GTx: ISYE6501x Introduction to Analytics Modeling

ANALYTICS AT DISNEY

Project report on use of analytics at Disney World

https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Industry-Profiles/Disney

Table of Contents

1. Objective:

Walt Disney World is the most visited vacation resort in the world with average annual attendance of 58 million customers. What most guests don't see is the careful planning taking place "behind the scenes" to run the operation smoothly. The case study I have picked up mentions how important analytics is at Disney. In this report, I describe few of the analytics tools that might have been used to run the operations smoothly and ensuring the guest experience is maximized, despite the humungous volume. The sheer volume of data the team must deal with, presents its own challenge. The main factor to improve guest experience at the resort is to handle the demand efficiently at all parts of the resort. In case of Disney world, the demand is the attendance/visitors at the parks and forecasting the attendance is foundation to all the subsequent planning at each phase. For example, if the expected attendance is more than capacity, then park will extend the hours by opening early and closing late. Increase the number of food carts throughout the park. Prepare for more entertainers to be employed on busy days to handle the extra crowd.

As mentioned in the case study, analytics is used in every aspect of the customer experience like staffing decisions at the hotel on arrival, number of "FAST PASS" tickets that can be issued, customer engaging queue systems when people are waiting in lines for the rides and many more. Also, I was fascinated by how Disney runs so many real time analytics and wondered how companies run real time analytics. If the intelligent decisions don't come on time, there is little value to the whole analysis setup. In my current job, as an application developer of lot of customer facing applications, I get to work real time data and tools to synthesize real time data. I have tried to present that knowledge as well along with analytics. The main areas I am focusing in my report are :

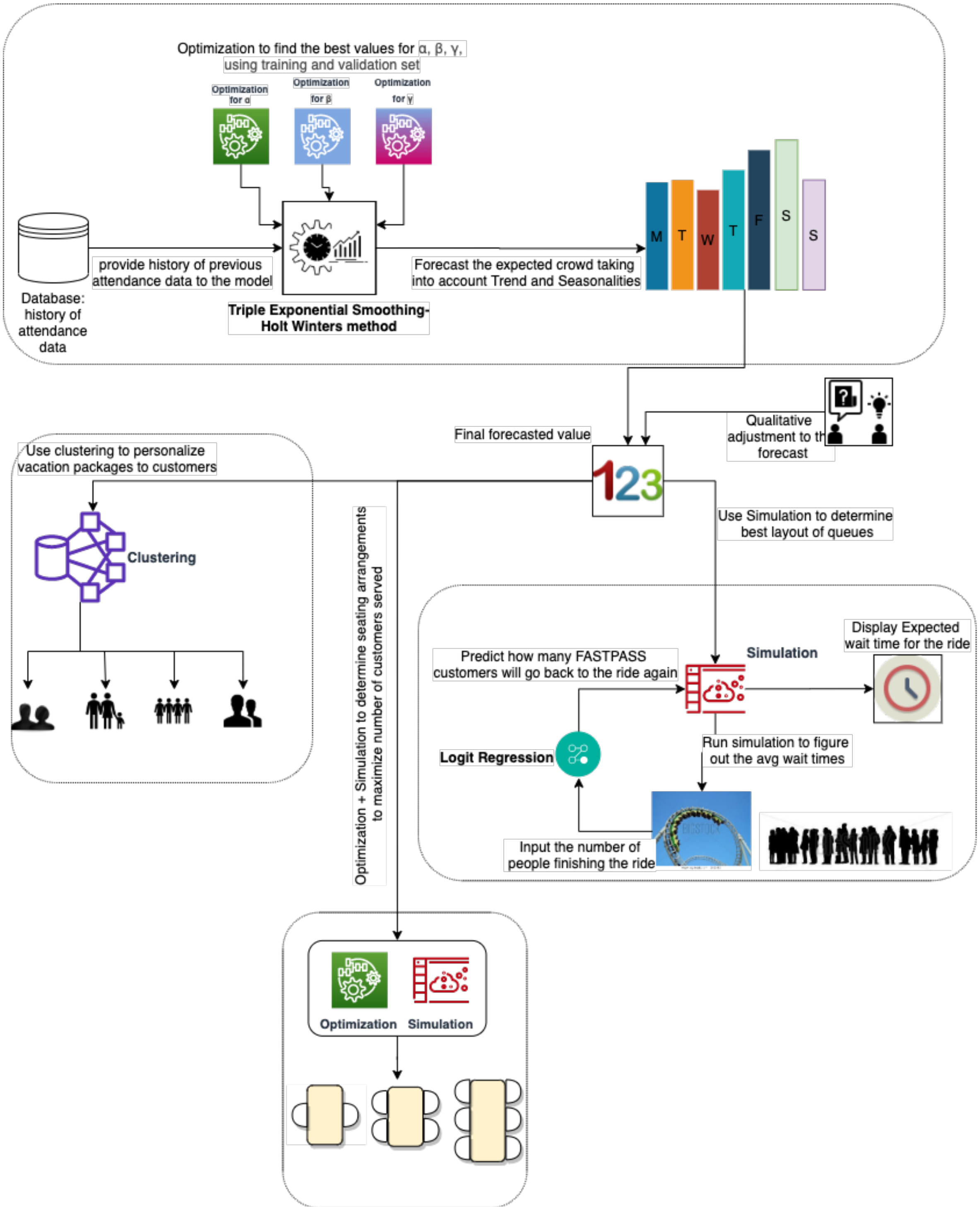
1) Forecasting daily attendance using Time Series models
2) Clustering to customize offerings and experiences to customers.
3) Simulation to identify bottlenecks and streamline the waiting process in the queues.
4) Logit Regression to predict if the customers with FASTPASS would go back to the ride once again. This number would impact the average wait times.
5) Optimization and simulation for restaurant seating arrangement to maximize the number of people served.
6) Real time data processing for analytics.

