Midterm Project

Restaurant Visitor Forecasting

Name: Young, Brent

Kaggle Name: Brent Young

Kaggle Score: Public: 0.585; **Private (Final):** 0.615

Predict 413 Section #: 55

Quarter: Winter 2018

Introduction

Problem

The purpose of the midterm project is to analyze restaurant visitor data using time series and forecasting methods to predict the total number of visitors to a restaurant for specified future dates. However, forecasting the total number of visitors to a restaurant for future dates can be problematic due to factors such as restaurant attendance, weather, competition from other restaurants in the surrounding area, or little historical data. The data also contains information from two systems: air and hpg, along with unique restaurant ID's as either air_store_id or hpg_store_id. However, it's important to note that not all of the restaurants are covered by both systems and that more data is provided than is needed for prediction purposes. As a result, per Dr. Fulton and for the purpose of this class, I will focus on the air_visit.csv file.

Significance

The problem is significant/interesting because restaurants need to know how many customers to expect each day so that they can ensure that they are staffed accordingly and have the right amount of ingredients available. Not only does this result in enhanced efficiency, but it allows them to focus on creating an enjoyable dining experience for their customers. Additionally, the fact that Recruit Holdings has unique access to restaurant, reservation, and visit data (*since they own products such as Hot Pepper Gourmet, AirREGI, and Restaurant Board)* makes this problem even more interesting.

Data Exploration

Description of Training and Test Datasets

The training dataset is from 2016 until April 2017. It's also important to note that the training dataset omits days where the restaurants were closed. The test dataset is from 4/23/2017 through 5/31/2017 and will contain predictions of visitors during this timeframe for the given set of restaurants with air_store_id's. This dataset is also split based on time (e.g., the public fold, then the private fold), covers a chosen subset of the air restaurants, and spans a holiday week in Japan called the "Golden Week." Golden week takes place on April 29th and May 3rd to 6th. There are also days in the test dataset where restaurants were closed and had no visitors. This is ignored in the Kaggle scoring and will appear as 0.

Structure and Size of air_visit_data

The structure of the training dataset has 252,108 rows and 3 variables: The first variable, air_store_id is a factor variable and represents each restaurant, visit_date is a factor variable and represents the date, and visitors is an integer variable and represents the number of visitors to the restaurant on the specified date.

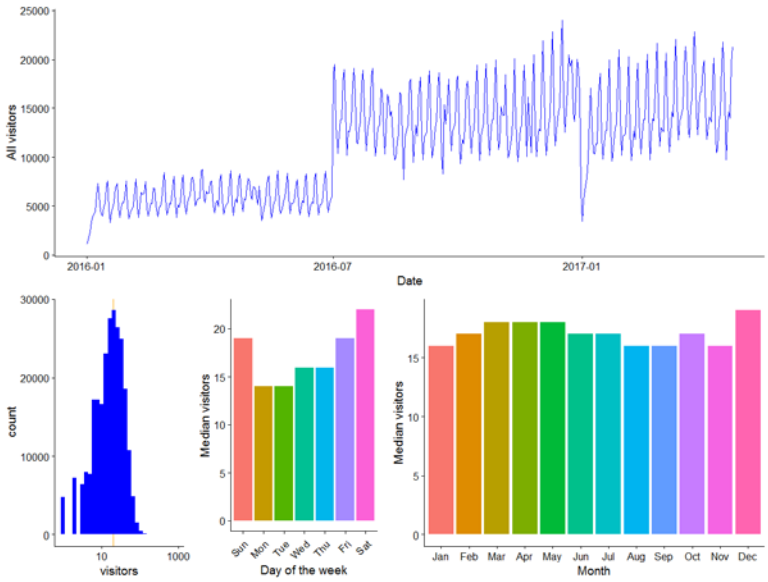
Figure 2: Descriptive Statistics of air_visit_data

Min	1st Quartile	Median	Mean	3rd Quartile	Max
1	9	17	20.97	29	877

	air_store_id	vist_date	visitors
Distinct	829	478	204

Observations: Figure 2 shows summary statistics so that we can check for missing values, outliers, etc. on the air_visits_data.csv file. The data shows that the mean number of visitors in the dataset is 20.97, the median number of visitors is 17, minimum is 1 and max number of visitors is 877. Additionally, there are 829 distinct stores, 478 distinct dates, and 204 distinct number of visitors on a specific date in the dataset.

