Recruit Restaurant Visitor Forecasting

Initial Findings

18th February, 2018

Foodie Analytics

Shipra Sethi
Dong Bing
Daniel Colvin
Subba Muthurangan
Erik Platt

Foodie Analytics
401 N. Michigan Ave.
Chicago, IL 60611

February 18, 2018

Dr. Donald Wedding, CEO
Recruit Holdings Co., Ltd.
Chuo-ku, Tokyo 104-0061 Japan

Dear Dr. Wedding,

Foodie Analytics has completed an in-depth statistical analysis of the data provided by your organization, including data cleansing and data transformations, gaining insights from historical business trends, and developing our initial predictive models. We have created this Initial Findings Report on the Recruit Restaurant Visitor Forecasting project and are pleased to share our progress and results with you.

In this report, we have provided a detailed section on exploratory data analysis (EDA), preliminary conclusions from our initial models, and sample data visualizations to see the results in action. As previously communicated, along with a web user interface, we also plan to deliver a mobile application for users on the go. We have provided initial screenshots of the mobile application showing the outcomes from our predictive model. Whether it is the web interface or the mobile application, users will be able to see the total number of visitors to the restaurant on a given future date. There is a separate section listing any assumptions and limitations associated with this project. Finally, a conclusions section is provided to highlight the main accomplishments of this engagement at the half-way mark.

We believe that you will find value in sharing this report with senior members of your organization. To ensure that this deliverable meets your expectations, we would be happy to review it with you. If you need any additional information or supplemental material, please do not hesitate to contact any team member.

Sincerely,
Foodie Analytics

Table of Contents

1. Problem Statement

Restaurant start-ups face a notoriously high failure rate. While the often advertised metric that 90% of restaurants fail within the first year is exaggerated, it is still far from an easy industry to find success. There are many factors that must coordinate to drive customers in; an appetizing menu simply isn't enough to guarantee traffic. However, increasing utilization of online reservation systems has offered an opportunity to gain a competitive advantage through customer data.

Recruit Holdings' partnership with Foodie Analytics will leverage this customer data. Using existing datasets within Hot Pepper Gourmet and AirREGI platforms, Foodie Analytics will build detailed analytics and predictive models to inform customer traffic patterns. Understanding when and where customers plan to dine allows the opportunity for efficient staffing, fresh ingredients, and an idealized experience.

2. Data Overview

2.1 Overview

The primary datasets that will be used for data insight generation and modeling purposes come from 2 separate data sources:

- Hot Pepper Gourmet (HPG): Similar to Yelp, users can search restaurants and make a reservation online. This data source contains 2 datasets.
- AirREGI / Restaurant Board (Air): Similar to Square, a reservation control and cash register system. This data source contains 3 datasets.

In addition to above datasets, 2 more datasets are provided, one identifying common restaurants between HPG and Air data sources and another indicating when Japanese holidays occur. We also used Japan Meteorological Agency (JMA) weather historical data. This dataset contains several weather factors including Precipitation, Average Temperature etc. A summary of the data provided can be seen in Table 1 and a description of each variable is summarized in Table 2:

Category	**Description**
Number of Sources	2
Total Number of Datasets	8
Number of Unique Records	963,705
Number of Fields	10
Number of Observations with Missing Values	57
Location of Restaurants	Japan
Time Frame of Observations	1/1/2016 - 4/23/2017

Table 1: Summary of Data

Variable	Description
air_store_id	Unique identifier for a restaurant from the “AIR” data source
hpg_store_id	Unique identifier for a restaurant from the "HPG" data source
visit_datetime	Date which the customer visited the restaurant
visitors	Number of visitors who visited the restaurant
reserve_datetime	Date which the customer made a reservation for the restaurant
reserve_visitors	Number of visitors for a reservation
genre_name	Restaurant food type
area_name	City where restaurant is located
latitude	Geographical latitude
longitude	Geographical longitude
holiday_flg	Indicator of days which were a local holiday

Table 2: Data Dictionary

2.2 EDA Summary

The steps involved in exploratory data analysis (EDA) are summarized below:

- Analyze the components of each dataset to better understand the business goals
- Merge the datasets into one, coherent object
- Identify missing values, potential outliers, and/or missing information
- Aggregate the data from hourly to daily observations
- Explore ways to splice the data to extract the most valuable information

2.3 Tools Used for EDA

R and **Tableau** were used to meet our EDA goals. R is a valuable tool for statistical analysis, model creation, and basic graphics capabilities. Tableau provides more advanced graphics for further insight into the data and a more aesthetical display for the end user.

2.4 Response Variable

The response variable is “visitors”. It is an indicator of how many people visited a restaurant at any given day. This information will be used to predict how many people will visit at a future date.

