Guided Capstone Project Report

1. Introduction

The Big Mountain Resort, a ski resort located in Montana, offers spectacular views of Glacier National Park and Flathead National Forest, and is serviced by 11 lifts, 2 T-bars, and 1 magic carpet for novice skiers. The longest run is 3.3 miles in length. The base elevation is 4,464 ft, and the summit is 6,817 ft with a vertical drop of 2,353 ft. Every year about 350,000 people ski or snowboard at Big Mountain.

Big Mountain Resort has recently installed an additional chair lift to help increase the distribution of visitors across the mountain. This additional chair increases their operating costs by $1,540,000 this season. The resort intends to charge a premium above the average price of resorts in the similar market segment, and they want to understand the limitations of this approach. In order to help the resort select a better value for the ticket price, we take into account a number of factors that will either cut costs without undermining the ticket price or will support an even higher ticket price.

2. Data Wrangling

The provided raw dataset has the information of facility features, ticket price, resort open days, etc. from more than 300 ski resorts across the US, which are considered part of the same market share in this study. However, the raw data has missing and incorrect values for some features, so data wrangling is needed to clean and transform the data to be used in the subsequent analysis.

For example, data cleaning and transformation were conducted for the following features:

- 'FastEight' data were removed because their values are mostly 0.
- Adult weekday ticket price was dropped because the weekend ticket price is almost the same as the weekday ticket price for most of the resorts (**Figure 1**).
- The 'SkiableTerrain_ac' value for one resort seems incorrect, and after careful investigation, this value was fixed.
- Additional information was added such as number of resorts in each state, how long the resorts open every year, etc.

Finally, the target value was identified as the 'adult weekend ticket price', and the values for the most features have a clean distribution after data wrangling (**Figure 2**).

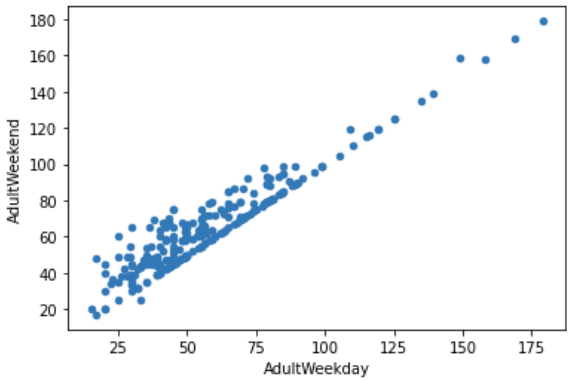

Figure 1. Relationship between adult weekday ticket price and adult weekend ticket price