Figure 3: Data Visualization



Observations: Figure 3 shows the total number of visitors per day across the full training data set together with the median visitors per day of the week and month of the year (R code: Kaggle Kernel borrowed from Heads or Tails). The timeplot shows that there is a "step" structure as time increases. This could be due to the fact that new restaurants/data were added to the database. The timeplot also reveals that there is strong seasonality that corresponds to a weekly cycle. The histogram of the number of visitors per restaurant on each day shows that it peaks at around 20 as indicated by the light orange line. This number corresponds to the number seen in our descriptive statistics. Furthermore, the histogram of median visitors per day of the week shows that Friday, Saturday, and Sunday are the busiest days, while Monday and Tuesday have the lowest number of visitors. Lastly, the histogram of median visitors per month of the year shows that March through May is consistently the busiest months, while December is the most popular month throughout the year.

Data Preparation (Processing of the Data and Cleaning)

Observations: In order to create time series based models, I reshaped/reformatted the dataset from long to wide format using the dcast function (R code: Kaggle Kernel borrowed from DSEverything). After doing so, my EDA analysis of the missing values showed that there were now 144,154 NA's. Interestingly, the histogram of missing data using VIM package (see figure 4 image on the right) showed a downward trend from 2016 until April 2017. For instance, a lot of NA's appeared mostly from January to June 2016 and this number steadily declined after June 2016. This coincides with the step level timeplot we saw earlier in our EDA. As a result, assuming that majority of these NA's were due to the fact that restaurants were lagging on reporting their visitor counts and/or new restaurants were being added to the database, I decided to fill all the missing data with 0, instead of using average, median, interpolation, na.kalman, or last observation carried forward/backward *(in which I tried using all of them…but did not improve my models)*. There are also days in the test dataset where restaurants were closed and had no visitors, so filling in the missing data with 0 also addresses this issue. Lastly, other modifications of the data contained reformatting/adjusting the visit_date to a "mdy" format using mutate.

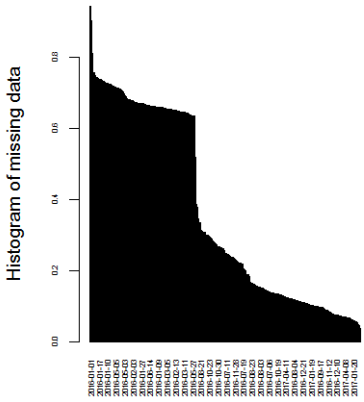


Figure 4

Types of Models, Reviews of Literature, & Formulation of Models

Types of Models

Given that the EDA revealed that there is strong seasonality that corresponds to a weekly cycle, I decided to use ETS (e.g., ANA) and auto.arima (e.g., ARIMA(1,0,0)(2,0,0)[7]) to build my models. I decided to use ETS and auto.arima because they both have the ability to handle seasonality, estimate the necessary parameters, and then select the appropriate model objectively for each individual time series (e.g., restaurant) based on AIC, which is perfect for a dataset that contains 829 different restaurants. I will also explore a mixed model approach as well. *See "Formulation of Models" section for more details on each model.*

Reviews of Literature *(see last page for references)*

There were many peer reviewed journals in the NU library database that used ARIMA or ETS to forecast tourist arrival/visitors, demand, or production which are similar to predicting the number of visitors. For instance, in the *Academy of Management Journal (1979)*, Berry, Mabert, and Marcus, use ETS to forecast teller window demand. Second, in the *Annals of Tourism Research* (1995), Dharmaratne uses ARIMA to forecast tourist arrivals in Barbados. Third, in the *Annals of Tourism Research* (2012), Gounopoulos, Petmezas, and Santamaria use ARIMA and ETS methods to forecast tourist arrivals in Greece. They found that ARIMA outperforms other models as a directional forecasting tool, but Holt's exponential smoothing model with trend is the best performing model. Fourth, in the *Journal of Animal and Plant Sciences* (2017), Karadas, Celik, Eyduran, and Hopoglu use ETS methods to forecast production of oil seed crops in Turkey.

