**Problem Statement:**

Big Mountain Resort is a ski resort located in Montana that has access to 105 trails and about 350,000 people ski or snowboard at the resort. However, the company has installed another chair lift increasing their operating costs to $1,540,000 this season. Therefore, the focused problem at hand would be to recommend a business strategy to Big Mountain resort for recouping the increased operational cost of $1.54MM for installing a new chair lift this season, while keeping the profit margins at 9.2% as well as provide insights on annual revenue for the season over the next year.

**Data Wrangling:**

To begin the analysis for the above problem, an initial “cleaning” of the dataset provided by the database manager Alesha Eisen was performed in order to properly assess the information provided.

The shape of the dataset has reduced from 330 rows with 27 columns to 277 rows with 25 columns. The change in the original shape was to remove any resorts or rows from the dataset that had null values within their Adult Weekend and Adult Weekday Price columns which were determined to be the main columns in order to compare and predict ticket pricing from different resorts.It seemed than from the two columns, the "AdultWeekend" ticket price column would be the target feature in order to predict ticket price as this had the least null values in its column. Another column dropped within the dataset was the fastEight column since over 50% were missing values. The Silverton Mountain Resort in Colorado's row data, after a wiki search, for the SkiableTerrain\_ac column was modified in order to accurately reflect its actual skiable area of 1819 instead of 26819 acres. Another row value was removed from the dataset that had a row value for the yearOpen column of 2019. This row value was assumed to be the year the park opened and not how many years the park has been in service. Also after "cleaning" the dataset, there are a few columns where the distribution is skewed where more values tend toward 0 namely fastQuads, fastSix, and Trams that will need to be explored further. Data from wikipedia regarding some state summary statistics such as state population, total state in square miles, etc. were also gathered to be further explored. After this initial “cleaning” and assessing for data quality, a more comprehensive exploratory analysis was performed.

**Exploratory data Analysis:**

From an EDA of the state summary statistics dataset generated from the previous step of "cleaning" the data, it was found that there seems to be no clear pattern or trend between the states categorical features of number of resorts and numeric feature of total population in the given state. For example, it would seem that New York would have more resorts given it being a densely populated state, however when taking the ratios of number of resorts given 100K population and the number of resorts given 100K square miles, New York is not in the top 5 for either of these ratios. Then in performing a PCA analysis of the state summary statistics, it was found that the first two components of the PCA analysis lead to about 75% of the variability of the data. Then further drilling down in terms of identifying the components, it was found these calculated ratios of number of resorts given a 100K population and number of resorts given 100k square miles were two of the largest positive driving factors wherein the states Vermont and New Hampshire stood. And specifically seems Vermont has the largest value for ratio of resorts given a 100k population. Therefore since there is variability in the state data, it would be in the best interest to equally consider all the states since there was no clear trend from the state summary statistics data and keep the state labels as the index, but it was found that the first two principal components, resorts\_per\_100ksq\_mile and resorts\_per\_100kcapita, were the larger positive driving factors leading to over 75% variance in the dataset.

From an EDA for the numeric data for the ski data set, a heatmap was initially constructed looking at the correlations for all the features within the dataset. From analyzing the correlations between the target feature, AdultWeekend ticket price, it seems there are positive correlations between the vertical drop distance of a resort, the number of trams at the resort, the number of fastQuads at the resort, the number of total chairs at the resort, the number of total runs at the resort, the LongestRun distance at the resort, the Skiable Terrain area at the resort, Snow Making\_ac or area the the snow making machines get to, the number of daysOpenLastYear for a resort, and also with the ratio of the resort's night skiing area within the state. A further assessment of the correlations through a scatter plot with the AdultWeekend ticket price column show that for Skiable terrain, when the value is low, there is variability in data, but it seems that largely the more area a resort has, the higher the price. Though there is a datapoint that has a skiable terrain of over 8000 square acres but lower price point. The resorts\_per\_100kcapita feature shows that a low value has high variability, but also largely as population increases, the price increases too. This same behavior is seen for the total chairs feature of the dataset. However, when taking a ratio of total chairs of a resort to total number of runs at a resort; it seems that more chairs per run leads to a lower price point. This might be due to an effect mentioned in this exercise of exclusive vs mass market where not having more chairs leads resorts to charge more for the tickets. Though a data point not seen and to be investigated is how many visitors each resort has per year. From an analysis of the correlations from the target feature of AdultWeekend, it seems that in some instances lower values tend to introduce high variability in the ticket price, which is something to be aware about when performing feature selection for modeling. Also in assessing the data quality, another feature to be wary about would be regarding a resort and its ratio of 100k capita as this might not be truly indicative of that specific resort generating revenue from only its particular “market” or state. States that are closer together for example in New England will likely be experiencing consumers from different states or market shares.