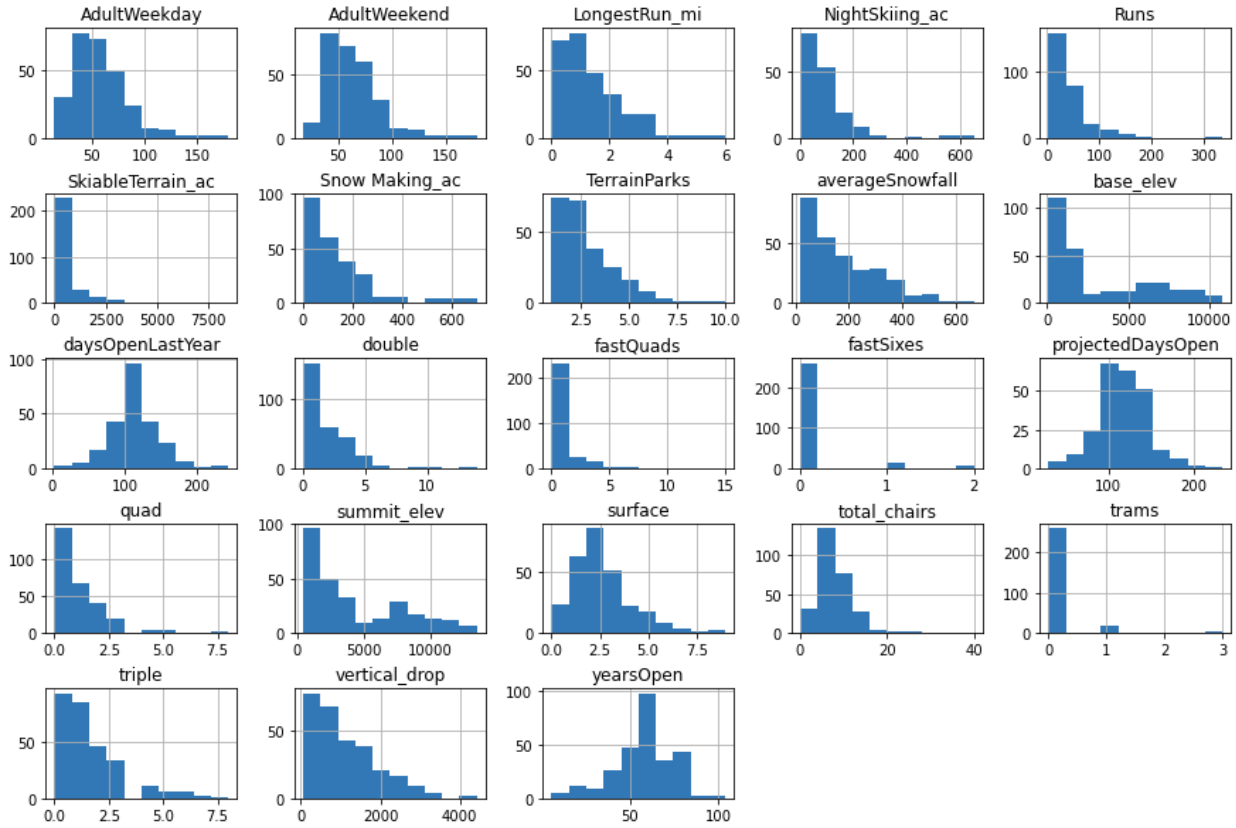

Figure 2. Histograms of most features after data wrangling

3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach for summarizing and visualizing the important characteristics and statistical properties of a dataset. Visualizing the data will help identify the potential relationships between the target value and other features and help form the hypotheses about the data.

In this study, the EDA was conducted to explore the top states by the order of total state area, total state population, resorts per state, total skiable area, total night skiing area, and total days open, as well as the resort density in each state by calculating the resorts per 100k capita, and resorts per 100k square miles. The results showed that some states are higher in some features but not in others, and some features are more corelated with one another than others. To disentangle these interconnected relationships, the principal components analysis (PCA) was applied to find the linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance they explain. After PCA transformation, the results suggested that there is not a clear relationship between state and ticket price, and therefore we can use the data from all states in the subsequent modeling.

Through the preliminary analysis, the ticket price has noticeable correlations with fastQuads, runs, snow making area, total chairs, vertical drop, etc. (**Figure 3**). Although there is not a clear relationship between state and ticket price, we can handle the states labels in the data using the additional features such as number of resorts per 100k capita, number of resorts per 100k square miles, etc.

