**Guided Capstone Project Report**

Ban Qu

**Problem Statement**

Big Mountain Resort has recently installed a chair lift that would cost $1,540,000 over the season and the company has always preferred a premium charge for the ticket. Based on the current condition, the company should increase the ticket price and cut costs from the additional chairlift cost. Considering this, to maintain a competitive strength in the market, the problem statement for Big Mountain is as follows: “How can Big Mountain Resort select a better value for its ticket price without undermining the ticket price or supporting an even higher ticket price over the year?”. However, one of the constraints is that there might be an issue of insufficient capitalization on facilities, which the company needs to tackle in the long run.

**Data Wrangling**

The dataset contains 27 columns and 330 rows, with categorical columns ‘Name’, ‘Region’, and ‘State’. Based on suspicious observations of the columns, one value is changed for ‘SkiableTerrain\_ac’, one row is removed for ‘SnowMaking\_ac’, the entire column is removed for ‘fastEight’, and ‘yearsOpen’ only contains those with less than 1000 years of history.

A summary variable containing ‘resorts\_per\_state’ and sum of ‘TerrianParks’, ‘SkiableTerrian\_ac’, ‘daysOpenLastyear’, ‘NightSkiing\_ac’ by states is created to store the state summary data in interest. And rows with no price data (i.e., no weekday and no weekend pricing data) are all dropped. Additional columns are added from an external wiki source that contains states population and states area via left merging summary table with the external website table.

Finally, ‘AdultWeekday’ column is dropped because weekend prices have the least missing values of the two, and weekend prices are higher than weekday prices seem restricted to sub $100 resorts.

The modified dataset contains 277 rows and 25 columns.

**Exploratory Data Analysis**

The absolute population and state size columns are replaced with ‘resorts\_per\_100kcapita’ and ‘resorts\_per\_100ksq\_mile’ respectively for better measures of resort density. Average price for ‘AdultWeekend’ is concatenated with PC1 and PC2 by states, along with quartile for the price. In the first two components, there is a spread of states across the first component. It looks like Vermont and New Hampshire might be off on their own a little in the second dimension, and New York and Colorado are a bit off in the first dimension.

To use state labels and explore the resort-level data in more detail, two datasets are merged (i.e., ski\_data left merging state\_summary on ‘state’). State resort competition features are added and the corresponding columns are replaced (i.e., 'resort\_skiable\_area\_ac\_state\_ratio', 'resort\_days\_open\_state\_ratio', 'resort\_terrain\_park\_state\_ratio', and 'resort\_night\_skiing\_state\_ratio' replacing 'state\_total\_skiable\_area\_ac', 'state\_total\_days\_open', 'state\_total\_terrain\_parks', and 'state\_total\_nightskiing\_ac'). Then a heatmap is created to get an initial overview of correlations between features. It is noticeable that ‘fastQuads’, ‘Runs’, ‘Snow Making\_ac’, and ‘total\_chairs’ are well correlated with the target feature ‘AdultWeekend’. From the scatterplots, ‘vertical\_drop’ is highly positively correlated with the ticket price; ‘fastQuads’, ‘total\_chairs’ and ‘Runs’ are also very important.

To add features of ratio of chairs to runs, 4 new columns are added: 'total\_chairs\_runs\_ratio', 'total\_chairs\_skiable\_ratio', 'fastQuads\_runs\_ratio', and 'fastQuads\_skiable\_ratio'. The data is missing an important feature which is the number of visitors per year. It also appears that having no fast quads may limit the ticket price, but if the resort covers a wide area, then getting a small number of fast quads may be beneficial to ticket price.

In sum, target features interested in the further modelling involves: fastQuads’, ‘Runs’, ‘Snow Making\_ac’, ‘total\_chairs’, and ‘vertical\_drop’.