A thorough analysis of the dataset and assessing the strength of correlations to our target feature, AdultWeeked ticket price, has provided some insights when designing and developing an appropriate predictive model.

|  |  |
| --- | --- |

\*Graphs showcasing PCA component analysis and a correlation heatmap from all features of the dataset

**Establishing a Predictive Model:**

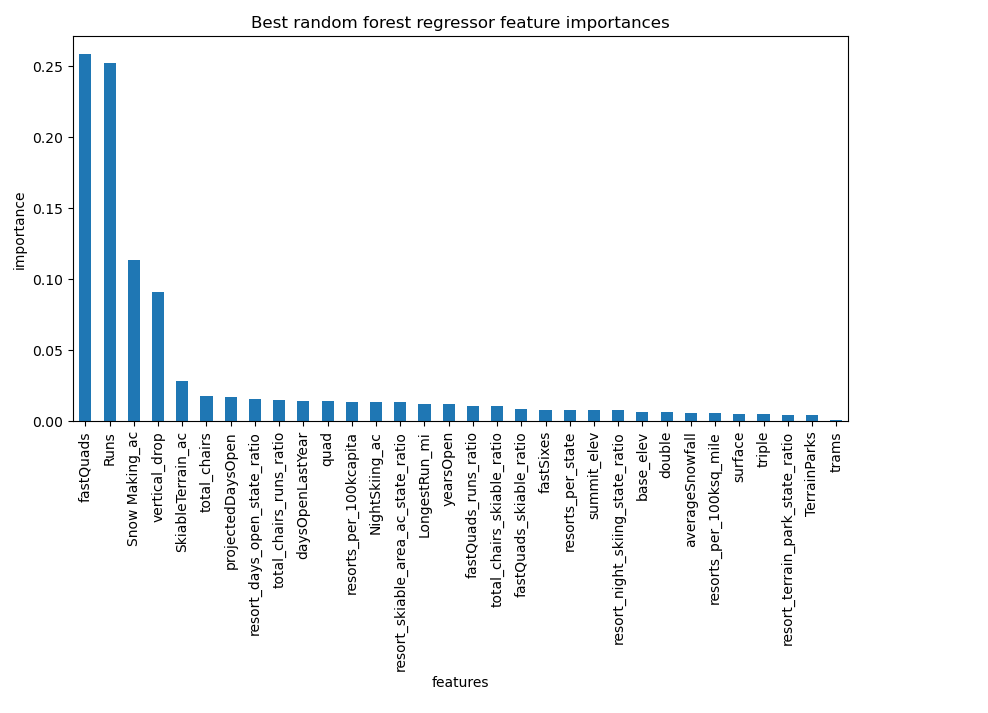
A modeling analysis was performed to compare the performance of the using the actual average, a linear regression model, and a random forest regressor model in terms of providing predictions of the target feature AdultWeekend ticket price column.