In the report, I have attempted to describe the application of few of the analytics topics we have learnt in the course. The link to the case study is at:

https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Industry-Profiles/Disney

1.1 Illustration:

I have created a flow diagram that captures the overall picture and how the models are all working together to provide highest quality customer experience at Walt Disney World resorts.


Optimization to find the best values for α, β, γ, using training and validation set
Optimization for α
Optimization for β
Optimization for γ
Database: history of attendance data
provide history of previous attendance data to the model
Triple Exponential Smoothing-Holt Winters method
Forecast the expected crowd taking into account Trend and Seasonalities
M
T
W
T
F
S
S
Final forecasted value
Qualitative adjustment to th forecast
123
Use clustering to personalize vacation packages to customers
Clustering
Optimization + Simulation to determine seating arrangements to maximize number of customers served
Use Simulation to determine best layout of queues
Simulation
Display Expected wait time for the ride
Predict how many FASTPASS customers will go back to the ride again
Logit Regression
Run simulation to figure out the avg wait times
Input the number of people finishing the ride
Optimization
Simulation

2. Exponential Smoothing for forecasting:

As mentioned in the above paragraph, the first thing to start off with is forecasting attendance at the park. We can use one of the time series models to forecast the expected attendance at the park. I am going with Exponential Smoothing time series model for the same.

Given:

- The past daily attendance data for a single day

Use:

- Triple Exponential Smoothing model

To:

- Forecast the expected daily attendance in the immediate future

Explanation: Walt Disney is a well run organization for so many decades, I am assuming they would have maintained the daily attendance data for past several years (at least 10 – 15 years). This gives good basis for us to use exponential smoothing model to predict immediate future. As we learnt in the lecture videos, using the simple exponential smoothing equation without trends and cyclical variations we can predict the today's attendance at the park:

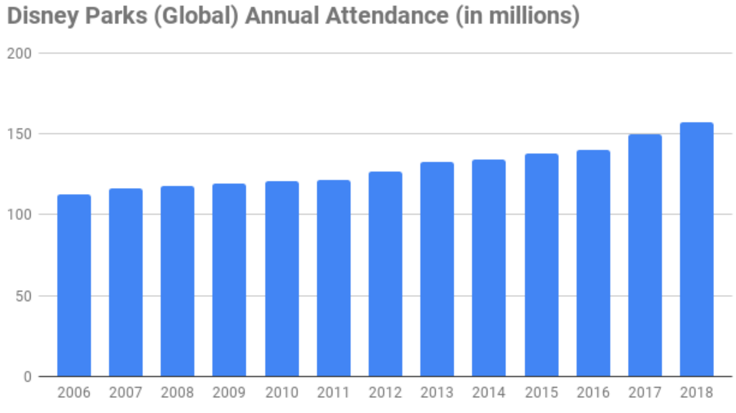
$F_{(t+1)} = S_{(t)}.$ Where $S_{(t)}$ is a latest baseline estimate calculated by $S_{(t)} = \alpha x_t + (1-\alpha)S_{(t-1)}.$,

x_t = observed daily attendance at the park yesterday,

$S_{(t-1)}$ is forecasted daily attendance for yesterday,

$0 < \alpha < 1$ helps us to trade-off between trusting our observation or trusting previous estimate.

As the park adds newer features and increases advertisements, the demand keeps rising. This added demand adds Trend to the exponential smoothing equation. If we look at the Disney's annual attendance numbers (pic below), we clearly see there is a growth in demand. When we forecast the attendance, we have to account for this increased demand by including a Trend factor in our exponential smoothing equation.



Source: https://disneynews.us/disney-parks-attendance/

Similar to trend there will be cyclical variations in the expected crowd turn out. Number of people visiting the parks will be higher during summer holidays when schools are closed. This number goes down when the schools re-open. Similar to trend term, we can also include a multiplicative seasonality to the exponential smoothing equation to forecast daily attendance. As we did in the homework, given the daily attendance numbers from the past years, analytics software can calculate the trend and seasonality. The software also can suggest ideal values of α, β, γ, but we will use optimization model to calculate them.

