**Guided Capstone Project Report**

The first step of analyzing the given data set is to determine how to handle missing data in the data set or data wrangling. Big Mountain Resort is in fact included in the data set which is good because we can compare summary statistics and other values to it. There are other methods to data wrangling, but several relevant variables have missing data, including adult weekday and weekend ticket prices. We will determine whether to combine the two variables as a target or treat only one later in the analysis but for now, we must decide how to treat missing values. Additionally, total days open, number of fast eight cars, Snow making area and Night skiing area also have a large number of missing values, but we do not know if these variables are highly correlated or related to the target. Approximately 14% of resorts have missing price data. We will decide to drop these records after a thorough data analysis. Another consideration is the variability of weekday and weekend pricing among states. Montana does not have any variability, but Nevada and Utah do.

Further investigation yields that Silverton Mountain has an incredibly large skiing area. This is erroneous and was corrected so that does not lead to an inaccurate model. Also, one resort inaccurately reported that they have been open for 2019 years when this is clearly the year 2019. This was also corrected.

Next, statewide summary statistics were created for state total terrain parks, number of days open last year, Snow making area and Night skiing area and rows that had no pricing data at all were dropped. Population and state area columns were now added to the data set. We know that we want the target to be pricing and in Montana, weekend and weekday pricing is equal. We also determined that there are more missing values for weekday pricing. This leads to a conclusion to drop resorts with only weekday prices and keep resorts that only have weekend prices. The number of rows were reduced from 330 to 227 and the number of columns has been reduced from 27 to 25. Due to the graph below, differences in weekday and weekend pricing appear with sub $100 resorts.

Chart, scatter chart

Description automatically generated

The next task for the data set is to perform exploratory data analysis (EDA). Most of the columns at this point are numeric and Montana possesses a large area for skiing as well as being a large state all together but is not one of the most populous states. Montana also did not have very high numbers in total days open and total night skiing area. Montana is reasonably high in resorts per capita.

There are many variables to make sense out of and principal component analysis (PCA) can be used to reduce dimensionality and order certain features by their variance after the data is scaled. The PCA chart below shows that the first two components explain about 75% of the data and the first four explain over 95% of the data.

A picture containing line chart

Description automatically generated

The result of the PCA shows that there is no real pattern or relationship between state and ticket price. The graph below shows that there are red points at the top, lower right and lower left indicating no real pattern. This could be good news as Montana is right smack dab in the middle of the chart and may have the ability for increased pricing, at least at Big Mountain Resort.

A picture containing chart

Description automatically generated

Next up is feature engineering which will help determine where a particular resort lies with regard to state level values for variables that we may be interested in. Examples of this engineering are as follows:

* Ratio of resort skiable area to total state skiable area
* Ratio of resort days open to total state days open
* Ratio of resort terrain park count to total state terrain park count.
* Ratio of resort night skiing area to the total state night skiing area

The result of creating scatterplots with all of these areas with weekend ticket price yield very similar results and are positively correlated as shown below. Resort night skiing state ratio may be the most correlated to ticket price.

A picture containing chart

Description automatically generated

Additionally, runs and total chairs appear similar and useful. There is a strong positive correlation with vertical drop and fast quads. Resorts per capita shows that there is a good deal of variability when value is low, but the range of value is high. These correlations are shown in the graphs below.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Now it is time to separate Big Mountain Resort data from the rest of the data and create training and test sets before creating what hopes to be an informative model. The training and test split will be 70/30 respectively, which is typical. This would place about 193 rows in the training set and about 83 in the test set. We will create a function to compute R-squared which represents the proportion of variance in ticket price predicted by the model, mean squared error and mean absolute error. The median values for all variables are now calculated for imputation, or in other words, to fill in missing values and is applied to both the training and test sets. The data is then scaled which is good to do with values that vary by orders of magnitude or variables that are measured in different units. A linear regression model is then trained on the training set and predictions are made on the training and test sets. An initial r-squared value of 0.818 and 0.720 are found on the training and test sets, respectively which both makes sense and is good. We want to see if we can to better, however as there may be overfitting of the data due to not selecting a parsimonious set of features or deal with multicollinearity. The mean absolute error indicates a variance of about $9 in ticket price. After imputing missing values, the R-squared, mean squared error and mean absolute error did not change much at all.

