Predicting Disney World Ride Wait Times

Catherine Magsino

I'm a huge Disney fan. I try to go to one of the Disney parks once a year, or at least every other year, with the goal of riding all the rides at least once. However, with the popularity of attending these parks comes long line rides that may last between a few minutes to a few hours. With my experience over the years, I've been able to come up with some of my own techniques to try to maximize my time. However, I've always wondered, is there a science to this?

Well, we just finished our third week of the Metis Data Science Bootcamp, and presented our 2nd projects last Friday. The learning goals of this project were both web scraping and regression, leading to the perfect opportunity to solve my question above.

Data Acquisition

As I scoured the internet to look for data that may help me solve my question, I decided to use two websites: Weather Underground and Touring Plans.

Since I was interested on how weather affects wait times, I web scraped hourly data from Weather Underground, using both Selenium and Beautiful Soup. I wanted to focus the analysis on the busy summer vacation months, so I chose to scrape data from May 1, 2018 to August 31, 2018. The pickling technique came in handy here, as sometimes there were unexpected errors that would occur as I scraped through each time period. Instead of having to run the entire period over and over again, I ran the web scraper for half a month a time, pickled each period, and combined them all at the end. The resulting data had to be cleaned up a bit, requiring me to strip unwanted characters, renaming the columns, and converting the data types.

Touring Plans provides wait time data for 14 rides across the 4 parks in Walt Disney World, along with related metadata pertaining to each particular day (these all came in separate files). For each ride, they provide the posted wait time every few minutes that the park is open. In addition, a Touring Plans employee actually has a job to wait in rides and record the actual wait times. I decided to combine these together into one Wait Time column. I also focused on only one ride for this project – Splash Mountain in Magic Kingdom. Since my weather data only included May-August, I limited this data to the same dates.

Once all the data was retrieved, I merged them all into one dataframe. I merged the weather data into the Touring Plans Wait Time data on a new column called "Date and Hour," which combined both the date and the hour of the day. I also merged the daily metadata based on the date column. Through exploratory data analysis, I also removed any outliers from the data, leaving with me with ~18K observations.

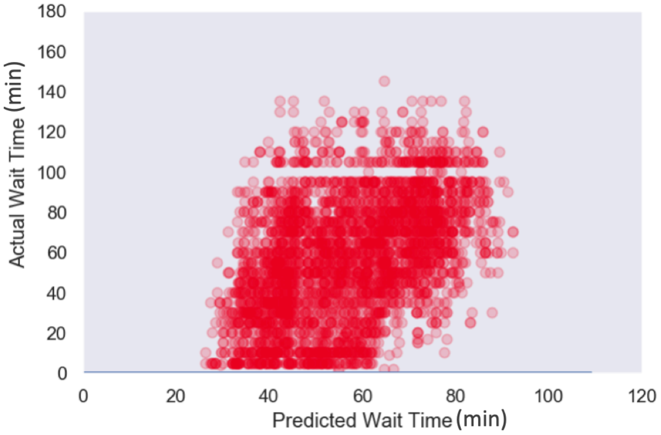
Multivariate Linear Regression

The multivariate linear regression model formula is:

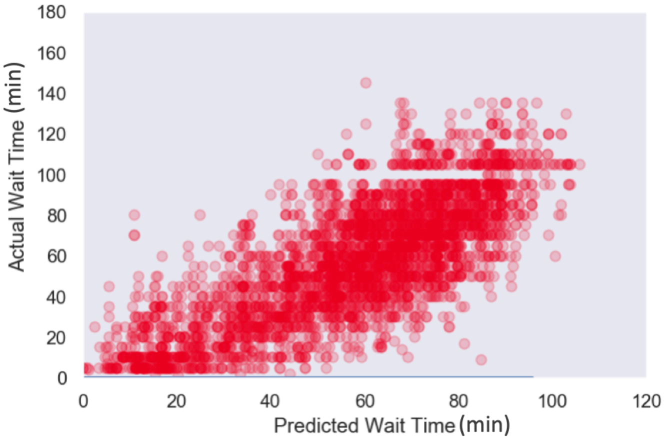
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n$$

Here, $\hat{y}$, is the predicted Wait Time, the x's are the features or columns in the data, β_0 is the intercept, and β_1, β_2, etc. are the coefficients of each feature.

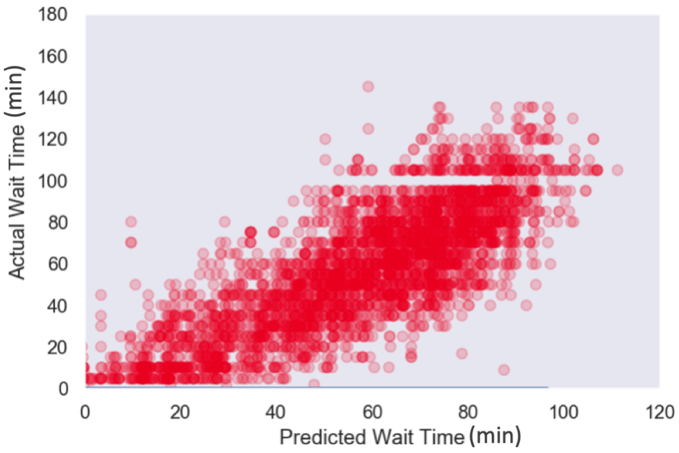
Using Ordinary Least Squares Regression, and the train/test split method (80% train, 20% test), I first focused on only using the weather features in my model: Temperature, Dew Point, Humidity, Wind Speed, Wind Gust, Pressure, and Precipitation. This resulted in an R^2 of 0.23 on my train set, and an R^2 of 0.20 on my test set, meaning the weather features alone explain 20% of the variance in wait times at Splash Mountain. My mean squared error (MSE) was 709.8 and my root mean square error (RMSE) was 26.6 minutes.



I then proceeded to add in my time features of Month, Day of the Week, and Hour of the Day. Since these were categorical features, I had to add them in as dummy variables. This resulted in an R^2 of 0.60, three times the R^2 of my first model. My MSE was halved to 354.0, and my RMSE was reduced to 18.8 minutes.



In my final model, I added the Ticket Season feature, which indicates if the One-Day ticket season is Peak, Regular, or Value. I also removed the Temperature and Pressure features from the model, since they both had a p-value of less than 0.05. This further increased my R^2 to 0.64, with my final MSE being 319.7 and my RMSE at 17.9 minutes. In addition, my residuals were normally distributed supporting my decision to use linear regression.



Conclusion

After examining the results, my regression model suggests the following guidelines for this time period for Splash Mountain:

Feature	Lower Wait Times	Higher Wait Times
Month (during this period)	May	June
Day of the Week	Sundays	Wednesdays
Hour	Before 10AM, After 8PM	12PM-5PM
Ticket Season	Value	Peak
Weather	More humid, rain	Windy conditions

Future Work

Future work to improve the model may include incorporating:

- Disney events – several events occur throughout the day, such as parades, fireworks, and other shows. Since many park visitors are watching these shows during these times, this reduces the number of people that are riding rides.
- New rides/attractions – Many rides and new lands have been opening at the Disney parks.

- Other rides and locations – Since this model was only based on Splash Mountain, I’d like to do the same for the other rides, as well as Disney World vs Disneyland, since each location brings different types of visitors.
- Full year/multiple years – I’d extend the time period to a full year to get the full annual picture.