Guided Capstone Project Report

Big Mountain Resort is a ski resort located in Montana that offers views of Glacier National Park and Flathead National Forest with access to 105 trails. The resort’s pricing strategy has been to charge a premium above the average price of resorts in its market segment; however, this does not provide the business with a good sense of how important some facilities are compared to others hampering investment strategy. The business wants some guidance on how to select a better value for their ticket price. The problem statement for this scenario is to recommend a strategy to Big Mountain Resort for recouping the increased operational costs of $1,540,000 for installing new chair lifts while keeping their profit margins at 9.2% and giving an insight on how they should set their ticket prices for next year.

With the ski resort data, we first reviewed the data and counted the number of missing values in each column. We found that the fastEight column had the most missing values with over 50% of the rows having missing values, as well as a desired target quantity, AdultWeekday column, with 15-16% of missing values. These 2 columns will be further investigated later, we then investigated some categorical features (Name, Region, state columns) and we found that we have unique records on each row. We also found that there are some non-state Regions included in the data.

A comparison of a number of states

Description automatically generated

The above graphs show New York accounting for most of the resorts. Since we do not have additional information such as the resort’s proximity to population or the # of visitors per year, it is hard to determine whether each state should be treated equally or differently. As mentioned in the scenario, we will treat each state as the same.

A graph of a number of tickets

Description automatically generated

Next, we look at the ticket prices for each state. As expected, weekend ticket prices are more expensive than weekday ticket prices. Montana, where our target resort is at, shows its ticket prices are on the lower end, being the 7th cheapest overall.

A graph of different colored and black lines

Description automatically generated with medium confidence

Aside from some relatively expensive ticket prices in California, Colorado, and Utah, most prices appear to lie in a broad band from around 25 to over 100 dollars. Some states show more variability than others. Montana and South Dakata both show small variability as well as matching weekend and weekday ticket prices. Nevada and Utah show the most range in prices. Some states, such as North Carolina and Virginia, have weekend prices far higher than weekday prices.

A screenshot of a graph

Description automatically generated

We then look at the distribution of features to get a feel for whether the values look sensible and whether there are any obvious outliers to investigate. The fastEight column had nearly half of its values missing, and the other half of the values were 0, this column was dropped since there was essentially no information in this column. The smallest number of years opened resort was dropped as well since the data may be inaccurate and it is not certain whether the number is a projection or actual. The rows where 2 price values are missing were dropped since these rows do not have any target information and, therefore, are of no use. Silverton Mountain has an incredibly large skiable terrain area, there was an error that occurred possibly through transmission or some editing or transcription stage. We modified the value of skiable terrain for this value to the observed value of 1819. As a result, we reran the features distribution, as shown above, and now the distributions lack much variance away from 0 and may have a small number of relatively extreme values.

A graph of blue dots

Description automatically generated

From the above plot, there is a clear line where weekend and weekday prices are equal. Weekend prices being higher than weekday prices seem restricted to sub $100 resorts. We also found that weekend ticket prices have the least missing values, so the weekday tickets prices were dropped and the rows with weekend prices were kept. In this Data Wrangling portion, we have dropped rows with no useful information and determine which features we want to focus on which is weekend prices.

Using outside population and area data for the US states from Wikipedia, we created another data frame of summary statistics for various states. We have come up with the following conclusions:

* Total State Area – Montana comes in at 3rd largest
* Total State Population – California dominates the state population figures despite coming in second behind Alaska in size. The resort’s state of Montana was in the top five for size but doesn’t figure in the most populous states. Thus Montana is less densely populated.
* Resorts per State – New York comes in top in the number of resorts in our market.
* Total Skiable Area – New York may have the most resorts but they do not account for the most skiing area. Montana is in the top five of the most skiable area. It is possible that New York has most smaller resorts whereas Montana has fewer larger resorts.
* Total Night Skiing Area – New York is the top of area of skiing available at night. In general, the top five of this feature are all the more northerly states. This raises a kit if questions and more data may be needed.
* Total Days Open – Total days open seem to bear some resemblance to the number of resorts, the more resorts open through the skiing season then the more days open we’ll see. New Hampshire is in the top five despite not being in the top five of resorts per state, it raises the question of whether location means resorts have longer seasons or stay open longer.

