Theodore Koby-Hercsky

March 2, 2022

DSC550-T301 Data Mining (2223-1)

Dr. Brett Werner

Introduce the Problem

My data mining project is on the reviews that are received from the guests that have visited the Walt Disney Theme Parks. As the focus point of my project is on the reviews that are received from the guests. With the problem of this project being how Disney employees can audit the mass number of reviews they receive from their guests, but I would like to further investigate these reviews and determine which parks have the highest rating and lowest. While also determining where these reviews are coming from by meaning where the guest is coming from, their age, how many are in their party and more.

Justify why it is important/useful to solve this problem

This problem is important to solve as it will indicate which parks are performing up to par and which parks need extra guidance to bring their reviews up. As by solving this problem it will be useful for the cast members to figure out what issues guests are facing and if Disney can fix the problem easily by adding more workers to an area or even retrain workers to better suit the guests need. While all the workers at Disney and guests that visit know that Disney aims to please and does their best to accommodate and the guests needs so they will continue to come back for years to come.

How would you pitch this problem to a group of stakeholders to gain buy-in to proceed?

Have you ever wondered how many reviews the Walt Disney Company receives a day? Well just to let you know it is a lot! As the problem I am working to fix is determining what parks receive the highest and lowest average reviews. Which will allow users to determine which parks are doing great and which parks will need some improvement. By auditing guest review cast members will be able to improve guests experience and increase the parks average review rating. But you might be thinking why does this matter and why would I care to invest money in this? Well for one by increasing guests' experiences and increasing the overall reviews guests will be more likely to visit for years to come. As one thing Disney needs is returning guests as with ought guests returning to the parks Disney's revenue would plummet. By further researching and continuing to research and improve on Disney's bad reviews raise their overall profit meaning Disney will be bringing more guests in and having them return.

Explain where you obtained your data

The data I used for this project is from Kaggle and includes 42,000 reviews for the three Disneyland branches such as Disneyland California, Disneyland Paris, and Disneyland Hong Kong. Which has two parks at each of the Disneyland branches but does not include the Walt Disney World branch in Orlando Florida and other parks over seas such as Disneyland Tokyo and Disneyland Shanghai. On another note, this data does include a unique review ID that distinguishes each review that can be used to determine if we have any duplicate review IDs. Next this data set has an initial rating that is on a scale from 1 to 5 with 5 being the highest satisfied a guest can be. As additional variable are the year/month and the review location of the

visitor. As these variables tell us when the guest visited and how far they traveled to get to the park. While the most important variables are the review text and the Disneyland branch. Which tell us which park the guest visited and the review they left allowing us to determine if it is good or bad.

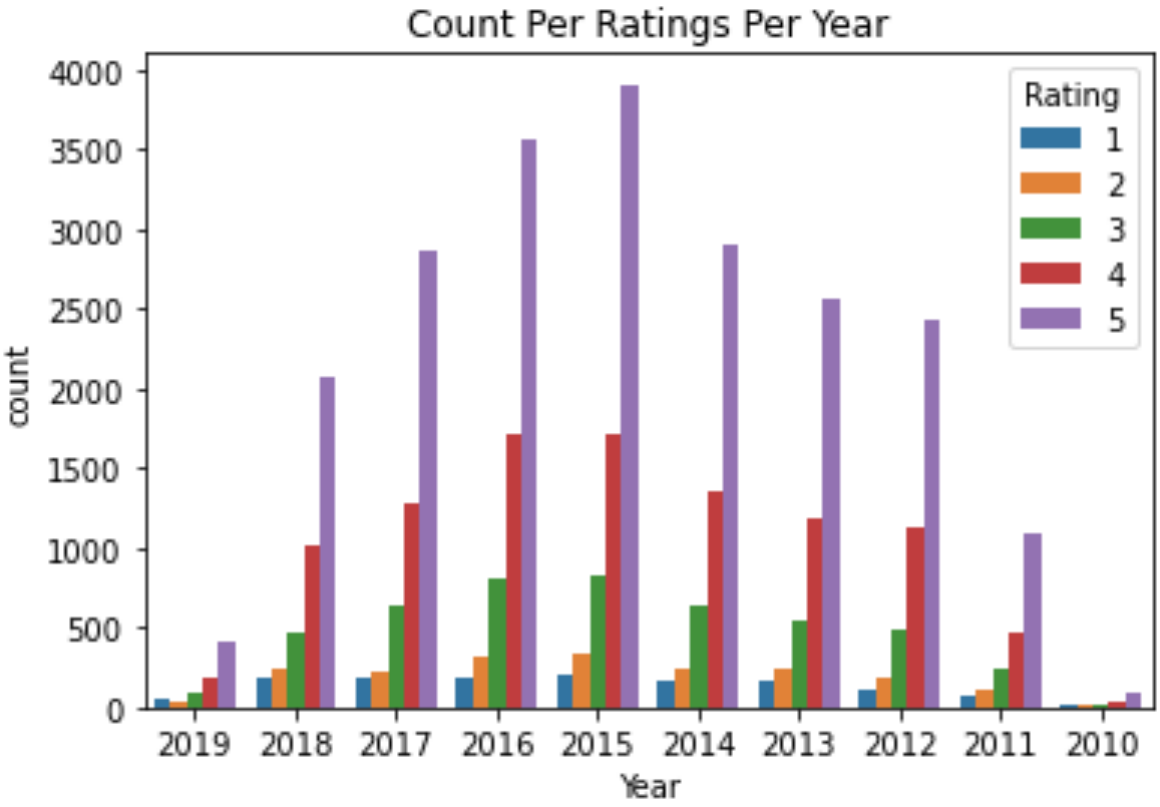
Link to data from Kaggle: https://www.kaggle.com/arushchillar/disneyland-reviews