$T_{(t)}$ is the trend term to account for increasing demand, calculated by $T_t = \beta(S_t – S_{t-1}) = (1- \beta)T_{t-1}$
$C_{(t)}$ is the Cyclic term to account for seasonality, calculated by $C_{(t)} = \gamma(X_t/S_t) * (1- \gamma)C_{t-L}$
Finally, the value to be forecasted is given by

$F_{(t+1)} = (S_t+T_t)\ C_{(t+1)-L}$
Forecasted Value = (estimated baseline + trend term)(seasonality term on the same day last year).*

2.1. Optimization for calculation of α, β, γ

As mentioned in the lecture, to calculate the alpha, beta and gamma values we can use the optimization model based on the mean square errors of all the previous predictions and the observed actual value.

Given:

- Predicted forecast of all the days upto today,
- Actual observed attendance of the customers visiting park each day

Use:

- Optimization method with following details:
  - Variables = α, β, γ
  - Constraints:
    - 0< α <1,
    - 0< β <1,
    - 0< γ <1
  - Objective function: minimize $f(x)=(\hat{x}-x)^2$

To:

- Determine values of α, β, γ which will give best forecast based on the history.

Data collection: The daily entry of the customers to the park can easily be tracked through turnstiles recording or ticket scanning at the entrance of the resort. This data has to be updated everyday for the forecasting of the next day's estimates. The trend, seasonal factors will be calculated daily to update the forecast for every day. In fact, in the case study they have mentioned the forecast calculations are run every 15 – 20 minutes to get the accurate estimate of crowd arrival at different points of interest.

Evaluating analytics accuracy: The daily recorded attendance at the end of the day gives the information on how good the model is performing. If the actual values are way off compared to predicted forecast, the analytics team can adjust the values of α, β, γ for future predictions. If there is consistent difference observed in predicted vs actual values, we can even run CUSUM technique to identify if there was any change in the demand. Once we detect the change, we can engage qualitative analyst to determine the reason for the change and accommodate that in the future estimations.

Considerations:

- Time series data models are good at predicting the immediate future. There is more uncertainty the farther into the future we go, the anticipated forecast error gets larger too. So this method will not be a good method for long term forecasting. We will have to use regression or combination of timeseries data and regression to predict the crowd attendance for the longer duration.
- Lot of macro level planning happens based on the long term forecast of attendance, like budgeting, adding new attractions/shows, adding new permanent restaurant, dynamic pricing for tickets. Long term forecasting can be done using Linear regression which will take into account factors like airline cost changes, foreign exchange (lot of international tourists visit the parks), weather patterns and whole lot factors from various domains.

2.2. Qualitative adjustment to the forecast:

The above method is a good quantitative method to forecast the customer attendance. This forecast purely based on what happened in the past. Even though the numbers suggested by this model are accurate most of times, there will be situations where knowledge of certain events will impact the forecasted numbers. Ex.: covid outbreak, inclement weather predictions can impact the turnout significantly. To accommodate for these unforeseen events, I would also provide an option to adjust this forecast with qualitative analysis if needed. An expert or group of experts in the field can adjust this number with factor. Experts can use other approaches like linear regression to come up with the adjustment parameter.

3. Customized customer experiences using clustering:

While accurate forecasting is important, the results are only valuable when utilized to make smarter decisions. As Disney analysts understand more about guests and their preferences, they can utilize analytics to customize offerings and experiences that better match resort guests' desires. As an example the analytics model like clustering can be run to identify the best matching vacation packages using the data collected at turnstiles, basic profile information of customers, like, a customer group of family with kids might be interested in package involving kids activities. Newlywed/engaged couples might be interested in a package that caters to adults.

Given:

- Number of members in the group,
- If kids present in the group?
- Type of group (categorical value like Family, Couples, Friends, Large group, Individual)
- Average age of the group
- Past purchase history of the customer
- Stay duration (1 day, 2 days, ..)
- Veterans/Local heros like Firefighters,police (for discounted packages)
- Household income
- First time visitors?
- Offers accepted in the past(they might want to try something else this time)
- Adventure affinity (Categorical like extreme, moderate, mild, none)
- Forecasted park/resort attendance.
- Percentage of customers that have entered the resort already.

-

Use:

- K-means Clustering

To:

Personalize the vacation packages/promotional benefits offering to the guests upon arrival.

Explanation: As mentioned in the case study, Disney analysts understand more about guests and their preferences, they can utilize analytics to customize offerings and experiences that better match resort guests' desires. We can run K-means clustering model to figure out which vacation packages suits the best for each customer type. The overall park/resort attendance forecast determined the first step, also has impact on availability of these packages. I opted to use unsupervised clustering instead of classification because we will not know the predefined groups under which we might want to classify the incoming customer. Instead, use clustering to determine the right cluster for the customer and offer the package relevant to that cluster. As mentioned in the case study, the analytics is performed every 15 to 20 mins using the newer data available as and when guests start coming into the park/resort. This information can be used to show the offers/packages at numerous kiosks located throughout the park and at the park entrance. I am guessing there will be integration with the mobile app where the promotional offers/ packages are advertised as and when the model runs every time with newer data.