Next, a function or pipeline will be created to impute missing values, scale features, train a certain model and calculate performance so that we do not have to go over the same procedure for each iteration of the model when parameters are changed. Select K best, which selects the k best features of the data. This method actually reduced the accuracy of the model of about 10% so selecting a subset of features may not be helpful at k=10 or at k=15. A way around this is to utilize cross-validation. The Grid Search CV method can be implemented to find the best parameters to use for optimum performance of a given model. The pipe that we utilized earlier is evaluated with this method and the graph below shows that the most appropriate value for k is 8 for k best features of the given data set.

A picture containing text, antenna

Description automatically generated

A picture containing text

Description automatically generated

The eight chosen features are listed above and most of these features are of no surprise after performing EDA. Longest run is only weakly correlated with ticket price, but it may be worth taking a look at. Number of trams and skiable terrain area are strongly negatively correlated with ticket price which is very interesting seeing how it appeared to be positively correlated non-linearly during EDA.

Now it is time to see if we can get a better model using a random forest algorithm and adjust some hyperparameters. The graph below shows that four of the five top features in the linear model are included after using a random forest model excluding total chairs.

Chart, histogram

Description automatically generated

Cross-validation shows that the mean absolute error and standard deviation for the random forest model is lower so that is the model we will choose to run with. It turns out that utilizing the median helped improve the model but scaling the data did not. This model is approximately $10 less than just utilizing the average ticket price!

Now that we have the model that we want to use, we will remove Big Mountain Resort from the data and use the model on that resort data to determine what the ticket price should be. The result is that if the adult weekend price for a ticket at Big Mountain Resort is $81.00, the resort could raise ticket pricing to $95.87 and increase profit. This is favorable assuming that other data is not relevant and other resorts are not really good at pricing strategy. We then create a function that created a histogram of features across the board for relevant features and assesses where Big Mountain is with respect to the distribution. The following graphs support the following results:

* Big Mountain is already the highest prices resort in Montana and is near the 75th percentile for all resorts in the United States.
* Big Mountain is doing well for vertical drop, but there are still quite a few resorts with a greater drop.
* Big Mountain is very high up the league table of snow making area.
* Big Mountain is among the highest number of total chairs, resorts with more could be outliers.
* Most resorts have no fast quads. Big Mountain has 3, which puts it high up the league table.
* Big Mountain has a large number of runs although there are some with more but not many.
* Big Mountain has one of the largest runs, but it is just over half of the longest run. The longer runs are rare.
* Most resorts have no trams similar to Big Mountain.

Chart, histogram

Description automatically generated

Chart, bar chart

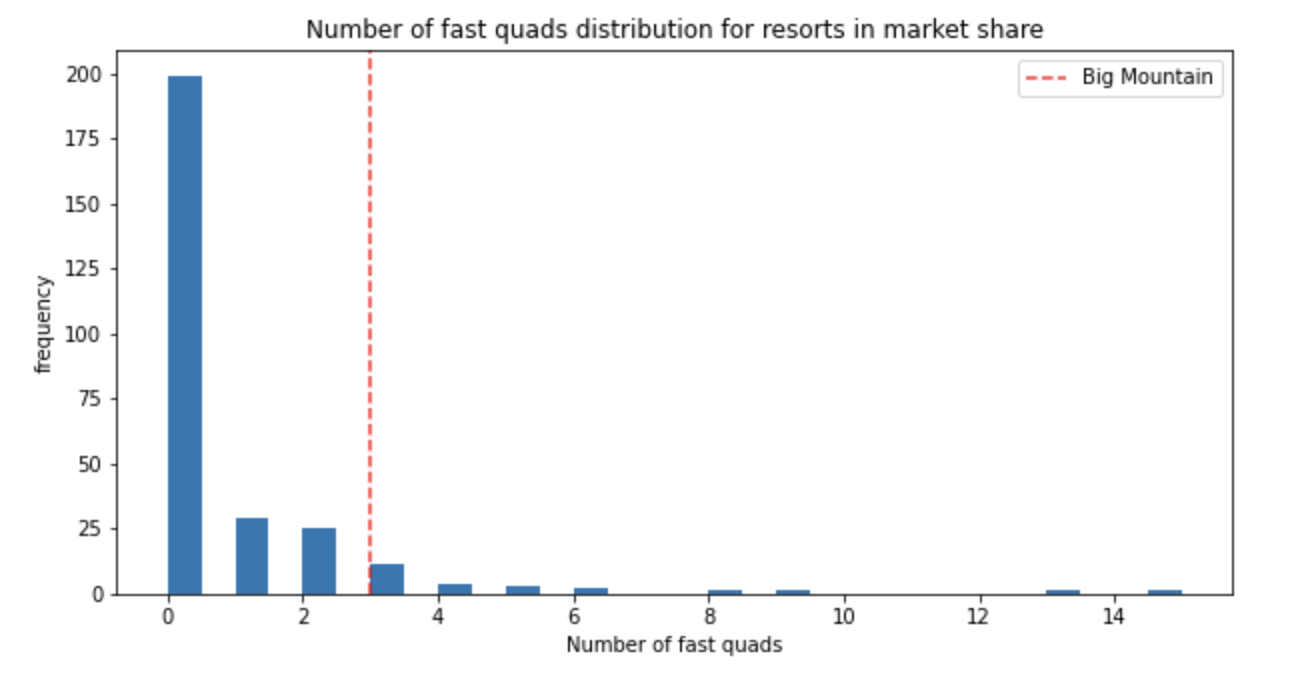
Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated



Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

A picture containing chart

Description automatically generated

Big Mountain Resort understands that the resort’s ticket prices are based upon whether certain facilities are being offered at the resort. The expected number of visitors over the season is 350,000 and visitors usually ski for five days on average. We know that the additional lift has already been financed and was recently installed. A function was developed called predict increase that will predict the amount of revenue increase based upon the given scenario.

Scenario 1 – Close up to 10 of the least used runs:

Chart, line chart

Description automatically generated

The graphs above show the results of the function for this scenario. Closing one run is inconsequential but closing 2 runs leads to a price and revenue drop. Closing 3 runs is the same as closing 4 or 5 but closing six runs leads to a big drop in ticket price and revenue.

Scenario 2 – Big Mountain Resort adds a run, increasing the vertical drop by 150 feet and installing an additional chair lift. This scenario leads to an increase in ticket price by $8.61 and leads to $15,065,471 in additional revenue! Now we are getting somewhere.

Scenario 3 – This scenario is the same as scenario 2 only adding 2 acres of snow making. This scenario leads to an increase in ticket price by $9.90 and leads to $17,322,717 in additional revenue. Is the added revenue worth further raising prices for 2 million in revenue? This is a business decision, and it may not be wise if visitors are not happy with a $10 increase in ticket price for a few acres of snow.

Scenario 4 – This scenario calls for increasing the longest run by 1/5 of a mile and adding 4 acres of snow making capability. This scenario showed no difference as the longest run does not influence this particular model significantly.

Conclusion

The chosen business models indicate that actually adding a run, increasing the vertical drop by 150 feet and installing an additional chair lift predicts a ticket price increase from $81.00 to $89.61 that would yield an increase in revenue by $15,065,471 in additional revenue. This scenario includes most of the features the random forest model predicted would have the most effect on ticket price. Adding more snow making area does raise revenue but also predicts a ticket increase of nearly $10.00 which may be a little high. Still, the business could opt to do this if it felt that marketing campaigns would lead to consumer interest. The 2 acres of snow does not seem to be that much however unless there was a good reason to do so like increasing the longest run, but this feature does not seem to be largely correlated to ticket price according to the random forest model. Increased operating costs of adding the additional chair is only $1,540,000 which is far less than the projected increase in revenue from the model. If Big Mountain Resort does add another fast quad, it may increase revenue even further by being able to further raise ticket price and would offset any operating costs of doing so. Or the resort could elect to raise ticket prices marginally since almost 10 times the operating cost is predicted to be obtained by the predicted price increase. This would be a good point to talk to business executives about other operating costs and why the resort has elected to underprice the resort up to this point, especially given the fact that the resort is already high on the league charts of facilities offered. This model should ultimately be provided to software engineers or business analysts to develop a business tool that would calculate predicted revenue increases based upon features that are most relevant to the model listed earlier.