GA Data Science Capstone Project: Predicting Attraction Popularity at Walt Disney World

Christopher Doughty

October 14, 2021

PROBLEM

- The Disney Parks branch of the Walt Disney Company wants to improve the guest experience by "imagineering" the best ride attractions possible – those that will have the most appeal to the broadest audience
- Accordingly, it is seeking to design a recommendation tool that will facilitate smoother park experiences from among existing attractions, as well as determine what types of attractions to develop for future projects

DATA SET

- Obtained Walt Disney World Ride Data from data.world: https://data.world/lynne588/walt-disney-world-ride-data

	Ride	Park_location	Park_area	Ride_type_all	Ride_type_thrill	Ride_type_spinning	Ride_type_slow	Ride_type_small_drops	Ride_type_big_drops	Ride_type_dark	...	Age_interest_teens	Age_interest_adults	Height_req_inches	Ride_duration_min	Open_date	Age_of_ride_days	Age_of_ride_years	Age_of_ride_total	TL_rank	TA_Stars
0	Alien Swirling Saucers	HS	Toy Story Land	spinning	No	Yes	No	No	No	No	...	Yes	Yes	32	1.5	2018-06-30	1197	3.277207	3 years 3 months 11 days	31.0	NaN
1	Astro Orbiter	MK	Tomorrowland	spinning, slow	No	Yes	Yes	No	No	No	...	Yes	Yes	0	1.5	1995-02-25	9723	26.620123	26 years 7 months 14 days	43.0	3.5
2	Avatar Flight of Passage	AK	Pandora	thrill	Yes	No	No	No	No	No	...	Yes	Yes	44	5.0	2017-05-27	1596	4.369610	4 years 4 months 14 days	9.0	5.0

- Contains features such as park location, ride type, duration, ride age, height req., and age interest group for each ride
- Pre-cleaning: 46 rows x 28 columns | Post: 45 rows x 24 columns
- Deleted 1 row (no review page), updated review data (data was originally compiled on October 23, 2019), renamed/lowercased columns, changed data types (string to Boolean), deleted 4 columns (irrelevant info, duplicated info, high correlations)

DATA SET

- Scraped individual review data from TripAdvisor for each of the attractions in the WDW Ride Data (through Oct. 9, 2021): https://www.tripadvisor.com/Attractions-g34515-Activities-a_allAttractions.true-Orlando_Florida.html

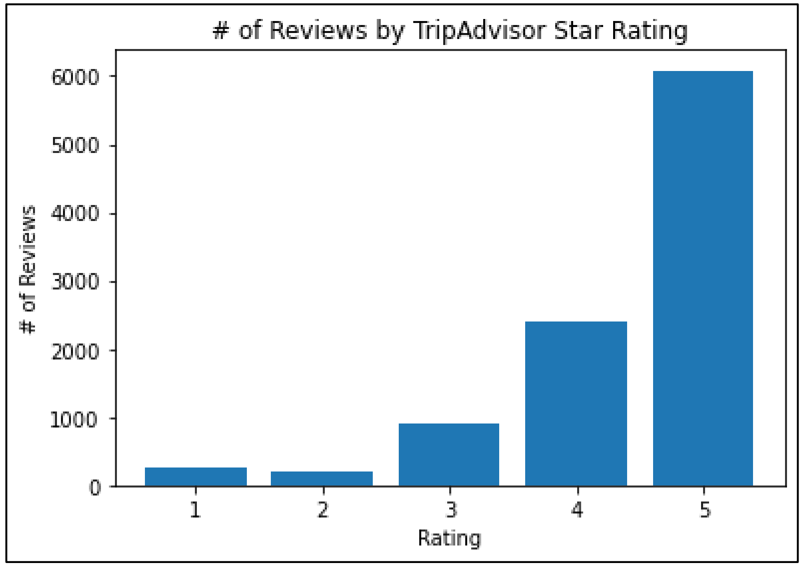
	ride	reviewer	reviewer_visit_group_type	review_date	review_title	review_text	rating
0	Astro Orbiter	Courtenay O	NaN	2020-12-20	Short & sweet ride to space!	This ride brings me closer to the sky than any...	5.0
1	Astro Orbiter	Love2Travel100	Family	2020-02-20	Fun both day and night	Recommend riding both in the day and at night ...	4.0