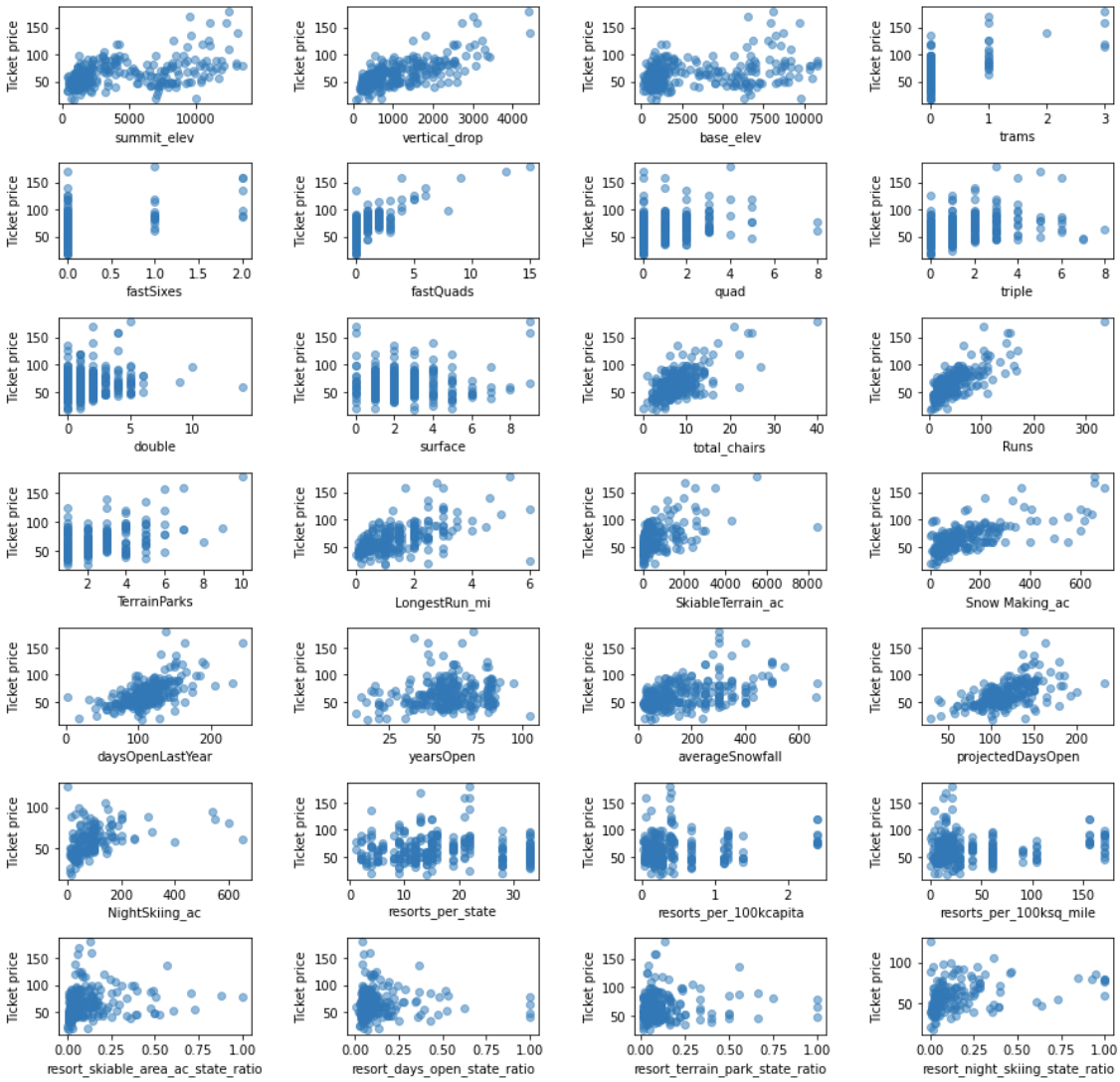

Figure 3. Relationships between adult weekend ticket price and other features

4. Pre-processing and Training Data Development

This step includes the process of splitting the dataset into testing and training subsets, which was adopted in this study to build the model to fit the training and test splits and then to use the cross-validation as a final check on the performance.

We conducted a simple linear model first to analyze the important statistics between the train split and test split, and then used the pipeline from sklearn package to train a linear regression model on the train split and apply to the test split for assessment. The results from the simple linear model and pipeline fit were about identical, which suggested the mean absolute error is around $9.

The linear model was then refined to prevent overfitting using SelectKBest from sklearn with f_regression. We created the pipeline with different values of k to train the model and selected the model with the best test set performance. After that, we used cross-validation to partition the train set

into k folds and train the model on k-1 of those folds, and eventually we built k models on k sets of data with k estimates of how the model perform on unseen data but without touching the test set. Using GridSearchCV, we determined the best value of k to be 8, and found out the following features are most useful, including vertical drop, snowing making area, total chairs, fastQauds, and total runs, etc.

Additionally, instead of repeatedly checking performance on the test split, the random forest model was applied to fit and assess the performance using cross-validation. We explored the hyperparameters with and without feature scaling and tried both the mean and median as strategies for imputing missing values. Using GridSearchCV, we also found the most important four features: fastQuads, total runs, snowing making area, and vertical drop (**Figure 4**), which are similar to those estimated from the refined linear model.

In the end, we selected the random forest model, because it has a lower cross-validation mean absolute error by almost $1, and it also exhibits less variability. In addition, both models have resulted in similar features which are important to the ticket price. In order to confirm that we have enough data for modeling, we used the learning_curve function for assessment and the results suggested that we have plenty of data.

