Recommendation on the Ticket Price Raise and Future Facility Changes for Big Mountain Resort

By Tingyin Xiao

Problem statement

Big Mountain Resort's price strategy has been to assess a premium above the average price of resorts in its market segment. It is currently charging 81 dollars for adults on weekdays and weekends. However, basing the pricing on the market average may have limitations, such as under-capitalizing its facilities. We have acquired the data on ticket prices and facilities of United States resorts, including Big Mountain, and built a model to predict a new ticket price based on Big Mountain's important facilities. The model can also aid the decision-making in future price strategies and cost cuttings under different scenarios of facility operations.

Data Wrangling

At first, we loaded the ski resort data with 330 rows, and our resort, "Big Mountain Resort," was present in the data. 13 columns had missing values. We conducted a preliminary exploratory analysis to inspect the distributions of the numeric variables and found several variables suspicious. For example, the skiable terrain area and snow-making area are clustered down the low end, "fastEight" only has one non-zero value, and "yearsOpen" has low values in general except for a maximum of 2019.

Therefore, we dropped the "fastEight" column as half of its values are missing, and only one value is non-zero. The one row's "yearsOpen" is 2019, apparently an erroneous entry because no resort could have opened for 2019 years. Therefore, we removed this row. During the data inspection, we found that some values were not correct. For example, the skiable terrain area for one resort, Silverton Mountain, was much larger than others. After searching for information about this resort, we found the correct number and replaced the original area with what we found.

After dropping the rows and columns mentioned above and fixing the wrong entry, we reviewed distributions and found that they are more reasonable than the original data. There are still some skewed distributions, such as for "fastQuads," "fastSixes," and "trams." They tend to lack much variance away from zero and have some extreme values. We should look out for these variables as some extreme values could overly influence models.

We found population information could be helpful, but the original data do not have this variable. Thus, we downloaded the US state data from Wikipedia, extracted each state's population and area, and merged them with the state summaries data.

We also dropped the rows with no price data as the price is our target. As weekend prices have the least missing values among the two price variables (weekday prices and weekend prices), we also dropped the weekday prices column ("AdultWeekday") and the rows that do not have weekend prices.

We have acquired the data we need for the project, examined the data shape, inspected the data quality, and cleaned the data. We can confirm that the target feature is the weekend prices ("AdultWeekend") as it has fewer missing values than the weekday prices. 277 rows and 25 columns are left in the ski resort data after the cleaning. We also create a new data file with state summaries. These data files are ready to be used in the following step of exploratory analysis.

Exploratory Data Analysis

To distill the information on relationships between state and ticket price, we first reduced the dimensionality of these statewide features through the principal components analysis (PCA). We found that the first two PCA components can explain 77.2% of the variance. Therefore, we extracted these two components, “PC1” and “PC2”, and plotted their scatter plot (Figure 1).

