Predicting the customer traffic of restaurants for future dates

Final Project

Baikara Ainur
Asawa Swapnil
aib25@pitt.edu
swa12@pitt.edu

ABSTRACT

Processing the time series data about the business operations can be useful if people could process them properly. We use restaurant business visitation data as the example to show how data mining techniques can be applied for such time series data. Based on those techniques we built a model to predict the future number of visitors to the restaurants of Japan. Having the number of future visitors restaurants to keep better management of their business. Moreover, our model would be useful in terms of efficiently planning the number of ingredients and staff management. We believe that proper services will give better customer satisfaction, and optimizing on resources will give better profit to the restaurants.

ACM Reference format:
Baikara Ainur and Asawa Swapnil. 2018. Predicting the customer traffic of restaurants for future dates. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 5 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Restaurants frequently face the problem of either wastage of food on one hand or devoiding their customers to provide enough choice because some items go out of stock. Sometimes, the customers have to wait a lot on busy days due to lack of staff and resources to handle traffic. The problem of the restaurant business is the lack of effective prediction about a number of customers they would have in advance. In order to help restaurants to keep better management in their place, we try to predict an approximate range of visitors will be visiting the restaurant for future dates using multiple datasets from different sources.

The goal of this project is to predict the number of future visitors for the restaurant. The novelty of the proposed idea is adding weather data to see whether it affects visitation. Using time series data is novel as compared to what we have studied in class and hence, it requires a different way of training and testing.

The features that make our project strong is that we have a good amount of data. Moreover, data is reliable, which is coming from Japan Meteorological Agency and Datasets from Recruit Holdings in Japan Applicable in real-world [1]. By having the knowledge about the number of visitors for future dates restaurant owners can efficiently order ingredients and improve restaurant maintenance

2 PRIOR WORK

Our project is related to business operation improvements. There is some research that has done the prediction using different models. One is using ARMAX model to forecast the hotel room and possible hotel revenue and performance [4]. They are using ARMAX model which is the extension of ARMA model by adding explanatory variables. Another research we considered is using ARIMA model to predict the future number of visitors for fast food restaurant chain for periodic time series data. Reading this research paper was very helpful for our project's regression model part as their area of focus was similar to ours [5].

Those research were using just regression model to predict the visitors for future dates. In our final project beside applying regression models, we also applied classification model. After that compared both models results.

3 DATASET

3.1 Data sources

The datasets are originally provided by Kaggle [2]. Provided datasets are coming from two sources:

- Kaggle Dataset from Recruit Holdings that owns restaurant review service and reservation log management software. The source of main datasets [2]
- The weather data (e.g. precipitation, humidity level) based on the restaurant?s locations from Japan Meteorological Agency [1]

Moreover, the main datasets have the following features:

- Location: longitude and latitude of restaurants (That information were very helpful to perform the geomapping)
- Dates: visit date for every restaurants
- Holiday information: a feature about the holidays in Japan
- Weather information: information about the precipitation, humidity level and more for each restaurants location based on our time range
- Number of bookings few days in advance: this is mostly reservation of the visitors. We have to note that reservation doesn't mean that those were actual number of visitors for the visit dates

Conference'17, July 2017, Washington, DC, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3.2 Brief Description of Recruiter dataset

Datasets (7 relational csv files) are collected from Japanese websites which are:

- Hot Pepper Gourmet (hpg), like Yelp (search and reserve)
- AirREGI / Restaurant Board (air), like Square: reservation control and cash register

3.3 Brief Description of Weather dataset

From Japan Meteorological Agency website, we received 1663 csv files on weather information. Each file corresponds to one city, and have daily weather information for the dates corresponding to our Main Data. Time ranges for our datasets:

- Training data: Jan 2016 - March (1st week) 2017
- Test set: March (last 3 weeks) 2017- April (first 3 weeks) 2017.

We have not used the test set time range as Kaggle was not giving this information for us. As our professor suggested, we split the dataset by ourselves. The split of the dataset looked like in above.