- Contains features like review date, review title, review text, and rating for each ride; initial cleaning in Excel, secondary in Jupyter
- Pre-cleaning: 9,843 rows x 7 columns | Post: 9,843 rows x 5 columns
- Dropped features not prepared to use for this analysis (nulls requiring imputation, time series data)

ANALYSIS

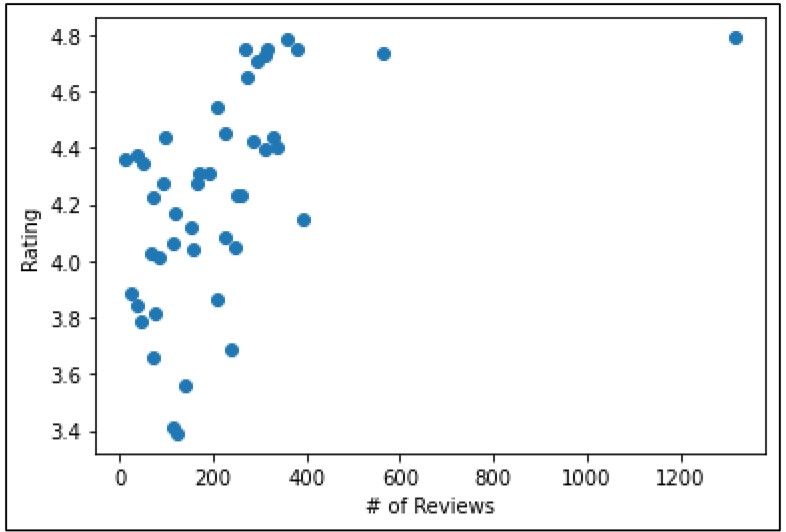
- Joined the data sets together (9,843 rows x 28 columns)
- Data set is very imbalanced – way more 4- and 5-star reviews than 1-3 star reviews

