Big Mountain Guided Capstone Documentation

Problem Identification and Context

Big Mountain Resort came to us with suspicions that they were not capitalizing on their facilities as much as they could. Their facilities included 105 runs, 14 ski lifts, a 3.3 mile long run, a base elevation of 4464 feet, a summit height of 6817 feet, and vertical drop of 2353 feet. They had recently installed a new chair lift at an operational cost of $1,540,000 this season. Big Mountain wanted guidance on how to better price their ticket and how to make some operational changes that would reduce costs without reducing ticket price or possibly support increasing ticket price.

Problem Statement

How can Big Mountain Resort better price their tickets based on the value provided by their facilities during the next ski season?

Data Wrangling

Importing the data, we checked for null values in the data at large and in our Big Mountain Resort information. We also checked that each row of data was unique. Inspection of some features was done using histograms. We found that 58 rows were missing some information on ticket pricing. We dropped 47 of those rows for missing all target price information. AdultWeekend ticket pricing was decided on as being the prediction target as it had fewer missing values left than AdultWeekday ticket pricing. All AdultWeekday pricing information was dropped. A column, FastEights, was dropped because only 1 value was non-zero and half the data was missing. Some rows were corrected for errors in skiable terrain or dropped for having an incorrect number of years open. We also created a second table of state data containing state name, total number of resorts, total skiable area, total skiable days, total number of terrain parks, total area of nightskiing, population and size in square miles.

Exploratory Data Analysis

We started exploratory data analysis by checking out the 5 top states for each summary statistic in our state data. We also created summary stats for resorts per 100K people and resorts per 100K square miles. With these 7 summary statistics for the states, we scaled them to have a mean of 0 and a standard deviation of 1. We ran a Principle Component Analysis on the data to find combinations of features that were uncorrelated with each other using the SKLearn library. We found 2 component combinations accounted for 77.2% of the variance in state data(see Figure 1)

When these two components were plotted with state-average ski ticket pricing information on a scatter plot, there was no clear pattern visible among the states(see Figure 2.) We concluded that we could run the rest of the analysis and modeling on data from all the states.

We then followed by merging the state summary statistics dataset to our ski resort dataset. We created ratios of each resort's area of skiable terrain, number of days open, number of terrain parks, and area of night skiing available to the total amount in each state. From this, we checked for correlation among the ski resort features using a heat map(see Figure 3. We saw some features that are positively correlated with ticket price: vertical drop, number of runs, number of fast quad lifts, and snowmaking acreage were key features that mattered with resorts. When looking at resort-to-state ratios, the amount of night-skiing the resort had compared to the state was the most positively correlated to ticket price. The ski resort dataset, having been cleaned and inspected, was saved for model preprocessing.

Model Preprocessing

We started preprocessing by separating Big Mountain Resort data from the rest of the dataset so it could be used in scenario modeling later. We then split the remaining data into 70% training data and 30% testing data.

We started modeling by creating a dummy regressor that predicts the mean. We found that using the mean of the ticket prices, we would be on average $19 off of the true value of the actual ticket price, according to the mean absolute error.

From there, we started making linear regression models, imputing missing values in our training dataset using the median, and scaling each feature to have a mean of 0 and standard deviation of 1. This linear model would be on average within $9 of the true value of the ticket price according to the mean absolute error. Trying a similar model with missing values imputed by the mean, we didn't find much difference.

We started wrapping the imputing, scaling, training, and assessing into a pipeline to speed up the process. As we weren't limiting anything in training, we were overfitting the data. We added into the pipeline process an option to select k number of features to train on. We found that using only 10 features performed worse than using all features. As we were currently assessing model performance against the test data, we were still sort of favoring models that best fit the test data as opposed to any data. To fix this, we also incorporated cross-validation into the pipeline.

