Car Traffic Exercise:

The model for the daily forecast of traffic in the parking lot at 1001 Company Ave is complete. From historical imagery, our team has constructed a model to count the daily number of cars in the parking lot for the past five years. In addition, by gathering relevant data on weather and sky condition, we were able to extract useful information for our forecast model. In doing so, we find evidence of a strong association between sky condition and car traffic. Additional insights are detailed in this report. The rest of the document will be organized in the following manner: 1.) What are the trends – seasonally, monthly, yearly – of car traffic in this lot? 2.) Which features contribute to car traffic and how strong are these relationships? 3.) What forecast models will we use? 4.) Model performance and prediction for July 5.) Further recommendations

I. Car traffic trends

We find that there are no noteworthy monthly trends for car traffic: that is, the average monthly vehicular traffic is not heavily skewed and is relatively constant across months. There may have been an expectation that car traffic would increase substantially in later months due to summer and more holidays, however this is not strongly supported by data. Although there is a slight statistically significant increase of around 5 additional car units for the months of July to December versus January through June – as confirmed by a t test – this effect is marginal as seen in Fig. 1. The main seasonal trends we see are across years and days. The first finding is that the average car count for the parking lot has decreased for the past five years. Car count reached its peak in 2012 at a daily car traffic of 137 and has seen a 30% reduction to a daily car count of 92 for the first 6 months of 2016 (Fig 3). In addition, the parking lot sees a slight increase in car traffic on the weekends (Fig 4). On average, the mean car count for Saturdays and Sundays is ~5 more than on weekdays.

Fig. 1

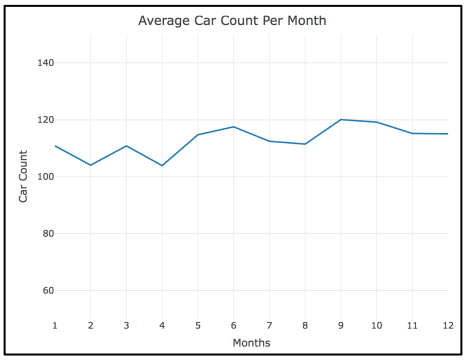


Fig. 2

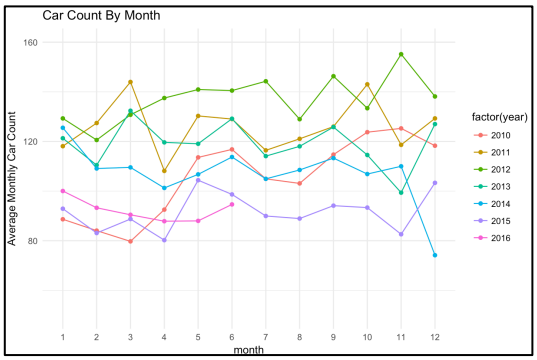


Fig. 3

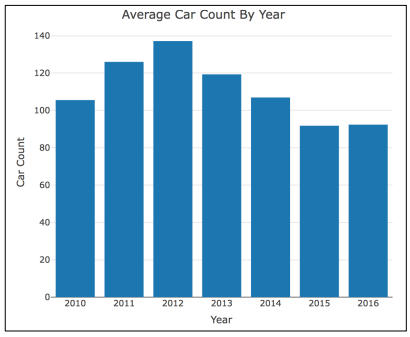
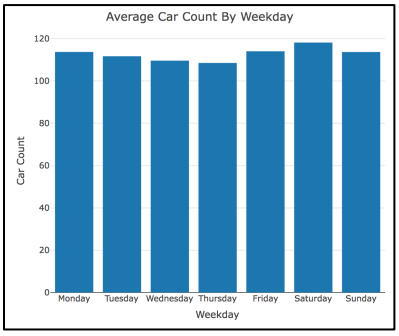


Fig. 4



II. Features' Importance

To investigate which features to include in our forecast model for car traffic, the team has used a combination of linear regression and exploratory analysis to determine which variables contribute to car traffic. From the previous section, we concluded that there are dependencies of car traffic on time variables: there is evidence that car count has daily, yearly, and seasonal trends. Fitting a regression model on car count on the categorical time variables of weekday, month, and year, has confirmed the hypothesis that these features help predict car count: their individual p values and

the F statistic are below .05 which indicate the rejection of the null hypotheses and an existence of a relationship between time and car count.