In just modeling the predictions of ticket price with the actual average of the AdultWeekend ticket price column, the predictions of ticket price would be off by about $19. Then when applying a train test split for a pipelined cross validated linear regression model, applying also the StandardScalar method to scale all the features to same magnitude and the SelectKbest and function GridSearchCV to select the best features of the data set in order to train the model, it was able to shorten this gap of predicting AdultWeekend ticket price to being off by 11.80 dollars. In applying a similar method of a train test split for a cross-validated pipelined and after scaling random forest regressor model, it was able to shorten this gap even further to being off by about 9.50 dollars. Therefore, the random forest regressor model from the models tested in this is the best in terms of predicting the price for the AdultWeekend column.

|  |  |
| --- | --- |

\*Algorithms respective linear regression model and Random Forest Regressor model

Some notes in terms of the steps included for preprocessing the above now trained models are for both the linear regression model and the random forest regressor model filling any ‘NA’ columns was done with the median price of the AdultWeekend column. A test was performed to determine if filling ‘NA’ columns with mean or median lead to better performance, but using one over the other seemed to not affect performance significantly. Key features assessed from the above models that lead to largely affecting the AdultWeekend ticket price column were for the linear regression model: vertical drop distance of a resort, Snow Making\_ac, total\_chairs, fastQuads, Runs, LongestRun\_mi; but it had a negative pull in trams, and SkiableTerrain\_ac. Key features in terms of affecting model performance of predicting price for the AdultWeekend price column for the random forest regressor model include: fastQuads, Runs, Snow Making\_ac, and vertical drop distance of a resort.



\*Key features identified from random forest regressor model

Finally, the key metric that adequately allowed for proper comparisons between model performance was in calculating the mean absolute error of the model's prediction from the training set versus the test set. It was seen that the mean absolute error of the random forest regressor model seemed to be within one standard deviation from the mean of the cross-validated mean absolute errors. Essentially, through the mean absolute error metric, it was seen that the random forest regressor model was able to predict the AdultWeekend price with better accuracy than the linear regression or just taking the actual mean of the AdultWeekend dataset. Cross-validating with the mean absolute error lead to consistent values with our estimate for the random forest regressor model, however the linear regression model seemed to generate slightly higher values for its prediction after cross-validating than the estimate from the dataset.

**Winning model and scenario modeling:**

Based on the previous step in pre-processing and training different models for the ski data set to predict prices in the AdultWeekend column, the random forest regressor model according the mean absolute error metric was determined as the best model given linear regression and just taking the actual mean from the AdultWeekend price column. The random forest regressor model was then applied taking all the data from other ski resorts in the nation in order to create a predicted price for the business in question, Big Mountain Ski Resort. From this analysis, it was seen that Big Mountain Ski Resort **seems to be underselling** their resort as the **predicted price is about $95** when the **actual price** in the AdultWeekend price column **is $81**. Though compared to just the market within Montana, it does have a higher ticket price.

Looking at how Big Mountain Ski Resort compares then with other resorts in the market share according to the key features deduced from the model:

Big Mountain has somewhat higher vertical drop distance though there other that have higher, Big Mountain has significantly larger area of the resort that is covered by snow making machines compared to largely other resorts in the market share, Big Mountain has higher number of runs than other resorts in the market share, Big Mountain has higher number of fastQuads where a large amount of resorts do not have any, Big Mountain has also larger number of runs than the amount seen in market share, Big Mountain also has one of the higher Longest Run length in miles than other resorts in the market share with some other being higher, but Big Mountain has no trams which is common to many of the resorts in the market share, and Big Mountain has one of the larger Skiable areas in acres for its resort compared to other resorts in the market share.

|  |  |
| --- | --- |

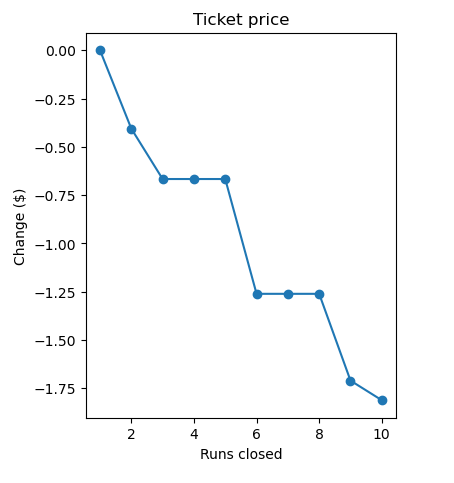
|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
|  |  |