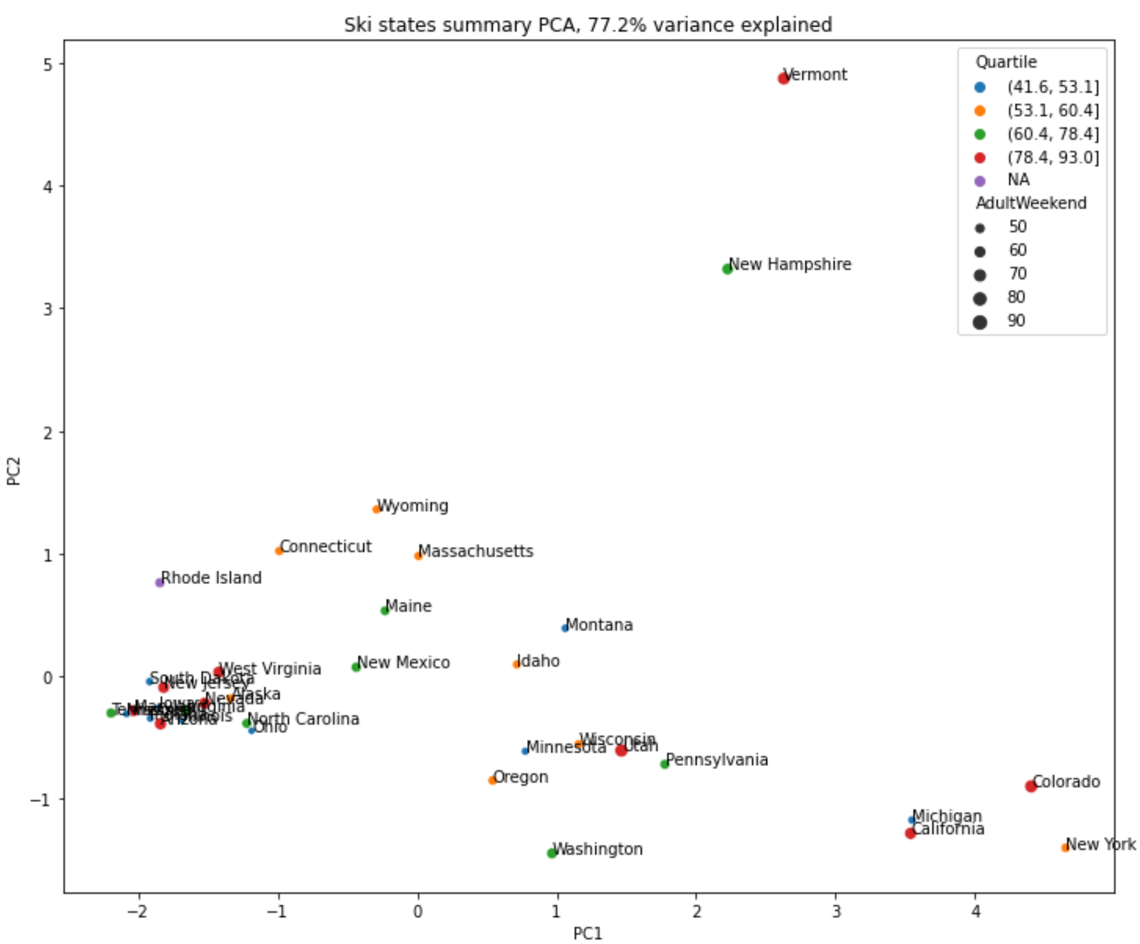

Figure 1. *Scatter plot of the first two PCA components, PC1 and PC2, of the state ski ticket price.*

On the other hand, we calculated each state's average weekend ticket price and added this information to the scatter plot. This scatter plot shows a wide spread of states across the first component, and Vermont and New Hampshire stand out in the second dimension. However, there is no obvious pattern with ticket prices in this scatter plot. To better understand why Vermont and New Hampshire are different for the second component, we checked how important each feature contributed to the second component. The result shows that features on the density of ski resorts, including “resorts_per_100kcapita” and “resorts_per_100ksq_mile”, contributed a lot in a positive sense. Looking into the original data, we can see that both states have particularly large values of “resorts_per_100ksq_mile” in absolute terms; and in the scaled data, we can see that they are more than three standard deviations from the mean of this value. Vermont also has an especially large value for “resorts_per_100kcapita”.

This observation offers some justification for building a pricing model that considers all states together, treating them equally. A couple of potential relevant statewide features emerged from this analysis, such as the resort density variables. Therefore, we merged the state data into the ski resort data and engineered some intuitive features, such as a resort's share of the supply for a given state. These 'state resort competition' features include the ratio of resort skiable area to total state skiable area, the ratio of resort days open to total state days open, the ratio of resort terrain park count to total state terrain park count, and the ratio of resort night skiing area to total state night skiing area.

Then, we plotted the feature correlation heatmap to gain a high-level view of relationships amongst the features (Figure 2). The heatmap shows multicollinearity between several features. For example, summit and base elevation is highly correlated, and the ratio features we calculated are negatively correlated with the number of resorts in each state. When selecting the features for modeling, we should be wary of the multicollinearity issue of the features.

The heatmap shows many reasonable correlations with our target feature, the weekend ticket price (“AdultWeekend”), including “fastQuads”, “Runs”, “total_chairs”, and “Snow Making_ac”. The last one indicates that visitors might value more guaranteed snow, which would drive up the costs and prices. Among the ratio variables, “resort_night_skiiing_state_ratio” seems to be the most correlated with the ticket price. If true, maybe supporting a larger share of night skiing capacity can contribute positively to the price a resort can charge.

To examine the relationships between each feature and ticket price, we plot their scatterplots against ticket price (Figure 3). The plots pick up many strong positive correlations, such as with “vertical_drop”, “fastQuads”, “total_chairs”, “Runs”, and “Snow Making_ac”. They also show some interesting patterns that cannot be seen from the correlation heatmap. For example, when the value of “resorts_per_100kcapita” is low, the variability in the ticket price is high, and its value can go very high. Ticket prices may drop slightly and then climb upwards as the number of resorts per capita increases. The reason is unclear, but it is interesting to explore further.