Summary of Milestones 1-3

EDA: include any visuals you think are important to your project

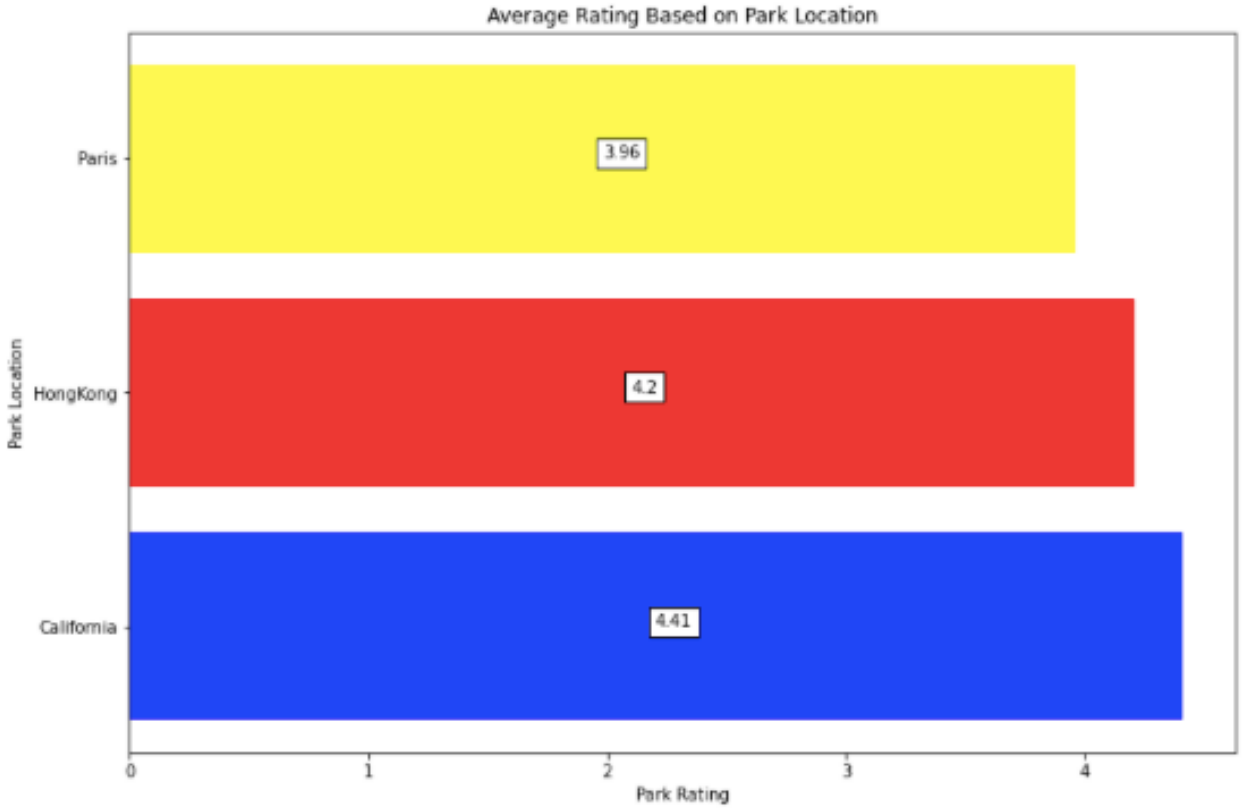
In this data mining project model, I will break down the guests' reviews using Data Visualizations using matplotlib and seaborn that will create graphs that focus on different aspects of the reviews Disney has received. When I first imported the Disney review data, I found using shape that the data set had 42,656 reviews from all the Disney parks included. Next, I determined that my year and month variable had 6.12% missing values in the column. As for this variable I ended up splitting my year and month variable into two separate variables. As seen below one of the data visualizations I created focused on the average rating per year for all the parks. When creating this visualization, I used seaborn count plot that shows as the years go from 2010 / 2011 to 2015 we see a rise in reviews for the theme parks. But as we go on from 2015 to 2019, we see a decline in reviews. In this visualization I updated it and added a title to my visualization and

