRegressor gives the lowest RMSLE score in both public and private leaderboards.

TABLE I
Root Mean Squared Log Error(RMSLE) on various algorithms

Algorithm	Public Score	Private Score
KNN	0.53015	0.54696
XGBRegressor	0.50740	0.52330
LGBMRegressor	**0.50231**	**0.51865**
Ensemble(xgb+lgbm)	0.50286	0.51916

Conclusion

This model predicts the number of visitors that a restaurant can expect on a given day to a reasonable extent. Due to lack of reservation data for many stores, we did not include these reservation features while training our model. When we later on included these features for the limited restaurants, there was a significant decrease in the error.

References

[1] Predicting Future Visitors Of Restaurants Using Big Data July 2018 DOI: 10.1109/ICMLC.2018.8526963 Conference: 2018 International Conference on Machine Learning and Cybernetics (ICMLC) Authors: XU MA,YANSHAN TIAN

[2] https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

[3] https://medium.com/eatpredlove/time-series-cross-validation-a-walk-forward-approach-in-python-8534dd1db51a

[4] https://www.kaggle.com/tunguz/surprise-me-2