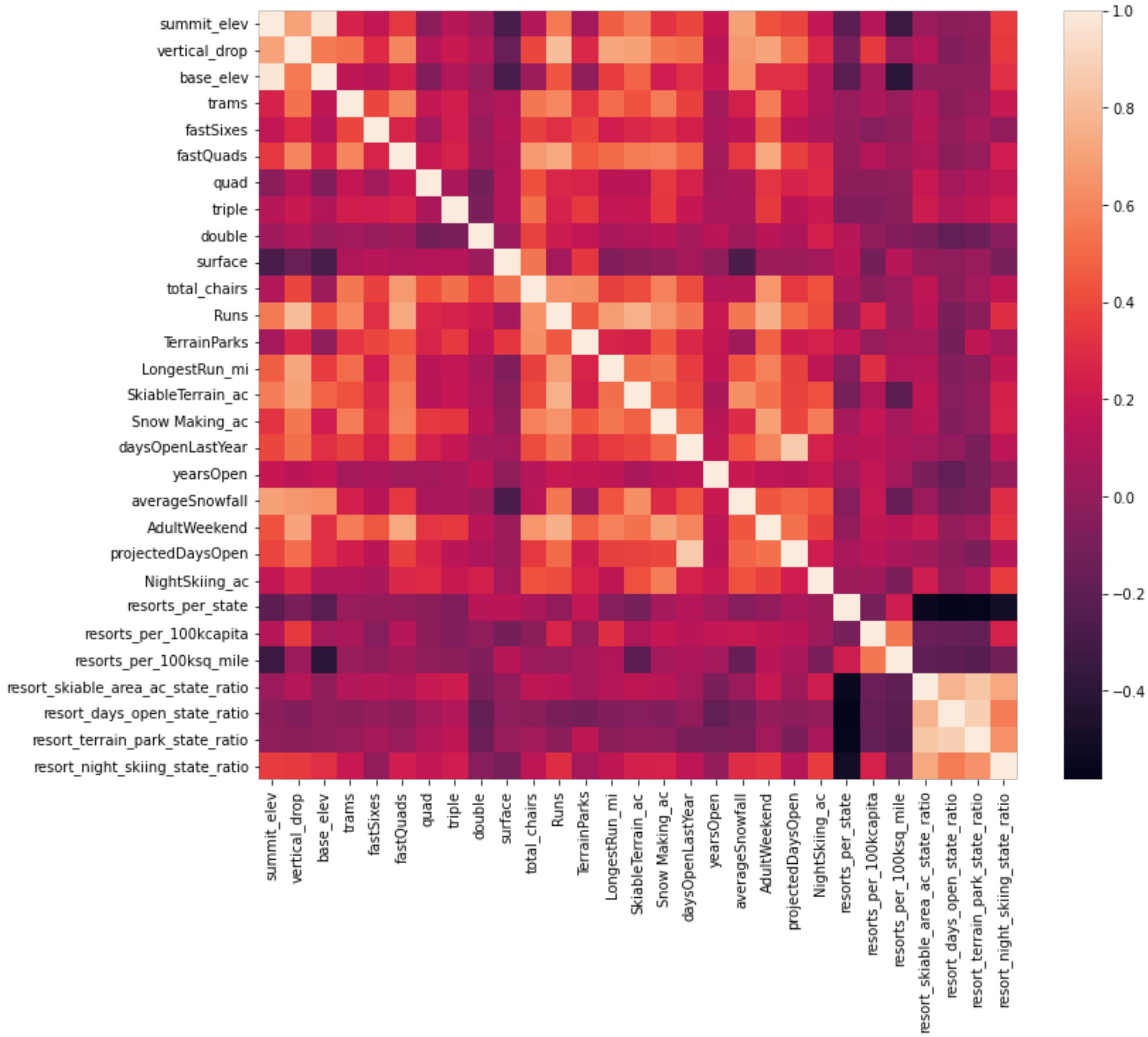

Figure 2. Feature correlation heatmap.