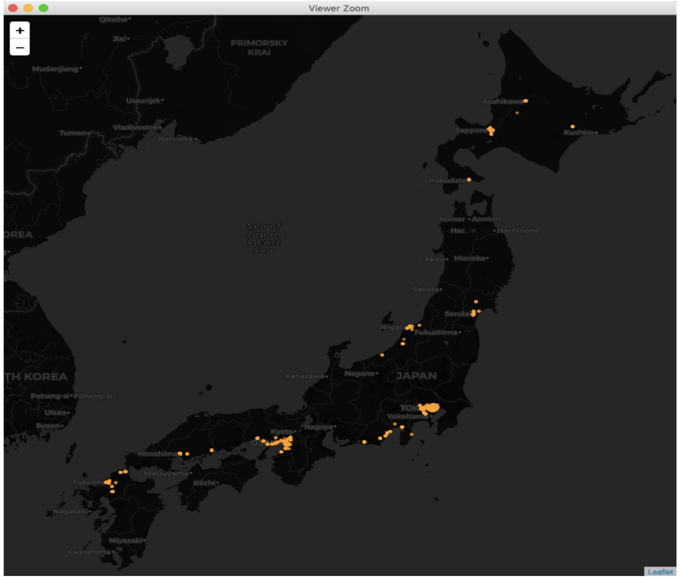
4 METHODOLOGY

Before applying the models we applied some data preprocessing.

4.1 Dataset Analysis and Preprocessing

We have changed the data information of all datasets as it doesn't make sense to make an hourly prediction. Figure 1 shows us that the regions of restaurants are very clustered to certain locations. After merging some datasets together in terms of relational features, we mapped it using leaflet library. This visual gives us a good representation of the values on the map. As can be seen from the map most of the restaurants are located in those three main cities: Tokyo, Kyoto, and Osaka.

Figure 1: Map with the distribution of the restaurants across Japan



Next visual that we would like to point out is the line graph which represents the visit date for the restaurants and the combined number of visitors. Figure 2 gives a good representation of the source of the datasets. It is clear that in the first half of the year the sum number of visitors was less than the rest of the part of the dataset. This means that there are some restaurants that have been added to the dataset. It made the sum number of visitors more starting from the second half of 2016. Another important part: we calculated cumulative distribution function and plotted it. It can be seen from the next figure:

Figure 2: Visit date for the restaurants versus the sum number of visitors for each date

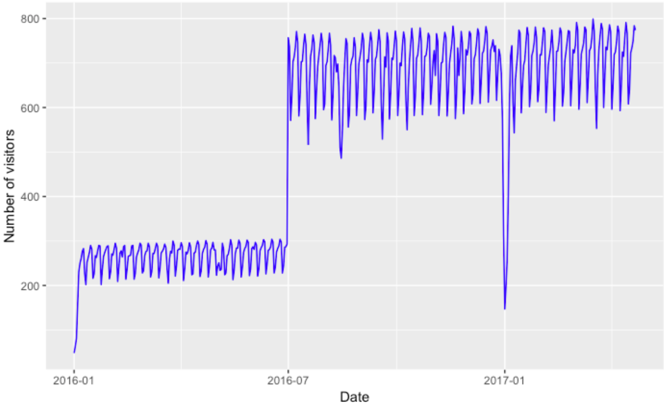
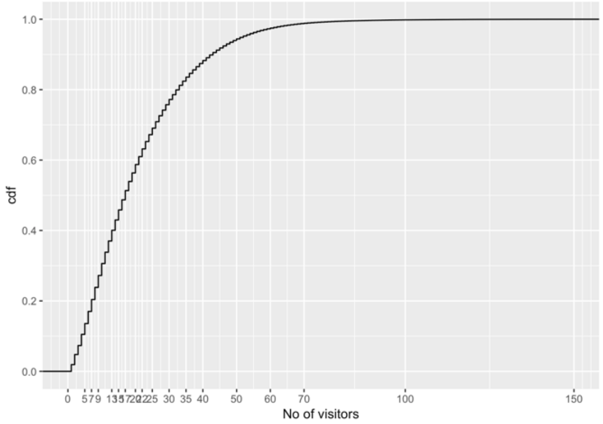


Figure 3: Visit date for the restaurants versus the sum number of visitors for each date

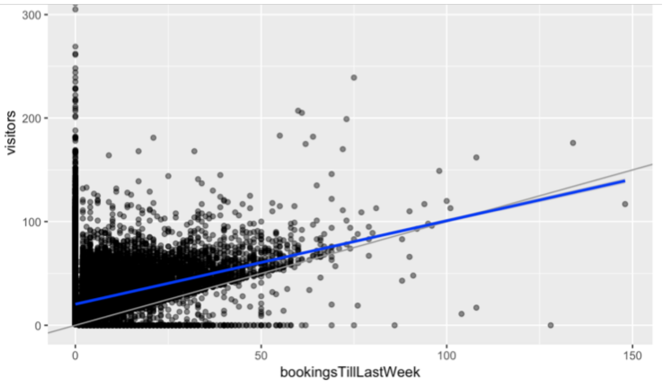


Based on Figure 3, if we divide the dataset into 5 classes, our classes would be:

Class1: no of visitors between 0-6, size of 7
Class2: no of visitors between 7-14, size of 8
Class3: no of visitors between 14-22, size of 8
Class4: no of visitors between 23-35, size of 13
Class5: no of visitors between 35 and more

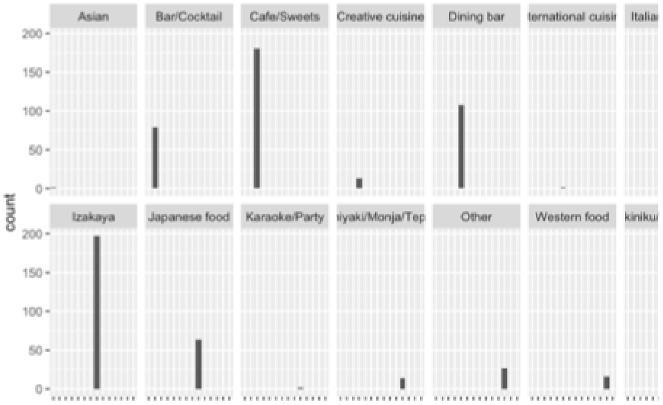
The median is 17. And we divide the data on 17. So we have got two classes, less than 17 (low), and more than 17 (high). We have done so to keep our data balanced Here are the more results of our data analysis part:

Figure 4: Booking till Last week VS number of visitors



Here are the observations from the plot. Points lower than grey line shows more people booked in advance than visited. Points upper than grey line shows less people booked one week in advance but visited

Figure 5: Cuisine types



- Cuisine: Asian, international, karaoke/party cuisine were removed as the main category
- weather data analysis: Weather source data had a lot of missing values for specific features, so we had to choose just those features that can be seen in most regions. Here are the final features we used for weather data: average temperature, minimum and maximum temperature, precipitation, hours of sunlight and wind speed.
- We have merged the datasets: reservation, visitation and restaurant information (location) datasets

As can be observed from this graph popular visitation days are weekend and Friday for all restaurants

In our project, we used both classification and regression model.

Figure 6: Density plot of number of visitors on different days of the week

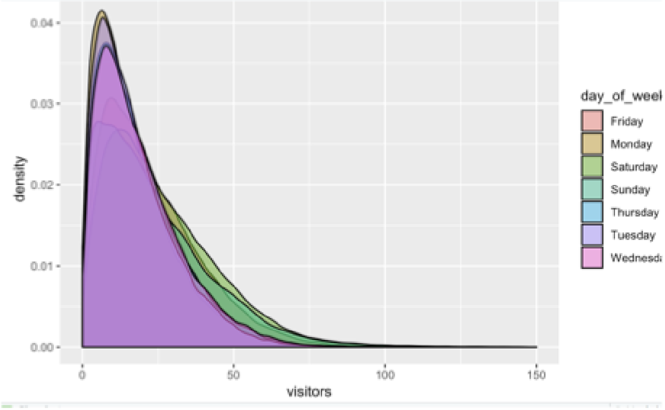
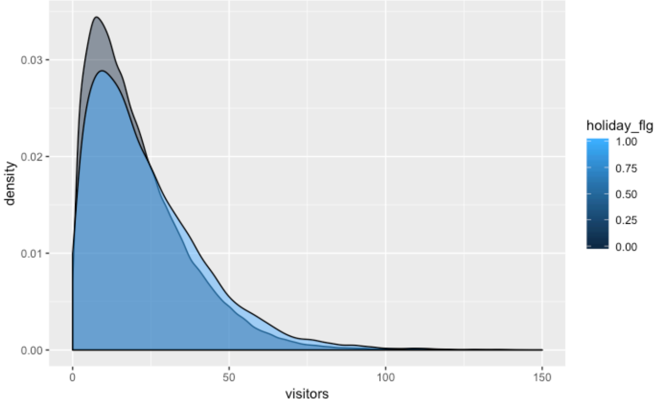


Figure 7: Density plot of number of visitors based on holiday vs no holiday



4.2 Classification Model

We applied various classification models like Logistic Regression, Decision Trees, Naive Bayes, KNNs, ada boost. We merged all the preprocessed data and started by converting the dataset output to 2 class classification problem by dividing the output column in 2 classes. If the number of visitors are greater than a threshold, we made it class 2 otherwise 1. The threshold is set to be median to keep the data balanced. We implemented time-series model to validate and test our results.