Rating	%	#
5	62%	6,081
4	24%	2,393
3	9%	902
2	2%	212
1	3%	255



ANALYSIS

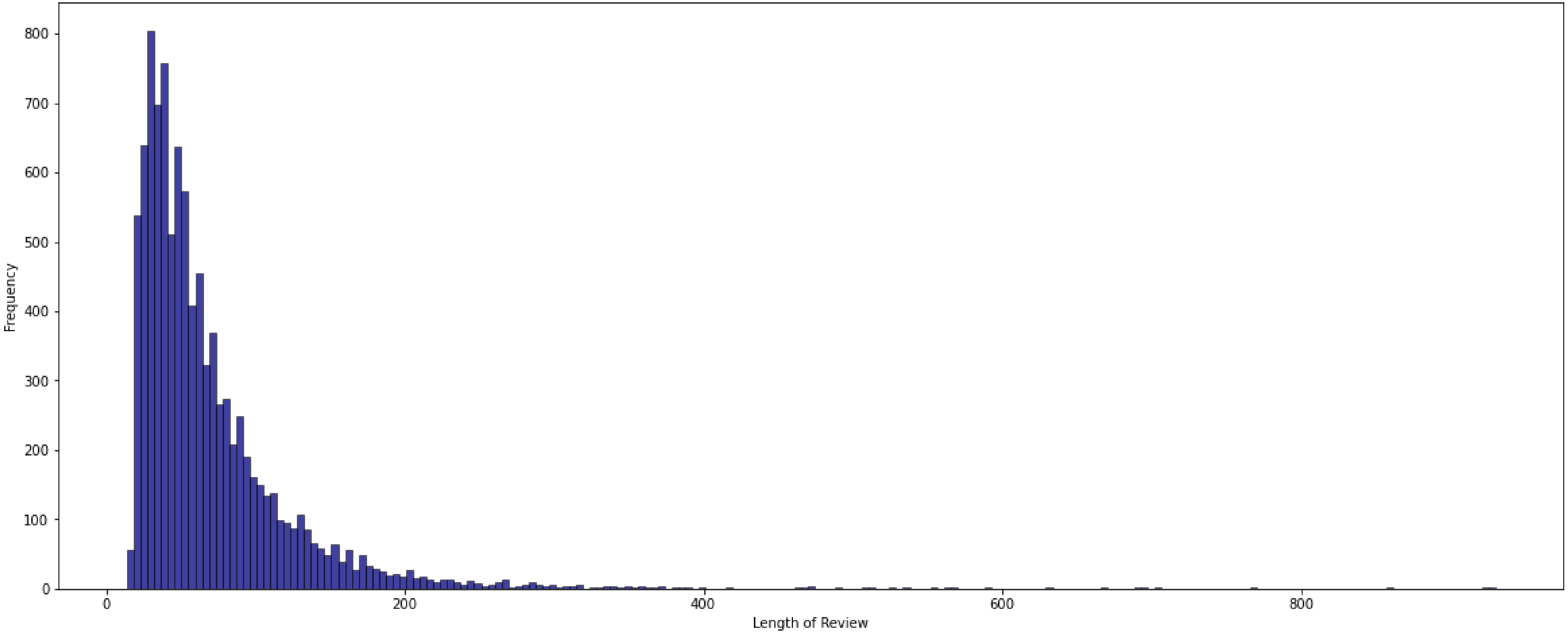
- Avatar: Flight of Passage has way more reviews than any other ride, but its rating distributions are reflective of the overall data set (mostly positive reviews)



ride	min_rating	max_rating	mean_rating	median_rating	average_rating	review_count
Avatar Flight of Passage	1.0	5.0	4.790274	5.0	5.0	1316
Soarin'	1.0	5.0	4.734982	5.0	4.5	566
Seven Dwarfs Mine Train	1.0	5.0	4.150895	5.0	4.5	391
The Twilight Zone Tower of Terror	1.0	5.0	4.748031	5.0	5.0	381
Expedition Everest	1.0	5.0	4.787115	5.0	5.0	357
Haunted Mansion	1.0	5.0	4.402367	5.0	4.5	338
Space Mountain	1.0	5.0	4.440729	5.0	4.5	329
Toy Story Midway Mania	1.0	5.0	4.748408	5.0	4.5	314
Pirates of the Caribbean	1.0	5.0	4.395498	5.0	4.5	311
Kilimanjaro Safaris	1.0	5.0	4.725806	5.0	4.5	310

