Problem statement:

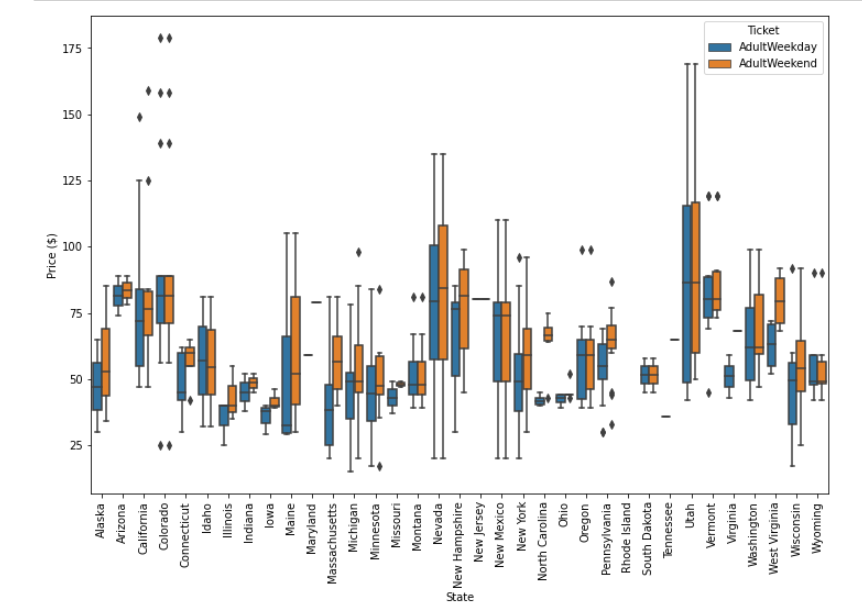
How can Big Mountain Resort (BMR) use their facilities to set their ticket prices and identify potential improvements in order to increase their profit during the next ski season?

Context:

BMR has recently installed an additional chair lift making this the twelfth and adding to their operating costs by $1540000. Currently their ticket sales price is solely based on adding a premium to the average price of the resorts in the same market segment. There is suspicion that BMR is not capitalizing on its facilities as much as it could. This hampers investment strategy. We need to get a good sense of the importance of the facilities compared to other resorts and how this can impact the ticket price as well as look into ways to cut costs.

Wrangling:

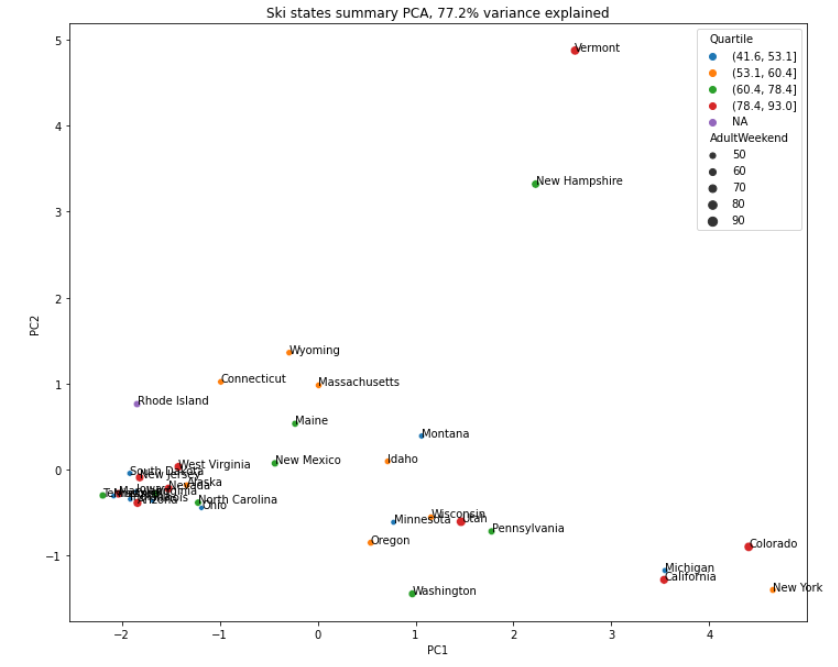
First thing in our project is to understand the data we have and consider adding more information as we proceed with this part of the project. Selecting the appropriate shape of a data is important as we proceed. First we looked for missing information especially if they lacked information about the ticket price or as in the case if fastEight, it had more than half missing values so we eliminated it altogether. In our attempt to identify possible duplicates we looked at the distribution of resorts per state as there was a resort with the same name but in different states so understanding this was important. In addition, we looked for unusual values for each variable by graphing histograms for each and also by calculating the descriptive statistics (mean, std. min, and the quartiles) and found the correct information in the resort’s website. Originally, we had 27 variables and 330 resorts, after this analysis we had 26 variables and 281 resorts. This is enough data to proceed with our analysis. This is a graph of the distribution of process for the resorts we will use for the rest of the project. As you can see there is a lot of variability amongst the resorts even though they correspond to the same market share and many have different ticket prices if it’s the weekend. However, in the state of Montana the ticket price doesn’t make this distinction.



EDA:

To begin to understand the relationship between facilities and ticket prices and knowing there is a significant variability between resort prices we created a couple of variables “resorts per 100k capita’’ and “resorts per 100k sq miles.” We also created a state summary data frame to further consider if the state had a strong relationship with the ticket price. As we considered the top states in terms of skiing area, night skiing, days open and state area, we didn’t find any clear patterns regarding ticket price.

We then used principal component analysis to help us identify clusters of resorts and we added the variable of price to the final graph.



