Big Mountain Resort Pricing Model Summary Report

By: Katia Lopes-Gilbert
April 1st, 2024

Abstract
Using the dataset provided, I built a model that leverages random forest regression to predict how much more Big Mountain Resort can charge for tickets for certain existing facility features, adding new features, and how certain cost-cutting measures might impact revenue.

Introduction
Big Mountain Resort operates a large ski resort in Montana that offers a breadth of facilities for snow sport adventurers of all levels. They recently installed a new ski lift that will cost $1,540,000 to run this current season. Currently, Big Mountain Resort pricing strategy is to charge above average prices compared to other resorts in the market. This strategy is not the most informative as it does not allow leadership to consider what part of their operations is actually sought after by snow sport people. In order to change their pricing model, Big Mountain Resort has provided the team with access to information about their facility offerings and a dataset of what other resorts offer (cost, facilities, location, etc.).

I was provided with the ski_resort_data.csv dataset containing information about 330 mountain ski resorts in the United States. Each row contains information about a resort, including name, geographical information, and facility features (ex: total skiable area, vertical drop, ticket prices, etc.)

Project Objective
The goal of this project is to use data science to determine what features are most valuable in Big Mountain Resort's ticket pricing strategy. Can Big Mountain Resort charge more for premium facility features? Can reducing or removing certain features cut costs without having a significant impact on revenue? What changes can be made to generate at least $1,540,000 in revenue to break even with the new lift installation?

Exploratory Data Analysis
The dataset was analyzed using Python libraries including pandas, matplotlib, seaborn, numpy, and scikit-learn.

After an initial look at the data, I found the following information:

- There were 330 records of the ski resort dataset, which contained no duplicate values.
- Big Mountain Resort was included in this dataset and had no missing values.
- Most resorts had information listed regarding features, however, some resorts were missing feature values, with one feature missing over 50%.

Table 1: Summary of Missing Data

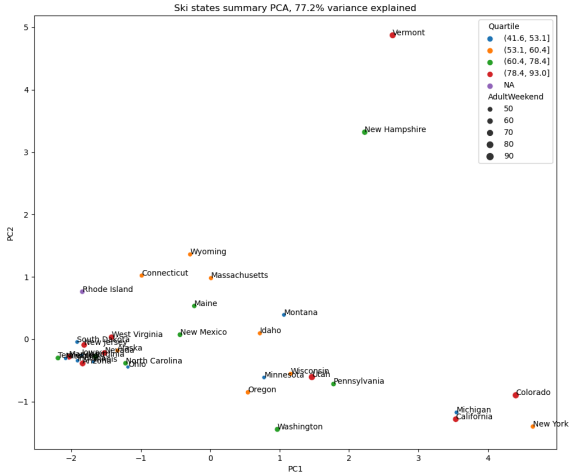
Feature	*Percent of Data Missing*
Fast Eights	50%
Night Skiing	43%
Adult Weekday	16%
Terrain Parks	16%
Days Open Last Year	16%
Adult Weekend	16%
Projected Days Open	14%
Snow Making	14%
Average Snowfall	4%
Longest Run (mi)	2%
Runs	1%
Skiable Terrain	1%
Years Open	0%

After reviewing the data's null values and outliers through functions and graphs, the following changes were made:

- Dropped the fastEight column because half of the values are missing and all but the others are the value zero.
- Weekday prices had more null values, so these were removed after performing some initial computations and comparisons between Weekday and Weekend prices.
- Created a states dataframe to add population and geographical area information that may be useful for future analysis.
- Removed resorts that were missing price data, but I did this after extracting other useful feature information for state-wide summaries.

Exploratory Data Analysis

One of the first things we did was conduct principal component analysis to find linear combinations of original features that are uncorrelated with one another for the state summary dataset. To do this, we had to scale our dataset and then calculate the PCA transformation. The first two components accounted for 75% of the variance, and the first four accounted for 95%. After analysis, there were no obvious relationships or groupings between states. Due to this finding, I did not treat any states differently in further analysis.