Moving forward, we investigated the association between weather and clouds on car traffic for the parking lot. The weather variable has a range of [-3.2, 4.3] and a standard deviation of 1. By binning weather to 15 buckets consisting of ranges of 0.5, we find that the car count is relatively constant across all weather brackets, except for the last bin where a high weather sees a dramatic increase in car traffic (Fig 5). This seems to indicate that an extremely high weather value may lead to more car traffic. However, when investigating further, we determined that there is more to the story.

Fig. 5

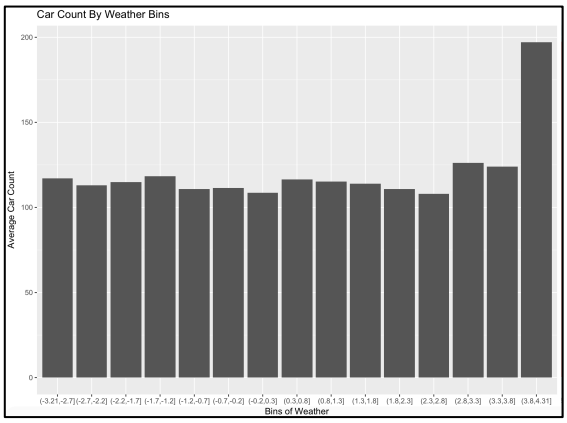
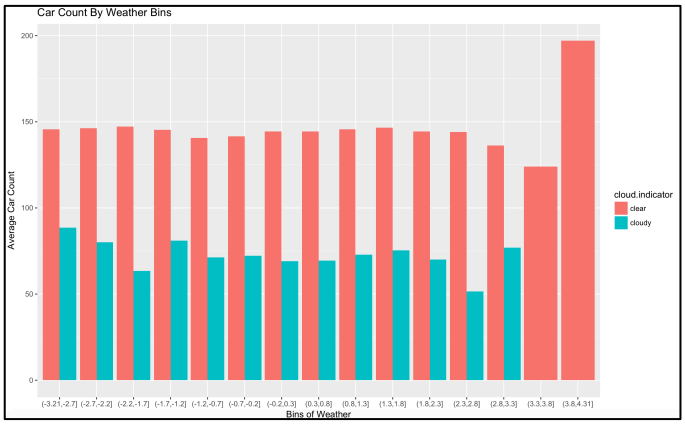


Fig. 6



The increase in car count for the extreme weather bracket is actually driven by the fact that the bracket has more clear days than cloudy days. In Figure 6, we see that what appears to be driving the increase in the car count is the fact that the last two weather bins from 3.3 to 4.31 consists of all clear days and that clearer days consistently has higher car traffic. Thus, we can conclude from these graphs that weather appears to not be associated with car traffic, while sky condition is highly related with car traffic. These results are confirmed by performing linear regression. When regressing on solely weather we find that its p value is 0.732, much larger than the 0.05 needed for an association: this suggests that weather is not related to car counts. When including the cloud indicator variable in the regression, we discover that 1.) cloudy or clear skies is a very strong predictor of car traffic with its p value of <2e-16 and 2.) the irrelevance of weather shoots up to 0.943 which suggests truly no relation when controlling for clouds. Throughout the 5 years of historical data, the mean car count during clear days is 141 while on cloudy days is 70. This represents a doubling of car traffic on clear days which cements sky condition as an important feature in our model. Particular attention would need to be spent on predicting whether a day is cloudy or not to forecast car traffic at this parking lot.

III. Forecast Models

From previous sections, we have determined that our final model for car count would consist of the relevant features of week day, month, year, and presence of clouds. For the new month of July 2016, a measure of whether a day would be cloudy would need to be predicted since this variable presents a strong association with car traffic for the parking lot. To investigate this, the team sought to find relationships with clouds and other variables in our data set. Our first discovery is that whether a day is cloudy or not depends on whether the days before were cloudy: historical data showed that cloudy days are often followed by more cloudy days. This suggests that there is an underlying conditional probability distribution within the phenomena of clouds: that is, the probability that a day will be cloudy is dependent on what occurred previously. To test this hypothesis, a logistic regression was run to predict whether the next day would be cloudy or not, based on whether the past 5 days was cloudy, as represented by lag variables. The final model is:

Eqn. 1

$$log\_odds(\textbf{cloud.indicator}) = \beta_0 + \beta_1 \times \textbf{month} + \beta_2 \times \textbf{year} + \sum_{i=1}^{5} \beta_i \times \textbf{Cloudy\_Lag}_\textbf{i}$$

It was discovered that the presence of clouds on a previous day helps predict the probability that the next day will be cloudy as gleaned from its corresponding p value below .05. This suggest a sequential modeling approach such as using standard time series models such as ARIMA or random walk processes to predict the cloudiness of future days based on the past. First, to examine the potential of ARIMA to predict clouds, the team examined the stationarity of cloudy days, specifically whether there are any differences in mean and variances throughout time. By converting the categorical variable to a proportion of cloudy days, a trend is observed.