**Pre-processing and Training Data**

The data is split into training and test for machine learning models. The ratio of split is 70/30. Training data involves all factors other than AdultWeekend and test data only includes AdultWeekend. Name, state, and Region are removed and only numeric factors are reserved for X\_train and X\_test.

Linear regression is applied first on the model prediction. Pipeline method is practiced in the following manners: imputing missing values (median and mean, respectively), scaling the features, training the linear regression model, and calculating model performance. The performance metrics for imputing missing values with Median and Mean are computed (Figure 1). Since the model always performs worse in the test data, it is suspected that the model is overfitting. To select the best features to fit the model, we refine the linear model with SelectKBest function along with cross-validation and GridSearchCV. As a result, k=8 (i.e., 8 features) gives the best performance. According to the model, vertical\_drop is greatly associated with pricing, followed by Snow Making\_ac, total\_chairs, fastQuads, and Runs, etc. It is similar to the results we did on EDA previously.

Similarly, Random Forest model is performed for the regression analysis. We fit and assess performance using pipeline and cross-validation again. The model suggests that the dominant top four features correspond with the linear model.

Finally, MAE is used along with cross validation to select the best model between linear regression and random forest. As a result, random forest model has a lower cross-validation mean absolute error which suggests it is a better model in this case (9.6231 for random forest vs. 11.7934 for linear regression).

**Modeling**

The current price (i.e., weekend adult price) for Big Mountain is $81. First, to see how the current price fits the market, a model is trained (random forest) inclusively with all other resorts other than Big Mountain. The result shows that the modelled price is $98.74, so even with a mean absolute error of $10.30, there is room for a price increase.

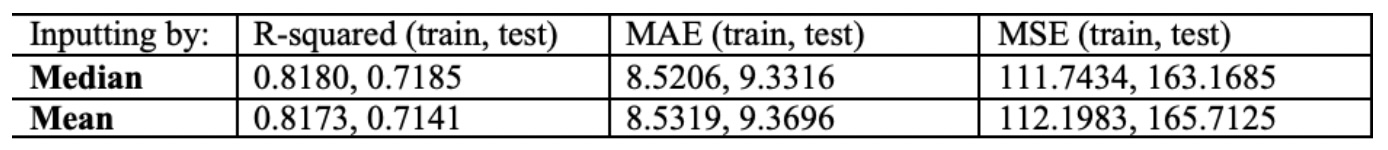
It is worthwhile to review the top features (i.e., features most likely to affect pricing) between Big Mountain and other resorts so that we would know where our resort sits in the market. Histograms are used for better visualizations. Overall, our resort compares very well for most important features. For instance, Big Mountain is among the top resorts with high vertical drop, or vertical change in elevation from the summit to the base (*Figure 2*), snow making, or total area covered by snow making machines (*Figure 3*), total number of chair lifts (*Figure 4*), the number of fast six person chairs (*Figure 5*), the number of runs (*Figure 6*), length of the longest run (*Figure 7*), and total skiable area (*Figure 8*).

With the options provided for discussion, I have the following recommendations to the board. First, regarding how many runs should be closed, the optimized number based on the model is 2 runs; closing 2 and 3 runs successively reduces support for ticket price and so revenue (*Figure 9*). From the illustration, increasing the runs over 6 will lead to a huge drop in ticket price and revenue. I would suggest that the business could start to close 2 runs and see how it affects the number of visitors and thus adjusting the ticket price. If doing well, the company may continue to close 1 and 2 more runs, respectively, to test for pricing. However, it should not close more than 5 runs based on the model. Second, if the resort is increasing the vertical drop by 150 ft, adding one run, with an additional chairlift, the support for ticket price will increase by $1.88, and revenue by around 3.3 million over the season. This option is highly recommended based on the profits it will bring as well as the fact that vertical drop is the top influential feature on pricing among all resorts. In fact, the company may consider increase vertical drop even further (e.g., by 200 feet or 250 feet) in the future to see the effect on pricing and revenue. Third, adding 2 acres of snow on top of the previous scenario has no effect on both ticket price and revenue, and therefore it is not recommended for the business. In addition, there will be no effect on ticket price and revenue if we increase the longest run by 0.2 miles and adding 4 acres of snow making. Therefore, the last option should be omitted from the discussion.