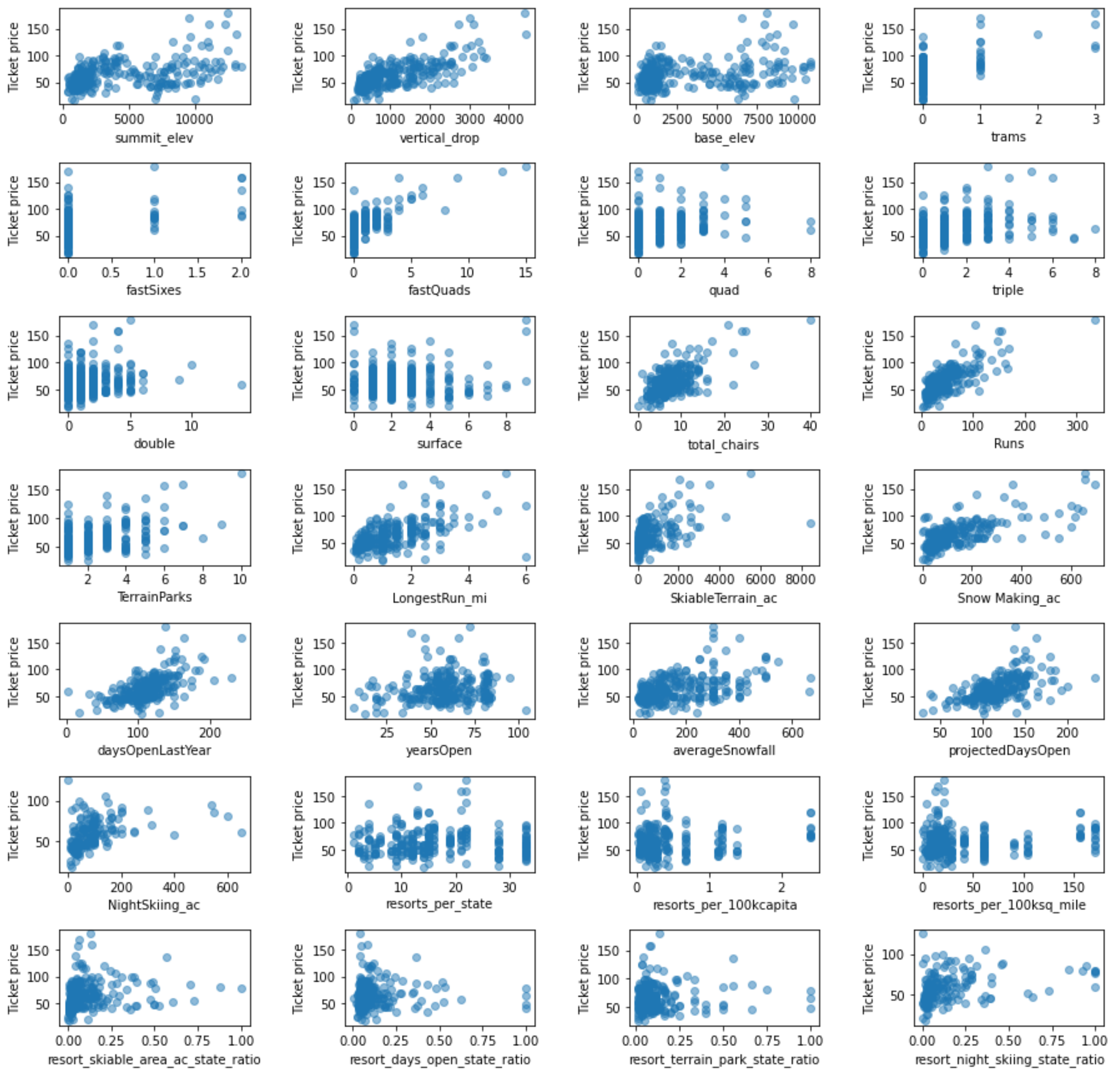

Figure 3. Scatterplots of each feature against ticket price.

The exploratory data analysis first inspected the data and engineered some features. It also provided a list of potential contributing features to the ticket prices and showed their relationships with each other. The data is saved and ready to be used in the next step.

Model Preprocessing with feature engineering

In the data preprocessing step, we firstly separate the data for Big Mountain, our target resort, from the rest of the data to use later. Then, we split the feature and target of the rest of ski resort data into training and test sets with a 70:30 partition.

Algorithms used to build the model with evaluation

Our first best guess of the ticket price is the average price, which is $63.8. As expected, the R-squared of this predictor is zero on the training set and negative (about -0.003) on the testing set. The mean absolute error (MAE) is 17.9 on the training set and 19.1 on the testing set, while the mean squared error (MSE) is 613.1 on the training set and 581.4 on the testing set. It can serve as the baseline of comparison for our prediction models, which should at least outperform the average price. The R-squared score is the larger, the better, while MAE and MSE are the smaller, the better.