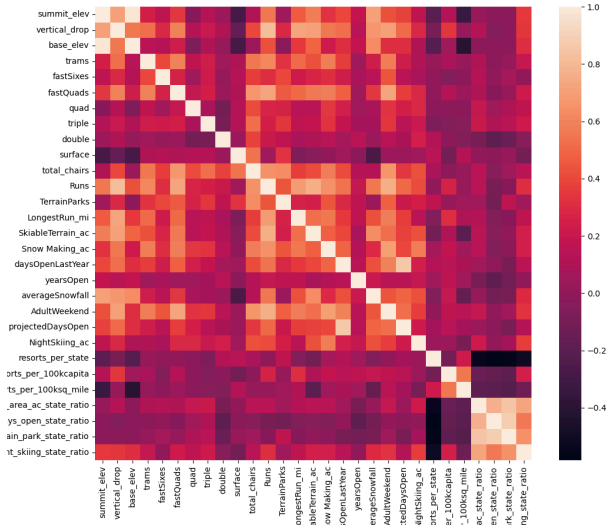


After merging the state summary data with the resort data, I created features to put each resort in the context of its state. The following features were created:

- 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 total state night skiing area

I created a heatmap to look at correlation between all available features in the combined dataset.

The biggest observations from the heatmap were that the ratio of resort night skiing in each state, number of runs, number of chairs, snow making, vertical drop, and fast quads were positively correlated with Adult Weekend ticket prices.



Processing and Training

I used a 70-30% train-test split of the ski resort data for modeling.

To start, I calculated the average price of tickets on our training data and ran some tests to see how it performed at predicting the prices of our test data. The mean price for Adult Weekend tickets is approximately $63 dollars.

After performing a R^2 test to identify the coefficient determination, we see that the mean performed worse than when trying to predict prices on values it hadn't seen yet with a score of -0.0031235200417913944.

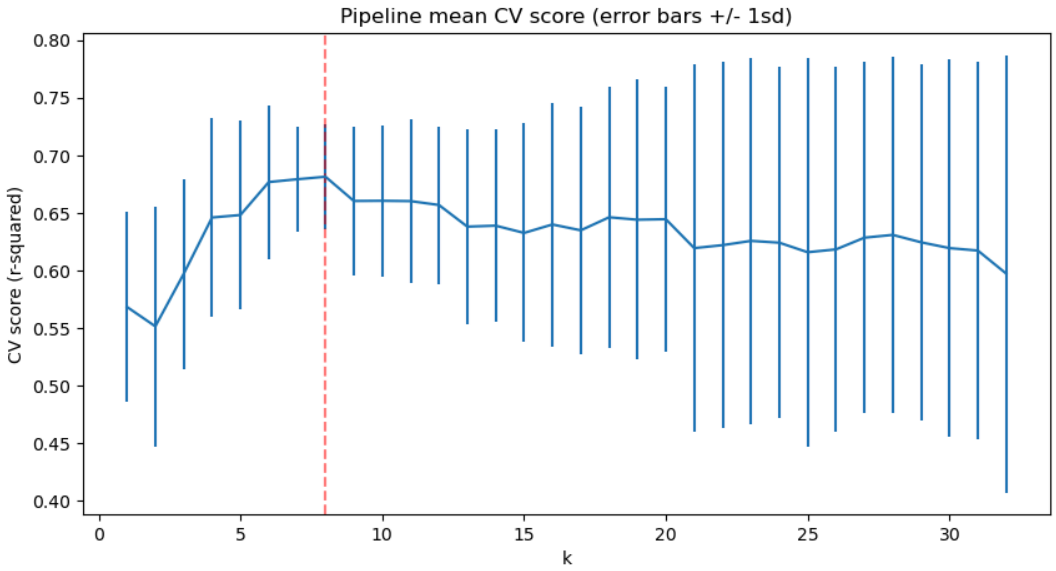
I moved on to calculating metrics that summarize the difference between predicted and actual values by calculating the mean absolute error and mean squared error. The mean absolute error tells us that, on average, we might expect to be off by around $19 if we guessed ticket price based on an average of known values. Our mean square error performed slightly better on the test than on the training data: (614.1334096969046, 581.4365441953483).