4.3 Regression Model

In the beginning, we ran our model for the whole dataset. Our models were performing poorly. Then, the resolution of the problem was to choose the top restaurants that have time series data for the whole dataset. What we mean is that the restaurant we chose for our model has the information of visits for every date (476 days). As it was pointed out in our prior work section we are using mainly ARIMA model for our project. On the top of ARIMA, we used Prophet, as to compare the results. For ARIMA, first, we checked whether time series data is stationary or non-stationary. Once we plotted the graph, it was clear that our time series data is stationary. Then we need to decide on how many days ahead we will be forecasting. After data smoothing technique we choose to predict a week ahead for the restaurant. After that, we deconstruct the time series so that we could comprehend the behavior of our dataset. This process would prepare our dataset as the foundation for forecasting model. ARIMA model we need to set some parameters. For ARIMA model we performed three iterations, by each iteration we changed the parameters of ARIMA model.

- first iteration we run the auto.arima model which will run the iteration with default parameters
- second iteration we added parameters that were suggested from auto.arima, which are ARIMA(1,0,1)(1,0,2)[7].
- third iteration: in order to get better results we checked our autocorrelation plots. ACF plots represent the correlation between the series and its lags. By looking at our ACF and Partial ACF plots, we decided to change the parameters into ARIMA(1, 0, 3)(1, 0, 2)[7]. Autocorrelation plots were suggesting the significant correlation at lag 3. Once we took into consideration this parameter, it gave us better results.

As was mentioned before, we also used Prophet for another autocorrelation model. It was giving good results too, but Prophet is the open source library developed by Facebook. We used this model because Prophet performs really well on stationary time series dataset. Among those two autoregressive models, we decided to use ARIMA, as ARIMA was performing well in our model and tuning the parameters went well. Even though Prophet was giving good results, we decided not to use it like the model doesn't allow us to perform the tuning process.

5 EVALUATION RESULTS

5.1 Classification model results

We developed a time series function which takes the data, train and test on validation or test test as asked by the user. It divides the data into number of folds asked by the user(16 in our case, one per month). We found it works better with 2 month as training for next one month as validation and next to next month as test. Starting from first 2 data subsets as training, 3rd subset as validate and 4th as test, it evaluates and shifts everything by one month. It calculates the accuracy, precision, recall and f-score for each of the val/test and model input by the user. Finally it calculates the average scores for the model. We validate for various models and found results were not good (accuracy of around 0.5). We tried removing various columns and validate. Finally, we only used data for one restaurant at a time and it gave us best results. Decision trees gave best results. Table 1 shows the evaluation results

D tree gives the minimum error and best F-score So, we choose this as our model and now apply on a test set.

5.2 Regression model results

We evaluated the performance of the autoregression model by the following measurements: AIC and BIC score and sigma$\hat{2}$ (model's assumed constant variance). Table 2 shows the evaluation results

Table 1: Classification Model evaluation results

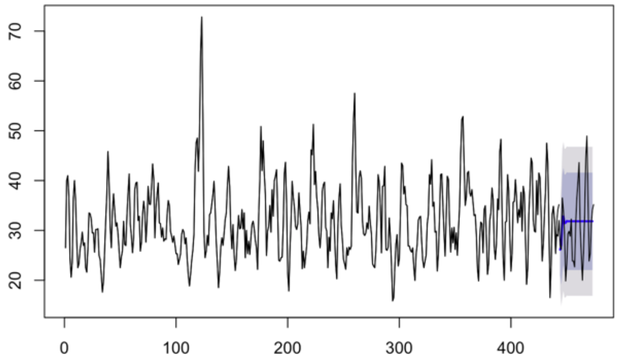
Measures	Linear Regression	Decision Tree	Naive Bayes	Final Test (Decision Tree)
Accuracy	0.8222812	0.8249337	0.7692308	0.8196286
Precision	0.826053	0.8249308	0.8396207	0.8196258
Recall	0.9933069	0.9999958	0.9075547	0.9999958
F-score	0.900907	0.9030203	0.8480525	0.8999587.

Table 2: Autoregression Model Results

Measurements	First Iteration	Second Iteration	Third Iteration
sigma$\hat{2}$	24.6	24.3	14.49
log likelihood	-1430.22	-1430.22	-1313.08
AIC	2874.45	2874.45	2644.15 .

As can be seen from the table the third iteration is giving the best results. Carefully tuning the parameters based on some correlation plots and other kinds of measures would give you better results. We tested our model on holdout data for the last month. Here is the plotted result from our rmd file.

Figure 8: ARIMA model prediction on unseen last month data



As can be seen from this plot. Black line is actual value, blue line represent the predicted variables. First week it is forecasting really well, but after that we can see the straight line which basically mean that our model is not predicting well a week after. It is very good result, as we have got the result we wanted. We initially aimed to predict a week ahead which basically our model doing that.

