Data Science Project Protocol

Recruit Restaurant Visitor Forecasting

March, 2019

Authors: ***Kate Brodiansky and Gali Brill***

1. Introduction

The aim of this challenge (outcome) is to predict the future numbers of restaurant visitors, in every specific day in each of the restaurants in the period between April 23 to the end of May, including the golden week. This makes it a Time Series Forecasting problem. The data was collected from Japanese restaurants. We have chosen this project because the data set is small and easily accessible.

The project is very important and interesting for us because it is a real data set and a real business problem.

If all of the restaurants know the number of visitors, this will help minimize costs and risks.

Restaurants need to know how many customers to expect each day to effectively purchase ingredients and schedule staff members. For online resources like “Hot Pepper Gourmet (hpg)” and “AirREGI (air)”, which make reservations and payment via the Internet, this will help determine the volume of traffic and the volume of users of the resource.

For new business, our forecast can help in understanding the market. We think that this is an amazing opportunity to use data analysis to enable a particular business to be more efficient.

Figure 1: Tokyo's Central Streets

Choosing this date set, we have an amazing opportunity to get acquainted with Japan. Traditions, holidays and food culture of this country. This is a nice opportunity to learn more about the country.

2. Data

2.1. Main source of our data is the base data set from the “Recruit restaurant visitor forecasting’’ project from the website Kaggle: https: //www.kaggle.com/ c/recruit-restaurant-visitor-forecasting/data .

Data is presented in the format of eight related tables. Which are exported from two internet separate sources:

- 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.

We use information about: booking, visits, genre, location and other information from these sites to predict the total number of visitors to a restaurant on a specific date.

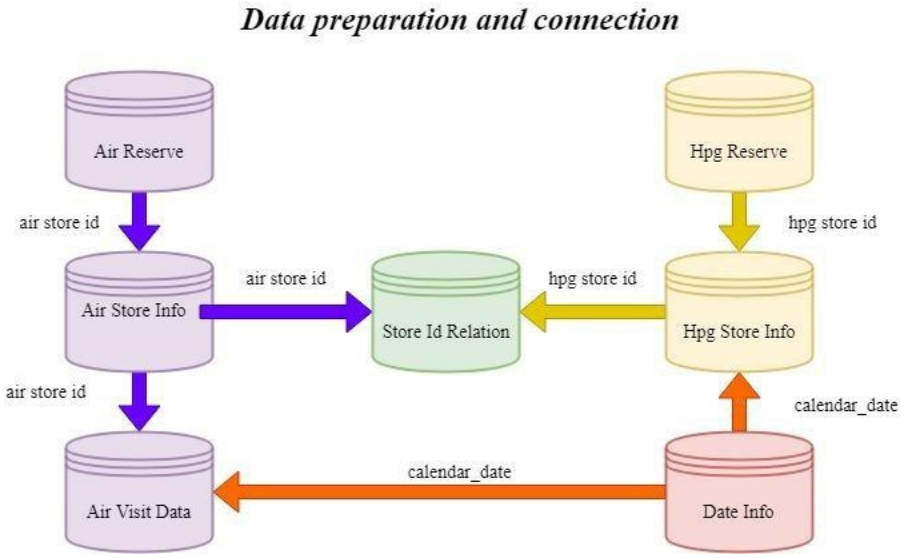

Figure 2: Data connection

So many different factors to effects the number of visitors to the restaurant: day of the week, season, geographic position, weather, culturally sensitive, competitors density. Part of the data we get in the source database another part of the data from other resources on the internet.

2.2. Geographical location of data

The first factor that affects our choice of restaurant is its location.