variables.



Another informative plot that I created was the average rating based on the park location using a horizontal bar chart. Which showed California had the highest rating out of the three parks with a 4.41 and Paris in last with an average rating of 3.98 out of 5. These scores can be justified as California is on of Disney's main parks, so they tend to have better customer service and reviews since the cast members at that location are trained by individuals that have been with the company since it started. Next, I will dive into the preparation I took to be able to build my

models from the Disney review data.



Data preparation

I started the data preparation process for the Disney review data by checking the data type of my variables and checking for missing values within my variables. In my EDA section of this report, I created two new variables to split out the year and month and decided to use front fill to replace the missing values. Next, I performed extraction and selection steps to transform my review variable. This was completed by removing the punctuation, inputting lowercase, number removal, creating tokens, removing stop words, and converting reviews to text. As this will condense the review and make it easier to determine if the review is positive, negative, or neutral. In the data preparation process I also created a new variable for the review polarity to determine if the review is positive or negative. While also creating the Vader rating and polarity

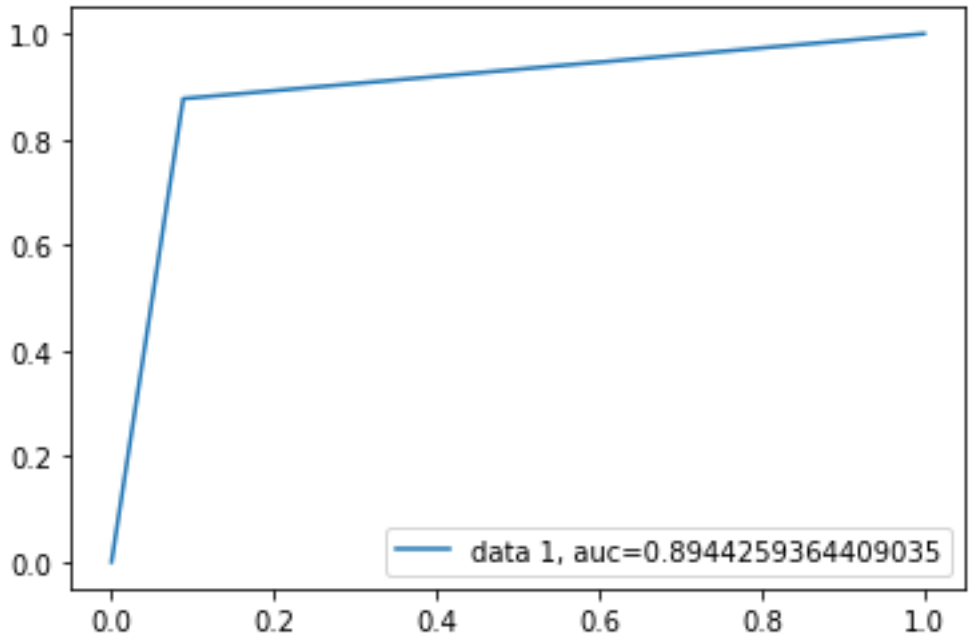
to show if the review is positive or negative along with a more in-depth rating. Last before the moving on to building my model I used get_dummies to create dummy variables for the three parks to be used in the model.

Model building and evaluation

In my model building for my Disney Park data, I created several models with one that focuses on the review polarity rating and the updated review. While first using label encoder on my review and Vader polarity which will allow me to then split my data into a target and predict variable with training and test sets. Which I used on my TFIDF Vectorization to the training data that shows the vector (34108, 33189) for x_train with 1774397 stored elements that are in a sparse row format. Next, I created a Random Forest Analysis that shows the accuracy of the training sets of 99.87% which is very high while decision tree and see the accuracy of the train sets is a little less accurate as it is 99.51%.