As we described before, Prophet is a good at finding patterns. So, it our model it found some patterns and keep repeating it. We haven't relied too much on this model as Prophet doesn't allow us to set the parameters by ourselves.

6 DISCUSSION

Future works: We are getting reasonable results with classification and autoregressive models. Also, time series cross-validation for

Figure 9: Prediction results for Prophet

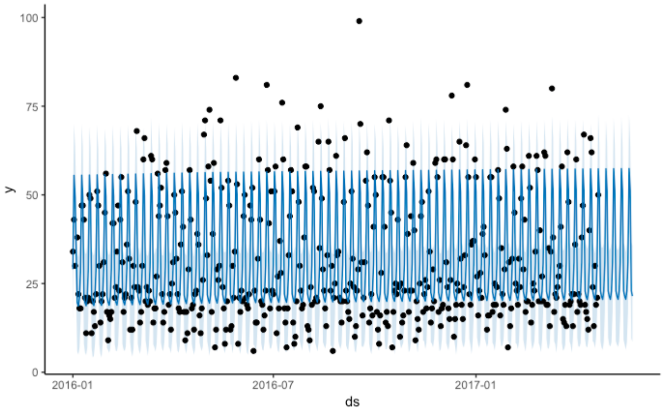
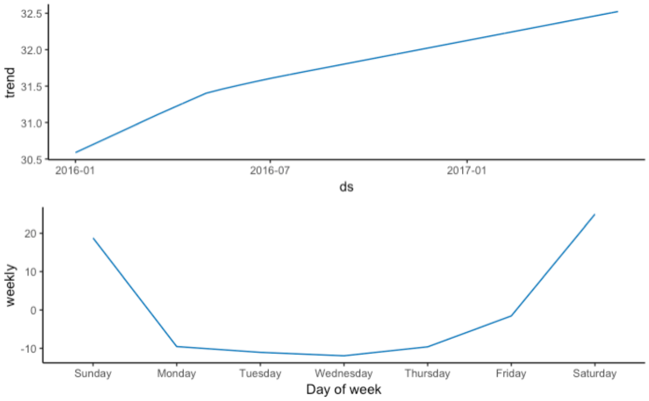


Figure 10: Prediction results for Prophet



the autoregressive models is yet to be implemented, like for classification we did. We later realize for the restaurant data we have used, the classes are not balanced i.e. their median is not 17, and if we run on their median of the hotel, we are getting bad results.The immediate future work needs fine-tune for classification models. We have worked on weather data a lot but the final manual change of names would take a lot of time before combining. Also, if we had more time, we would like to apply more autoregression models for our dataset like MA, ARFIMA, Holt-Winters.

7 CONCLUSION

We learned a lot from our final project. So we decided to choose this problem as it is a time series problem. We leveraged classification models knowledge and skills from we learned from our classes and applied it to the time series problem by converting our dataset into categories. Then we learned a lot about times series problem by applying autoregression models. Many features seemed to be very relevant for classification but turned out to be not that useful, and date information alone is a very good predictor for finding customer traffic. After spending some weeks, we realized that our model works better for each restaurant (separate restaurant) because before that we were applying the model for the whole datasets including all restaurants across Japan. We learned that simple autoregressive models can give good prediction results for one week ahead prediction (but there are still rooms for improvements). Overall this project gave us good learning of state-of-the art research as well as hands-on practices on those, with a business solution.

8 ACKNOWLEDGMENTS

God bless stackoverflow and Google's search algorithm. There were really helpful in terms of finding the sources. Thank you Yuru, for sharing the solutions for homework and class codes. Hillman Library for giving the opportunity to meet in nice group study areas on the forth floor. University of Pittsburgh's saferider shuttle was very helpful for giving us the ride after finishing the project in the middle of night.

REFERENCES

[1] Japan Meteorological Agency Data is retrieved from `http://www.data.jma.go.jp/gmd/risk/obsdl/index.php`

[2] *Recruit restaurant visitor forecasting* Kaggle, 2018 `https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting`

[3] Time Series Nested Cross-Validation ? Towards Data Science *Towards Data Science* `https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb91`

[4] Yang, Y., Pan, B., & Song, H. *Predicting Hotel Demand Using Destination Marketing Organization?s Web Traffic Data.* Journal of Travel Research, 53(4), 433?447 `https://doi.org/10.1177/0047287513500391`

[5] Lon-Mu Liua, Siddhartha Bhattacharyyaa, Stanley L. Sclovea, Rong Chena and William J. Lattyak *Data mining on time series: an illustration using fast-food restaurant franchise data* Computational Statistics & Data Analysis 37 (2001) 455?476