Optimizing a Disney Day: Final Milestone

Sophia Weidner

Bellevue University

DSC 630: Predictive Analysis

Professor Andrew Hua

March 4th, 2023

Introduction and Business Problem

As a frontline cast member who was hired in the midst of the COVID-19 pandemic, I have been able to witness firsthand some of the major changes the Walt Disney Company has implemented into its parks. Whether or not these decisions are popular is up to the public, but regardless, going to a Disney park is more expensive than ever and budgeting a trip is becoming a guest's first and foremost priority. If a majority of consumers aren't getting the "bang for their buck", public opinion will be negatively impacted and Disney is at risk of losing revenue. Some of the factors that may influence a guest's opinion could have to do with special events, extra "free" perks, the time of year the guest visits, the amount of other guests in the park, etc. A lot of these factors fall within the grasp of Disney to control, especially monetary factors, but something as simple as the weather could have such a huge impact on someone's trip; for example, there's a difference between visiting during hurricane season and a chilly Florida winter.

Because things continue to get more expensive, and the new features in the park aren't increasing at the same rate, many guests feel frustrated with their Disney experience. To maintain a loyal consumer base, it is important more now than ever to focus on optimizing a guest's day to ensure their return to the parks. If you just hike prices without any incentive to return, you simply won't make any money. The one thing about Disney, though, is that it relies heavily on a guest's ability to plan their day from beginning to end. A lot of research goes into a seamless Disney trip, and even then, a guest can encounter unexpected circumstances. Disney has made steps to assist their guests in optimizing their day, but there are still so many influences that Disney does not include in their *My Disney Experience App*. My goal through this analysis is to

analyze trends and provide guests with more information on those previously stated unexpected circumstances.

Data

The data was obtained from https://touringplans.com/walt-disney-world where anything from ride times to crowd levels are recorded to help guests plan their trips. The dataset obtained contains anything from the day of the year, to the type of "season" it is (e.g. spring break, holiday, etc.), and it even includes the temperature it was that day. These seemingly trivial details can severely impact a guest's trip; and since the goal is to optimize a Disney World vacation to the guest's liking, all of these little things matter. The more details we can include, the more we can specifically cater to the guest planning their trip.

Preparing the Data

I have worked with this data in the past, and the unfortunate attribute to this dataset was that it contains a lot of NaN values. In my approach, I decided to first split the dataset into categorical and numerical variables and either fill or delete the empty values as I see fit. After I split and cleanse the dataset, I would then explore the data further.

I first went through the categorical value dataset and discovered that nearly every row had some sort of empty value. The categorical data points are all events of some sort, such as "WDW Race", so for the NaN values, I replaced them with "No event". After replacing all of the NaNs, I created dummy variables so that a proper analysis could be run on the data.

After cleansing the categorical data, I moved my cleansing process to the numerical data. Before replacing NaN values, I removed a few columns. Three of the four Walt Disney World parks currently do not have any parades. Since I am basing my analysis on planning a Disney trip today, I deemed it was necessary to remove those aspects so it would not impact a guest's

decision regarding their vacation. After those data points were removed, I replaced all NaN values with the median. I chose the median because there were very few NaN numerical values, so replacing the empty values with the median would have little to no impact on the model. I rejoined the numerical and categorical datasets back together to form a new, cleansed version of the original data.

Visualizations

I primarily work in Attractions in the Walt Disney World parks, and one of the most common comments I receive from guests is usually about the crowds. Larger crowds means longer waits in lines, which means less time for attractions, meals and other activities. I thought it would be interesting to do a deeper dive into the capacity for the parks, especially before, during, and "after" the COVID-19 Pandemic. The following examples are the visualizations created for the Magic Kingdom Park.