After these tests, we reviewed the data before imputing information for missing values. First I imputed missing values using the median value for each variable. I then scaled the data using

StandardScaler() and the `.fit()` method to fit the scaler. I then called the `.transform()` method to apply the scaling to the training and testing data.

After creating a linear model, I see that our simple linear regression model explains over 80% of the variance on the train set and over 70% on the test set. Clearly there is something to this, however, the much lower value for the test set suggests the model was overfitted. I then calculated the mean absolute error and mean squared error. Using the linear model, on average I'd expect to estimate a ticket price within $9 or so of the real price. This is much, much better than the $19 from just guessing using the average. The mean squared error also improved with the following result: (111.89581253658483, 161.7315645119226). I compared these results to imputing for the mean and didn't find any significant differences. I cross-validated these results by using sklearn's pipeline method. The results confirmed that the pipeline is doing what is expected and the results were identical to the earlier steps.

I modified the pipeline to to select a different number of features. I started out with 15 but then used GridSearchCV to identify the best number of features to use.



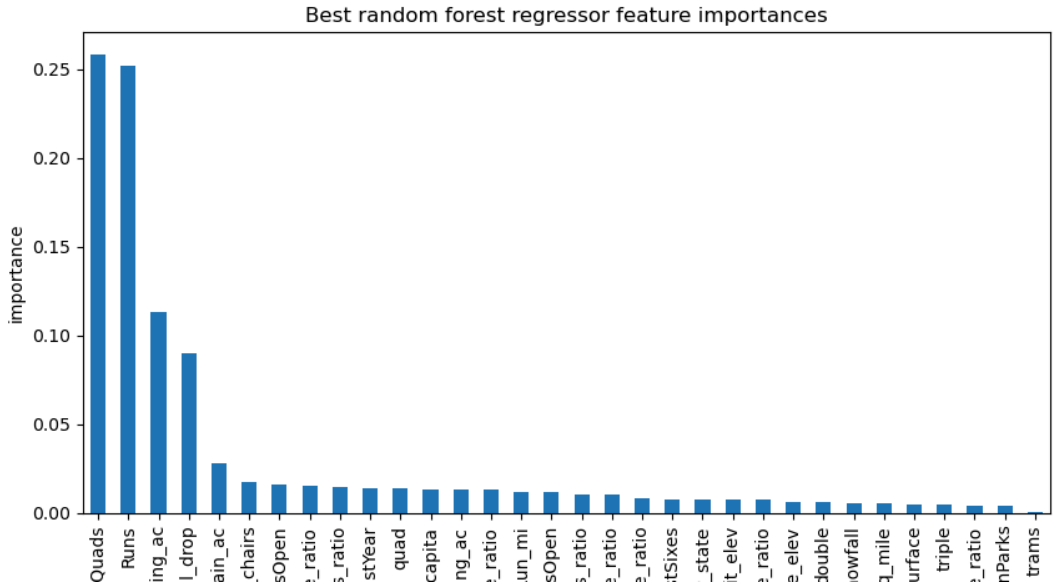
The chart shown above suggested a good value for k is 8. There was an initial rapid increase with k, followed by a slow decline. Also noticeable is the variance of the results greatly increased above k=8. As we increasingly overfit, expect greater swings in performance as different points move in and out of the train/test folds. We also identified which features were most useful, and got the following results:
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

This suggests that vertical drop is your biggest positive feature. Also, the area covered by snow making equipment is a strong positive as well. The skiable terrain area is negatively associated with ticket price! This seems odd. People will pay less for larger resorts? There could be all manner of reasons for this. It could be an effect whereby larger resorts can host more visitors at any one time and so can charge less per ticket.

I continued to try a random forest regression model and identified which features were best for predicting the Adult Weekend price. Encouragingly, the dominant top four features are in common with the linear model:

- Fast Quads
- Runs
- Snow Making
- Vertical Drop



I finally cross-validated results and computed metrics for the linear model and the random forest regression model. The random forest model had a lower cross-validation mean absolute error by about $1.