Data collection and prep: I have listed few of the data in the definition above that can be used to find the cluster to which the incoming guests might belong to. The types of data are spread across multiple domains like economic, personal preferences, current snapshot of the park capacity. Gathering this data can be done at different points in the parks and at different stages of the customer arrival. First, we can get lot of group related details like number of people travelling in the group, at the time of the booking itself. We can ask the users to create account profile with basic information about themselves and link that information to the ticket id. Usually a group that travels together will book the tickets together. We can extract the travelling party details from the ticket booking website where they enter the names of all the people travelling with them and their ages. Using the ages of all the travelers, we can figure out if the travelling party are all same aged like group of friends or family with kids. Any data that is not available can be obtained upon arrival using automated kiosks or survey.

Evaluating analytics accuracy: We can keep track of the number times the clustering predictions are not accurate when customer opts for a package other than what the models suggested. We can run a quick survey after the whole process carefully designing the questions to extract the key information that made the customer select something other package. Example: Based on criteria that customer has already gone on a particular cruise, model might put the customer into a cluster whose package suggestion was not a cruise. But the customer's satisfaction level of the previous trip was so high, that customer may want to do it again. If the number of occurrences of this behavior increases beyond a threshold, we can include this as one of the questions prior to clustering.

Refreshing model and execution:

The case study mentions that the company runs the algorithm every 15 to 20 mins to keep updating the analytics system. Some of the data is stationary and will not change as the day goes by. But some of the data like availability of promotional packages, amount of forecasted customers already entered the park, time of the day can be changed and updated every 15 to 20 mins cycle. My guess is that a model is executed every 15 to 20 mins but new models might be built may be once a week or fortnight to adjust the clustering parameters.

Considerations:

- The number of clusters and the type of clusters can keep changing throughout the day. When the first batch of analysis is executed it might result in fewer clusters. As day goes by and more customers walk through the gates, the mix of customers keep changing and so does the availability options for packages and events. I believe we need one more analytics model like logit/regression model to predict the promotional package as cluster characteristics change or newer clusters start getting created with newer characteristics.
- We can also start with exact number of clusters based on different type of promotional events/packages that are available. As the crowd rolls in and if the events/packages start reaching their capacities, we can remove them from the equation resulting in lesser number of clusters.
- There will be some missing data as people might not be willing to provide their income details or family details at the time of the booking. This information at each interval if is less than 5% of the overall data for that interval, we can impute the missing data based on the other factors.
- The forecasted turn out of customers also helps in labor planning. If the number is high, we might need more people at the sales and front office who present these promotional packages/events. This in itself is an optimization and scheduling model to determine the number of personnel needed and how they should be distributed across different parts of the park.

4. Simulation:

Rides and shows are main attractions that draws huge crowds every day to the Walt Disney World resorts and theme parks. If the crowd is not managed well then lot of time might go off waiting in lines to get into the rides. That will be a big let-down to the customers. As mentioned in the study article, Disney uses analytics tool to keep the crowd engaged during the whole process. One analytics tool that I think would be very beneficial is Simulation. It can be used to design the queue structure, movement and keep the crowd engaged while they wait for their turn to get into the rides. During the busy days the arrival rate of crowd into an attraction will also be high causing longer wait times. We can use simulation to design the queue with lot of customer engaging interactive stations along the queue. It will also be beneficial to display expected wait times for the attractions at the entrance based on the simulation findings and queuing process.

Given:

Arrival rate of customers to each attraction
Processing times for each interactive stations
Processing times to load and unload rides
Processing times for the actual ride

Use :

