The Happiest Hotel on Earth:

A Disney Hotel Recommender System

By Noah Johnson, Annie Phillips, Maddie Preston, and Rebecca Wilhelmi

With Dr. Chris Alvin, Dr. Liz Bouzarth, and Dr. Kevin Hutson

Introduction

The objective of this project was to construct a recommender system that, based on user input and existing hotel data, returned a Disney hotel personalized to the respondent’s party, preferences, and previous experiences.  The ultimate goal was to automate this process using Python code and eventually open up the opportunity to repeat the process for Disney World restaurants and attractions to create a computer-generated touring plan personalized to TouringPlans clients.  By experimenting with various machine learning and vector-based models, we have found a model that is implemented in Excel code with the option of eventually translating these calculations into Python.

Data Cleaning

* Cleaned:
  + **Trip Start/End Date**- edited improperly formatted and “impossible” start and end dates (such as trips starting in 1902 and ending in 2026)
  + **Hotel Trip Other Name**- inconsistently formatted strings, edited for uniformity (for example, Boardwalk vs. Disney’s BoardWalk Inn)
  + **Hotel Past Other Name**- inconsistently formatted strings, edited for uniformity; also made eight columns to separate all of the hotels so that there was one hotel per cell
  + **Hotel Selection Other String**- made similar responses uniform; for example, a general “monorail” category
  + **Hotel Room Number**- ensured that all room numbers are valid
  + **Hotel Room Connecting Room Number**- ensured that all connecting room numbers are valid
  + **Hometown**- fixed capitalization and spelling, divided this information into three columns
    - Hometown- when states/provinces were not provided, could only sometimes assume where cities were (for example, we could assume Boston was in Massachusetts but not that Greenville was in South Carolina)
    - State/Province- only provided for US, Canada, and the UK
    - Country
  + **How Many Trips**- eliminated unrealistic values, set a limit at 100 trips
* Removed:
  + **Hotel Selection Spa Fitness Center -** no data in entire column
  + **Hotel Satisfied Other String** - no information, all 0’s
  + **Hotel Satisfied Other Rating** - because satisfied strings were empty, ratings didn’t mean anything
  + **Duplicates**- eliminated consecutive entries that contained all of the same data
    - These entries were often submitted within seconds of each other
    - Number of identical entries ranged from 2 to 14
    - If included, would have weighted duplicate data points more than others
* Wrote a Python program to:
  + Separate the ages and genders of party members from the given party string and calculate new columns from them
    - **Number in Party**
    - **Age Categories**- calculated the number of guests in a party who were:
      * Ages 1 year and younger
      * Ages 2-4 years
      * Ages 5-8 years
      * Ages 9-12 years
      * Ages 13-17 years
      * Ages 18-20 years
      * Ages 21-25 years
      * Ages 26-35 years
      * Ages 36-50 years
      * Ages 51-65 years
      * Over 65 years
    - **Number of Males**
    - **Number of Females**
  + Return the season- Spring, Summer, Fall, and Winter
  + Return the Disney season- what festivals and events were happening at Disney while the party was there
  + Return a view given a room number- if a correct room number was entered, we looked up what type of room they stayed in
* In Excel, we calculated:
  + **Length of stay**- number of days between trip start date and end date
  + **Hotel Category Number**- Value = 1, Moderate = 2, Deluxe = 3, Deluxe Villa = 4
  + **Percentage of Males in the Party**- number of males / number in party
  + **Percentage of Females in the Party**- number of females / number in party
  + **Percentage of the party in each age category**- number of each category / number in party
  + **Percentage of the party under 18**- number of guests under 18 / number in party
  + **Percentage of the party over 18**- number of guests over 18 / number in party
  + **Normalized Selected**-(selected value from user - average of that attribute for all users) / standard deviation of that attribute for all users
  + **Normalized Satisfied**- (satisfied value from user - average of that attribute for all users) / standard deviation of that attribute for all users
  + **Combined Satisfied Attributes** - average of relevant satisfied attributes for a given selected attribute; used later for intensity calculations