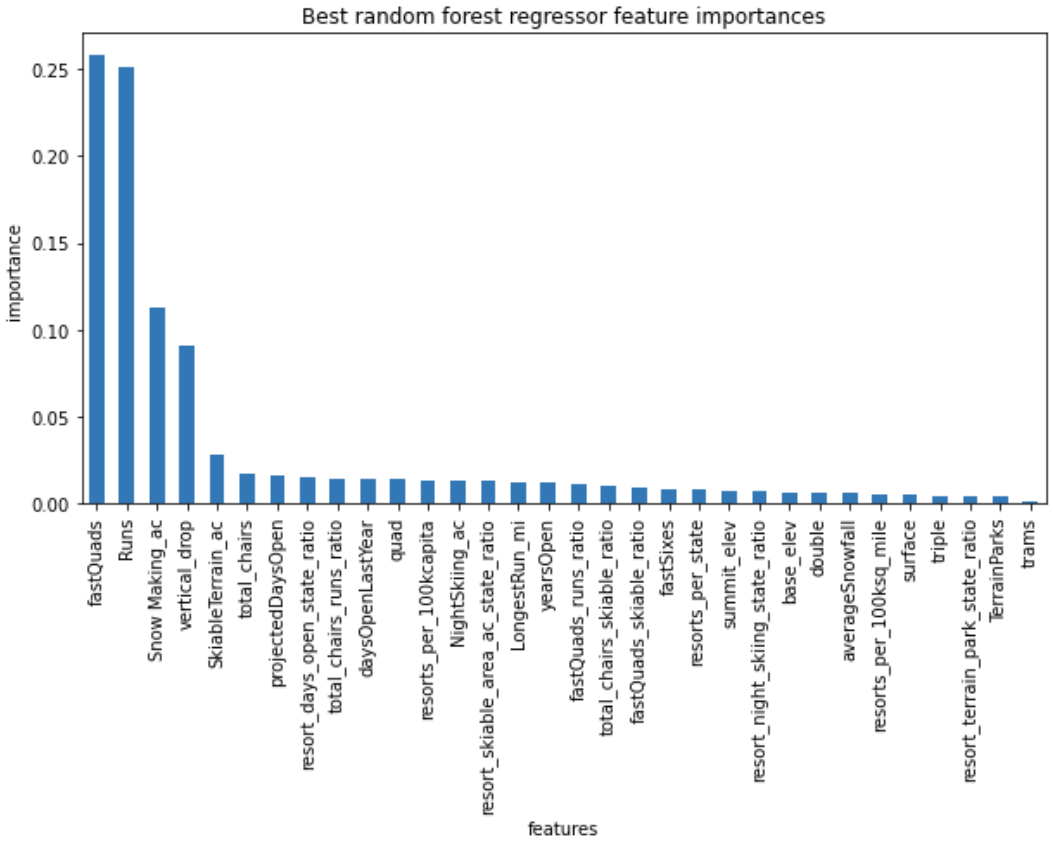

Figure 4. Importance of best random forest regressor features.

5. Modeling

With the model selected in the previous step, we can apply the model for ski resort ticket price and leverage it to gain some insights into what price Big Mountain's facilities might actually support as well as explore the sensitivity of changes to various resort parameters.

The current adult weekend ticket price is $81 for Big Mountain resort. After refitting the model with all the available data (except Big Mountain data) using the random forest model, the modelled price is $95.87 with a mean absolute error of $10.39, which suggested that the Big Mountain resort might be undercharging currently. Big Mountain is fairly high on some of the league charts of facilities offered, such as vertical drop, snow making area, total chairs, fastQuads, total runs, etc. (**Figure 5**); therefore, increasing the ticket price has the support due to those important features.