Simulation

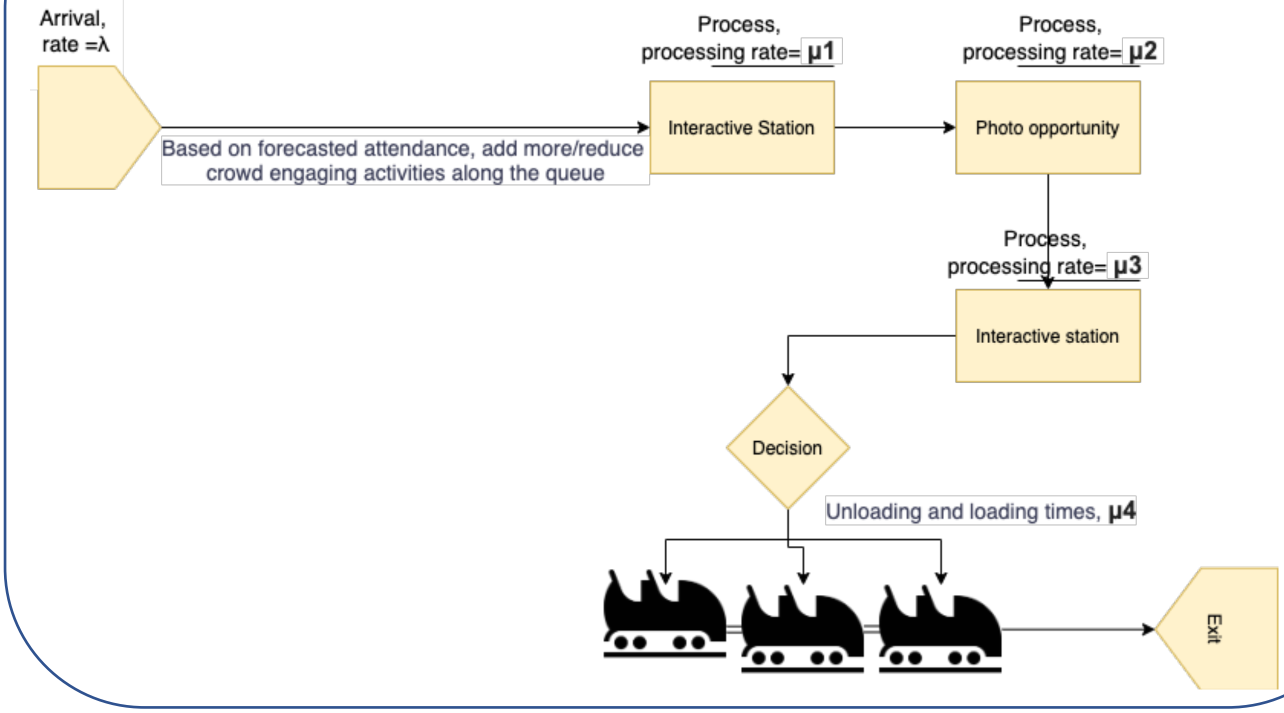
To:

Design queues structure
Determine number of interactive stations
Identify any bottlenecks in the queuing
Determine if the capacity of the rides should be increased/decreased.

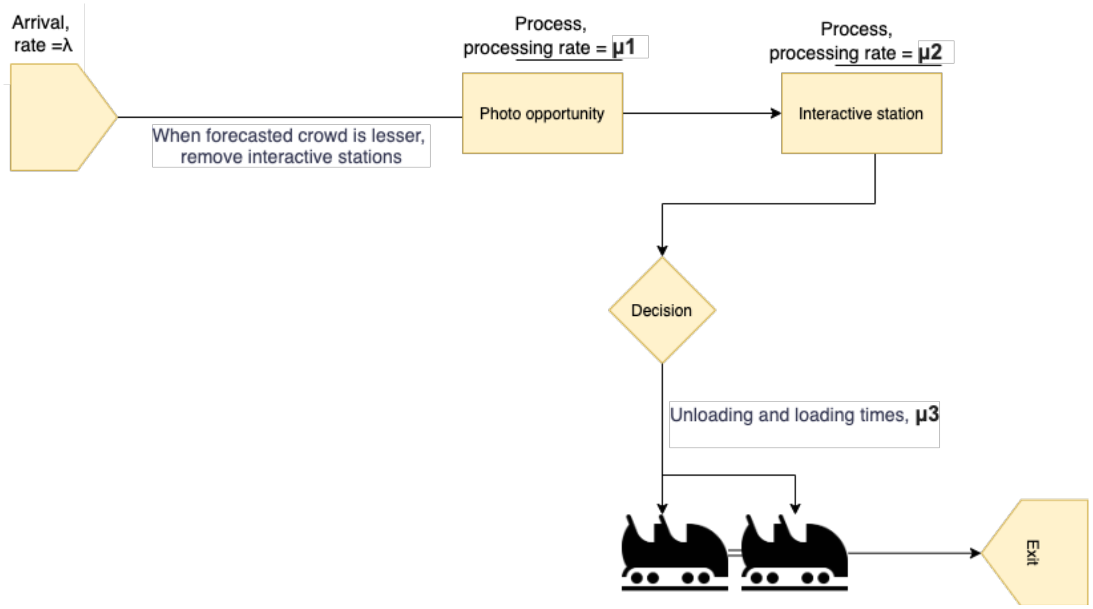
Explanation: Based on the forecasted turn out of customers, we can predict the arrival rate of customers to queue for rides. If there are new rides added to the park, there will be high number of people who would want to check it out. We can run simulations to design the queues, setup crowd interactive stations along the queue to keep them engaged while they wait in line. We can identify if there are any bottlenecks through the process and avoid them. We can run multiple scenario simulations ahead of time for different arrival rates of customers. Using the simulation output combined with other queuing techniques, we can estimate the expected wait time for a ride and display it in the front. Similar to the assignment we did on simulation, I created 2 simulation models for different scenarios.

Data: To run simulation a simple simulation mentioned above, we will have to know the arrival rate of customers, processing times at each interactive station, processing time during loading and unloading of the ride cars and the time duration of the ride itself. The historical data of the processing times of stations and ride should be available in the database. If not, setting up sensors to track start of a ride and end of the ride should get the information. The crowd arrival rate at the attraction will be related to the overall forecasted attendance and so is the number of people who have already entered the theme park. The turnstile count, ticket scanner at the gates can give us that information real time.