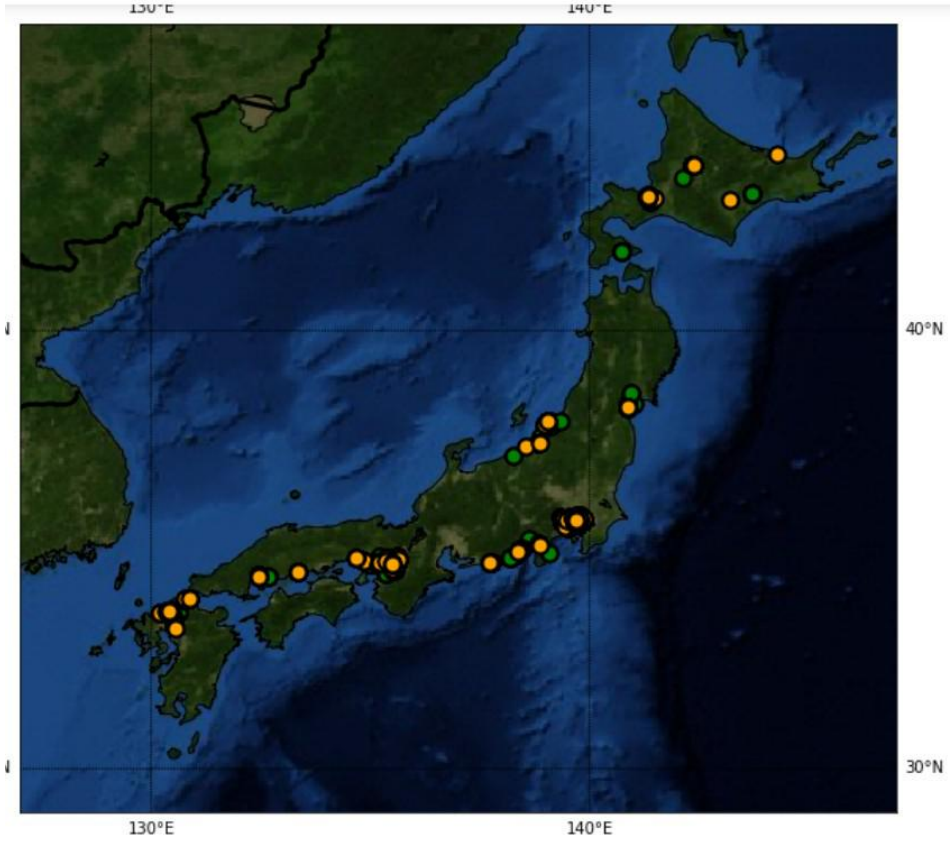

Figure 3: Map of restaurant distribution.

How we can see on the map (Figure 3) restaurants from our data set are located in all regions of Japan. The map shows us there are more popular regions in which there are more restaurants and visitors. And there are not-popular regions in which no restaurants. This is mainly due to the population density in Japan. Using cluster analysis we divide our restaurants into nine classes by geographical location (longitude, latitude). We compare the cluster number with the name of the prefecture in Japan. For verification, we calculate what percentage of restaurants are in a particular region.[1]

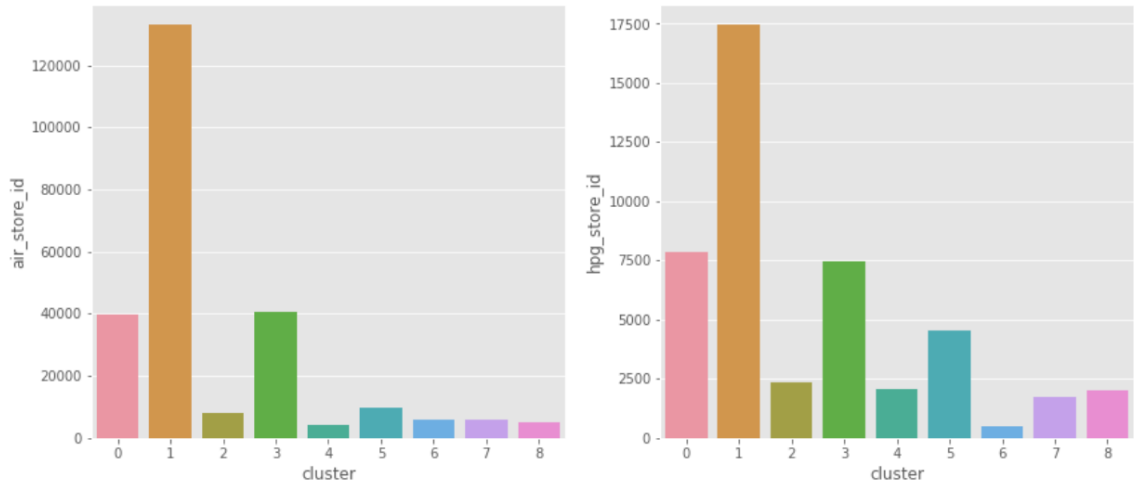

Figure 4: Distribution by clusters

We relate the identified groups with the biggest cities:

- Cluster 0 refers to Fukuoka-ken
- Cluster 1 refers to Tokyo- to
- Cluster 2 refers to Hokkaido
- Cluster 3 refers to Hyogo- ken
- Cluster 4 refers to Niigata- ken
- Cluster 5 refers to Hiroshima- ken
- Cluster 6 refers to Miyagi- ken
- Cluster 7 refers to Shizuoka- ken
- Cluster 8 refers to Osaka- fu

2.3. Time-frame of data

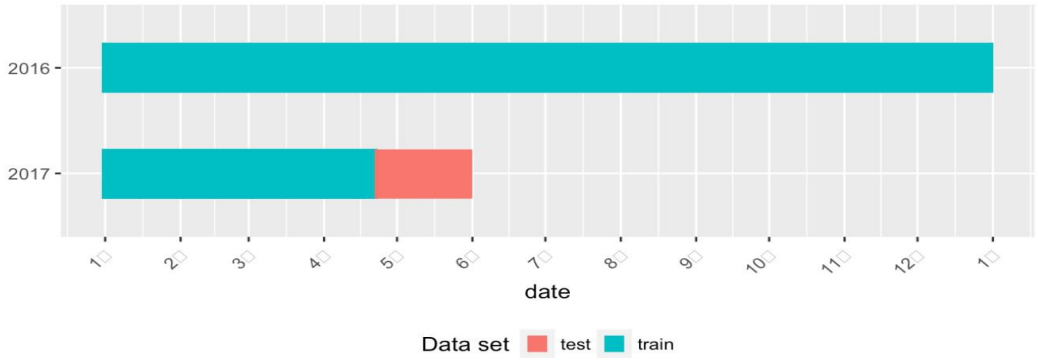

Figure 5: Time-frame of data.

The training data is based on the time range of Jan 2016 -most of Apr 2017, while the test data-set includes the last week of Apr plus May 2017.

The test data “intentionally spans a holiday week in Japan called the “Golden Week”. The data description further notes that:”There are days in the test data-set where the restaurant were closed and had no visitors. These are ignored in scoring. The training set omits days where the restaurants were closed". [2]

It is important for us in this time frame to know how the number of visitors to restaurants is distributed. We are interested in how the number of visitors during this period is distributed.

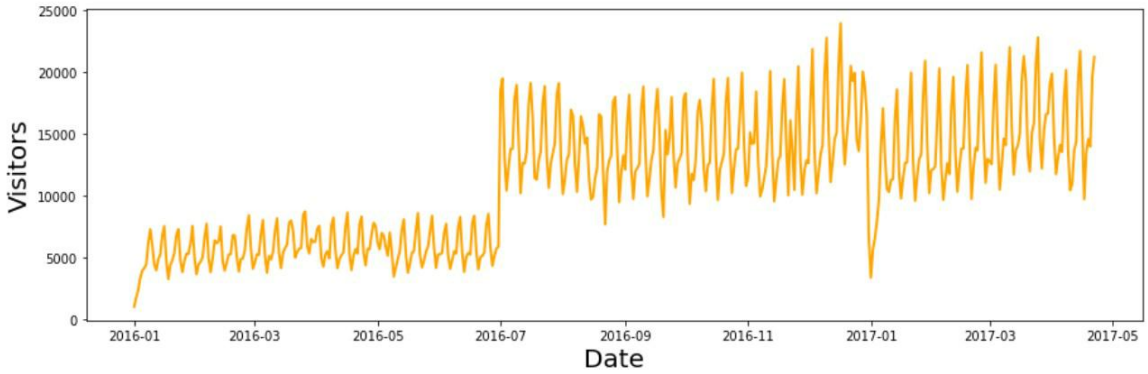

Figure 6: Time-frame of distributed visitors.

We group data on restaurant visitors by date and by Air ID. And we also create a graph of the average number of visitors.