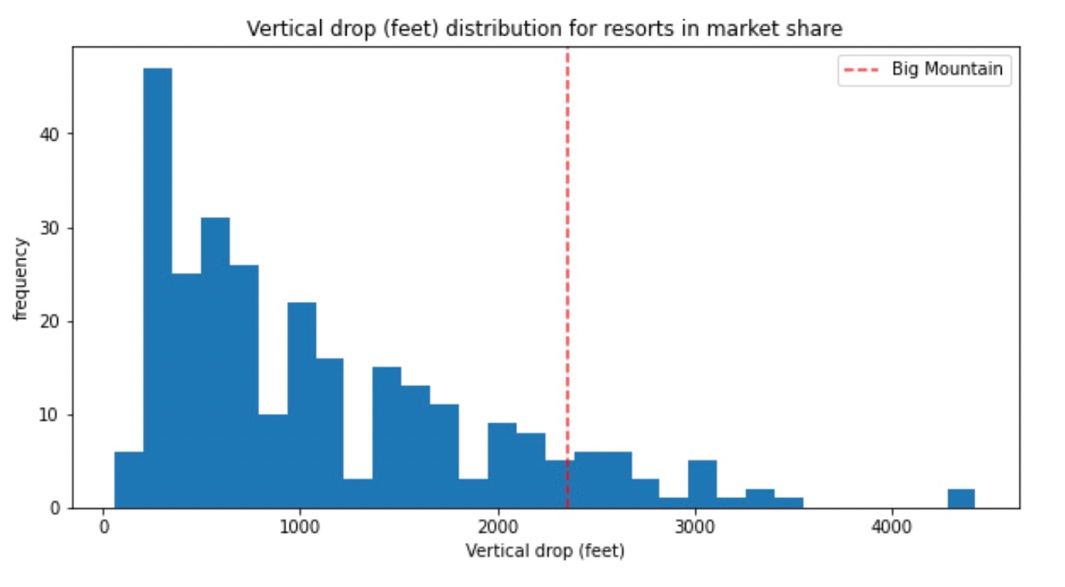
For further suggestions, the company may consider how adding/dropping the number of fast quads would affect the number of visitors and ticket price. Since most resorts have no fast quads, and fast quads is among the top influential variables, I would recommend another modelling scenario in the future regarding adding or dropping fast quads. If adding or dropping 1 or 2 fast quads has no makable impact on ticket price and revenue, the company might need to consider dropping fast quads to cut related expenses.

In sum, to increase pricing for a better revenue, as starters, Big Mountain could start to close 2 runs and increase the vertical drop by adding a run from 150 feet lower down with an additional chair lift. Based on the progress and results, the business could start to test on secondary options to further increase pricing, such as further increasing the vertical drop, closing more runs up to 5 (based on the usage frequency), and adding/dropping fast quads.