4.1. Simulation 1: when the number of guests expected is high, more interactive stations and ride carts.


Arrival, rate =λ
Process, processing rate= µ1
Interactive Station
Process, processing rate= µ2
Photo opportunity
Based on forecasted attendance, add more/reduce crowd engaging activities along the queue
Process, processing rate= µ3
Interactive station
Decision
Unloading and loading times, µ4
Exit

4.2. Simulation 2: when the number of guests expected is not high, less interactive stations and ride carts.



Evaluating analytics accuracy: Using the expected arrival rate, our simulations can predict average wait times for each arrival rate. The ticket/tag scanner at the entrance of the queue, at the start of the ride and at exit can calculate the observed wait times. This will give us a good comparison of actual vs estimated to adjust the simulation. For different values of arrival rates, the simulation scenario can be different. The queues should be constructed in such a way that it has provisions to dynamically add/remove interactive stations along the queue without disrupting the flow. We can combine simulation with queuing and use distributions to calculate the average wait times for the rides and display the values at the queue entrance.

Refreshing model:
Simulations might not be run every 15 to 20 mins like other analytics models mentioned in the article. This could be run few times a month based on the average predicted attendance. If the crowd wait time goes beyond a critical threshold, we can set up notifications to concerned teams who can create alternate events like parades around the busy attraction to spread out the people gathering.

5. Logit Regression for Queue Entry and Re-entry:

Disney World has a feature of FASTPASS where customers pay extra to get a FASTPASS which will allow customers to skip the lines in the queue and can do it any number of times. This adds more complexity to estimate the estimated wait time for the rides. One option here is to predict how many people with FASTPASS are most likely to show up for a ride or take a ride once again right after they finish. The output of this model can be a valuable input to the previous model to establish the arrival rate.

Given:

- Number of FASTPASS passes sold for that day.
- Number of people near the vicinity of the ride.
- If the customer has already done ride once.
- Number of people already entered the park.
- Time of the day.
- Current wait times for the ride.
- Ride popularity index.
- Number of days a customer is vacationing.
- Has the customer used FASTPASS for re-runs on their previous visits?
- Any open event that is happening nearby the ride.

Use:

- Logit Regression

To:

- Predict the probability a customer might use FASTPASS to enter or re-enter the ride.

Explanation: Estimating the number of people that can use FASTPASS to enter or re-enter the queue to enjoy the ride will be a key factor in estimating the expected wait time for the rides. Disney publishes these wait times across the park and on the mobile apps. We can use a logit model to estimate the probability that a customer will be entering the ride for first time or even use it to re-enter the ride for more than once. This model is coupled closely with the previous model to provide effective insights into the ride design process. We can set a threshold probability percentage (ex.: 0.65) beyond which that customer is likely to enter the ride. The sum of all the people whose probability is beyond the threshold percentage can give us count of people who might enter the ride.

Data Collection and Prep: The sources of data I listed above, are many and change faster. Sales data and forecasted values can provide the values for forecasted attendance, number of FASTPASS sold until that point. Most of the FASTPASS customers wear a bracelet with microchip in it that records if the customer has already used the ride. This can also give us the information about the number of days the customer is vacationing in the resort and what is the current day out of that day(Ex.: 3rd day out 4 days vacationing). The turnstiles at the park entrance will provide the information about the number of people that have already entered the park until that point. The scheduling data in the company's database will give the information about the open events like parades, street shows near the ride which can draw good number of crowd, easing the congestion on the ride queue. We can establish a relative ride popularity index for all rides based on historical data on ride usage and also customer surveys. Mobile geo location data and the location info from the bracelet can give us the information about the number people in the vicinity of a given ride.

Evaluating analytics accuracy: The model prediction has to be cross checked timely against the actual numbers which can be obtained by the scanners at the rides. One more measure to assess the quality of the model output is accuracy of the expected wait time that is published. Since this model's output is an input to the previous model, if the prediction is incorrect here, it will reflect on the prediction of the expected wait time. If the numbers cross beyond a threshold, then the model and factors need to be re-adjusted. Near past data, like last 6 month's data, can be taken and split into training, validation and testing set. The accuracy of the model can be measured by building multiple models using training data and verified using validation data set. The best model can be chosen based on the accuracy against the test data set.

Execution and Refreshing the model: The process has to be run frequently throughout the day at 30 mins to 1-hour interval to keep the updates as latest as possible. The customers on whom the model has to be run can be identified by the bracelet information near the vicinity of the ride. The models themselves might be built or refreshed once a month or a quarter but they have to be executed every 30 mins to 1-hour.

6. Simulation and Optimization for Seating Arrangement

Walt Disney world has numerous restaurants throughout the entire park and resorts. Similar to the queue management in rides, it is important to manage the restaurants to quickly move the customers in and out during peak time. The resort's table-service restaurant operation might leverage optimization and simulation to maximize the number of customers served. Statistical analyses helps the company understand the patterns around party sizes, arrival times and table turn times. This knowledge is incorporated into mathematical models that determine the right mix of tables to best meet guest demand.