On this graph we can clearly see that ticket price is not something that sets resorts apart from each other because we don’t see the resorts clustered by ticket price.

At this point we are safe to say that the state a resort is in, doesn’t seem to be a differentiator to set the ticket prices and recommend that we consider them altogether. There must be other features in the resorts that customers value more important than the state.

To include features related to the state resort competition we added the following fields and we dropped the state columns. At this point we now hade a dataframe with 29 columns:

* 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 counts to total state terrain park count
* ratio of resort night skiing area to total state night skiing area

To look for a correlation between the variables and the ticket price we considered a correlation heatmap and the scatterplots of each variable to the ticket price. We found that there is a strong positive correlation (>0.65) between ticket price and vertical drop, fastQuads, runs, total chairs, and snow making. Meaning that as the number of these increase, the ticket price also increases. We also found that the resorts per 100k capita shows that when the value us low, there is a lot of variability in ticket process, then there is a drop in ticket price before it increases again. This could possibly be explained by the popularity of the resort.

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

Model Preprocessing and training the model

We defined that a model would be trained with 80% of the data and tested with the remaining 20%. Our data frame has 277 resorts with 36 variables; leaving Big Mountain resort out of this process allows us to use the model to predict the price we are looking for. This means that there I enough data to train the model (193) and test with the remaining 83.

We explored two models, a linear regression using SelectKBest to reduce the number of features and the Random forest tree. For each model we used cross-validation to help us decide with one is better. We’d choose the model whose scores have the least std deviation.

The features recommended to use in the linear regression model are shown below in a descending order list from the greatest contribution (these are the coefficients of our linear model).

Per the cross validation results the random forest tree is a better because the standard deviation of the scores is less and there are also less features.

|  |  |  |  |
| --- | --- | --- | --- |
| **Linear regression features** | | **Random forest tree top features** | |
| vertical\_drop  Snow Making\_ ac  total\_chairs  fastQuads  Runs  LongestRun\_mi  trams  SkiableTerrain\_ac | 10.767857  6.290074  5.794156  5.745626  5.370555  0.181814  -4.142024  -5.249780 | fastQuads  Runs  Snow Macking\_ac  Vertical\_drop | |
| **Cross validation n=5** | | | |
| Mean | 0.6976 | Mean | 0.71 |
| Standard deviation | .071 | Standard deviation | .065 |

Other metrics to consider:

Cross validation Mean Absolute Error (mae). To use this metric we’d select the model with the smallest value as well as the smallest mae standard deviation.

|  |  |  |  |
| --- | --- | --- | --- |
| **Linear regression mae** | | **Random forest tree mae** | |
| **mae** | $11.8 | **mae** | $9.54 |
| **Mae mean** | $10.5 | **Mae mean** | $9.64 |
| **Mae std** | $1.62 | **Mae std** | $1.35 |

Both the mae and the standard deviation of the mae are smaller for the random forest tree. Therefore we recommend to use this model to predict BMR’s ticket price.

Modeling:

Now that we have decided to use the random forest tree, we will use all 276 resorts training the model and we will use this model to predict BMR’s ticket price. Doing this gives us the following metrics:

|  |  |
| --- | --- |
| **Random forest tree mae** | |
| **Mae mean** | $10.39 |
| **Mae std** | $1.47 |

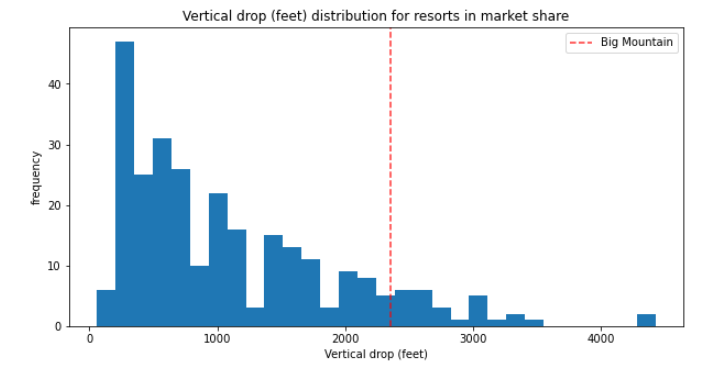
Now using this model with BMR’s data we get a price of $95.87 from the current $81. Because the model’s mean mae is $10.39 and std is $1.47, there is some room to increase the price.

Conclusion and recommendations:

Assuming that the expected number of visitors over the season is 350,000 and, on average, visitors ski for five days. And that the provided data includes the additional lift that Big Mountain recently installed. We are recommending that BMR’s ticket price should increase to $96 or $97.5 The increase in price is based on the current facilities with the addition of the twelfth chairlift. Some of the features that seem important to skiiers and BMR is on the top resorts are: total number of chairs, fastQuads, Runs, Longest run, skiable terrain, snow making area and vertical drop.

Of the proposals made to make improvements and decrease costs, closing one run, the least used will decrease some of the costs without compromising ticket price. For this we need to investigate the costs of the run to understand the impact in revenue.

Increasing the vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up, without additional snow making coverage. This would be beneficial because although we are above the average, we could see a benefit on the ticket price and increase by $1.99 with and expected revenue of aprox. $347,000. This is 40 cents more per ticket. According to the distribution of vertical drop notice the BMR is on the top resorts so further cost-benefit analysis would be advised.



Going forward additional data on the resorts operating costs, numbers of skiers that buy day tickets and multiple day passes could potentially give more information to make improvements to the model.