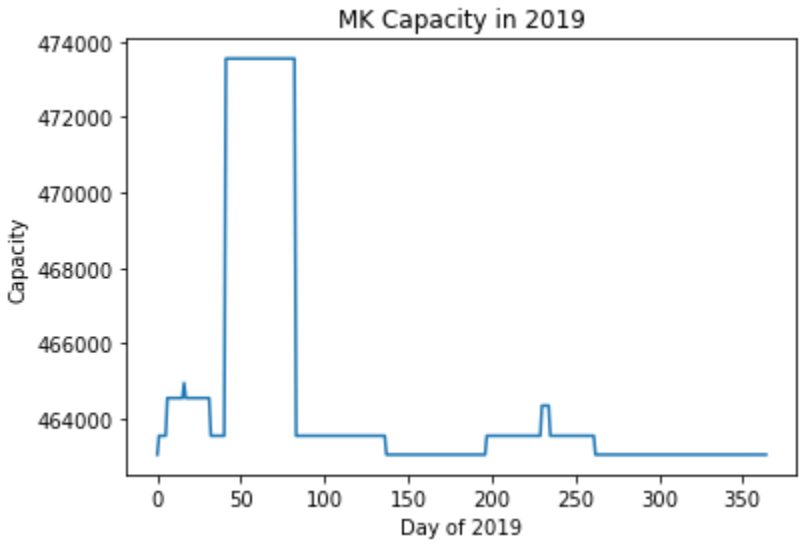

Figure 1. Magic Kingdom Capacity throughout 2019.

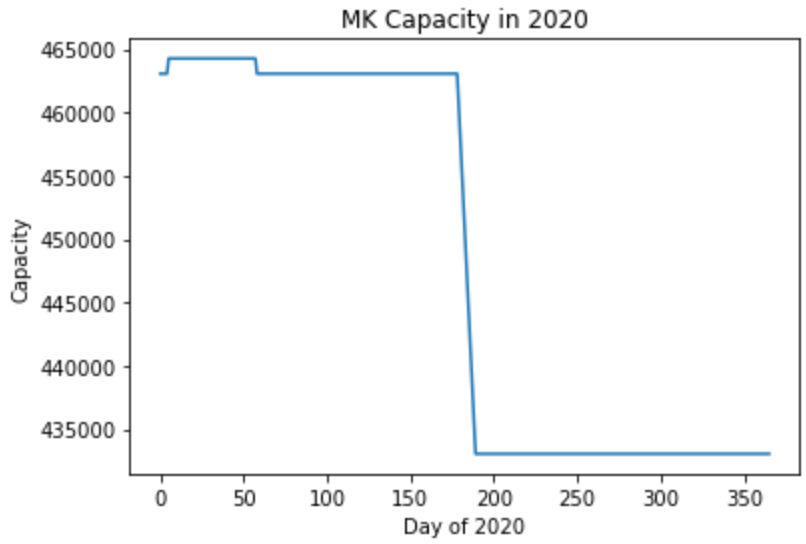

Figure 2. Magic Kingdom Capacity throughout 2020.

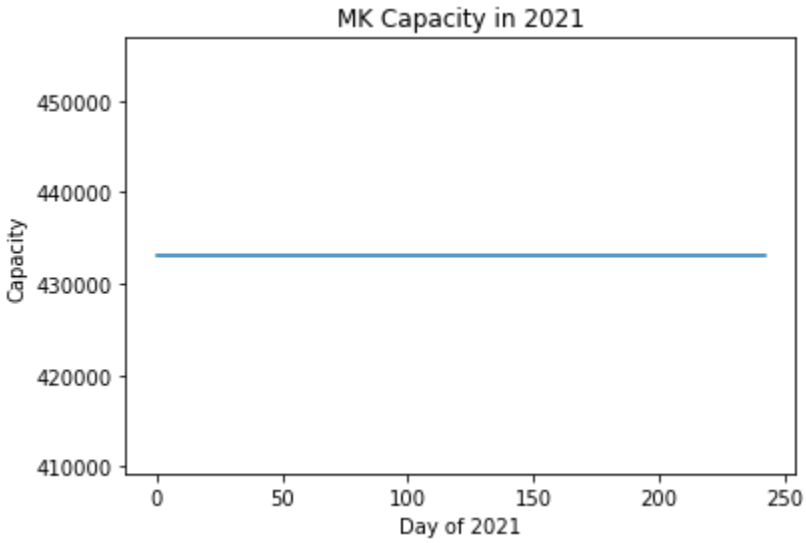

Figure 3. Magic Kingdom Capacity throughout 2021.