Given:

- Average arrival of customers of party size =1, during peak periods
- Average arrival of customers of party size =2, during peak periods
- Average arrival of customers of party size =2 to 4, during peak periods
- Average arrival of customers of party size =4 to 6, during peak periods
- Average arrival of customers of party size =6 or more, during peak periods
- Turnaround times for each party size
- Maximum building capacity of the restaurant
- Maximum number of single seater tables
- Maximum number of double seater tables
- Maximum number of 4-seater tables
- Maximum number of 6-seater tables

Use:

- Optimization along with simulation

To:

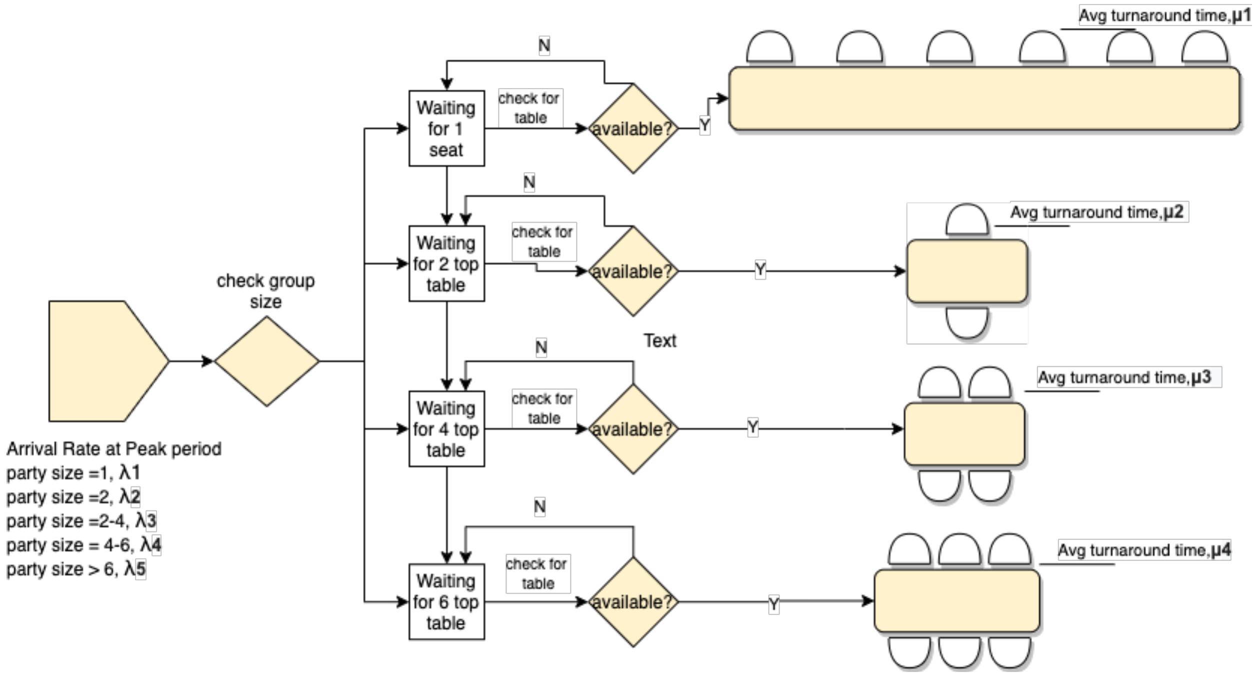
- Maximize the number of people served
- Minimize the wait times for customers.

Explanation: Using the forecasted numbers and the past history, we can predict the arrival rate of customers according to the party sizes. With the knowledge of the restaurant capacity we can run optimization solution to get the right number of tables of all types to be used to maximize the number of customers turned around. As Prof Sokol mentioned, optimization can become a hard problem to solve. So I chose to club the optimization with simulation to determine the right number of tables to be used to handle the peak traffic optimally. Based on weekly forecasts, the team can run this model when the expected attendance is going to be unusually high, ex.: during long weekends.

Data: The data of the arrival rate for each party size can be derived from the estimated attendance and past history of the breakdown of the party sizes at a given restaurant. The restaurant capacities are established numbers, if there is an expansion, this number can be updated. Table turnaround times can also be obtained from the past data. If there is no past data, we can setup the scanners to scan the entry and exit of the groups using their arm bands.

