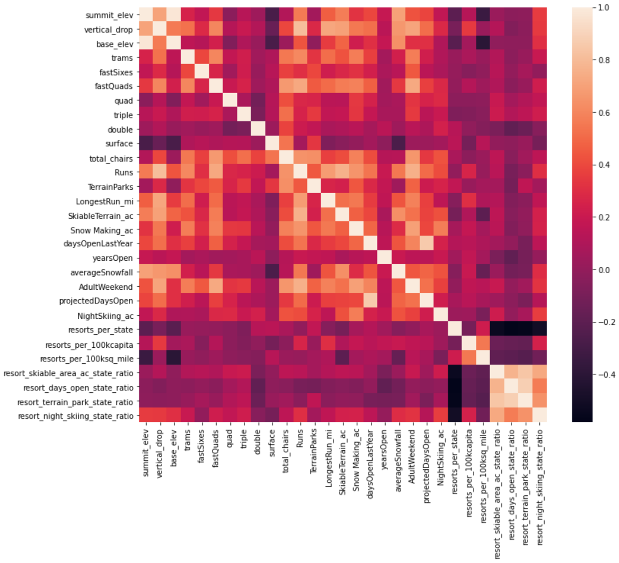
Guided Capstone Project Report

Big Mountain Resort currently determines their ticket prices based on the market average; essentially charging more than what other resorts in its market segment charge. However, it is believed that determining ticket prices based on the facilities Big Mountain has on the resort instead would lead to better value for their ticket prices. As a result, the focus was on using data, provided by Database Manager Alesha Elsen, about facilities and ticket prices in Big Mountain Resort and other resorts to determine the best ticket prices.

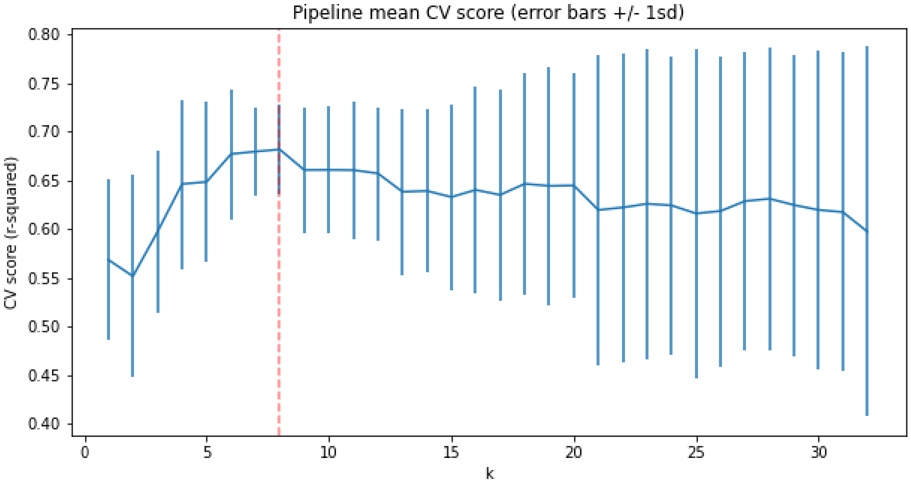
An exploratory analysis of the data was undertaken to find relationships amongst the features that may give an insight into solving the problem. To help, a heatmap of correlations between the features was created as seen below. The lighter the square between two features the higher the correlation between those two features are and the vice versa is also true. Of course, the white diagonal line just shows when a feature is being compared to itself. From this heat map, a lot of relationships can be seen at a high level. When looking at adult weekend ticket prices, labeled ‘AdultWeekend’, we can see some very interesting correlations. For example, the labels ‘fastQuads’, which is the amount of fast four person chairs a resort has, and ‘Snow Making_ac’, which is the total area in acres covered by snow making machines in a resort, have a positive correlation with weekend ticket prices. The last one is interesting because it means visitors may be willing to pay more if there was more guaranteed snow in the resort, however more snow machined would drive up operating costs. There also seems to be a positive correlation between the ratio of resort night skiing areas to the total state night skiing area and the price of tickets. This perhaps indicates that when a resort takes a larger share of the night skiing area within the state, it may lead to visitors adopting a greater willingness to pay higher prices for tickets. The heatmap also shows that the labels ‘Runs’ and ‘total_chairs’, total number of runs and total number of chair lifts respectively, have a positive correlation. This makes sense seeing as more runs, which require more chairs, can attract more visitors.