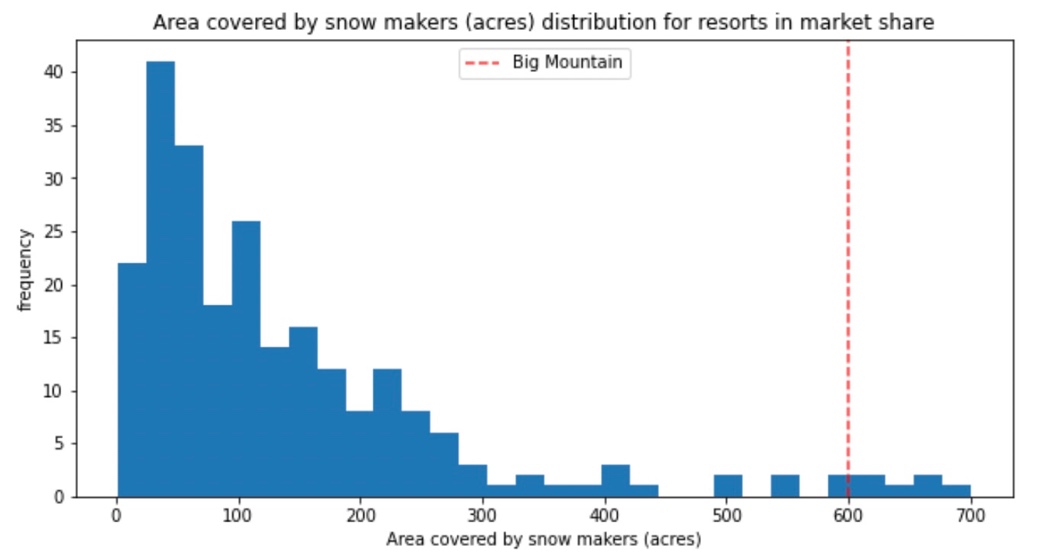
***Figure 1***

******

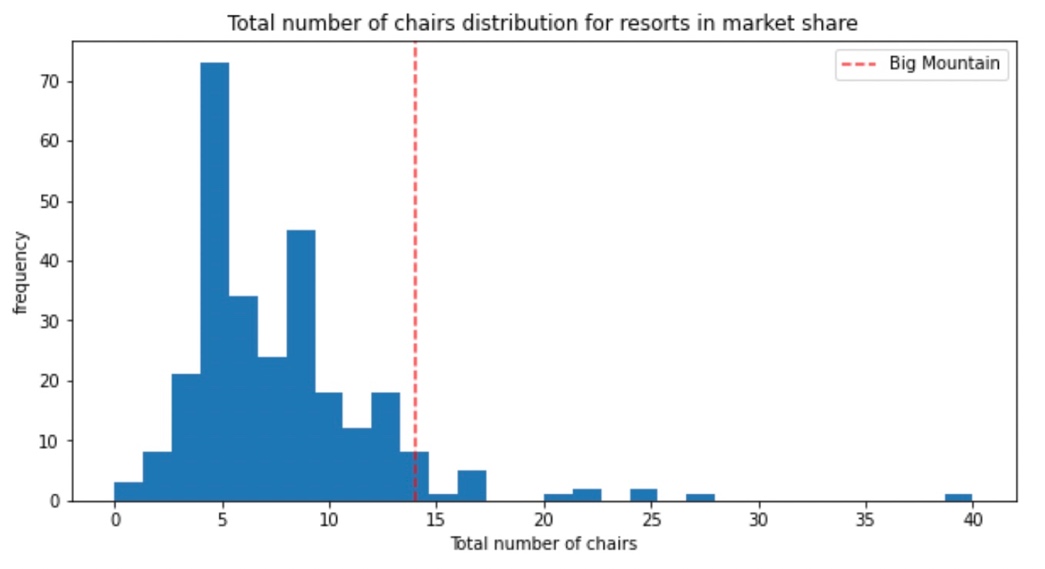
***Figure 2***