Machine Learning Models

* **Clustering**
  + Used k-means clustering to determine the number and location of the clusters
  + Determined the most important variables on which to build relationships
  + At first, “selected” data were the only variables used
    - Then, introduced party characteristics, but no significant change
    - Eventually introduced all the information we had, but no characteristics stood out as significant
  + Began trying to create a cluster for each hotel in hopes of finding a single type of party that would stay at that hotel
    - Found there were multiple types of parties that would stay at some hotels
    - Also found that party dynamics were very similar across all hotels
  + We spent time trying to cluster within hotel categories
    - Found that frequently certain hotels that had specific differences from the rest of the category (like Art of Animation or Cabins at Fort Wilderness) would clearly differentiate themselves
    - Found it was hard to separate the majority of hotels
  + We also spent time trying cluster the category numbers. From this we found that generally the most important selection factors (in order of importance) were:
    - Value: cost and food court
    - Moderate: food court and cost
    - Deluxe: sit down dining, fine dining, and distance to parks
    - Deluxe Villa: in room kitchen, multiple bedroom suites, and room size
  + We also found the factors that people were generally most satisfied with (in order) within the different resort categories:
    - Value: airport shuttle and find your way around
    - Moderate: room quietness and airport shuttle
    - Deluxe: sit down dining, recreation amenities, and child services
    - Deluxe Villas: recreation amenities, sit down dining, and room size
  + Spent a lot of time clustering, but found we could not clearly differentiate parties who stayed at particular hotels, but clustering helped us understand the data and gave us some general knowledge about the hotels and categories
* **Decision Trees**
  + Needed a model that accounts for multiple paths to a given hotel
  + Binary decision trees for simplicity and ease of processing by software
  + Most common top-level feature is “selected in-room kitchen,” that separated villas, Art of Animation, and the Cabins at Fort Wilderness
  + Most common second-level feature is “selected distance to parks,” that separated deluxe resorts and deluxe villas from value and moderate resorts
  + Accuracy for all hotels not grouped by hotel category was consistently 20% (highest 25%)
  + Then divided by hotel category
    - Highest accuracy for value resorts was greater than 50%
    - Conversely, highest accuracy for deluxe resorts was just over 30%
  + Decision trees easily isolated some hotels (Cabins at Fort Wilderness and Art of Animation), but some could not be separated (All-Star Resorts)
* **Neural Networks/Ensemble Models**
  + Since surface-level data analysis did not reveal high-accuracy models, decided to implement the more complex, “black-box” methods of neural networks
    - Multilayer perceptron had average accuracy relative to other models (15-20% typically) when all hotels were included
    - However, when only deluxe resorts were analyzed, had 71% accuracy
  + Tried to combine the positive aspects of decision trees, clustering, and other models such as Naive-Bayes functions, logistic functions, and neural networks to create a more robust model
  + Accuracy did not increase significantly when the models with the highest accuracies were combined via voting and bagging mechanisms
    - Often decreased the overall accuracy
* Due to unsatisfactory accuracy for most machine learning models, these methods served as exploration and a means to further develop our intuition about the data.

Intensity

* **Defining Intensity**
  + Goal: take average satisfied attributes from each of the 32 hotels and intensify them with user selected attributes
  + Since the selected and satisfied attributes do not map one-to-one, we used the attributes that did or were closely related.
  + Below is the mapping of the 7 attributes that we chose:

|  |  |
| --- | --- |
| **Selected Attribute** | **Satisfied Attribute(s)** |
| Airport Shuttle | Airport Shuttle |
| Food Court | Food Court Overall and Food Court Value |
| Kids’ Activities | Recreation Amenities and Child Services |
| Park Shuttle | Park Shuttle |
| Pool | Pool Size, Pool Crowds, and Pool Cleanliness |
| Room Size | Room Size |
| Sit-Down Dining | Sit-Down Dining |