Execution and Refreshing the model: We can run a simulation using the output of an optimization model to simulate a real-world scenario. I am seeing this would be mainly used during peak hours on busy days. We write a logic in a decider to move the customers to the next available sized table if the wait time has exceeded a certain amount. Ex.: If a party of 2 have been waiting for more than 20 mins for a table and a table of 4 is free, and there are no party of 4 waiting for more than 20 mins, the party of 2 is assigned a 4 top table. A simple simulation might look as below:

Scenario simulation for seating arrangement with different configuration:



7. Real Time Data Analytics:

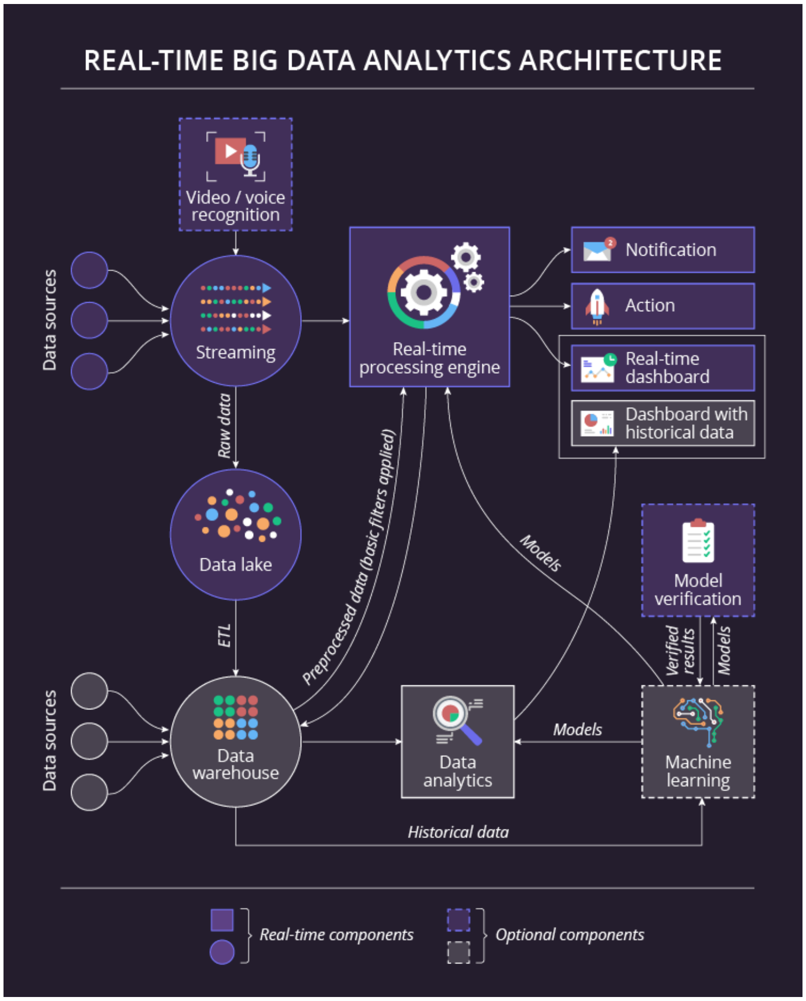
As I was reading the case study of Disney, I was fascinated by the scale of the data they handle and also the real time analysis they run. It is mentioned in multiple places in the case study that they run analytics every 15 to 20 mins once. Many times, they develop new models that frequently. I was wondering how one would do real time data analytics which is a key to most of the operations. Real time data analytics is the glue that is holding everything together and enabling such quick execution and decision making possible. In the preparation of this project report, I started looking for how companies do real time analytics and was pleasantly surprised to see lot of overlap with tools I am currently using in my job as an application developer for multi-channel interface. We use lot of real time streams like AWS

Kinesis , Apache Kafka , Confluent Kafka to stream real time live data on to websites and mobile apps. The websites and mobile apps react to any change in the data immediately.

Data stream is like a live tunnel that is connected to all the other components in the ecosystem like databases, decision engines, dashboards etc., We can imagine streams as how water flows coming from multiple sources into a destination. Similarly, streams are used for continuous, never ending data originating from multiple sources and provide constant feed that can be analyzed/acted upon without needing to be downloaded first. Compared to traditional way of handling data, where the data arrives in scheduled batches. The program that needs this data have to download the data, process it and provide it to the next component in batches again. With streams, the data arrives when it is ready and consumer of the data has to react to the new data, process it and put its response back to the stream for downstream systems to consume.

I believe that understanding how real time data analytics work is as important as understanding various models. I found a very good illustration of architecture of the real time data analytics and have included it below. It is clearly not my work, and I have credited the source which has lot of interesting information about real time analytics.

Source: https://www.scnsoft.com/blog/real-time-big-data-analytics-comprehensive-guide



8. References:

1) Disney attendance statistics: https://magicguides.com/disney-world-statistics/#:~:text=With%20an%20average%20annual%20attendance,and%20innovating%20is%20simply%20astounding.

2) Disney case study reference: https://www.informs.org/Impact/O.R.-Analytics-Success-Stories/Industry-Profiles/Disney

3) Realt time data analytics: *https://www.scnsoft.com/blog/real-time-big-data-analytics-comprehensive-guide*

4) Optimizing restaurant table configs: https://scholarship.sha.cornell.edu/cgi/viewcontent.cgi?article=1739&context=articles

5) Walt Disney World Attendance statistics: https://disneynews.us/disney-parks-attendance/

6) Real time streaming: https://www.confluent.io/learn/data-streaming/