Using a Grid Search Cross Validation program, we found that using the 8 best features to train the data worked best, having the highest average cross validation score with the least variation(see Figure 4.) Out of interest, we inspected the 8 best features and their coefficients(see Figure 5.) These features were ones we expected from our earlier EDA.

After inspecting Linear Regression Models, we also tried a Random Forest approach. Running a grid search cross validation of random forest regressor model with parameters(scaling vs. no scaling, impute with mean vs. impute with median, number of trees in a logarithmic range from 10 to 1000), we found that the best parameters were to have no scaling, 69 trees in the random forest, and impute with the mean. Asking the random forest model to provide the most important features of the best model, we found 4 similar features to our EDA and to our linear regression model: number of fast quads, number of runs, total snowmaking area and total vertical drop(see Figure 6.)

The Mean Absolute Error for our Linear Regression Model was $11.79. The mean absolute error for our Random Forest Model was $9.53. Based on the performance difference, we opted to save the random forest model to a pickle file to make use of during scenario modeling.

Scenario Modeling & Price Recommendations

Retraining the random forest model one last time on all data excluding Big Mountain Resort, we ran the model to predict Big Mountain Resort's ticket price considering Big Mountain's features. Big Mountain's modeled price was $95.87 while its actual price was $81. This price difference was outside the expected mean absolute error of $10.39. This suggested there was room for a price increase.

Scenarios:

Big Mountain, while interested in pricing based on its facilities, also wanted guidance on how to make operational changes that would reduce costs or support a higher ticket price.

Scenario 1: Closing down the 10 least used runs.

Our model suggested no price change was necessary for closing 1 run, but did suggest dropping the price further as more runs were closed. For closing 10 runs, the model suggested dropping the ticket price by at most $1.75.

Scenario 2: Adding a run, increasing vertical drop, needing to add a chair lift.

Our model suggested a price increase of $1.99 for the scenario. An estimated increase of revenue from this is $3.4 million, covering the new operational cost of the chair lift.

Scenario 3: Adding a run, increasing vertical drop, adding 2 acres of snow-making capacity, needing to add a chair lift.

Our model returned the same result as scenario 2, suggesting that adding snow-making capacity did not return any benefit in ticket price.

Scenario 4: Increase the longest run by .2 miles, adding 4 acres of snow-making capability.

Our model suggested zero change in ticket price for the scenario.

Pricing Recommendation, Conclusion and Future Scope

We would recommend scenario 2 as a starting approach for Big Mountain. Customers would be more receptive to a ticket price increase if they could relate it to the construction of a new ski run and an additional lift. The estimated ticket revenue increase for scenario 2 would cover the operating costs of two ski lifts. After the ticket price increase to the model suggested value, then Big Mountain could engage in scenario 1; closing some of the least used runs to further save on costs. They could drop the price of their ticket as they close the runs if they want. Big Mountain could raise their ticket price to $97.86 under scenario 2, drop the price back to $96.11 under scenario 1, and still have a higher ticket price than now with no change in features.

For future modeling considerations, operational costs for snow making equipment would be useful. Snow-making equipment could be very valuable on the most used trails making it an essential cost, or snow-making equipment could be wasteful on little used trails making it costly.

Operational costs of maintenance for ski runs could be useful. I could imagine that runs need to be inspected and occasionally snow plowed to better maintain the trail. This cost could be independent of the amount of trails/total length of trails or highly dependent.

Big Mountain might have based its current price on the prices supported in Montana and other nearby states. Its price was the highest in Montana and the model didn't really take into consideration the ski resort’s proximity to other ski resorts. It could be that 3 to 5 ski resorts are in the same town as Big Mountain, pressing prices down to compete for skiers.