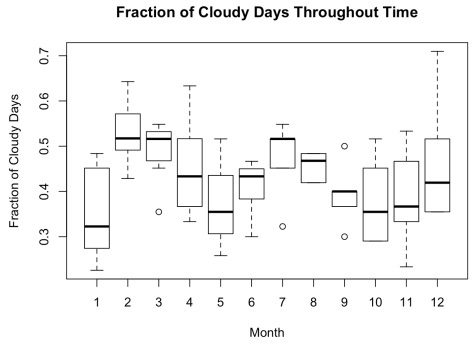


Fig. 7

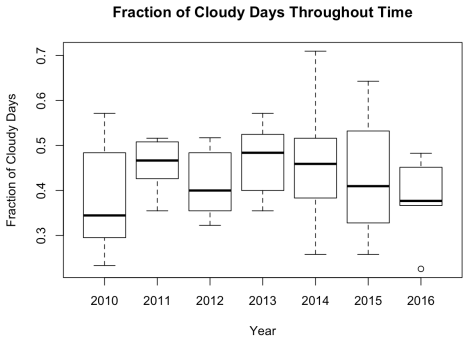


Fig. 8

Figure 7 illustrates a distinct sinusoidal pattern between month and fraction of days that were cloudy for that month in the parking lot of the past. This suggests that there may be a seasonal trend in which the current month has an effect on how many cloudy days can be expected. Thus, when forecasting the amount of cloudy days for the next month, this monthly dependency needs to be accounted for. Secondly, it appears that the fraction of cloudy days throughout years is inconsistent, and that the yearly mean fraction of cloudy days see a variation across time (Fig 8.).

This relationship appears to be more of a random process as we would not expect the proportion of cloudy days across an entire year period to affect what would be expected next year. From the previous logistic regression, the time dependency of the underlying weather system that govern clouds is more localized in nature and depends heavily on what occurred in the previous days, thus across years would be too wide of a scope. However, by treating year as its own independent system that have a distinctly defined mean proportion of cloudy days and creating an interaction term between sky condition and year, we can integrate the differences in mean proportion of cloudy days across different years periods. When, integrating this interaction term in the model, we find that this interaction and statistically significant and boosts the model fit with an increase in R^2.