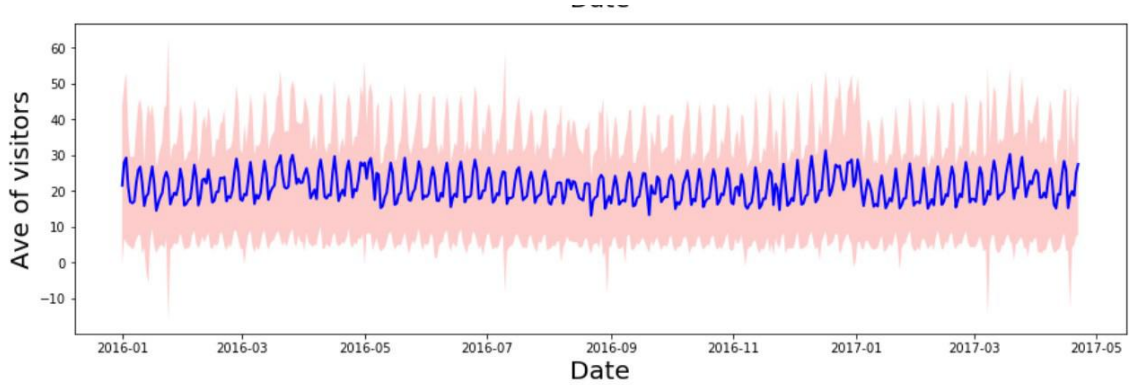

Figure 7: Time-frame of Ave visitors.

2.4. External data sources

To enrich our data we use other data sources that can help us make a more accurate forecast.

In this project, we use data on holidays in Japan and we add a new variable with holiday and weekend indices. Information about the holidays we download from the official website of tourism and recreation in Japan: https://www.officeholidays.com/countries/japan/index.php.
We also add weather data from the official meteorological site. Which indicates that the data is verified and approved by the Institute of Meteorology in Japan. We combine the data by date and cluster of the region.

2.5. Enrichment and influence of periodic variables

As we see in the graph(figure 6, figure 7) the variable changes depending on the period. We create new variables: day of the week, month, day of the year, day. And we also calculate variable indices for weekends and holidays.

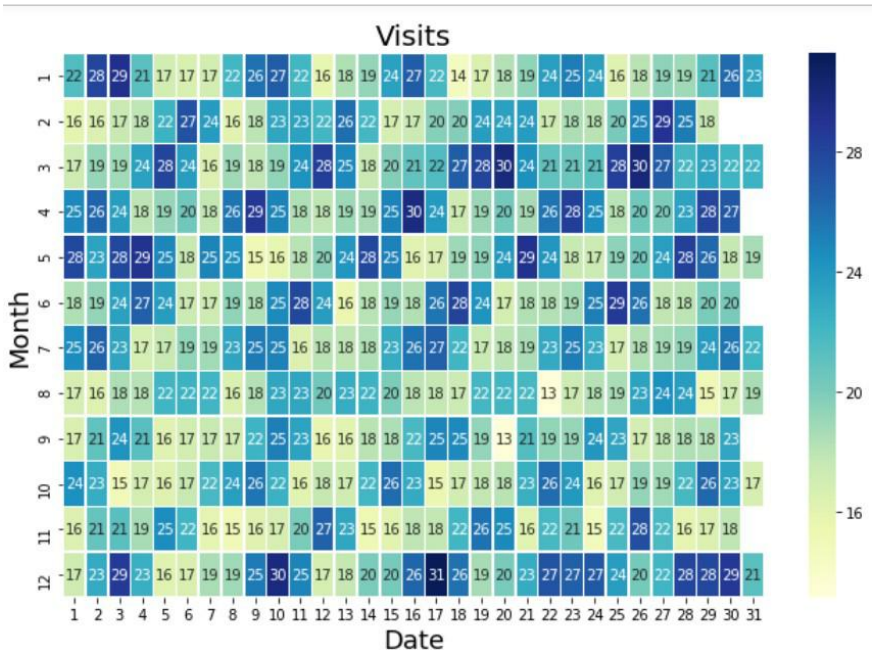

Figure 8: Heat map of the visits calendar.

And we also calculate variable indices for weekends and holidays.[3]

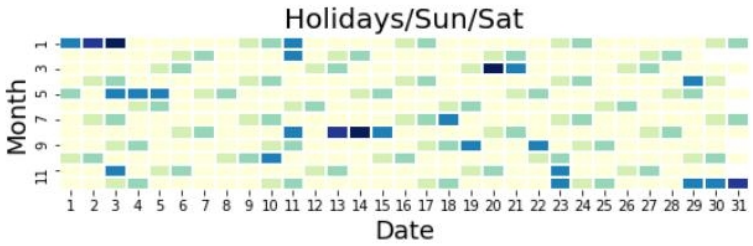

Figure 9: Heat map of the holidays and weekends.

2.6. Cleaning and preparing data

After merging all the tables and creating new variables, we need to check our final table for errors and extreme values. We are checking: missing values and outliers. And decide what to do with these values.