\*AdultWeekend Ticket price along with Key feature analysis with Big Mountain Ski Resort Comparison

Given a few business strategies, from tuning the predictive random forest regressor model, a potential business strategy Big Mountain could focus their attention on when adding another chair lift to their resort, would be to also increase their vertical drop distance by adding 150 feet, as well as adding 2 more acres for their snow making coverage. This leads to a ticket support price increase from the model to 9.90 dollars making the resort still be markedly competitive of 90.90 dollars while still below the modeled predicted value of about 95 dollars and hopefully recoup the costs of adding another chair lift to the facility. This assessment was made with the assumption that Big Mountain Ski Resort would continue to experience 350000 visitors over an open season each paying for 5 days at the resort.

If Big Mountain Ski Resort wanted to focus their attention as well on potentially closing runs at the resort, it seems a safe margin to do so would be between 3-5 runs being closed leading to ticket price drop of about 67 cents. Any lower in terms of run closures, leads to significant price markdown where closing 10 runs means marking down the ticket price to about 1.81 dollars.



\*Plot describing markdown for ticket price support depending on number of Big Mountain Ski Resort run closures

**Pricing Recommendation:**

Therefore, the recommendation, based on the random forest regression model, would be that Big Mountain Ski Resort in adding another chair lift to their resort also increase their price to $90.90 per ticket. This is also under the guidance that they increase their vertical drop distance by 150 feet while also increasing their snow making machine coverage by 2 acres.

**Conclusion:**

The above analysis and the random forest regression model desgined in a general sense adds value by providing a tuning structure where when modifying the key features identified provides a general sense of how the specific tunings will have an affect on the overall ticket price for a ski resort. This provides an understanding of where resources can largely be shifted in order to increase revenue as a whole for Big Mountain ski resort. The proposed adjustment of business case from scenario will increase Big Mountain Ski Resorts price by about 12% while still keeping Big Mountain Ski Resort within a competitive price point according to the total market share to hopefully recoup the costs of adding another ski lift for the resort.

**Future Work:**

The price column used for this dataset and model prediction was the AdultWeekend column for price. The AdultWeekday price column was dropped in their analysis because it had more null values than the AdultWeekend column for price. However, if the business scenarios in terms of predicting behavior of ticket price for Big Mountain Ski Resort is under the assumption that visitors stay 5 days at the resort, then it be in the best interest to analyze the AdultWeekday column for price as well to see if this would change the behavior of ticket price for Big Mountain Ski Resort as well as compare with entire market share of ski resorts. Other cost or information that would have been helpful in terms of fine tuning a predictive model would have been assessing the number of visitors the entire market share of resorts experienced within an open season. Also knowing the operational costs for maintaining a snow making machine for each resort as well as how much it costs to increase a vertical drop distance as well as maintaining fastQuads or other chair lifts might have influenced the model predicted price value.

The underpricing seen from the model predicted value versus Big Mountain actual ticket price value might come as a surprise to executives, but again as the predictive random forest regression model again did not assess ticket pricing in the Adult Weekday price column. This pricing column which might provide additional insight as per a proper pricing margin for the resort especially if assuming business scenarios wherein customers pay for 5 days at the resort. However, the model still gives an understanding in a general sense where Big Mountain ski resort stands in terms of the key features that were used to predict the model pricing. And provides an understanding of what key features can be dialed or changed that significantly affects their ticket pricing. This would enable the business to be able to have a more defined focus on elements of the resort that they can be wary about over others. The model gives at least an understanding of the top key features that affects a ski resorts pricing and then other analysts can explore with tuning these parameters what combination of factors will significantly affect the ticket price without leading to generally overselling or underselling the resort.