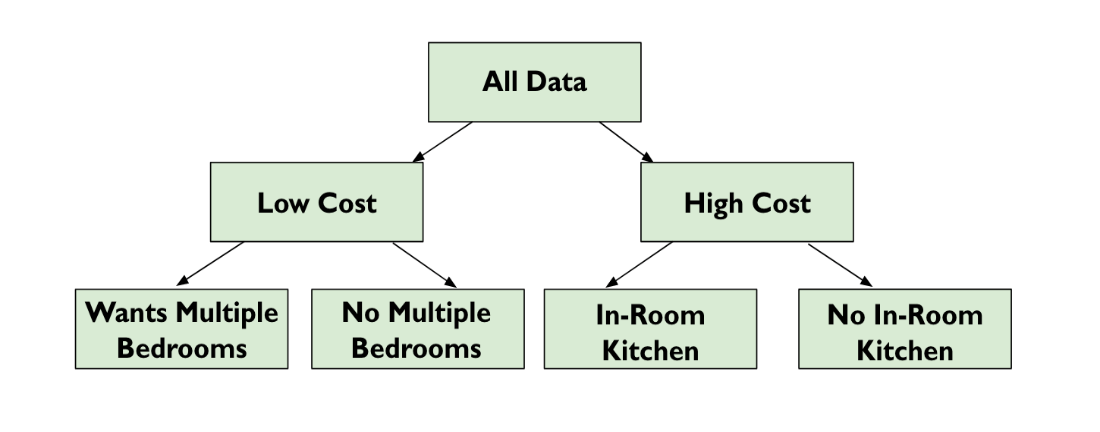
* **Calculating Intensity Vectors**
  + Each user will receive a seven-dimensional vector based on their specific selection preferences of the above attributes and each hotel will receive a seven-dimensional satisfaction intensity vector
  + To fill the user vectors, we calculated z-scores (in an attempt to normalize the data) for each attribute based on the selection preferences provided by the user. A new vector will be calculated for each individual user.
  + User z-score = xUser Input- All UsersAll Users
  + To fill the hotel vectors, we calculated z-scores (in an attempt to normalize the data) for each attribute based on the average satisfaction of each hotel provided by our data. Thus, these hotel vectors will not change, they are meant to be constant.
  + Hotel z-score = xHotel Average- For all HotelsFor all Hotels
  + Next, the z-scores were shifted such that the lowest possible value is 0 and all possible values are positive that way when calculating a dot product, two negative values would not produce a positive value.
  + So, for each attribute:
    - User z-score = xUser Input- All UsersAll Users+xMin User Input
    - Hotel z-score = xHotel Average- All HotelsAll Hotels+xMin Hotel Average
    - These values become the new components of the selected and satisfied vectors, respectively
* **Determining Intensity Scores and Recommending Hotels**
  + For each user vector, to calculate intensity scores for each hotel, we performed a dot product between the individual hotel vector and the user vector.
  + This produces 32 scalar values, resulting in an “intensity” score for each hotel.
  + The top 3 “intensity” scores are recommended to the user

Issues with Data

* **Satisfied and Selected attributes are not one-to-one.**
  + When we want to compare if the user is satisfied with the attributes they claimed to have chosen on, it is hard to see if they enjoyed that part of the hotel.
  + Had to compensate by averaging similar satisfied attributes together
    - This gives an inexact approximation of how satisfied the user really was with that attribute.
* **Incomplete/inconsistent survey information**
  + Many people did not enter their entire party.
    - For example, one zero-year-old female was the entire party.
  + The survey is formatted so that each party member has to be added manually, so people will naturally forgo this step.
    - Or they may enter one party member and forget to add the others
  + Some respondents reported staying at the Riviera Resort
    - Not currently open so the option should not exist on the survey
  + Hometown data was regularly inconsistent
    - Some respondents provided only the city which may have multiple instances around the world (for example, Greenville)
    - Respondents from international locations often provided only their country, and respondents from the US provided only their state
    - Comma separation and capitalization differed among respondents
      * Potential solution: create three separate boxes for the user to provide hometown, state/province (if applicable), and country
* **Inconsistent answering**
  + Selected responses were not answered in the same way each time
  + As a result, we had to split data into four different Excel sheets
    - Answered all of the selected - provided a ranking 1 through 5 for all selected attributes
    - Chose five and ranked by most important - provided their five most important factors on which they selected their hotel (1 being the least important and 5 being the most important)
    - Answered some of the selected but not all of them - provided a ranking 1 through 5 for some but not all selected attributes
    - Didn’t answer any selected - did not provide any selected attributes
  + This lowered our sample size by more than half
* **Misinterpretation of “Food Court”**
  + Respondents seem confused between quick-service options and food courts
  + Only value and moderate resorts have food courts, but deluxe resorts and deluxe villas often have high preference and satisfaction for food courts
  + Thus, respondents should only be able to rank selected and satisfied attributes pertaining to a food court if the hotel they stayed at actually has a food court
  + Similarly, adding more specialized dining categories such as quick-service and table-service alongside fine dining may eliminate confusion also
* **Distance to Parks and Location on Property very similar**
  + “Location on Property” may refer to how good a hotel’s location on Disney World’s campus
    - Clear parallel with “Distance to Parks” as a good location is often predicated upon how close the hotel is to certain parks