Lastly, in the *Tourism Economics* (2012), Nowman and Van Dellen use ARIMA methods to forecast overseas visitors to the UK.

Formulation of Models

In order to test our predictions for both auto.arima and ETS, we will forecast for an individual restaurant (e.g., air_04341b588bde96cd) and then compare the results for this restaurant to other restaurants to see how these methods perform. We will also test our predictions by splitting the data into a train and test set. For instance, the training set will be from 1/1/2016 to 3/14/2017 and the test set will be from 3/15/2017 to 4/22/2017 (39 days) to mimic the period from 4/23/2017 to 5/31/2017 (39 days). We will also fill all the missing values with 0 so that we can conduct time series analysis (weekly) (R code: Kaggle Kernel borrowed from Heads or Tails).

Auto.arima: ARIMA(1,0,0)(2,0,0)[7] with non-zero mean

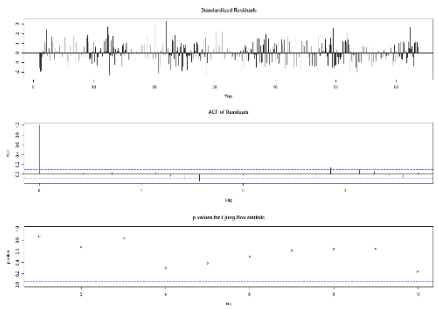


Figure 5

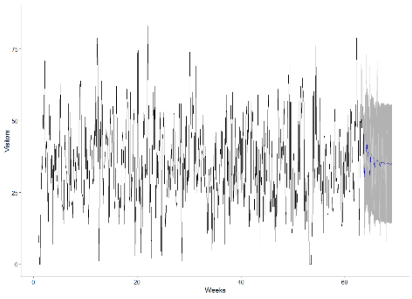


Figure 6

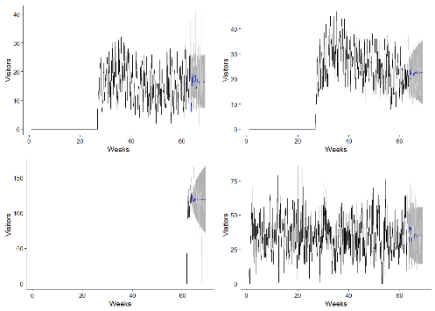


Figure 7

Observations: Figure 5 shows a tsdisplay of the residuals, ACF of Residuals, and p-values for Ljung-Box test. The ACF of the residuals and the Ljung-Box test (p-value = 0.07649) confirmed that the model contained white noise. Additionally, the ADF (p-value=0.01) and KPSS Test for stationarity (p-value = 0.1) also confirmed that the data was stationary and differencing is not required. As a result, ARIMA(1,0,0)(2,0,0)[7] with non-zero mean, which means 1 order of autoregressive, 0 degrees of first differencing, and 0 order of the moving average part for non-seasonal, and 2 order of autoregressive, 0 degrees of first differencing, and 0 order of the moving average part for seasonal is valid. Figure 6 shows the forecasted visitor counts in dark blue, while the light gray lines indicate the test set or actual visitor counts. The results show that the forecast fits quite well for the first days, but begins to deteriorate as the days increase. For instance, the forecast is unable to capture the large spikes. Additionally, the forecast also shows very wide prediction intervals. Figure 7 shows the forecasted visitor counts for air_04341b588bde96cd (lower right hand corner) compared to other restaurants: air_f3f9824b7d70c3cf (top left), air_8e4360a64dbd4c50 (top right), and air_1c0b150f9e696a5f (bottom left). The results show a similar story that we saw with air store 04341b588bde96cd, the forecasts fit quite well for the first days, but begin to deteriorate as the days increase.