Missing Values Now that we have the correct column data types, we can start analysis by looking at the percentage of missing values in each column. Missing values are fine when we do Exploratory Data Analysis, but they will have to be filled in for machine learning methods. We wrote a function that shows us columns in which there are null values and their number in percentage terms.

	Missing Values	% of Total Values
reserve_visitors_hpg	238558	94.6
hpg_date_diff	238558	94.6
total_snowfall	228577	90.7
deepest_snowfall	227302	90.2
reserve_visitors_air	224044	88.9
air_date_diff	224044	88.9
hpg_store_id	206089	81.7

Figure 10: Missing Values.

We decide to remove three variables: ‘’Total snowfall’’, ‘’Deepest snowfall’’ because It's snowing in japan only in winter. This is less than ninety percent of our data and we also make a forecast for the spring.

We do not want this to have a negative impact on our model.
And ‘’Hpg store id’’ this variable was needed only for joining tables and it was possible to delete it earlier.

In the original data booking is much less than visits, this causes null values in columns: ‘’HPG date diff’’, ’’AIR date diff’’, ‘’reserve visitors air’’ and ‘’reserve visitors hpg’’. We do not know what provoked it but we decided to divide the data into two tables to analyze in the variables selection.

We also detecting outliers with Z-scores and chose the number of visitors that does not get into the interval from 25 percent to 75 percent.[4]

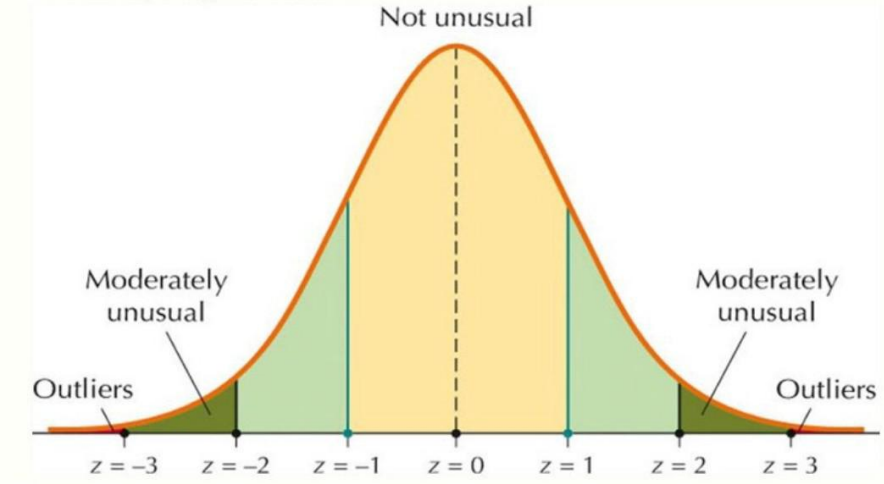

Figure 11: Outliers with Z-scores

And after analyzing them, they found out that a large number of visitors are connected with the popular region and the genre of restaurants and it is advisable to keep these values.

2.7. Variable selection

In this chapter we want to analyze all our variables and choose the most effective.

For this goals we choose three regression models: Lasso, Random forest and Gradient boosting model.

```
result = varSel[varSel['Sum'] == 0]
```

```
result
```

	Variable	Lasso	RandomForest	GradientBoost	Sum
7	sunday	0	0	0	0
25	Other	0	0	0	0
29	Bar/Cocktail	0	0	0	0
30	Creative cuisine	0	0	0	0
31	Western food	0	0	0	0
33	Asian	0	0	0	0
34	International cuisine	0	0	0	0
36	Karaoke/Party	0	0	0	0
37	Wednesday	0	0	0	0
41	Monday	0	0	0	0
42	Tuesday	0	0	0	0
43	Sunday	0	0	0	0
44	Tōkyō-to	0	0	0	0
45	Ōsaka-fu	0	0	0	0
46	Hyōgo-ken	0	0	0	0

Figure 12: Variable selection zero results.