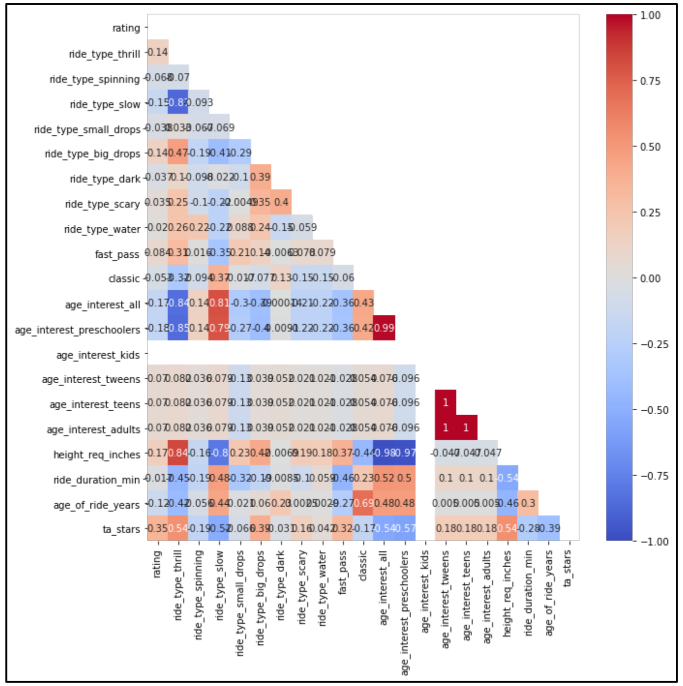
ANALYSIS

- Most reviews are 200 words or less



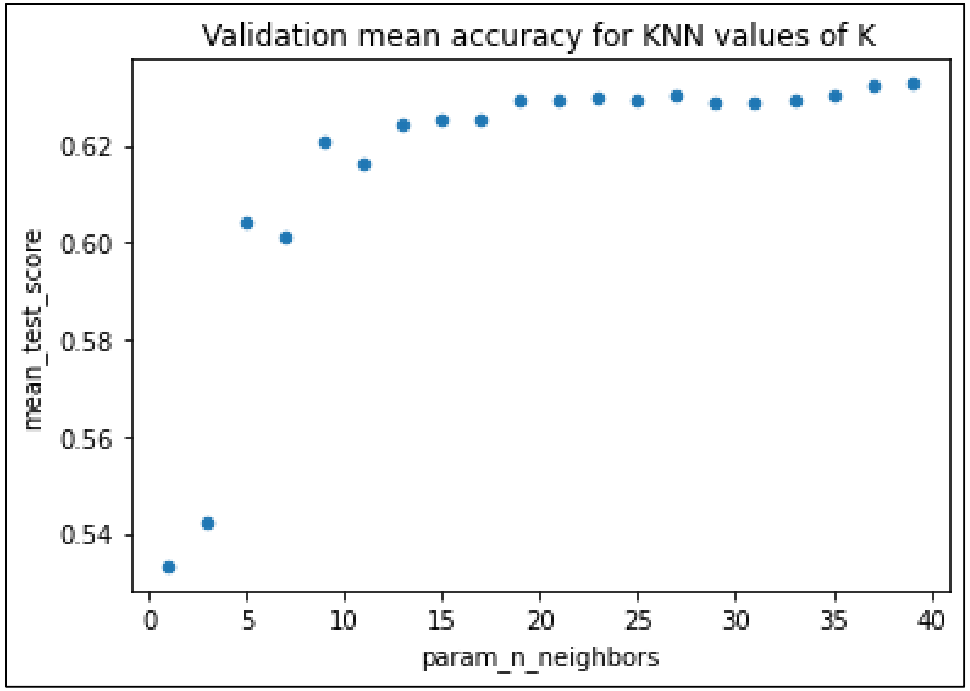
MODELING

- Adjusted which columns I used or dropped depending on the model (dropped to 17 columns for linear/lasso/ridge regressions – 34 after one-hot encoding)
- After adding/removing columns, trying different levels of regularization (including with grid search), could not get a linear regression model based on ride characteristics to explain more than 10% variance



MODELING

- Predicting ratings can be treated as either a regression or classification problem (https://towardsdatascience.com/1-to-5-star-ratings-classification-or-regression-b0462708a4df)
- Tried K nearest neighbors, random forests, and random forests with gradient boosting, but couldn't come up with a model that did better than the null model (guessing all 5-stars would be 62% accurate)