ETS: ANA Model

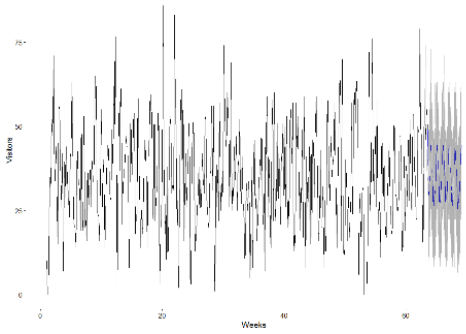


Figure 8

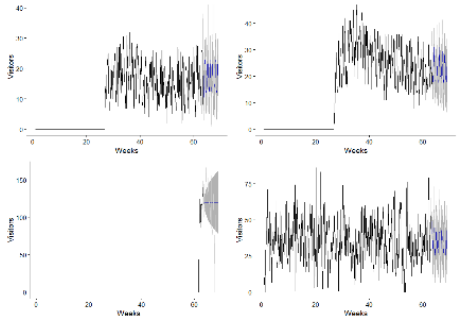


Figure 9

Observations: Figure 8 shows the forecasted visitor counts using ETS (ANA), which indicates addictive errors, no trend, and additive seasonality. The additive seasonality makes sense given that the magnitude of the seasonal fluctuations or the variation around the trend cycle does not vary with the level of the time series. The results of the forecast shows that the forecast does a much better job compared to auto.arima in predicting the large spikes. Additionally, the forecast also shows narrower prediction intervals. Figure 9 shows the forecasted visitor counts for air_04341b588bde96cd (lower right hand corner) compared to other restaurants: air_f3f9824b7d70c3cf (top left), air_8e4360a64dbd4c50 (top right), and air_1c0b150f9e696a5f (bottom left). The results show a similar story that we saw with air store 04341b588bde96cd, the forecast seems to predict the large spikes quite well, but seems to breaks down when a lot of missing data exists (bottom left).

Mixed Model (Best)

Given that auto.arima and ETS model produced somewhat different results and are diverse, I decided to combine the two models and create a mixed model. In other words, I used an ensemble modeling approach by taking the average forecasted visitors from the two models. By doing this, my hope was to create a more stable and robust model that would improve my predictions/accuracy. I implemented this model for the Kaggle competition by first changing the date format using mutate, reshaping the dataset from long to wide using dcast, and then creating a time series on the train data and a forecast interval of 39 days. Next, I created a matrix for the models, used parallel processing (DoMC), created the forecasts and applied it to all 829 restaurants, post processed the forecast table *(which involves making anything less than 0 = 0)*, changed the forecast data frame back from wide to long format for final submission, and then processed the sample submission file (R code: Kaggle Kernel borrowed from DSEverything).

Performance/Accuracy

Model Name	Train/Test	ME	RMSE	MAE	MPE	MAPE	MASE	AIC	AICc	BIC
ARIMA	Training	0.0590897	15.53463	12.48023	-Inf	Inf	0.7434168	3664.7	3664.84	3685.12
	Test	3.9624006	3.962401	3.962401	10.16	10.16	0.2416165			
ANA	Training	0.0939037	14.75522	11.69426	-Inf	Inf	0.6965983	5054.318	5054.832	5095.163
	Test	9.8869478	9.886948	9.886948	25.35115	25.35115	0.6028795			
MIXED	Test	3.62	14.76	10.62						

Figure 10: Accuracy Metrics for air_04341b588bde96cd

Kaggle - RMSLE		
Model Approach	Public	Private (Final)
Auto.Arima	1.010	1.000
ETS	1.017	1.013
MIXED	0.585	0.615