2.5 Predictor Variables

There are 18 predictor variables used for initial data modeling, as described in Table 3 below.

Predictor Variable	Description
air_res_visitors	Date which the customer made a reservation for the restaurant
air_mean_time_ahead	Number of visitors for a reservation
air_genre_name	Restaurant food type
air_area_name	City where restaurant is located
latitude	Geographical latitude
longitude	Geographical longitude
holiday_flg	Indicator of days which were a local holiday
day_of_week	Day of week
rank	Store rank based on number of visitors
min_visitors	Min visitors from visitors variable
mean_visitors	Mean visitors from visitors variable
median_visitors	Median visitors from visitors variable
max_visitors	Max visitors from visitors variable
count_visitors	Total count of visitors from visitors variable
month	Month of year
day	Day of month
precipitation	Rain fall on given day
avg_temperature	Avg temperature on given day

Table 3: Predictor variables with descriptions

3. Exploratory Data Analysis

3.1 Overview

Foodie Analytics has conducted extensive data review to ensure the quality and integrity of the data. Our process looks to ensure the data used for model development is complete, accurate, and reliable. This assessment includes determining the number of missing values, detection of any outliers and their potential impacts, as well as any anomalies within the historical data. The goal is to address and resolve any data quality issues that would have significant adverse impacts on model feasibility.

3.2 Merging Datasets

Even though the given data came from two different sources, HPG and Air, there were commonalities in the fields which made it possible to merge them together. Three of the datasets contained the “store_id”, “reserve_datetime”, “visit_datetime”, and “reserve_visitors” fields. Two more datasets contained the “store_id”, “genre_name”, “area_name”, “latitude” and “longitude” of many restaurants. One dataset contained information on when holidays occurred, and another dataset indicated which restaurants were in both data sources (but had different IDs). Finally, we used weather historical data that contained “precipitation” and “avg_temperature”.

Using “store_id”, we were able to combine these 7 datasets into one coherent object. For the weather data, we used the nearest location to merge with all observations. In the training dataset, we only kept the records which contained a value for the “visitors” variable (i.e. 213,510 records). The final object contained the following columns: “air_store_id”, “hpg_store_id”, “visit_datetime”, “visitors”, “reserve_datetime”, “reserve_visitors”, “genre_name”, “area_name”, “latitude”, “longitude”, “holiday_flg”, “precipitation” and “avg_temperature”.

3.3 Missing Values

Once we merge the HPG and Air data sources together, we perform missing value analysis on this single data source. The goal of this exercise is to understand the nature of the missing values and assess their impacts on model performance. We want to make sure missing values do not compromise the integrity of the data or introduce significant bias in prediction results. Table 3 below shows the data fields with the number of missing values in the combined dataset.

Variable	**Missing value count**
hpg_store_id	51,232
reserve_datetime	19
reserve_visitors	19

Table 4: Missing Values

The field with the most missing values is “hpg_store_id”. Unlike the Air data source, the HPG data source only has reservation information and not the actual visitor information. Since our response variable is the number of visitors and not reserve visitors, most of the records in HPG data source can not be tied back to the Air data source by store ID when we combined the data sources. In other words, records that has missing “hpg_store_id” will not be considered during the model training phase due to its lack of actual visitor information.

We also see a small number of missing values in “reserve_datetime” and “reserve_vistors”. Since the count is less than 0.005% of the total sample size, removing these values is the best course of action.

3.4 Outliers

In this phase we examine extreme values that deviate from other observations and could have undue influence on model performance. We took the univariate approach by looking at the distribution of the target variable.

The histogram in Figure 1 below shows there are a number of potential outliers in our response variable, as extreme values peak at a visitor count of 239. The top five highest values are 189, 199, 205, 216, 239. While these values are questionable and worth further investigation, it is plausible to expect an extreme number of visitors in certain instances such as company outings in holiday seasons. For this reason, we will not remove these outliers from the dataset for the initial model. The density distribution shows that 90 percent of the samples consist of party sizes of 50 or less, indicating that there are low occurrences of extreme values.