As you can see, the capacity has fluctuated throughout those few years. In Figure 1, the peak capacity was around 474,000 guests. In 2020, that number sharply decreased during the closure of the parks. Once the parks reopened, the capacity remained steady for 250 days in 2021 at around 450,000 people. The reason capacity is important is especially highlighted in Figure 1;

the capacity changes depending on the time of the year. If someone wants to go at a time with lower crowds, they would not want to visit Magic Kingdom during the day 50-100 mark.

Model

Planning a vacation involves many decisions. Because of this, I believed that a Decision Tree Model was the best fit for both my business problem and my dataset. According to MindTools.com:

> (Decision Trees) provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options. They also help you to form a balanced picture of the risks and rewards associated with each possible course of action. (*Decision Tree Analysis*)

Since several different aspects go into preparing a trip, details such as the weather, the crowds, and even an event like Fireworks can provide different outcomes depending on what a customer wants. I proceeded with the Decision tree analysis.

Analysis

I began with splitting my data into testing and training sets. I split the data based on the data point “Week of the Year”. The reasoning for this was that a “week”, or around seven days, is an average amount of time for a family to go on a Disney vacation. When I fit the model to the data, it returned an accuracy of 0.79. The accuracy was a little lower than I would have preferred, but the number does show that the model was neither over nor under fitted. I also calculated the R squared factor as well as the RMSE. The R squared value came back at 0.99, which can be an indication that the quality of the statistical measure may be unfit for the model (Yse, 2020). The RMSE was 0.47, which shows the Decision Tree model can predict for the dataset accurately (ResearchGate).

Conclusion

Overall, I believe that the dataset that I obtained was the root of many issues. There were a plethora of NaN values which I believe is what caused the Decision Tree Model to be as inaccurate as it was. For this type of analysis to be more meaningful, I believe we would have to either split the data set by years, or even months. However, the work that was done was still meaningful. It does show that there is a way to optimize a guest's Disney World vacation based on the series of decisions that goes into planning.

Ethically, this dataset and analysis do run into some challenges. Again, there were a lot of NaN values in the original data frame, and that causes a lot of issues. A lot of values were replaced, especially in the categorical dataset, which has an impact on the overall accuracy of the Decision Tree model. And, if those values weren't replaced, they were removed, which was completely up to my discretion. I tried to remain as impartial as possible and only remove categories that are not occurring in Walt Disney World today, such as the parades in the three parks, but that is still data removed and not analyzed. These same issues would persist if the project were live in production today, because there are days where there are no events, which would cause a NaN to appear in the data. To combat this, data columns could be combined or removed entirely, since details like "Capacity" are a catch-all for some smaller categories. For example, a race event means more people will be in the parks, which is reflected in the capacity. By manipulating how the data is input into the frame, this tool could become more accurate.

References

Decision Tree Analysis. MindTools. (n.d.). Retrieved March 4, 2023, from https://www.mindtools.com/az0q9po/decision-tree-analysis

What's the acceptable value of root mean square error (RMSE), sum of … ResearchGate. (n.d.). Retrieved March 4, 2023, from https://www.researchgate.net/post/Whats-the-acceptable-value-of-Root-Mean-Square-Error-RMSE-Sum-of-Squares-due-to-error-SSE-and-Adjusted-R-square

Yse, D. L. (2020, August 31). *Modelling regression trees*. Medium. Retrieved March 4, 2023, from https://towardsdatascience.com/modelling-regression-trees-b376e959d02e