I also decided to create three different splits for each park and the updated reviews that started with a TFIDF Vectorization and set the training data in a logistic regression that told me that Hong Kong had the highest accuracy at 93.76% then Paris at 92.01% and last California with 91.93% accuracy. Next, I pulled the model accuracy for the test set for each park and found that the highest accuracy was from Hong Kong again at 92.64% then Paris as 90.02% and California in last again at 89.59%. Following that I created a Confusion matrix for each park and found that the true class for Hong Kong had the highest negative reviews coming in at 6,516 while California had the lowest true negative reviews at 4238. On another note, on the predicted side, we saw that the highest positive reviews came from Paris at 597 and the lowest positive was California at 477 predicted. A classification report ran on all three parks and found the highest recall score at Hong Kong coming in at 0.99 for negative reviews while the highest precision

came from Hong Kong as well at 0.94 for positive reviews. Last the ROC curve told us that the accuracy of each park was in the upper 80% with California being the highest at 89.44%. Overall, I felt these models showed a lot and wasn't what I was expecting as we learned that Hong Kong had the highest accuracy in most of our models. While our ROC curve showed California to be the most accurate as seen below.



Conclusion

What does the analysis/model building tell you?

When working through my analysis and model building, I learned that regarding my review's variable and my Polarity review rating when used with Random Forest Analysis I found the highest accuracy at 99.87%. While also deciding to create several other splits for each theme park location such as Disneyland California, Paris, and Hong Kong that in a logistic regression Hong Kong had the highest accuracy at 93.76%. While the model accuracy for the test set found that the highest accuracy was from Hong Kong again at 92.64% then Paris as 90.02%. While the confusion matrix showed and told all regarding each of the theme park locations as the true class

for Hong Kong had the highest negative reviews coming in at 6,516 while California had the lowest true negative reviews at 4238. Which told us that either Hong Kong has a lot more visitors that leave reviews or they have more guests that where not satisfied with their experience. As California on the other had over 2,000 less negative reviews as this makes sense as California parks is the birthplace of Disneyland meaning for the most part know how to please their guests and what to change. While on the predictive side of the Confusion Matrix I learned that the highest positive reviews came from Paris at 597 and the lowest positive was California at 477 predicted. Which is a very large decrease from the number of negative reviews but is interesting that California would not have the highest positive reviews being that Disney's main hub is in California. Overall, I found out that Hong Kong had the highest accuracy between all locations which is interesting but can be explained as their language is like a code like coding as they must learn sentences in sequences which is what I learned from working with Hong Kong cast members over the years. As their coded language makes it easier to determine if the review is leading in a positive or negative direction. While the ROC curve that was created on all park locations showed that the highest accuracy came from the California park location.

Is this model ready to be deployed?

I would say this model is not completely ready to be deployed as I would like to gather more data on the Disney theme park locations that where not included such as the Walt Disney World branch in Orlando Florida and other parks overseas such as Disneyland Tokyo and Disneyland Shanghai. As the inclusion of these parks would allow us to further review all locations and determine which parks are suffering from negative reviews and which parks have positive reviews. As this would allow upper management to deploy additional resources to areas that need assistance in raising their review rating and determine what positive review areas have

done to receive and continue to gain positive reviews. While I would also like to further investigate and create models that use the test data to further evaluate Disney's reviews.

What are your recommendations?

I would recommend that Disney will need to implement further resources to Disneyland Hong Kong as they had the highest number of Negative reviews. Which can be justifiable as that location does see a high number of visitors each year but being that their language is so easy to interoperate and has such high accuracy in detecting positive and negative reviews, Disney should have an easier time determining their issues. Which would allow them to implement change very well within the park. As Disney has learned that in Hong Kong their guests are all about taking photos and making memories and are not as interested in buying toys. So, Disney needs to be able to follow their reviews and implement new experience that draw in positive reviews. While we also saw that Disneyland California had the lowest number of positive reviews between all park locations which needs to change as this location is a staple for all Disney goers and needs to be improved on by increasing guests experience by responding to their reviews.

What are some of the potential challenges or additional opportunities that still need to be explored?

A potential challenge I see is the language barrier between each park location as the parks that are stateside will take different approaches to fixing their negative reviews while the parks that are located outside the states might have a harder time communicating with the changes they need to make in comparison to California. While another challenge I see is constantly checking reviews and determining what actions can be taken to increase guests' positivity. While an opportunity that I see is including all park location reviews in the next pull of data to further

understand our park reviews and see what other parks have done to increase positivity. I would also say another opportunity would be to implement some automation so this can be pulled automatically, and Cast Members can easily dive in and analyze the reviews.