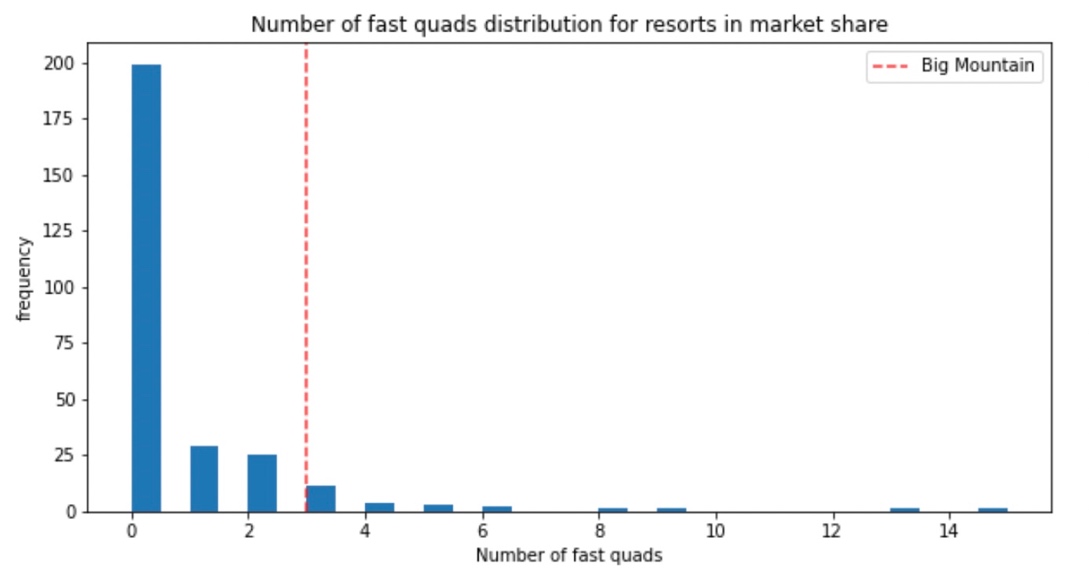
***Figure 3***



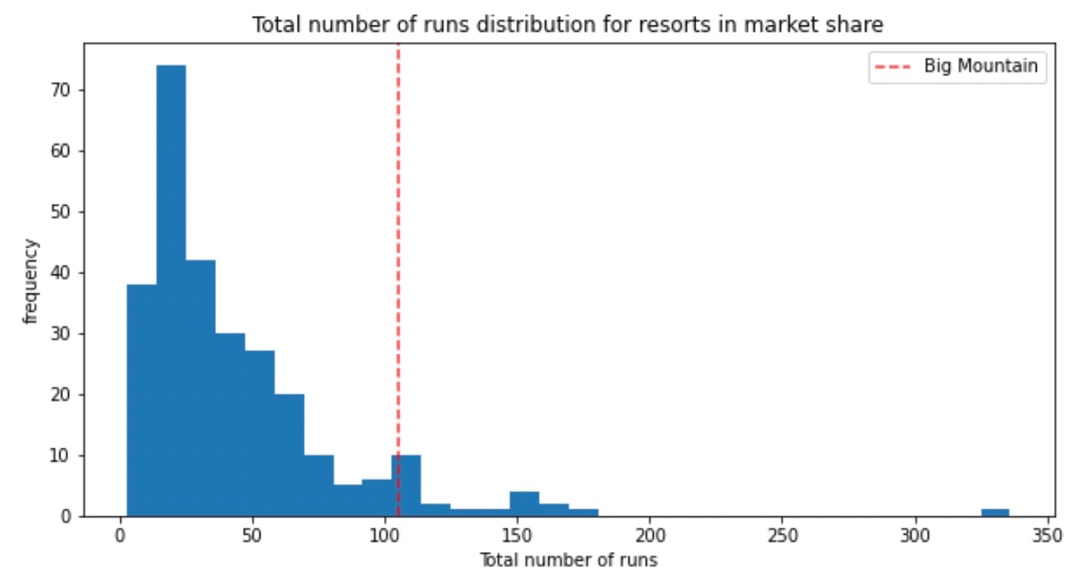
***Figure 4***



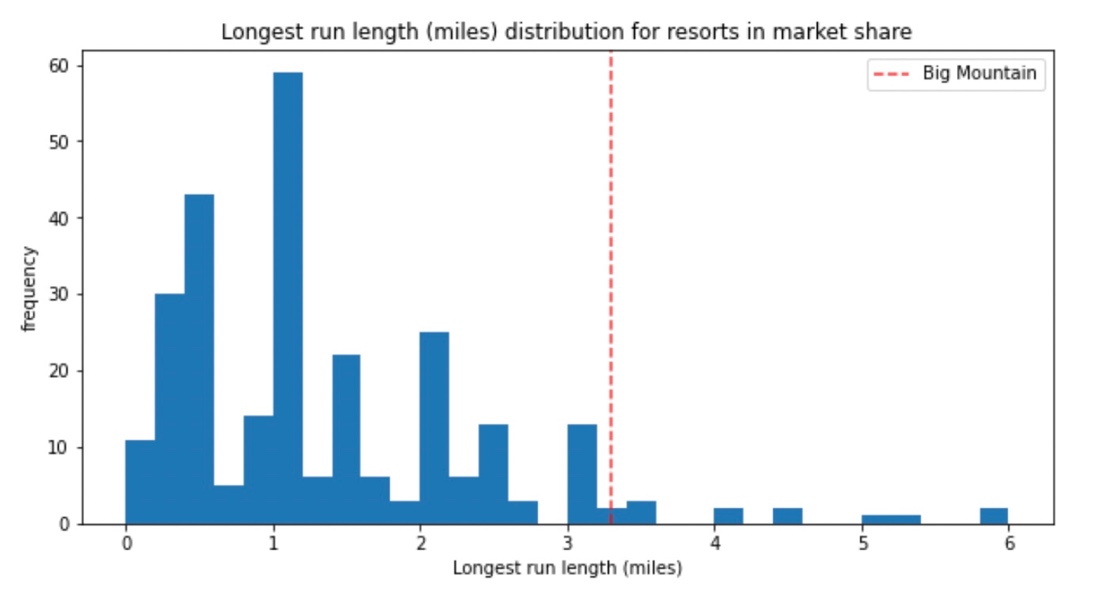