To find and develop the best model for this problem a baseline performance had to be established. First, the data, without the Big Mountain part, was partitioned into training and testing splits with about 70% of the data put into the training split. The baseline model was simply using the mean as a predictor. The average price under the ‘AdultWeekend’ column in the training split was calculated to be about $63.81. In order to assess how good the mean is as a predictor, the R^2, MAE, and MSE were calculated when comparing the mean to the ‘AdultWeekend’ prices from the test split. The calculated values were about -0.0031, 19.14, and

581.4, respectively. This shows that the mean is not a good predictor since it is on average about $20.00 off from the actual price. However, it does establish the baseline performance.

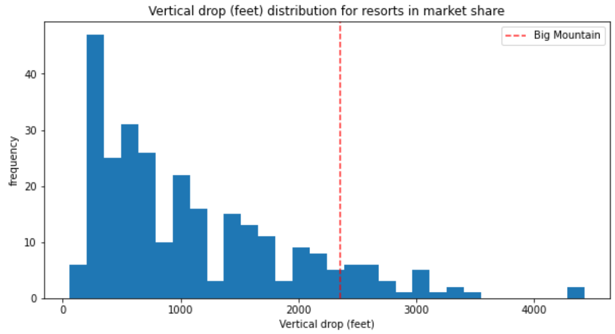
One of the first models I tested was a linear regression model. Using five fold cross-validation with GridSearchCV, the best number of features to train this model on was determined. That number was 8 features and those features were the vertical drop, the area covered by snow making machines, the total number of chair lifts, the number of fast 4-person chair lifts, the number of runs, the distance of the longest run, the number of trams, and the area of skiable terrain. Curiously, the number of trams and the skiable area were negative features but were still important in determining the best price.These results come from GridSearchCV having tested every possible amount of features from using just one to using all 32 features. As can be seen in the graph below, at 8 features, represented by the red dashed line, the mean of R^2 is at the peak and the standard deviation of R^2, represented by the vertical error bars, is relatively small.

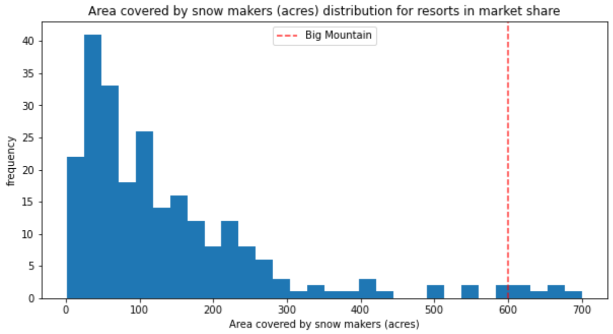


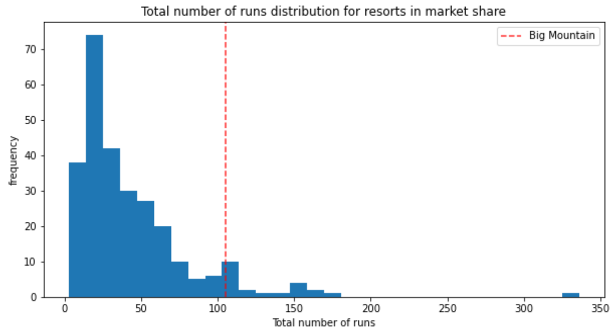
Then a Random Forest Regressor Model was found using a similar method. GridSearchCV was used in order to find the best number of estimators, whether using feature scaling or not was better and whether using median or mean as imputing strategies works best. It determined that the best number of estimators was 69, using the median is the best imputing strategy, and it is better to not scale the features. This model produced a mean of R^2 that was about 0.710 when used on the training split which was higher than the linear model.

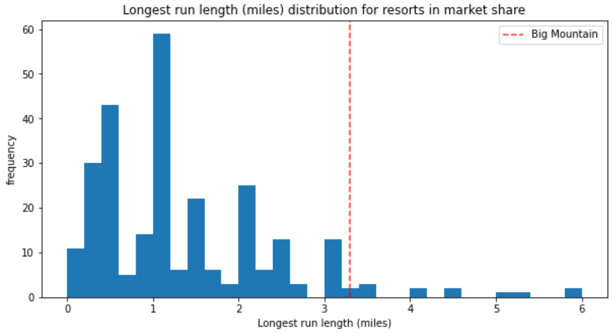
Using the linear regression model on the training split produces an MAE mean of about 10.50. The standard deviation was about 1.622. Doing the same thing with the random forest regressor model produces an MAE mean of about 9.644 with a standard deviation of about 1.352. This model has a better and a better deviation then the linear regression model. When both of these models were tested on the testing split, the linear regression model produced an MAE of about 11.79 while the random forest regressor model produced an MAE of about 9.538 which is also better. As such, the random forest regressor model was chosen.