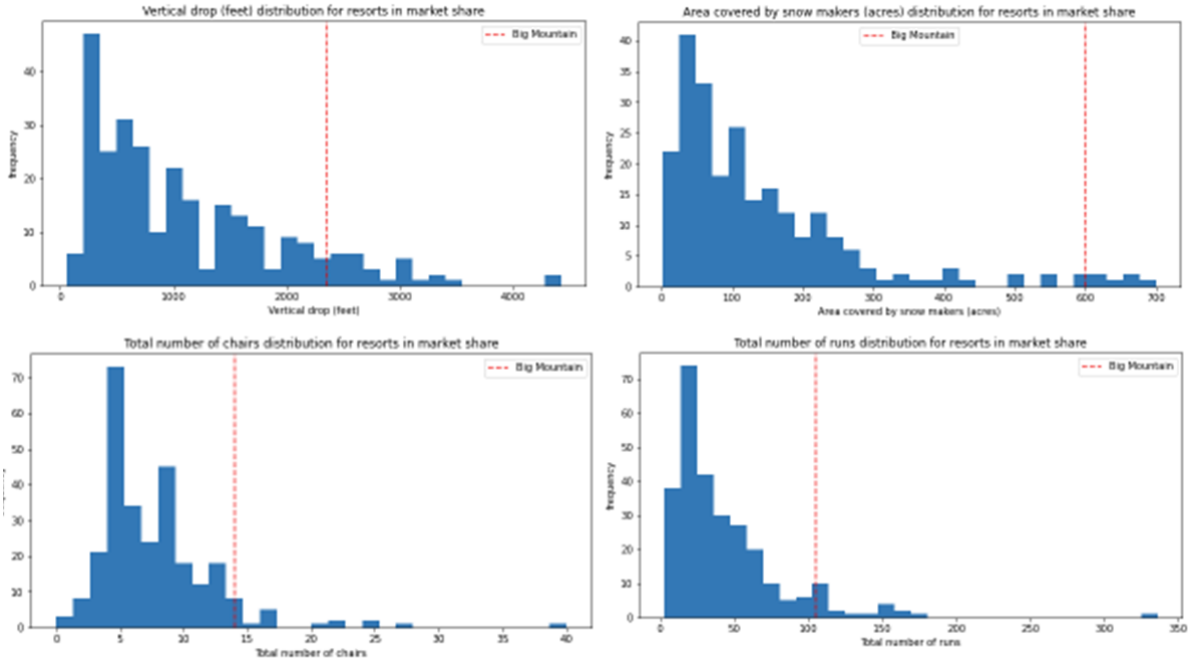

Figure 5. Histogram of vertical drop, snow making area, number of chairs, and number of runs

Four sensitivity scenarios were conducted to explore the impacts on price increase support with either cutting operating cost or improving some facilities, which assumed that the expected number of visitors over the season is 350,000 and the data has already included additional lift that Big Mountain recently installed. The results from the four scenarios suggest that:

- Closing several runs can reduce support for ticket price and revenue.
- If Big Mountain is adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift, then the ticket price increase is supported by $1.99.
- In addition to the changes in the scenario above, if Big Mountain is adding 2 acres of snow making, then the ticket price increase is supported by $1.99.
- If Big Mountain is increasing the longest run by 0.2 miles and adding 4 acres of snow making, it doesn’t support any price increase.

All these scenarios provided the potential changes to balance the revenue and operating cost by either cutting operating cost from other facilities or increasing the ticket price with improvement for some facilities.

6. Conclusion and Recommendations

To help the Big Mountain resort select a better value for the ticket price, we take into account a number of factors that will either cut costs without undermining the ticket price or will support an even higher ticket price. We performed data wrangling on the raw dataset to clean and transform the data to be used in the subsequent analysis, conducted EDA and pre-processing and training data development to visualize the important characteristics and statistical properties of the dataset, and to determine the relationships between the target value and some important features. Finally, we selected random forest model to conduct predictions on the ticket price for the Big Mountain resort using all available data (except Big Mountain data itself), which is predicted to be **$95.87** with a mean absolute error of $10.39.

The recommendations for the Big Mountain resort include:

- The Big Mountain can potentially increase the ticket price from **$81** currently to **$95.87** according to model prediction, which suggests that the Big Mountain resort might be undercharging currently. Big Mountain is fairly high on some of the league charts of facilities offered, such as vertical drop, snow making area, total chairs, fastQuads, total runs, etc.; therefore, these important features do provide support for the ticket price increase.
- If Big Mountain is adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift, then the ticket price increase is supported by $1.99.
- In addition to the changes in the scenario above, if Big Mountain is adding 2 acres of snow making, then the ticket price increase is supported by $1.99.
- If Big Mountain is increasing the longest run by 0.2 miles and adding 4 acres of snow making, it doesn't support any price increase.
- If the Big Mountain moves forward to close some runs, they should monitor how the operating cost decreases due to the run closure and if the number of visitors decrease significantly. As noted previously, the Big Mountain is doing well for several important features and thus increasing the ticket price has the support, so in addition to run closure, the resort can increase the ticket price slightly before any additional improvements to the facilities are made.
- The model applied in this study can be tested with any combination of parameters to predict the ticket price change. A future web-based interface tool based on this model is suggested to be created so the business analyst can use it to conduct scenario modeling.
- There are limitations in this study. For example, we only considered the ticket price, but there is a lot of other areas which can increase the revenue such as equipment rental fee, membership fee, and others. In terms of the operating cost, we only considered the cost of the new chair lift, but other operating costs should be included as well in the full-scale data analysis such as snow making cost, maintenance cost for chairs, quads, etc.