Figure 1

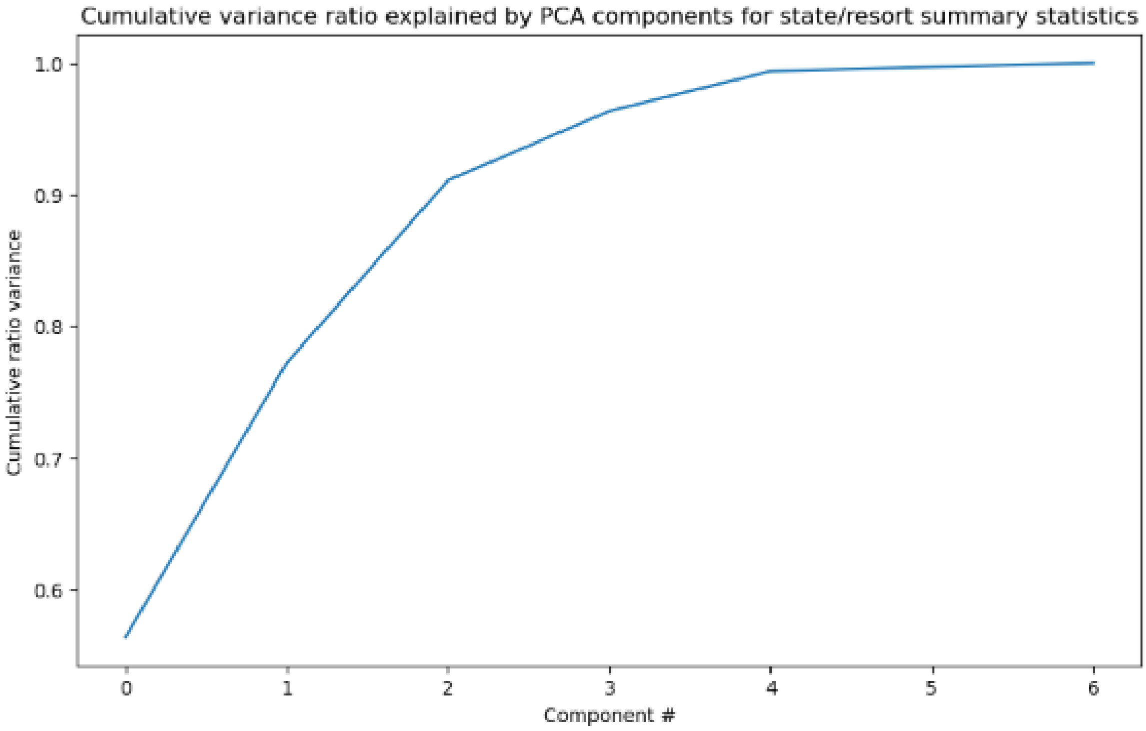


Figure 2

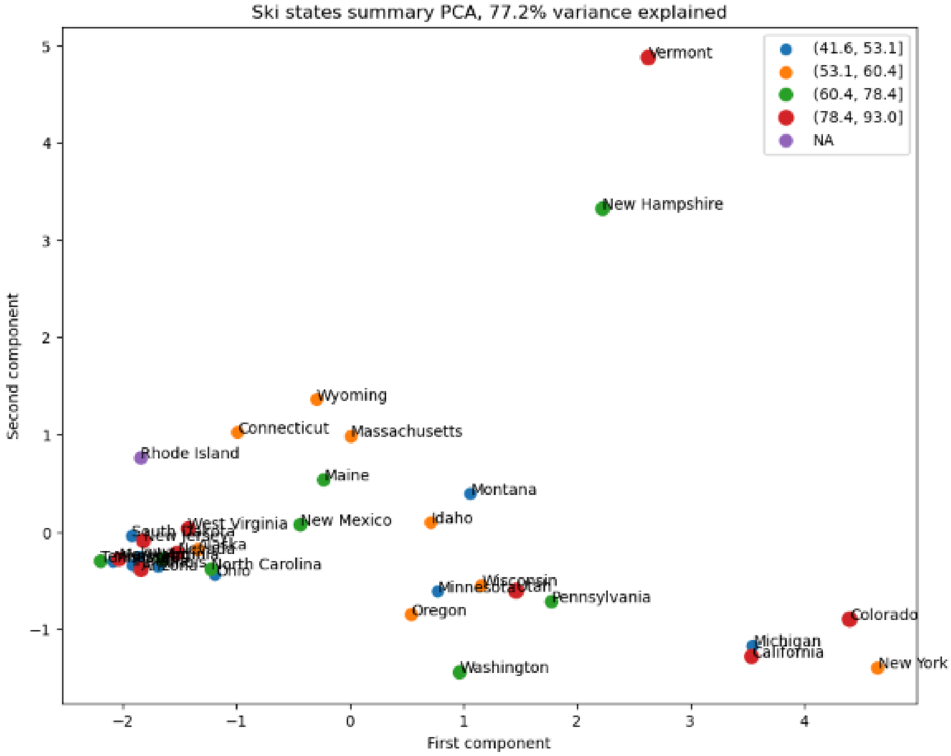


Figure 3

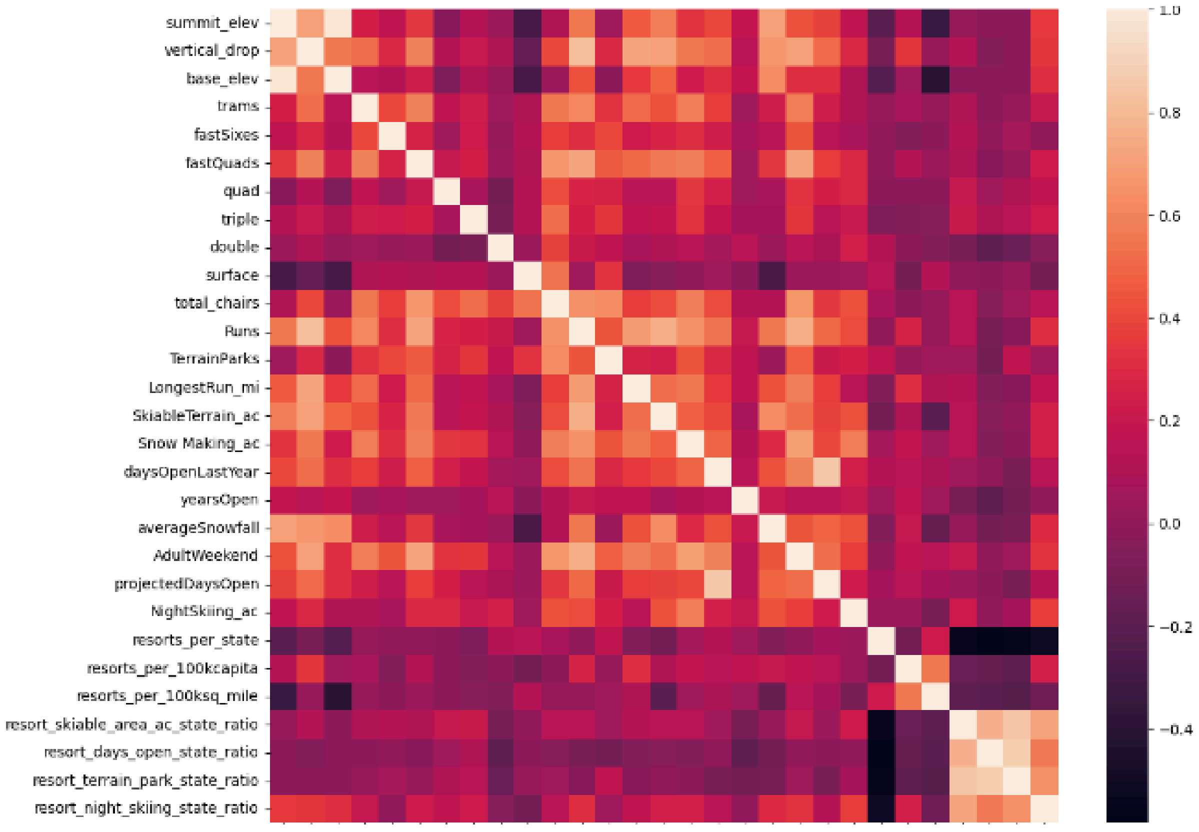


Figure 4

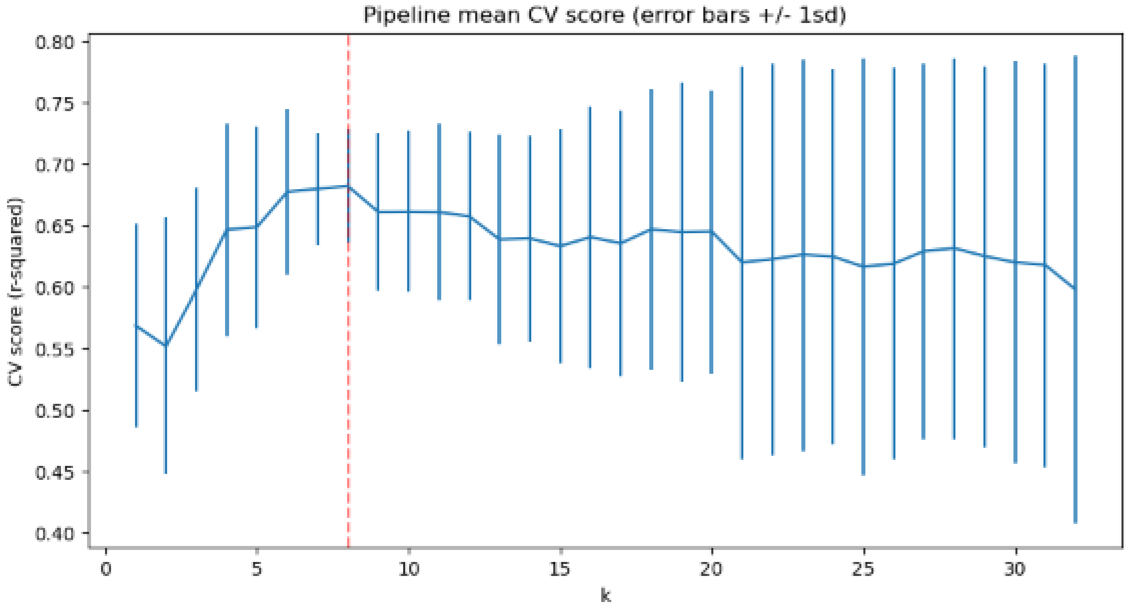
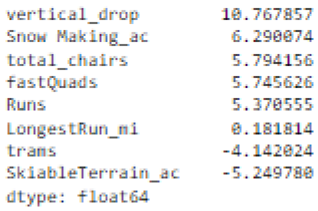
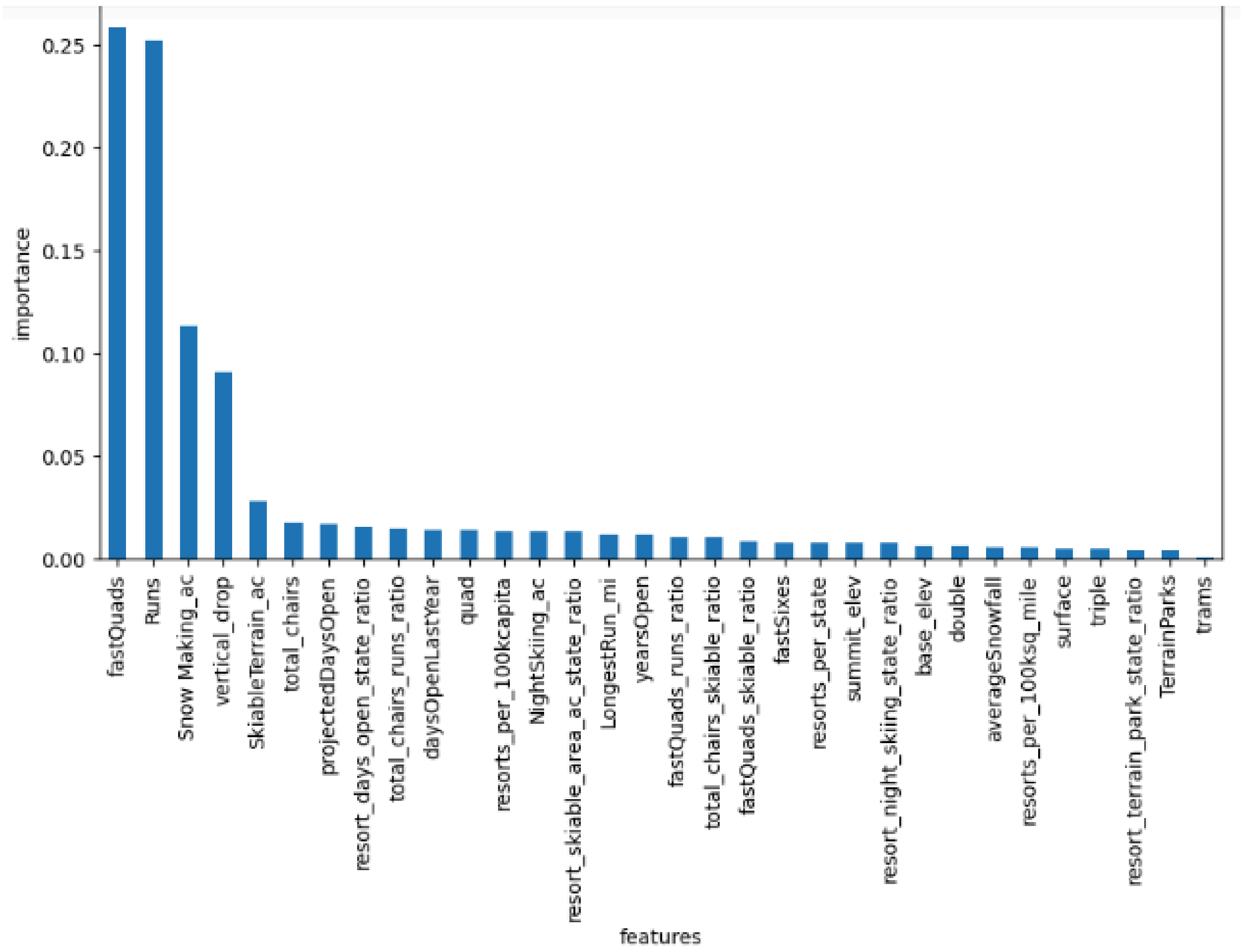


Figure 5.



```
vertical_drop        10.767857
Snow Making_ac        6.290074
total_chairs          5.794156
fastQuads             5.745626
Runs                  5.370555
LongestRun_mi         0.181814
trams                -4.142024
SkiableTerrain_ac    -5.249780
dtype: float64
```

Figure 6


importance
fastQuads
Runs
Snow Making_ac
vertical_drop
SkiableTerrain_ac
total_chairs
projectedDaysOpen
resort_days_open_state_ratio
total_chairs_runs_ratio
daysOpenLastYear
quad
resorts_per_100kcapita
NightSkiing_ac
resort_skiable_area_ac_state_ratio
LongestRun_mi
yearsOpen
fastQuads_runs_ratio
total_chairs_skiable_ratio
fastQuads_skiable_ratio
fastSixes
resorts_per_state
summit_elev
resort_night_skiing_state_ratio
base_elev
double
averageSnowfall
resorts_per_100ksq_mile
surface
triple
resort_terrain_park_state_ratio
TerrainParks
trams
features