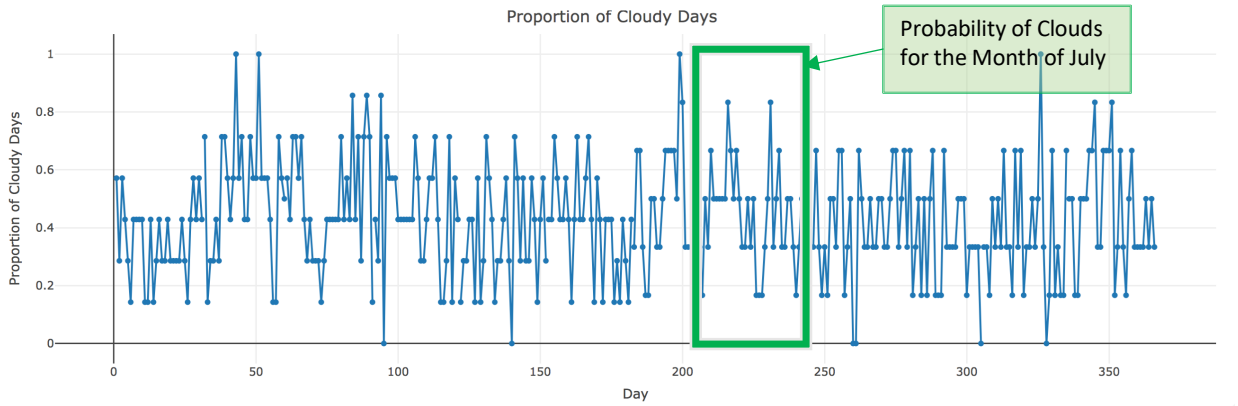


Fig. 9

To predict the cloudy days for the next month, the team has decided to forego a Markov or time series analysis due to added complexity brought from stationarity of cloudy trends, in favor of a model that predicts the probability of clouds on a given day. This is an especially important decision, as weather is a complex atmospheric system and to predict which days might be cloudy and not is particularly error-prone. Rather than having the cloud indicator to be a categorical variable as before, the team has changed its encoding to be quantitative in the range of 0 to 1. This new variable will represent the probability of clouds on a given day. Historical data will still only have values of 0 and 1 since the probability that a past day was cloudy is already known. However, new data for the next month will be obtained by using historical data of the proportion of cloudy days for the specific month-day combination. Figure 9 illustrates the section of the plot that we will use to determine a baseline estimate of the probability of cloudy days in July. In this figure, we also see a confirmation of the time-dependent pattern of the presence of clouds. Even averaged across means for the past 5 years, we find that probability of clouds on a particular day is a function of the day before, as seen by the sequential increasing and decreasing probabilities. The final forecast model is seen below.

Eqn. 2

$$\textbf{Car Count} = \beta_0 + +\beta_1 \times \textbf{weekday} + \beta_2 \times \textbf{month} + \beta_3 \times \textbf{year} + \beta_4 \times \textbf{Pr(Clouds)} \times \textbf{year} + \epsilon$$

Encoding the cloud indicator as a numeric value averages out the effects of clouds on car count. Hence, although the total monthly car count of the forecast is relatively accurate, each day's forecast can be interpreted as a weighted averaged across days due to this numeric encoding. To obtain a model that predicts cloudy days throughout time in a more absolute manner, a step wise function was added to Pr(clouds). Any value below 0.25 is treated as zero probability of clouds, while any value greater 0.6 is set to a probability of one. These thresholds were set using cross-validation where the dev set was populated by 2016's historical weather data. Step functions for a range of the values cause the model to have the advantages of both flexibility and rigidity by allowing for more extreme daily dips and peaks. An alternative model to predict clouds would be to iterate the previous logistic regression model (eq. 1) throughout time with just one lag variable.

IV. July Forecast

To assess model performance and perform model selection, the team used the cross-validation approach of forward chaining. The data was trained 6 times for all data up to January until June 2016. The trained models were used to forecast the car count for the next month. A depiction of the predicted and actual car counts for June is shown below. As you can see, this model benefits greatly with the use of a step function. The mean absolute error of our final model is 30 car units, and the RMSE error is 37 car units. On average, we expect the actual car count on each day to differ from the forecast by up to 37 units.

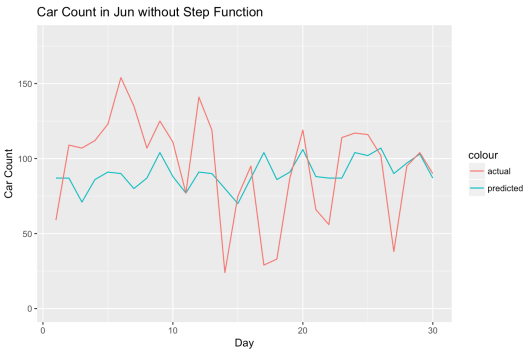


Fig. 10

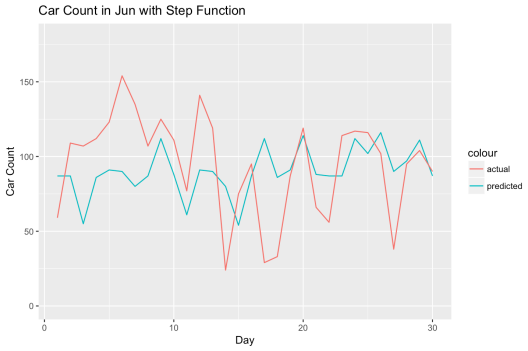


Fig. 11

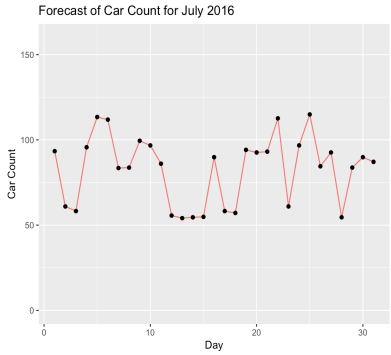


Fig. 12

V. Recommendations

Since the presence of clouds is highly related with car traffic, the aforementioned model can be made more robust if monthly weather predictions were obtained through local sources and inputted in the model directly. By adjusting the cloud indicator variable using real-time data from weather forecasts, the model would benefit from more advanced numeric models that integrate various atmospheric factors. This leans to a more dynamic modeling approach. In addition, the team plans to fine-tune the threshold values monthly as new data is inputted in order to boost model performance.