We calculated the resort density by dropping the absolute population and state size columns. With the removal of these 2 columns that only had state-specific data, we now have a Dataframe that shows skiing competitive landscapes of each state. It has the # of resorts per state, total skiable area, and days of skiing. We found that the top states by resort density turns out to be New Hampshire and Vermont at the top of the chart and New York does not appear in the top five.

A graph with a line

Description automatically generated

After performing PCA to find linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance they explain, we found that the first two components seem to account for over 75% of the variance and the first four components account for over 95%.

A graph of states with numbers and names

Description automatically generated with medium confidence

We took the average ticket price by state and added it to the scatter plot, there is not an obvious pattern. We concluded that the ski summary for each state, which accounts for 77% of the variance, does not have a pattern with price. In the first two components, there is a spread of states across the first component. Vermont and New Hampshire show extreme variability in the second dimension and New York and Colorado show extreme variability in the first dimension. Vermont and New Hampshire have large values of resorts\_per\_100ksq\_mile in absolute terms, they are more than 3 standard deviation from the mean. Vermont also has a large value for resorts\_per\_100kcapita. New York does not seem to be a stand-out for density of ski resorts in terms of state size or population count.

A screenshot of a computer screen

Description automatically generated

We merged the state summary features into the ski resort data and added “state resort competition features” to find the ratio between various features. The above heatmap shows that summit and base elevation are highly correlated and the ratio features are negatively correlated with the number of resorts in each state. If the number of resorts in a state increases, the share of all the other state features will drop for each. There is a positive correlation between the ratio of night skiing area with the number of resorts per capita. When resorts are more densely located with population, more night skiing is provided. In regards to AdultWeekend, there are a few correlations including fastQuads, Runs, and Snow Making\_ac. We concluded that visitors seem to value more guaranteed snow, which would cost in terms of snow making equipment, which would drive prices and costs up. resort\_night\_skiing\_state\_ratio is the most correlated with ticket price. If this is accurate, seizing a greater share of night skiing capacity is positive for the price a resort can charge. Runs and total\_chairs are also correlated with ticket price; it could be possible that the more runs there will require more chairs which may count for more than the total skiable terrain area. The total skiable terrain area is not as useful as the area with snow making. It seems there is more value in guaranteed snow cover rather than more variable terrain area. The vertical drop seems to be a feature that may raise ticket prices as well.

A group of blue dots

Description automatically generated

In the above scatterplot, there is a strong positive correlation with vertical\_drop. Also, fastQuads, Runs, and total\_chairs seem useful. Resorts\_per\_100kcapita shows that when the value is low, there is a variability in ticket price, but it is capacity of going high. Ticket prices may drop before increasing as the number of resorts per capita increases. Ticket prices may increase with the number of resorts serving a population since it indicates a popular area for skiing with plenty of demand. The lower ticket price when fewer resorts serve a population may be due to the less popular state for skiing. The high-ticket price for some resorts when resorts are rare may indicate areas where a small number of resorts can benefit from a monopoly effect.

A group of blue dots

Description automatically generated

The more chairs a resort has, relative to the # of runs, ticket price decreases and stays low. We may be seeing a exclusive vs. ass market resort effect. If a resort has fewer chairs, they can charge more for tickets, but fewer chairs will serve fewer visitors. Price per visitor is high but # of visitors may be low. Having no fast quads may limit the ticket price but if the resort covers a wide area, then a few fast quads may be beneficial to ticket price.

First partitioning the data into training and testing splits then measuring the coefficient of determination (R2), mean absolute error (MAE), and mean square error (MSE). We found that the mean absolute error is the most intuitive of all the metrics, it showed that, on average, the expected ticket price is around $19 based on an average of known values. After creating a linear regression model, it explains over 80% of the variance on the train set and over 70% on the test set. The lower value for the test set suggests we may be overfitting. Using this model, we expected to estimate a ticket price around $9 of the real price. This is more accurate than $19 from using the average. We will be using the median for filling missing values due to the skew of many of the predictor feature distributors, the results do not differ from using the median for imputing missing values. Possible reasons include overtraining dominates, other feature transformations may help, or can try with a subset of features rather than using all of them as inputs. To perform the median/mean comparison, we will create a pipeline. The results suspected that the modal was overfitting, and a subset of features would generalize better; however, selecting a subset of features has an impact on performance and SelectKBest defaults to k=10. To get the best k, we will create a new pipeline with a different value of k. Rather than turning the model to the arbitrary test set that fails to generalize new data, we will use cross-validation to build k models on k sets of data with k estimates of how the model performs on unseen data but without having to touch the test set. By using cross-validation for multiple values of k and use cross-validation to pick the value of k that gives the best performance, we generated the below figure:

A graph with blue lines

Description automatically generated

This suggests a good value for k is 8, there was an initial rapid increase with k then a slow decline. The variance of the results greatly increases above k=8. As overfit increases, there will be greater wings in performance as different points move in and out of the train/test folds.