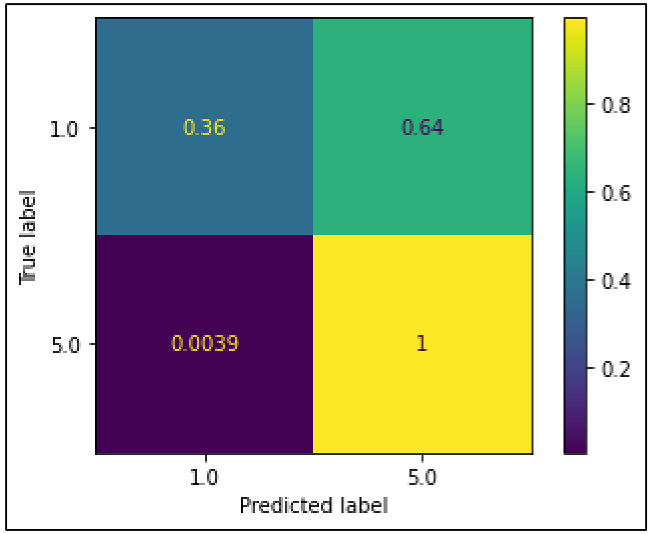
MODELING

- Determined that perhaps the ride attribute data wouldn’t work, and wanted to try looking at the textual data instead
- Used natural language processing to tokenize the body of the review text and use only those tokens to predict the rating (dropped the ride characteristic features)
- Also decided to simplify prediction from multi-classification, to just a binary 1- or 5-star prediction

MODELING

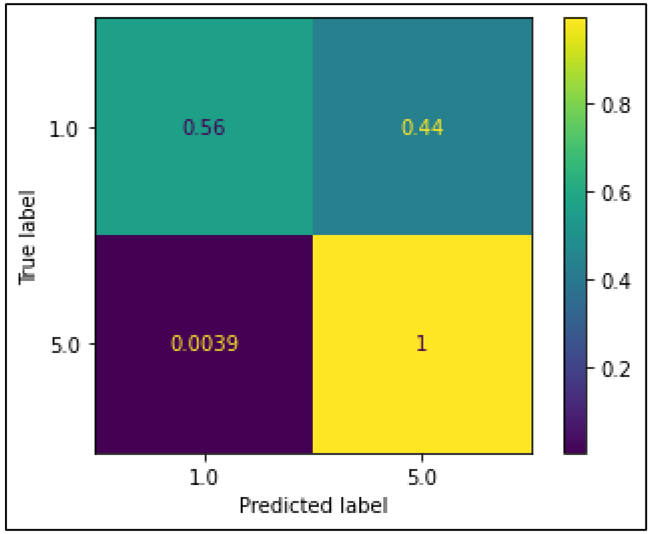
(1) Logistic Regression Grid Search Pipeline with stop words excluded

Precision: 79%
Recall: 36%
F-1 score: 49%



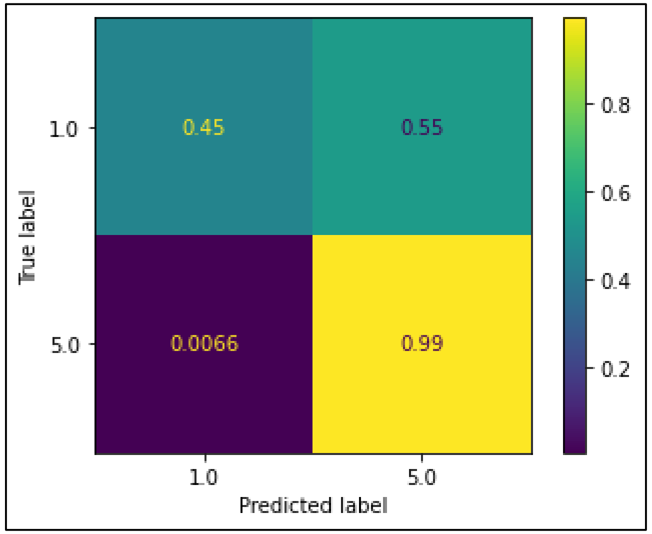
(2) Mulitnomal NB Grid Search Pipeline with stop words included