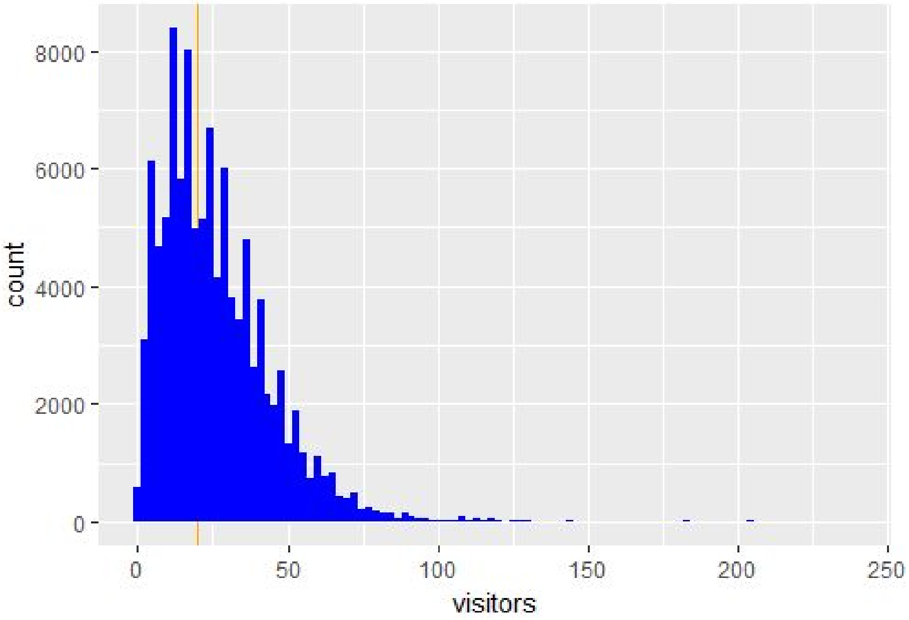


Figure 1: Histogram of response variable

3.5 Data Trends and Business Insights

Foodie Analytics looks for important indicators in historical trends that can potentially improve model performance. Figure 2 below shows the historical trend of the number of reservations in the data source. As the trend line indicates, the number of reserved visitors is clearly influenced by seasonality. The fluctuations are in an orderly fashion, based on day of week. We also see a large increase in reservations leading up to Jan 2017, potentially related to the seasonal factor of New Year's custom in Japan. There also seem to be higher level of activities in 2017 compared to 2016, which can be attributed to the increase in the usage of the reservation system. While many of the restaurants were present throughout the entire dataset, new restaurants entered the dataset at a steady pace. Starting October 2016, the rate at which new restaurants entered began increasing, helping to explain the artificial growth we are seeing in "all_visitors" counts.

The bottom left chart shows the overall distribution for time of the reservation; it is no surprise that most reservations are made for dinner in the evening hours. There is also a curious relationship between reservation time and visit time, as shown in the bottom right chart. There is a rough 24 hour pattern to be identified between the reservation and the visit time, while it is still most common to book few hours right before the dinner.

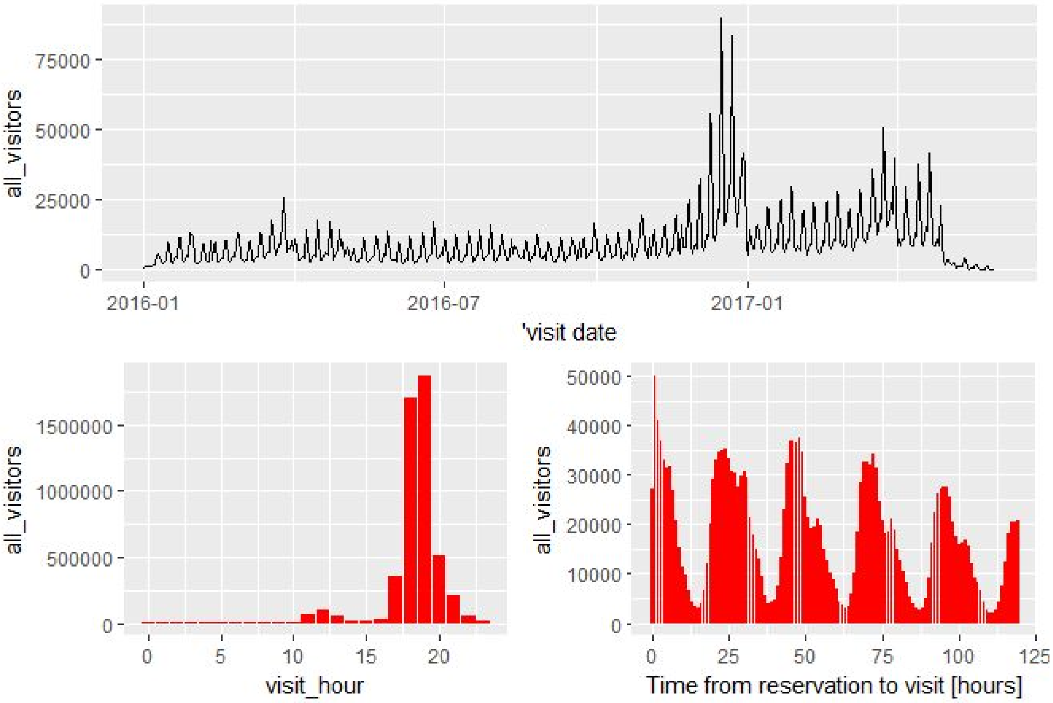


Figure 2: Trends of reserved visitors

Next, we examined the relationship between reserved visitors and actual visitors through the lens of a scatter plot, given these are the only two numeric variables in our dataset. As one can see from the Figure 3 below, most of the points fall above the line, indicating there are more actual visitors than reserved visitors on a given day. This observation makes sense because walk-in visitors who did not make a prior reservation are normally accepted in restaurants. The data points that fall below the line indicate that people made reservations in advance but did not end up visiting. Again, this is within our expectations. There seems to be a low level of correlation between reserve and actual visitors, as a result we will use reserved visitors as one of the predictor variable in our model.