We were surprised by the results. Most inefficient variables were derived from the conversion of class variables. We tried a different type of conversion. One hot encoding or Label encoding are both forms of feature engineering where a data scientist is trying to represent categorical information (country, ID, region, genre) as an input vector. But this unfortunately did not improve our result.[3]

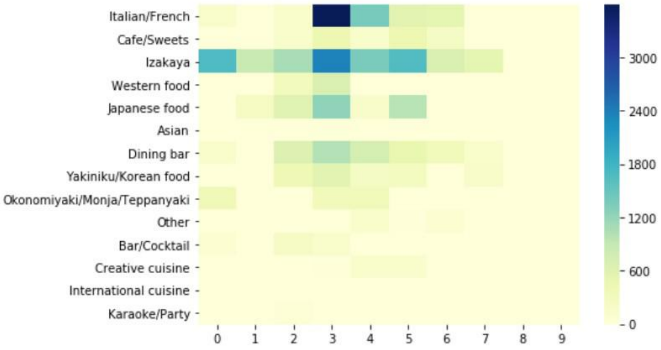

Figure 13: Heat map of the genre and location.

3. Models

3.1. Model selection

For best results, we try different models and their ensembles. Our final decision will give us the best result and the smallest error is not standard. We know that usually the Gradient boosting model is combined with other models, but in this case, the best solution was the following:

Generalized Boosted Models (GBM) This model uses gradient boosting which builds an additive decision-tree model in order to predict the outcome in a regression. The model greedily adds base learners from a select hypothesis class, and attempts to find a weighted combination of them that minimizes the training error. Though the performance of single base learners is generally poor, they can be combined to form a very strong learner – this process is known as boosting. GBM is a flexible model that can be adapted to a wide range of distributions, including the Poisson distribution, which again turns out to give the best generalization results out of the different distributions. [3]

The K-Nearest-Neighbors (KNN) method of classification is one of the simplest methods in machine learning, and is a great way to introduce yourself to machine learning and classification in general. At its most basic level, it is essentially classification by finding the most similar data points in the training data, and making an educated guess based on their classifications. Although very simple to understand and implement, this method has seen wide application in many domains, such as in recommendation systems, semantic searching, and anomaly detection. [1]

Extreme Gradient Boosting (XGBoost) is one of the most popular machine learning algorithm these days. Regardless of the type of prediction task at hand; regression or classification. XGBoost is well known to provide better solutions than other machine learning algorithms. In fact, since its inception, it has become the "state-of-the-art” machine learning algorithm to deal with structured data.[4]

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the modeof the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of over-fitting to their training set.

Linear Regression Model (LM) in statistics, linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

3.2. Selection of data separation

We decided to separate our test date set into two parts: Train and Dev. To maintain balance and compliance, we performed the division by date. Training data is based on the time interval from January 2016 to March 2017, and the Dev data set includes March and two weeks in April 2017.

We know that most data is balanced. Because these are teethe restaurants that are in the same places and have the same attributes. But we also know from the first chapter(EDA) that the number of visitors (outcome) changes depending on the season and weather conditions, which also depend on the season.

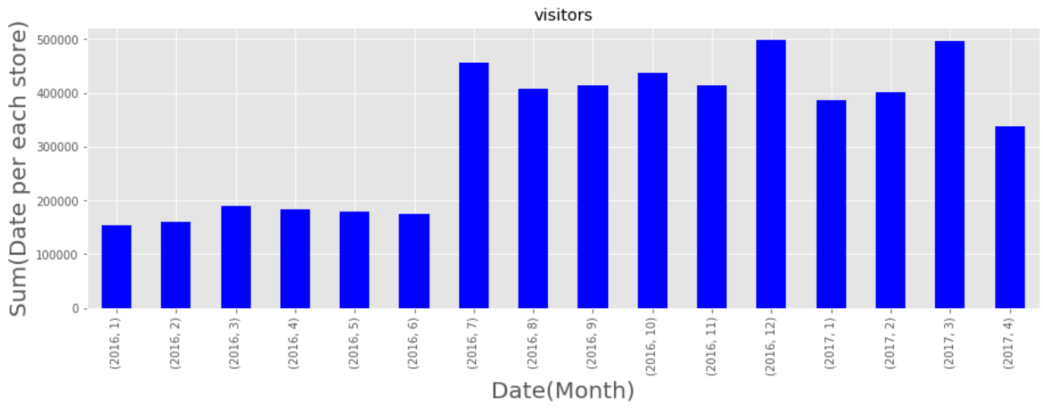

Figure 14: Number of visitors by months.

We see that the number of visitors in March and April, for example, is very different. This is the meaning of our task.

3.3. Metrics selection

Our target optimization metric is the Root Mean Squared Logarithmic Error (RMSLE), which is computed as follows:

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

Where “n” is the number of observations, pi is our predicted count, and “ai” is the actual count. We seek to identify the models that result in predictions which minimize this error.

