Predicting the customer traffic of restaurants for future dates.

Members:
Ainur Baikara, aib25@pitt.edu
Swapnil Asawa, swa12@pitt.edu

Problem:
Restaurants frequently face 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 restaurant business is the lack of effective prediction about number of customers they would have in advance. In order to help restaurants to keep better management in their place, we try to predict approximate range of visitors will be visiting the restaurant for future dates using multiple datasets from different sources.

Why is it interesting problem?
This is interesting problem for the following reasons:

1. Good application in real world: This will help restaurants to know how many customers will be visiting restaurants. Hence, restaurants can efficiently buy ingredients and do some other useful stuff with it.
2. Proper services will give better customer satisfaction, and optimizing on resources will give better profit to the restaurants. That is why, this information would be very useful for the restaurants.
3. Doing this prediction and analyzing how customer traffic gets impacted with weather, holidays, locations, will be fun.

The Approach we Planned:

The first step is to reformat the dataset in the most efficient way for querying. Because we have multiple datasets from different sources, this involves merging most frequently used information (restaurant ids in multiple datasets) with least redundancy, and separating (deleting) less useful information in different tables.

Restructuring the output: Predicting actual number of visitors is actually a regression problem (or classification problem with classes with all natural numbers) but that's less effective we plan to convert this using classification techniques for the following reasons:

- The range of number of visitors gives sufficient information to plan management for the day, and a little difference in the number doesn't matter much if the number of visitors are many.
- Secondly, applying k means, and other algorithms will give better results when the output is grouped into bins, rather than each number as one class.

We plan to divide the classes range on the basis of distribution of our data (positively skewed). For example, Class 1: 1-3 visitors, Class 2: 3-10 visitors, Class 3: 10-20 visitors, Class4: 20-40 visitors. We will refine these values more by analyzing the p-values from dataset.

The next step will be divide the dataset in training, validation and test set. Because the data is temporal in nature (time series data), we will use nested cross validation method Forward Chaining (3).

The third step is to try various machine learning algorithms on our training set and improve using validation sets, and finally getting results from test sets. We expect boosting/ensembling will be very useful because all features of our dataset are highly imbalanced (like holiday(few) vs no_holiday(many)), so boosting can help in giving more weight to minorities. While evaluation, we will use various measures like precision, recall, F-scores to find the best model with statistical significance.

We will use the best model to run on the test sets and report the result.

Datasets:

We are using dataset from the two sources:

Dataset from Recruit Holdings that owns restaurant review service and reservation log management software. The dataset is taken from Kaggle competition(2). This gives details about number of reservations, number of visitors, the restaurant locations, their cuisine types, holiday information.

We believe that weather information will be very important for us to predict results accurately because weather conditions affects the visitation status of customers. Yang et.al (3) has also used weather data to predict the number of guests for the hotel visitation, which is a bit similar problem. We are taking weather data (e.g. precipitation, humidity level) based on the restaurant's locations from Japan Meteorological Agency (1)

References

1. Japan Meteorological Agency Retrieved from http://www.jma.go.jp/jma/en/Services/indexe_services.html
2. Recruit restaurant visitor forecasting (2018). *Kaggle*. Retrieved from https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting
3. Time Series Nested Cross-Validation – Towards Data Science (May 19, 2018). *Towards Data Science*. Retrieved from https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9
4. Yang, Y., Pan, B., & Song, H. (2014). 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