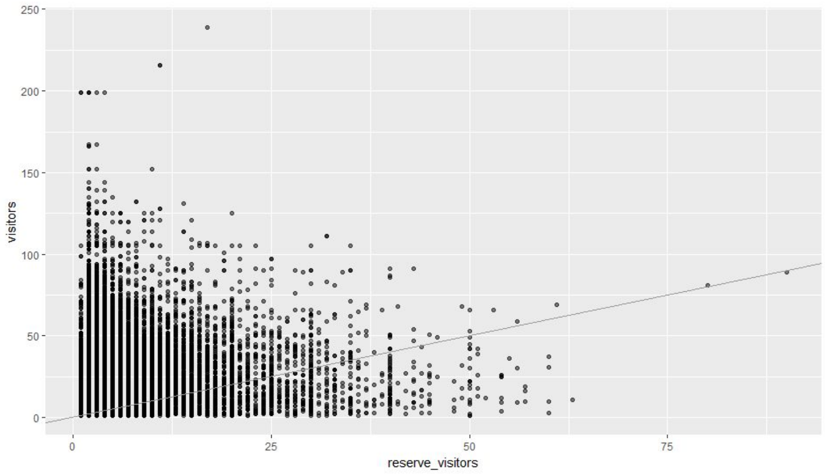


Figure 3: Reserved vs Actual Visitors

4. Data Transformation

Although the data consists of hourly observations, the goal is to predict the total daily visitors. To do this, a new variable, “day_of_week” was created to represent the day of each record, and the response variable was then summed for each day. The resulting object contains a date, the number of daily visitors, the day of the week, and an indicator of whether there was a holiday on that day or not. The following derived variables were created from “visitors” response variable: min visitors, mean visitors, median visitors, max visitors and count visitors.

We also provided the option of specifying a specific restaurant, genre, or area, and then aggregating the data to show daily visitors for the specific criteria. The resulting object would therefore contain the total number of visitors on a given day only for the specified restaurant, genre, or area, the date, and day of week which they visited, and the holiday flag indicator.

5. Initial Model Results

The model development started with understanding the basic model building process:

1. Model selection
2. Model fitting
3. Model validation

The customer data suggested the use of machine learning technology because of the following reasons:

1. Since the number of records is close to 1 million, we can use machine learning for better results.
2. There is no linear relationship between response and predictor variables, except we see a low level of correlation between reserved visitors and actual visitors (as shown in Figure 3 above). The correlation map in Figure 4 below suggests there is no correlation between variables other than derived variables from “visitors” response variable.
3. The big advantage of using machine learning is to capture all patterns beyond any boundaries of linearity or even continuity of boundaries.