Our results:

Model	Train	Dev
GradientBoosting	0.6044	0.5999
KNN	0.6657	0.6653
XGBoost	0.5748	0.5703
RandomForest	0.7764	0.7745
ExtraTrees	0.7641	0.7623
Linear	0.7586	0.7581

Figure 15: Errors table.

We also improve our results by fitting parameters**.**

3.4. Final decision

We used an ensemble of models: 30% Gradient boosting, 30% Kneighbors and XGBoosting. We chose this solution because using our best regression model “XGBoosting” gave us the worst result. The percent of value in our final decision conforms to the error value.[5]

4. Golden Week

The Golden Week is a collection of four national holidays within seven days. In combination with well placed weekends, the Golden Week becomes one of Japan's three busiest holiday seasons, besides New Year and the Obon week. Trains, airports, restaurants and sightseeing spots get very crowded during.

Figure 14: Golden Week.

The test set intentionally spans a holiday week in Japan called the "Golden Week." There are days in the test set where the restaurant were closed and had no visitors. These are ignored in scoring. The training set omits days where the restaurants were closed. To solve this problem we used “Holiday trick” based on decision: https://www.kaggle.com/h4211819/holiday-trick.

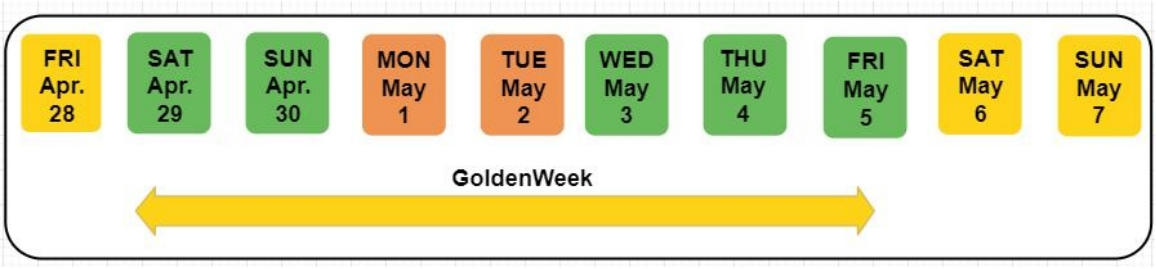

Figure 15: Golden Week calendar.

The first of may restaurants are closed this is public holidays. The following days are holidays too. But restaurants are already open to visitors. And we expect a large number of visitors. Which we calculate by the formula:

$$\text{Visitor}(\text{May } 3,4,5) = \text{Sqrt}(\text{Visitor}(\text{Apr } 29) * \text{Visitor}(\text{May } 13))$$

$$\text{Visitor}(\text{May } 2) = \text{Sqrt}(\text{Visitor}(\text{Apr } 28) * \text{Visitor}(\text{May } 12))$$

Figure 16: Holiday trick.

5. Conclusion

Using different combinations of models and combinations of variables, we found that latitude, longitude, date, reserve visitors and day of the week is by far the most predictive feature for our problem, whereas weather variables play a much smaller, but still noticeable effect in contributing to accurate predictions.[5]

For categorical variables we used one hot encoding. We also combined the union and grouping of variables. The summation of variables: longitude + latitude also gave good results.

Averaged predictions from three models: GradientBoostingRegressor * 0.3, KNeighborsRegressor * 0.3 and XGBRegressor * 0.4 in the ensemble were able to best fix the relationships in the data, which led to the best model performance. And the smallest error. [6]

References

[1] Belur V. Dasarathy, ed. (1991). Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. ISBN 978-0-8186-8930-7.

[2] Pedregosa. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2011.

[3] Prasad P., (2018)What is Exploratory Data Analysis?

[4] Friedman, J. H. (February 1999). "Greedy Function Approximation: A Gradient Boosting Machine".

[5] Abraham, B. and Ledolter, J. (1983). *Statistical Methods for Forecasting*, Wiley, New York, NY.

[6] Chen, Tianqi; Guestrin, Carlos (2016). "XGBoost: A Scalable Tree Boosting System". In Krishnapuram, Balaji; Shah, Mohak; Smola, Alexander J.; Aggarwal, Charu C.; Shen, Dou; Rastogi, Rajeev. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. pp. 785–794.