***Figure 5***



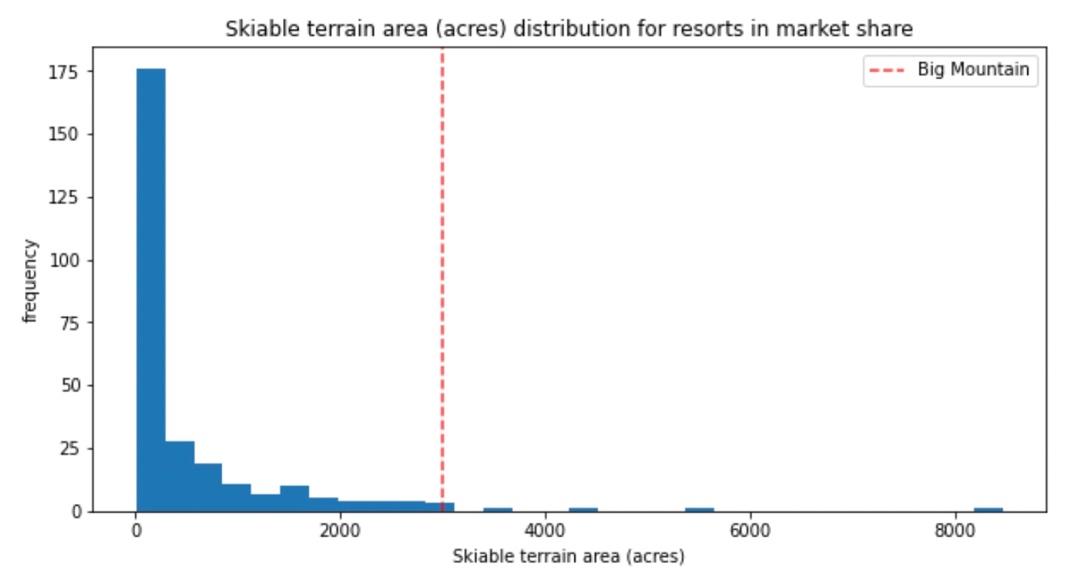
***Figure 6***



***Figure 7***



***Figure 8***



***Figure 9***