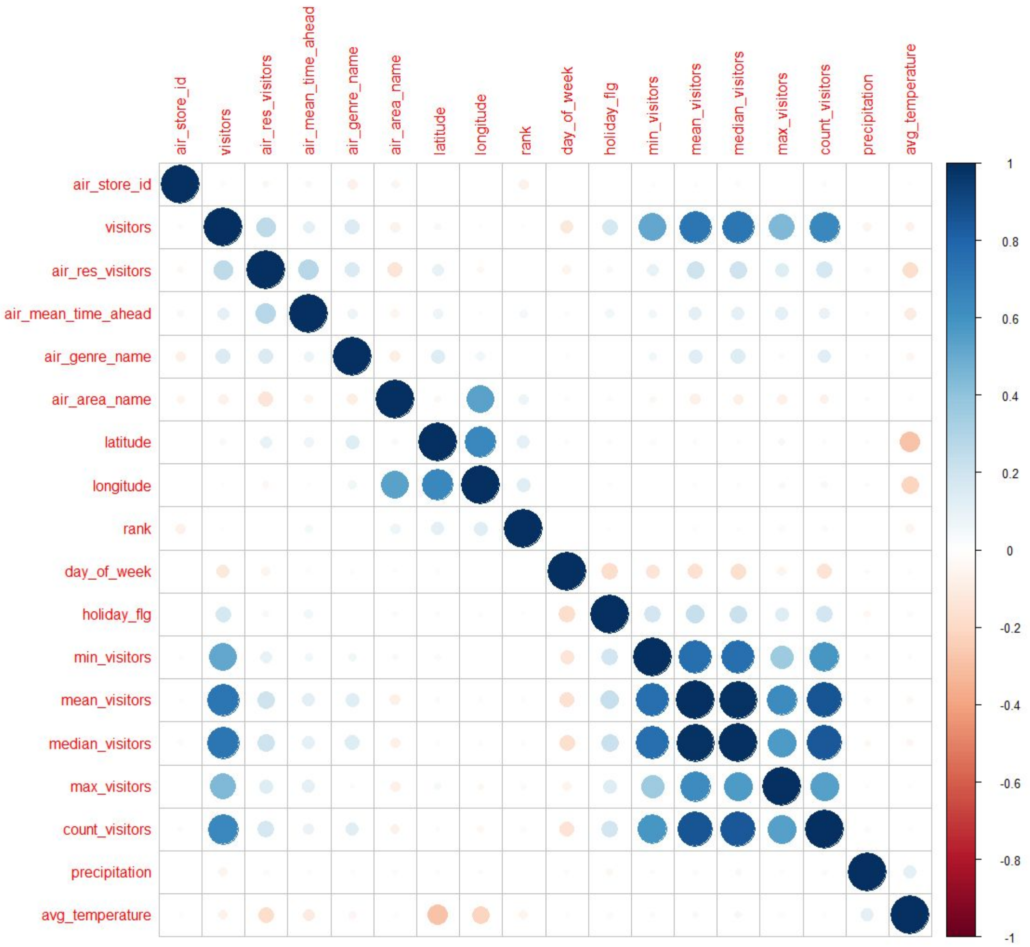


Figure 4: Correlation map

The chosen models to predict the number of visitors are XGBoost, H2O, LightGBM and K-Means with XGBoost. We additionally attempted several other time series forecasting methods such as ETS, NNETAR and ARIMA. Table 5 below provides the model names and descriptions.

Model Name	**Description**
XGBoost	Extreme Gradient Boosting with XGBoost
H2O	H2O is open-source software for big-data analysis
LightGBM	Light GBM is a boosting framework based on decision tree algorithm
K-Means with XGBoost	k-means clustering is a method to partition n observations into k clusters
ETS	Econometric Time Series
NNETAR	Neural Network Model
ARIMA	AutoRegressive Integrated Moving Average

Table 5: Model algorithm details

Our initial model development approach involves multiple levels of validation. We split the data into a training set (from 01-01-2016 to 03-08-2017 : 213,510 records) and a validation set (from 03-09-2017 to 04-22-2017: 44,321 records). We used the same data with different models to measure our prediction validation via RMSLE (Root Mean Squared Logarithmic Error). Table 6 below shows the RMSLE values for each model which we used to predict the total number of visitors for a restaurant from 04-23-2017 to 05-30-2017 (test set).

Model Name	RMSLE(Validation Set)	RMSLE(Test Set)
XGBoost	0.52	0.52
H2O	0.512	0.512
LightGBM	0.525	0.525
K-Means with XGBoost	0.47	0.47
ETS, NNETAR, ARIMA	0.61	0.61

Tabe 6: Initial Models used for prediction

We had superior results from K-Means with XGBoost due to the following reasons:

1. K-Means attempts to find discrete groupings within the data, where members of a group are as similar as possible to one another and as different as possible from members of the other groups. Restaurants visitors are highly volatile based on holidays so we need to cluster data based on high volume of visitors.
2. XGBoost "continue training" feature helped to further boost an already fitted model on new data. This is an iteration process where the first set of data is analysed and then it is merged with next set of data for further analysis.
3. We used the "elbow" method to find out appropriate number of clustering for this data set, which identifies the point where adding a new cluster does not greatly increase the prediction accuracy.

6. Preliminary Visualizations for Web/Mobile Application

6.1 Technical Design

The following sequence diagram shows our mobile/web application flow:

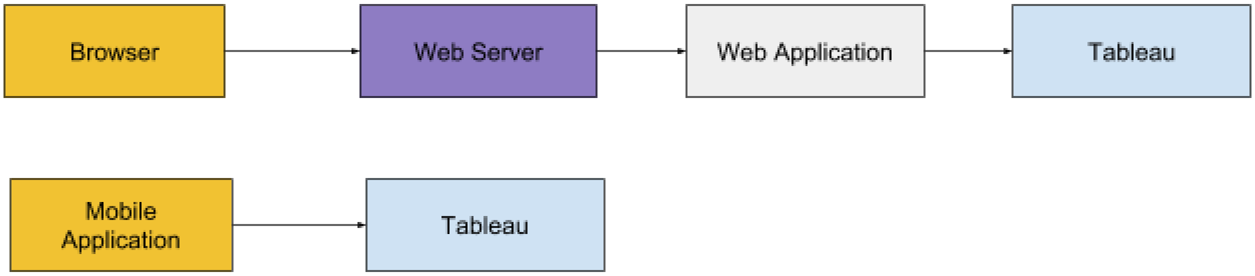


Figure 5: Application flow

The client’s web browser initiates a web request to the web server, and the web server identifies the request and forwards it to the web application. The web application gets the dashboard page from Tableau and sends it back to the client. The same applies to the mobile application, however we don’t need the web server and web applications in the middle to connect to Tableau. There will be some other features developed for the mobile application and web application apart from the Tableau dashboard which are:

1. Predict the number of visitors on any given date
2. The web interface and mobile interface will provide components for users to select the date and restaurant for prediction
3. The web and mobile interface will provide filtering functionality to see area and genre specific data

6.2 Dashboard Design

The intent of the dashboard is to give restaurateurs, such as Recruit Holdings, the power to understand what is happening within the competitive market. This will be accomplished via two functions.

The first function is advanced EDA based on historical data. The self service dashboard allows drill downs from a country level aggregation for total visits down to location and day specificity. In Figure 6a below, the end user is presented with a map of Japan; the size of the dots indicate total visitor volume, while the color indicates average store traffic volume. Figure 6b below shows the list of restaurant genres, with the respective visitors and traffic. By interacting with the map, the list is updated, and vice versa. It takes one click to understand that the Asian genre has the highest traffic, and no restaurants exist outside Tokyo. Likewise, it takes two clicks to understand that Izakaya is the most popular genre with Tokyo driving the most visitors, but Miyagi has nearly 50% more per-store traffic.

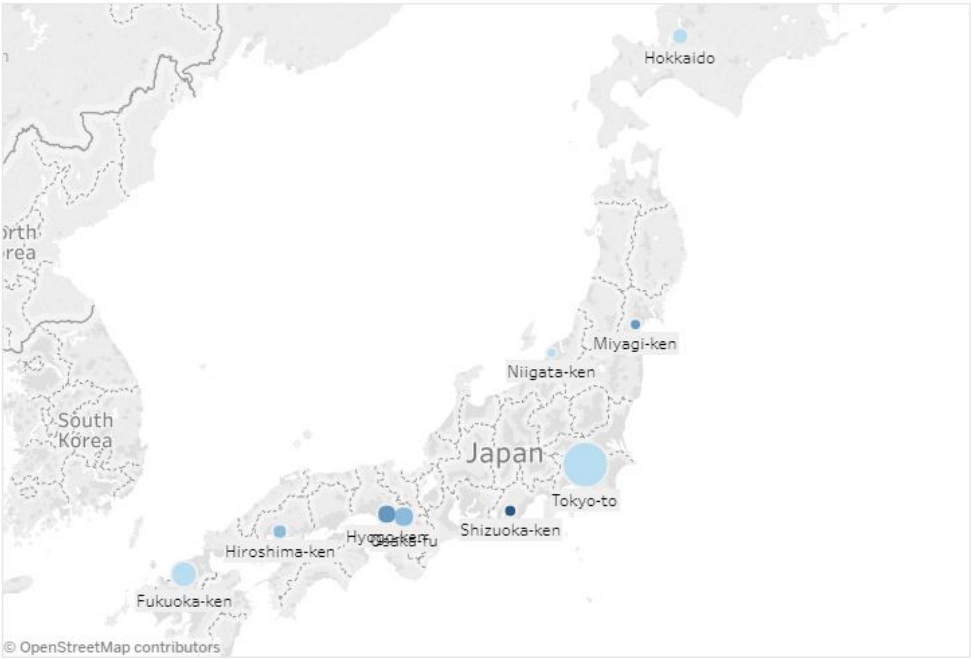
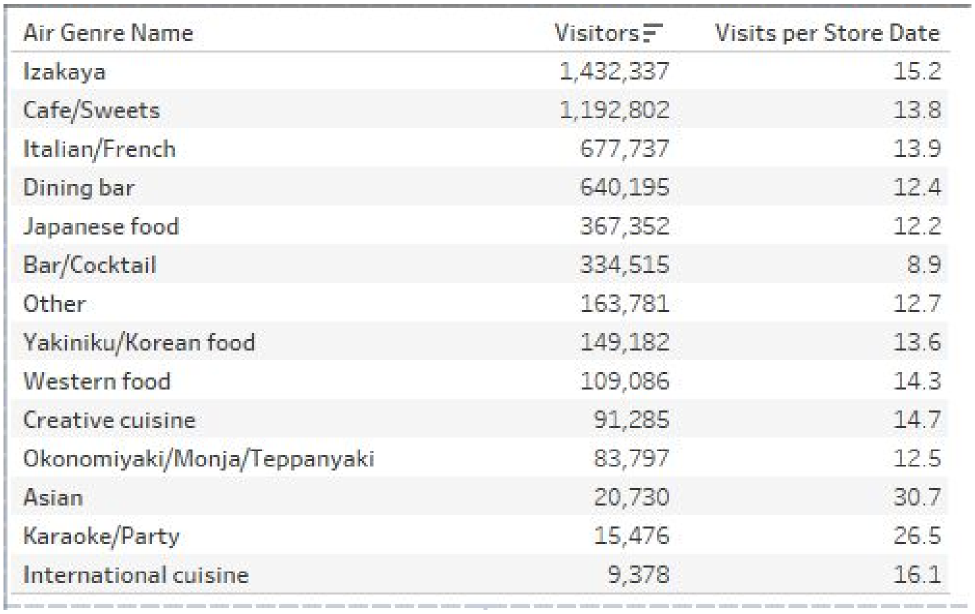