We then built a linear model and found "vertical_drop", "Snow Making_ac", "total_chairs", "fastQuads", "Runs", and "LongestRun_mi" are the features that contribute most positively to the ticket price, which are all reasonable. The features "trams" and "SkiableTerrain_ac" contribute most negatively to the ticket price. The reasons why these features would contribute negatively to the price are unclear, but we should pay attention to them in the future modeling building. We imputed missing feature values using the median or the mean and scaled them to bring them all on a consistent scale. The estimate of the model performance is much better than the average price. Imputing missing feature values using the median or the mean does not make much difference. The linear regression models explain over 80% of the variance on the train set and over 70% on the test set. The lower value for the test set is expected and may suggest we are somewhat overfitting.

We did cross-validation by partitioning our training set into k folds, trained our data on k-1 of these folds, and calculated performance on the fold not used in training. We cycled through k times with a different fold held back each time. In this way, we built k models on k sets of data with k estimates of how the model performs on unseen data without seeing the test set. Using k = 5 in this case, we got the performance estimates with a mean R-squared of 0.63 and a standard deviation of 0.095. The mean absolute error of the cross-validation with k = 5 has a mean of 10.50 and a standard deviation of 1.622. Its performance on the test split was consistent with this estimate, with the mean absolute error being 11.79.

We also tried a random forest regressor. With hyperparameter search, we found that imputing with the median is the best, but scaling the features does not help. The estimate of its performance has a mean R-squared score of 0.71 and a standard deviation of 0.065. The mean absolute error has a mean of 9.64 and a standard deviation of 1.353. Its performance on the test set was consistent with this estimate too, which has a mean absolute error of 9.54.

Winning model and scenario modeling

Based on the above analysis, we chose to use the random forest because it has lower cross-validation mean absolute error than the linear regression model by almost one dollar. It also shows less variability in the estimate of performance.

Pricing recommendation

Using the random forest model, we found that Big Mountain Resort has several advantages regarding facilities influencing ticket prices. For example, its vertical drop is greater than the majority of the resorts; its snow-making area is more than almost all the resorts; its number of total chairs is among the highest; it has three fast quads while most resorts have none; it also has more runs than almost all the other resorts and has one of the longest runs; it is also among the resorts with the largest amount of skiable terrains (Figure 4). These advantages in the market context can justify a substantially higher price than the market average. Our model suggests Big Mountain charge 95.87 dollars, with the expected mean absolute error of 10.39 dollars, assuming other resorts set their prices accurately according to the market's support.

Big Mountain Resort also has recently installed an additional chair lift to help increase the distribution of visitors across the mountain. This extra chair increases the operating costs by 1,540,000 dollars this season. Assuming the number of expected visitors for this season is 350,000, and each visitor buys five-day tickets on average, we expect Big Mountain to sell 1,750,000 tickets. Therefore, raising the ticket price by 0.88 dollars can cover these additional operating costs.