Using the linear regression model and grabbing the model coefficients via its coef\_ attribute, the results suggest that vertical drop is the biggest positive feature. The area covered by snow making equipment is a strong positive as well. The skiable terrain area is negatively associated with ticket price, this may be because larger resorts can have more visitors and can charge less per ticket.

A graph with blue and white text

Description automatically generated

Using the random forest regressor, we found that the top four features that are in common with the linear model is fastQuads, Runs, Snow Making\_ac, and vertical\_drop. The random forest model has a lower cross-validation mean absolute error by almost $1 and less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.

A graph with a line

Description automatically generated

The above graph shows there is an initial rapid improvement in model scores, but it is levelled off by around a sample size of 40-50.

We will train the model to predict Big Mountain’s ticket price based on data from the other resorts and calculate price based only on its competitors. The results show that the Big Mountain Resort modelled price is $95.87 while their actual price is $81.00. Even with the expected mean absolute error of $10.39, this suggests there is room for an increase. We assume that other resorts set their prices based on the market. Since Big Mountain Resort is charging less, this suggests that they may be undercharging. Some important features that came up on the modeling include vertical\_drop, Snow Making\_ac, total\_chairs, fastQuads, Runs, LongestRun\_mi, trams, and SkiableTerrain\_ac.

A graph of a line graph

Description automatically generated with medium confidence

A graph with blue rectangles

Description automatically generated

A graph of a drop

Description automatically generated with medium confidence

Big Mountain Resort is doing well for vertical drops but there are other resorts with greater drops.

A graph of a graph

Description automatically generated with medium confidence

Big Mountain Resort is high up in the snow making area.

A graph of chairs distribution

Description automatically generated

Big Mountain Resort has the highest number of total chairs, other resorts with more seem to be outliers.

A graph with numbers and a number of quads

Description automatically generated

Most resorts do not have fast quads while Big Mountain Resort has 3 which puts it higher on the scale.

A graph of a number of runs

Description automatically generated

Big Mountain Resort compares well for the number of runs, only a few resorts have more.

A graph of a running graph

Description automatically generated

Big Mountain Resort has one of the longest runs even though it is over half the length of the longest.

A graph with numbers and lines

Description automatically generated

Majority of the resorts including Big Mountain Resort does not have trams.

A graph of a bar graph

Description automatically generated with medium confidence

Big Mountain Resort is one of the resorts with the largest amount of skiable terrain.

After reviewing the potential scenarios for either cutting costs or increasing revenue, the business has shortlisted a few options which we will test each option. The expected # of visitors over the season are 350,000 and visitors, on average, ski for five days. We also assume the provided data includes the additional list that Big Mountain Resort recently installed.

A comparison of a graph

Description automatically generated with medium confidence

Scenario 1: The model shows that closing 1 run does not have any difference. Closing 2-3 runs reduces support for ticket prices and revenue. Closing 3-5 runs does not further reduce the ticket price. Closing 6 or more runs signification reduces ticket prices.

Scenario 2: Big Mountain Resort is adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift. The model shows that this increases the support for ticket price by $1.99. Over the season, this could be expected to amount to $3,474,638 in profit.

Scenario 3: Big Mountain Resort is adding a run, increasing the vertical drop by 150 feet, installing an additional chair lift, and adding 2 acres of snow making. The model shows that this increases the support for ticket price by $1.99. Over the season, this could be expected to amount to $3,474,638 in profit. This is the same result as Scenario 2. Therefore, a small increase in the snow making area does not make a difference.

Scenario 4: Big Mountain Resort is increasing the longest run by .2 miles and guaranteeing its snow coverage by adding 4 acres of snow making capability. The model shows there is no difference.

Conclusion: We suggest Big Mountain Resort implement Scenario 2, this strategy supports the increase in ticket price by $1.99 and generates revenue.