Figure 6a: Sample Dashboard visualization



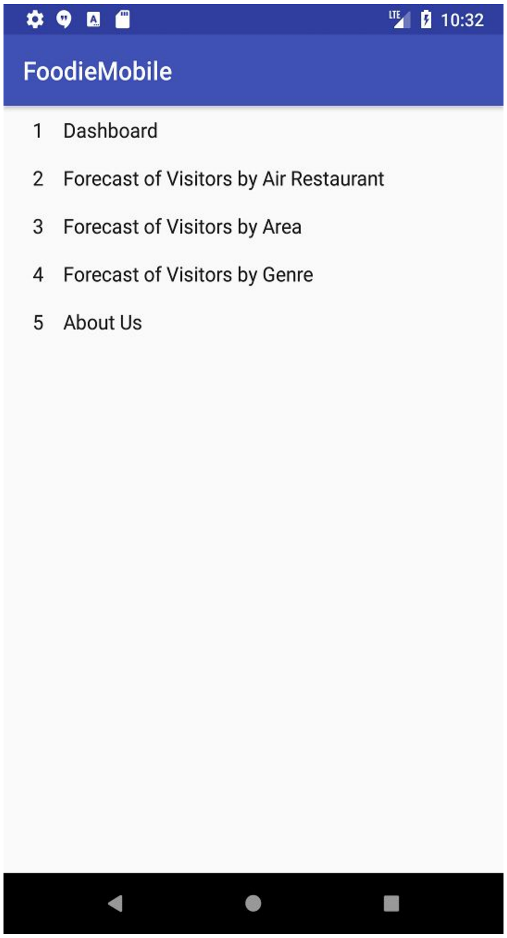
Air Genre Name	Visitors	Visits per Store Date
Izakaya	1,432,337	15.2
Cafe/Sweets	1,192,802	13.8
Italian/French	677,737	13.9
Dining bar	640,195	12.4
Japanese food	367,352	12.2
Bar/Cocktail	334,515	8.9
Other	163,781	12.7
Yakiniku/Korean food	149,182	13.6
Western food	109,086	14.3
Creative cuisine	91,285	14.7
Okonomiyaki/Monja/Teppanyaki	83,797	12.5
Asian	20,730	30.7
Karaoke/Party	15,476	26.5
International cuisine	9,378	16.1

Figure 6b: Sample Dashboard visualization

The second function leverages the predictive model to understand future traffic patterns. The intent of this functionality is to provide restaurateurs with an understanding of where and when customers will visit. This will provide data down as granular as store date, allowing individual restaurants to manage staffing and food purchasing to minimize unnecessary expenditures and waste.

6.3 Mobile Application Design

We developed a working prototype for our mobile application. Figures 7-9 below show screenshots of the mobile application with real data. This is an initial application prototype, which still requires testing for high quality control.