Figure 11: RMSLE for all 829 restaurants

Observations: Figure 10 shows the performance/accuracy metrics of the ARIMA, ETS (ANA), and MIXED model for air_04341b588bde96cd. The results show that ARIMA produced better quality scores than ETS as indicated by the lower AICc in the training dataset. In regards to the error statistics, ETS produced better error statistics (lower RMSE, MAE, MASE) compared to ARIMA on the training set. However, ARIMA produced better error statistics (lower RMSE, MAE, MASE) compared to ETS on the test set. As a result, although it seems like there's some mixed accuracy results between ETS and ARIMA, ARIMA has the slight edge since it had a lower AICc. In regards to the MIXED model, it did not perform as well compared to ARIMA and ETS for this particular air_04341b588bde96cd restaurant. However, when using the MIXED model across all 829 restaurants, it seems to be the better model according to RMSLE and my scores on Kaggle (see figure 11). This, combined with the advantages mentioned above for using a mixed model approach made this model the best.

Conclusion

Limitations

There are two primary limitations from the MIXED model that I designed. First, is that it only takes into account two methods: ETS and auto.arima. As a result, there could be other possible methods and/or combinations that could help increase accuracy and performance. Second, there are some cases where an ETS or auto.arima model would have performed better than the MIXED model. This was seen in our comparison of performance metrics for air_04341b588bde96cd.

Future Work

In regards to future work, there are three areas that could help improve the model. First, it would be beneficial to continue to explore different mixing and weight combinations. For

instance, instead of just using ETS and auto.arima, using SES and neural net could help make the model better to enhance model diversity (similar to diversifying a stock portfolio). Second, exploring different transformation techniques such as Box Cox or log can help stabilize the variance of each time series. Third, figuring out a way to handle the huge outliers and possibly use a mix of imputation methods could also improve the model, especially since using 0 to fill all NA's could be problematic for days where data was missing due to no data available (e.g., stores being added to the system) and not because the stores were closed.

Learnings

In the end, I learned five primary things from building these models. First, I learned how to forecast a large dataset that contained multiple different sets of time series and how challenging it can be to accomplish this feat. Second, I learned that using one modeling approach is not sufficient and that mixing other modeling approaches can yield more accurate results by adding stability and robustness. In other words, taking a "one size fits all" approach such as just using ETS or auto.arima will not produce great results for a dataset that contains multiple different set of time series. Third, I learned that there are some cases where ETS performs better and other cases when auto.arima may perform better. We can never assume that one modeling approach will do better than the other. Fourth, I learned how to use the dcast function to transform a dataset from long to wide format and how to incorporate parallel processing. Lastly, I learned that having a large number of missing values within a data set can really alter our forecasts. Knowing how to address these missing values accordingly by conducting a proper missing value EDA can help make our forecasts more accurate.

References

1. Berry, W.L., Mabert, V.A., Marcus, M. (March 1979). Forecasting Teller Window Demand with Exponential Smoothing. *Academy of Management Journal*, 22, 129-137. [Peer Reviewed Journal].
2. Dharmaratne, Gerard S (1995). Forecasting tourist arrivals in Barbados. *Annals of Tourism Research*, 22(4), 804-818. [Peer Reviewed Journal].
3. Gounopoulos,Dimitrios., Petmezas, Dimitris., Santamaria, Daniel (April 2012). Forecasting Tourist Arrivals in Greece and the Impact of Macroeconomic Shocks from the Countries of Tourists' Origin. *Annals of Tourism Research*, 39(2), 641-666. [Peer Reviewed Journal].
4. Karadas, K., Celik, S., Eyduran, E., and Hopoglu, S. (2017, Oct 31). Forecasting Production of some oil seed crops in Turkey using Exponential Smoothing Methods. *Journal of Animal and Plant Sciences*, 27(5), 1719. [Peer Reviewed Journal].
5. Nowman, K.B., Van Dellen, S. (2012). Forecasting Overseas Visitors to the UK Using Continuous Time and Autoregressive Fractional Integrated Moving Average Models with Discrete Data. *Tourism Economics*, 18(4), 835-844. [Peer Reviewed Journal].