Precision: 86%
Recall: 56%
F-1 score: 68%



(3) Mulitnomal NB Grid Search Pipeline with stop words excluded

Precision: 74%
Recall: 45%
F-1 score: 56%



MODELING

(1) Logistic Regression

Positive

journey
thrilling
wet
fastpass
nice
stand line
quite
interesting
experience
kids
straight
relaxing
worth wait
wonderful
fast
miss
make
queuing
morning
incredible
favorite

loved	1.942935
amazing	1.846825
best	1.696945
great	1.587399
classic	1.575069
enjoyed	1.533735
awesome	1.481636
love	1.391646
beautiful	1.299003
great ride	1.227289
moving	1.212137
fantastic	1.182848
fun ride	1.157767
flight	1.145366
fun	1.127956
little	1.122434
ages	1.096186
best ride	1.089211
exciting	1.071515
saw	1.069654

Negative

boring	-3.667130
poor	-2.953340
waste	-2.871824
money	-2.776897
worst	-2.596002
breaking	-2.470273
55	-2.141151
awful	-2.088805
recover	-2.063510
skip	-1.969993
needs	-1.963728
disappointed	-1.961874
ride large	-1.804217
hours	-1.760666
overhaul	-1.735599
waited hours	-1.661210
badly	-1.644463
waste time	-1.643827
planning	-1.612293
horrible	-1.603740
stupid	-1.600445
outdated	-1.596391
uncomfortable	-1.544295

(2) Multinomial NB Alt. 1

Positive

fast pass
do
like
line
this is
if you
pass
it is
there
time
disney
great
wait
can
your
fun
all
be
in the
have

the	-3.259037
and	-3.951733
it	-4.068430
ride	-4.098153
to	-4.146492
you	-4.231904
is	-4.315316
this	-4.445875
of	-4.457257
in	-4.681042
for	-4.934960
we	-4.984389
on	-5.066830
was	-5.111990
but	-5.244806
that	-5.296012
are	-5.350884
as	-5.370018
at	-5.381118
so	-5.417106

Negative

torture	-15.356973
boring ride	-15.356973
waste of	-15.356973
dangerous	-12.959078
stupid	-12.959078
wasted	-12.959078
very disappointed	-12.959078
for such	-12.312451
felt sick	-12.312451
breaking down	-12.312451
how bad	-12.312451
boring and	-12.312451
badly	-12.312451
degrees	-12.312451
overhaul	-12.312451
to fall	-12.312451
recover	-12.312451
rip	-12.312451
instructed	-12.312451
sick and	-12.312451

(3) Multinomial NB Alt. 2

Positive

going
definitely
magic
different
year
people
favorite
make
got
went
family
feel
way
fastpass
roller
attraction
did
minutes
kids
amazing

ride	-3.020367
fast	-4.563174
fun	-4.694459
wait	-4.776102
great	-4.789625
disney	-4.789625
time	-4.804868
pass	-4.925526
line	-4.959006
like	-4.964396
fast pass	-5.018058
long	-5.181903
really	-5.207993
just	-5.211447
rides	-5.309518
love	-5.310791
worth	-5.393054
experience	-5.429673
times	-5.451425
best	-5.454362

Negative

torture	-14.279187
wasted	-11.881291
stupid	-11.881291
dangerous	-11.881291
speakers	-11.234664
boring ride	-11.234664
badly	-11.234664
instructed	-11.234664
felt sick	-11.234664
line 30	-11.234664
recover	-11.234664
puke	-11.234664
degrees	-11.234664
nightmare	-11.234664
overhaul	-11.234664
panic	-11.234664
rip	-11.234664
jerking	-10.845200
spun	-10.845200

NEXT STEPS

- Might be helpful as part of a park recommendation tool in an app to help guests plan their visit
- Sentiment analysis of the most important words used to predict ratings as a way of identifying best/worst ride characteristics
- Would like to build upon the model by adding in the text from the review titles, as well as the dates, as seasonality may also impact enjoyment of different attractions – could also try a multi-classification model using natural language processing
- Could revisit a model based on ride attributes at some point with additional data/new features about the rides

QUESTIONS?