FoodieMobile

1 Dashboard

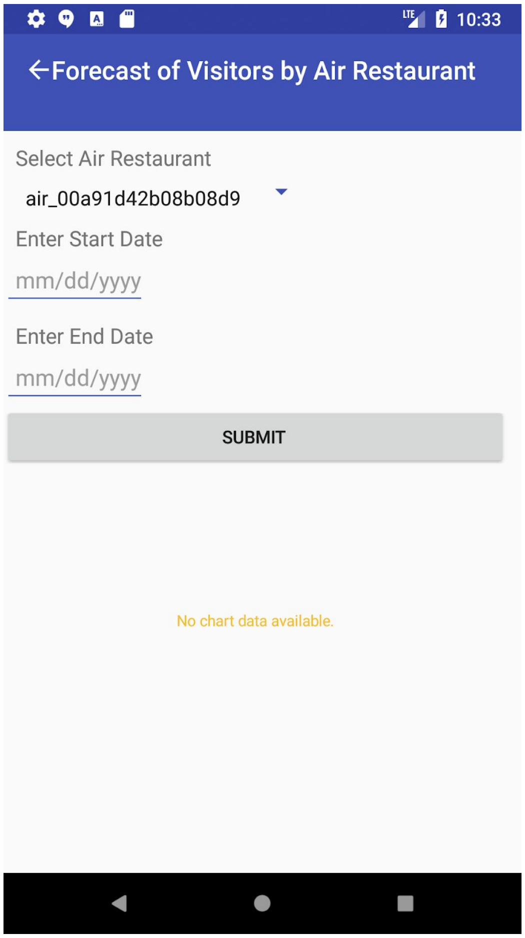
2 Forecast of Visitors by Air Restaurant

3 Forecast of Visitors by Area

4 Forecast of Visitors by Genre

5 About Us

Figure 7: Mobile application initial screen



Select Air Restaurant

air_00a91d42b08b08d9

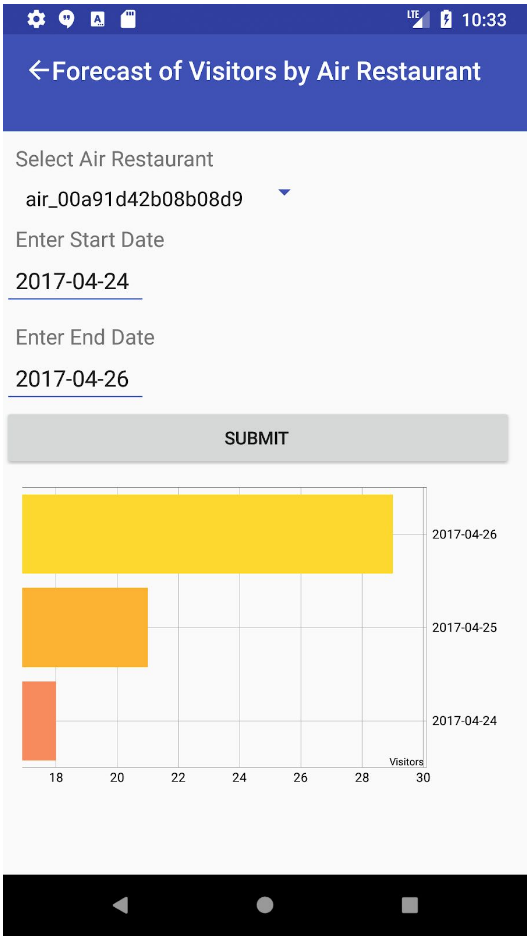
Enter Start Date

mm/dd/yyyy

Enter End Date

mm/dd/yyyy

Figure 8: Mobile Application Selection Screen



Forecast of Visitors by Air Restaurant

Select Air Restaurant

air_00a91d42b08b08d9

Enter Start Date

2017-04-24

Enter End Date

2017-04-26

Figure 9: Mobile application showing visitor prediction for a date range

7. Assumptions and Limitations

7.1 Assumptions

- iOS users will leverage the web application until an iOS application is created
- Tableau will be accessed from the web application as well as the mobile application using the same URLs
- The look and feel will not be the same for the web application as well as the mobile application through a responsive design
- There will not be any authorization or authentication for the web application as well as the mobile application

7.2 Limitations

- Since we don't have restaurant names, restaurant IDs will be supplied instead
- Predictions can be made for 39 days following the last day of the dataset, 04-23-2017; sample data is only available until 5-30-2017.
- Within the 10 week engagement, we will be developing the mobile application for Android based mobile phones

8. Conclusions

Given the highly competitive nature of the restaurant industry, there are many challenges in maintaining a successful enterprise. Foodie Analytics hopes to resolve some of these challenges through leveraging the vast and detailed data collected by Recruit Holdings via Hot Pepper Gourmet and AirREGI platforms.

Foodie Analytics has completed an in depth exploratory data analysis to ensure that all the intricacies and nuances of the data have been captured and understood. We have identified and analyzed the response variable and each predictor variable used for data modeling. We identified commonalities within the datasets and merged them into train and test datasets, then iterated through several rounds of modeling. After identifying five promising modeling algorithms, we chose to proceed with a K-Means with XGBoost model given its high accuracy of the Root Mean Squared Logarithmic Error. Finally, Foodie Analytics has developed draft formats of visualizations and dashboards to help Recruit Holdings in actioning our analysis and beginning to drive benefit.

Ultimately, this effort will provide Recruit Holdings with the necessary tools and information to make informed decisions in the adverse conditions of the restaurant industry.