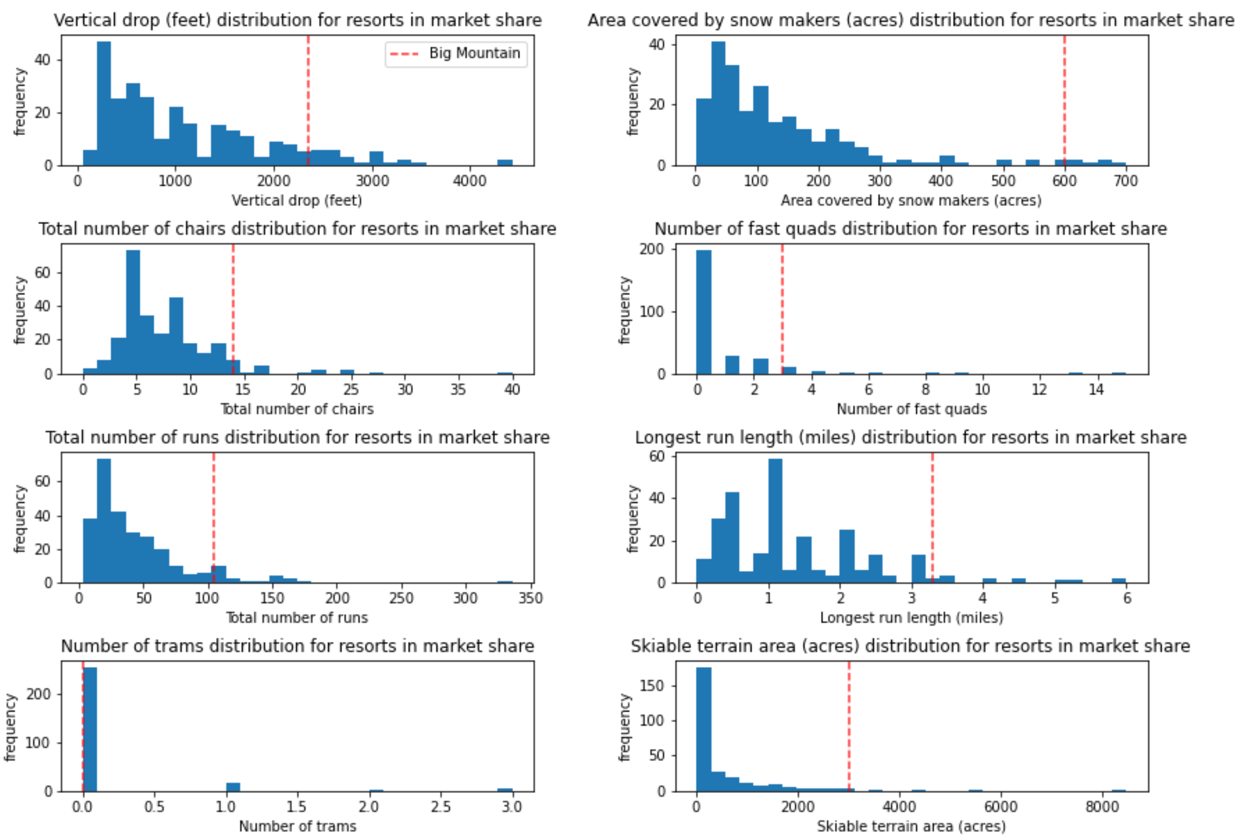

Figure 4. *Position of Big Mountain Resort regarding facilities in market share*

Conclusion

For future improvements, we explored several scenarios of facility changes that could help support higher ticket prices or cut costs without undermining the ticket price too much. We found adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift can increase support for ticket prices by 8.61 dollars and could amount to $15,065,471 this season. The model also suggests cutting costs by closing the least used runs. It shows that closing one run seems to make no difference; closing 2 and 3 runs reduce support for ticket price and revenue and closing 3 runs has the same effect as closing 4 or 5 runs; closing 6 or more runs will lead to a significant drop (Figure 5). Therefore, we suggest closing the least used run first and then testing the closing of the second least used run to see the effects; if the outcome is manageable, close the third, fourth, and fifth least used runs together and see the impact.

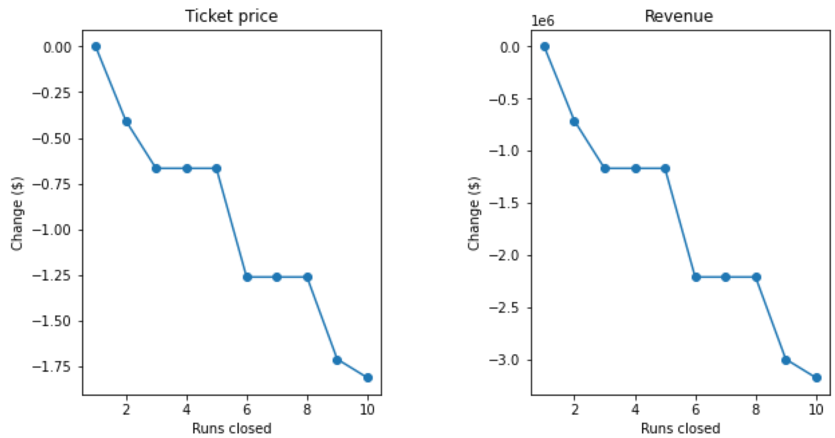

Figure 5. *Predicted ticket price (left) and revenue (right) change in dollars (y-axis, note that for revenue it is in 1e6 dollars) after closing number of runs (x-axis).*

Future scope of work

The scenarios we have tested are limited. Suppose Big Mountain finds our model helpful and wants to test a new combination of parameters in a scenario. In that case, we can write a function so that their business analysts can easily explore different scenarios and make decisions on their own.

On the last note, the above discussion of the scenarios is mainly based on the information on the ticket prices alone. The other component of profitability, the costs, are also essential in decision making. In addition to the additional operating cost of the new chair lift that we know, it will be helpful to have other cost information to make a more accurate prediction for pricing strategy design.