Capstone Project : Recruit Restaurant Visitor Forecasting

Domain Background : This project is derived from time series forecast. The data holds the following information for 829 restaurants in Japan which includes restaurant's id which is unique for each restaurant, date, No of visitors on those dates, restaurant genre and area name and its location using latitude and longitude. This data are time series forecast because here Input is Time(it can be date and time or onlydate) and output is no of visitors. In this, order of data are important ,change in order will result in poor forecast. For example if the order of no of visitors in that column is changed and train the model with those un ordered data will result in wrong forecast on those days. So ordering is important that is crucial in time series data. Solving this visitors forecast problem makes restaurnat to effectively plan on scheduling staff members and purchasing ingredients, which in turn will reduce loss and it will help restaurants to be much more efficient and allow them to focus on creating an enjoyable dining experience for their customers. The reason I chose this capstone because when I owned a restaurant, and not knowing the approximate no of visitors at the intial year had led to ineffecient staff planning and buying ingredients which leads to loss. But later years I have used historical data and information about holidays and was able to guess the no of visitors which helped me to reduce the loss significantly. Building models using historical data, location, genre, holidays, would help to better approximate no of visitors.

Problem Statement: In this project, I would forecast 'No of visitors a restaurant will recieve in future'. The problem here is time series where time as input and no of visitors as Output. But I would convert the time series problem into supervised learning problem where instead of giving time as input to model I would use previous month's day of week average on no of visitors as input and no of visitors on each day as output which is strongly correlated than giving just time as input. This conversion still maintains the order by replacing time with with average no of customers. This conversion leads to approach this problem as supervised machine learning regression problem.

Dataset and Inputs: Dataset for this problem is obtained from competion organized in Kaggle. In this competition, I was provided a time-series forecasting problem centered around restaurant visitors. The data comes from two separate sites:

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 You must use the reservations, visits, and other information from these sites to forecast future restaurant visitor totals on a given date. The training data covers the dates from 2016 until April 2017. The test set covers the last week of April and May of 2017. The test set is split based on time (the public fold coming first, the private fold following the public) and covers a chosen subset of the air restaurants. Note that 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.

File Descriptions This is a relational dataset from two systems. Each file is prefaced with the source (either air *or hpg*) to indicate its origin. Each restaurant has a unique air_store_id and hpg_store_id. Note that not all restaurants are covered by both systems, and that you have been provided data beyond the restaurants for which you must forecast. Latitudes and Longitudes are not exact to discourage de-identification of restaurants.

air_reserve.csv This file contains reservations made in the air system. Note that the reserve_datetime indicates the time when the reservation was created, whereas the visit_datetime is the time in the future where the visit will occur.

air_store_id - the restaurant's id in the air system visit_datetime - the time of the reservation reserve_datetime - the time the reservation was made reserve_visitors - the number of visitors for that reservation hpg_reserve.csv This file contains reservations made in the hpg system.

hpg_store_id - the restaurant's id in the hpg system visit_datetime - the time of the reservation reserve_datetime - the time the reservation was made reserve_visitors - the number of visitors for that reservation air_store_info.csv This file contains information about select air restaurants. Column names and contents are self-explanatory.

air_store_id air_genre_name air_area_name latitude longitude Note: latitude and longitude are the latitude and longitude of the area to which the store belongs

hpg_store_info.csv This file contains information about select hpg restaurants. Column names and contents are self-explanatory.

hpg_store_id hpg_genre_name hpg_area_name latitude longitude Note: latitude and longitude are the latitude and longitude of the area to which the store belongs

store_id_relation.csv This file allows you to join select restaurants that have both the air and hpg system.

hpg_store_id air_store_id air_visit_data.csv This file contains historical visit data for the air restaurants.

air_store_id visit_date - the date visitors - the number of visitors to the restaurant on the date sample_submission.csv This file shows a submission in the correct format, including the days for which you must forecast.

id - the id is formed by concatenating the air_store_id and visit_date with an underscore visitors- the number of visitors forecasted for the store and date combination date_info.csv This file gives basic information about the calendar dates in the dataset.

calendar_date day_of_week holiday_flg - is the day a holiday in Japan The description about data is from following link https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data (https://www.kaggle.com/c/recruit-restaurant-visitor-forecasting/data)

Solution Statement: Steps to solve this problem 1.Time series problem can be framed as supervised learning problem. This will opens up a lot of linear and non linear supervised machine learning algorithm on my problem.This can be done using the previous time steps as input variable and next time step as output variable. The number of previous time steps is called the window width or size of the lag. This method referred as sliding window method for time series data.

1. Once my training data is reframed , similarly I should reframe my testing data for the forecast.
2. The training data covers the dates from 2016 until April 2017. The test set covers the last week of April and May of 2017. So I have decided to use previous month day of week visitors average as input and no of visitors on training data as output as well as on testing data but no of visitors have to be predicted.
3. Once my training and test data is prepared, then I would build model using supervised machine learning regression algorithm for forecast.

Benchmark Model The problem is supervised regression problem. The model suits for this dataset is Knearest neighbor regressor. In ths model the prediction based on average of nearest neighbors. The nearest data for no of visitors on specific day would be previous days no of visitors, and its previous month day of week average

which are more appropriate for predicting the current day visitor count. The no of Neighbors is hyperparameter and it is given by user while training the parameter. The right number will make the model less biased and variance.

Evaluation Metrics: Root mean square error have chosen to measure the performance of Bench mark model and other models like support vector machine and Random Forest. For the regression problem the metrics can be Mean absolute Error as well. But RMSE gives high weight to large error even if the those large errors are few in number. The large errors are not desirable in most cases. It gives weight to large errors irrespective of their count.

Project Design: The following steps are planned to solve this problem

1. Importing the dataset
2. Merging relevant files with training data and convert the data to relevant datatypes.
3. Perform Feature Engineering - for instance: extracting day of week, month, week of the year information from date column
4. Filter the no of visitors on training data with no holidays and compute every month day of week average using groupby in pandas.
5. Performing time series to supervised learning using groupby pandas and helper method so that previous month day of week average would be computed and merged with training dataset.
6. Preparing train and test data set which includes following fields as input restuarant id, month,dow, previous month average(which we got by helper method), genre,areaname,holiday_flag(bcoz the previous month average computed on days with no holiday),latitude,longitude as inputs and no of visitors as output. Same with testing data but no of visitor column will be zero need to be predicted. 7.Split the training data into train and test split to evaluate the model. 8.Using sklearn libraries, implement KNN, SVM and Random Forest model. 9.Measure the performance of model with test set of training data using RMSE metrics. If the value is closer to zero or negative means the model is better model.Based on RMSE value decide the better model. 10.Once the model is finalized then train the model using the whole training data and perform prediction on the test data where no of visitors are zero and need to be predicted.