Best Model

* Found our best model was a combination of a decision tree, intensity model, and equivalence classes
* **Decision Tree**

****

* + First division was based on cost
    - Value and moderate resorts on one side, deluxe resorts and deluxe villas on the other
  + Second division on the low-cost side was the presence of multiple bedrooms
    - Separated the Cabins at Fort Wilderness, part of Art of Animation, and part of All-Star Music from the rest of the value and moderate resorts
  + Second division on the high-cost side was the presence of an in-room kitchen
    - Separated the deluxe villas and Animal Kingdom Lodge from the rest of the deluxe resorts
  + Four total divisions of our data
* **Intensity Model**
  + Once the user has been classified into one of the four categories, we then do the intensity calculations on each of the four leaves.
* **Equivalence Classes**
  + Some hotels were so similar that they were hard to distinguish
  + Someone could have stayed in a hotel very similar to the hotel we recommended to them, so we wanted to include this in accuracy calculations
  + We created 13 equivalence classes of similar hotels.
    - Grouped them first on cost then on location, and amenities
  + The equivalence classes are:
    - Art of Animation & Pop Century
    - All-Star Sports, All-Star Music, & All-Star Movies
    - Caribbean Beach & Coronado Springs
    - Cabins at Fort Wilderness
    - Port Orleans - French Quarter & Port Orleans - Riverside
    - Grand Floridian, Contemporary, Polynesian Village, & Wilderness Lodge
    - Animal Kingdom Lodge - Jambo House & Animal Kingdom Lodge - Kidani Village
    - Beach Club, Yacht Club & BoardWalk Inn
    - Swan & Dolphin
    - Bay Lake Tower, Polynesian Villas and Bungalows, & Villas at Grand Floridian
    - Copper Creek Villas & Boulder Ridge Villas
    - Beach Club Villas & BoardWalk Villas
    - Old Key West, Saratoga Springs, & Treehouse Villas
  + If the model predicted a hotel that was in the same equivalence class as the hotel where the user stayed, we considered our prediction to be correct
* **Results of Implementing Model**
  + Our final model was splitting the testing data into the four leaves, then performing intensity scores, and testing accuracy with equivalence classes.
  + Accuracy:
    - Low Cost, No Multiple Bedrooms: 60%
    - Low Cost, Multiple Bedrooms: 48%
    - High Cost, No In-Room Kitchen: 40%
    - High Cost, In-Room Kitchen: 23%
    - Overall: 42.75%

Further Questions/Explorations

* Should we consider the fact that most people staying at a deluxe villa are Disney Vacation Club members and will likely not answer this survey as much?
  + Because of this, we don’t want to over-recommend deluxe villas, especially considering the values in their satisfied vectors are typically higher than value and moderate resorts.
  + So it would be helpful to ask future respondents if he or she is a DVC member
* Could we create a “heat map” that colors WDW hotels based on their intensity score?
  + Using an actual map of WDW property, this would create a user-friendly and possibly interactive visual interface of our model.
* One idea we had towards the end of this project was to create an expert model.
  + Since the data we currently have is inconsistent and, consequently, inaccurate, we would use data-driven intuition to determine multiple sets of rules.
  + These rules would comprise a master list that would alleviate inconsistencies within the provided responses.
    - For example, if a user demands an in-room kitchen, any hotels that do not have a path that involves an in-room kitchen will be immediately eliminated from consideration.
  + We did not have a sufficient amount of time to explore this idea fully, but we felt as though an expert model would be more beneficial to make appropriate recommendations from than data-dependent models.