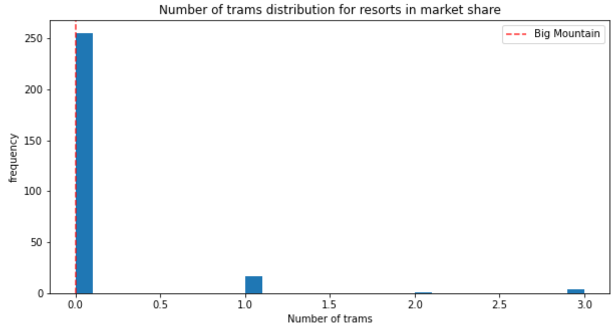
Using the model on all available data excluding the Big Mountain data, that is, on the combined training and testing splits, produces a mean MAE of 10.39 with a standard deviation of 1.47. Big Mountain currently charges $81 for an adult weekend ticket. Applying the model to the Big Mountain data indicates that the resort should charge $95.87. Even with an MAE of $10.39, this suggests that there is room for an increase in price. This leads to the possibility that Big Mountain may be undercharging. Looking at Big Mountain within the context of the market in terms of features shows it ranks really well in many of the features that were determined to be important by not only the random forest regressor model but the linear regression model as well. The graphs below reflect this.

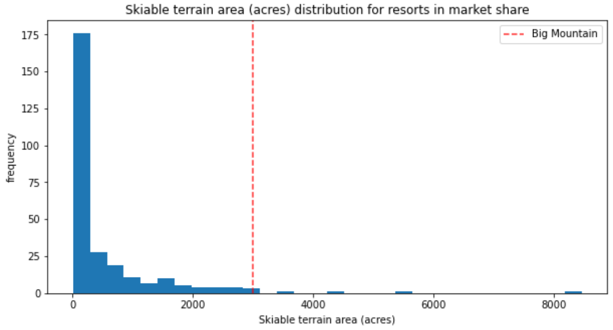

Vertical drop (feet) distribution for resorts in market share
Big Mountain
frequency
Vertical drop (feet)

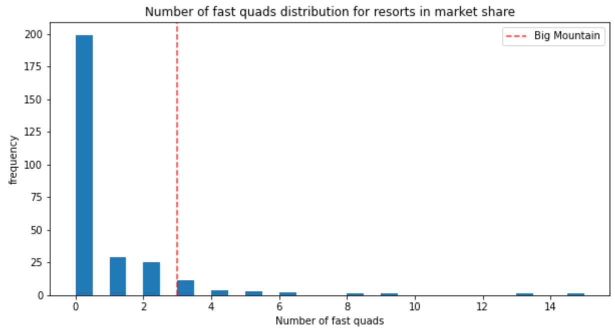

Area covered by snow makers (acres) distribution for resorts in market share
Big Mountain
frequency
Area covered by snow makers (acres)

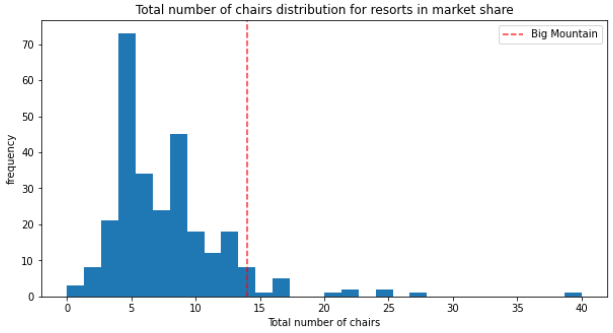

Total number of runs distribution for resorts in market share
Big Mountain
frequency
Total number of runs


Longest run length (miles) distribution for resorts in market share
Big Mountain
frequency
Longest run length (miles)


Number of trams distribution for resorts in market share
Big Mountain
frequency
Number of trams


Skiable terrain area (acres) distribution for resorts in market share
Big Mountain
frequency
Skiable terrain area (acres)


Number of fast quads distribution for resorts in market share
Big Mountain
frequency
Number of fast quads


Total number of chairs distribution for resorts in market share
Big Mountain
frequency
Total number of chairs