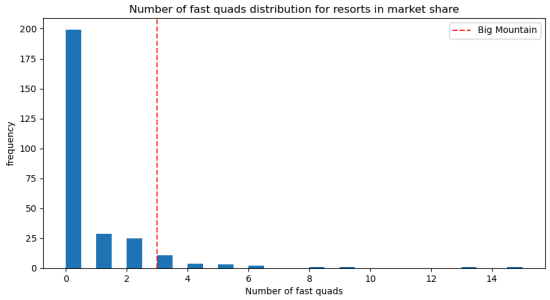
Model Selection and Predictions

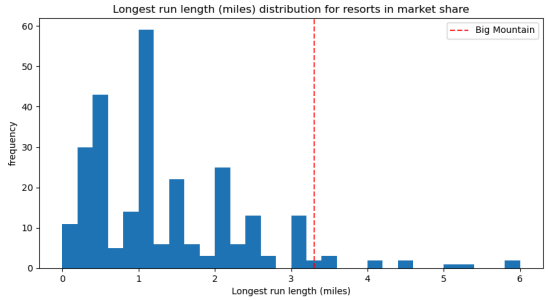
Big Mountain resort currently charges $81 for Adult Weekend tickets. First I fit the model to Big Mountain's data to calculate the expected price. The model returned a price of $95.87, suggesting that there is room for an increase in ticket price but not necessarily a ten dollar increase.

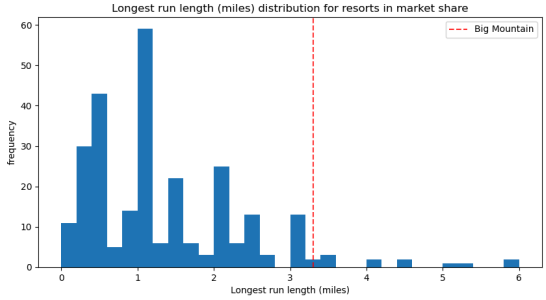
Features that came up as important in the modeling (not just our final, random forest model) included:

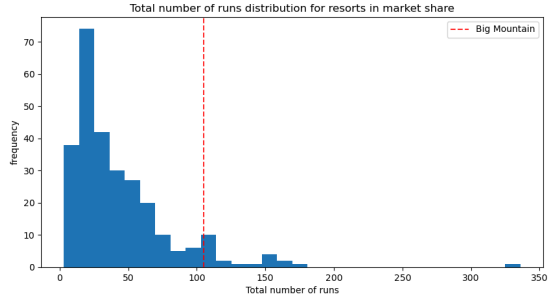
- vertical_drop
- Snow Making_ac
- total_chairs
- fastQuads
- Runs
- LongestRun_mi
- trams
- SkiableTerrain_ac

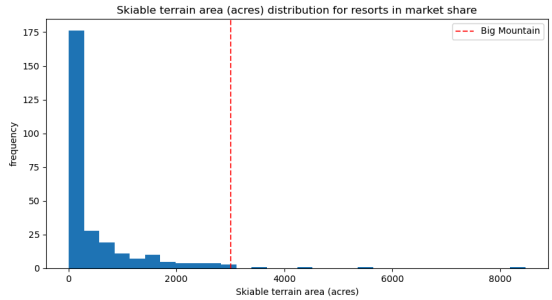
I then plotted histograms to compare Big Mountain's key features against these features for other resorts. For many of these features, Big Mountain offers more than most resorts, yet some still do have even more like higher vertical drop, longer and more runs, etc.

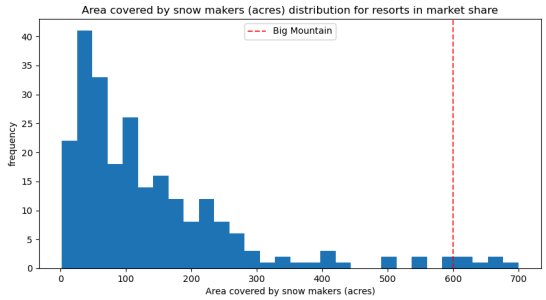


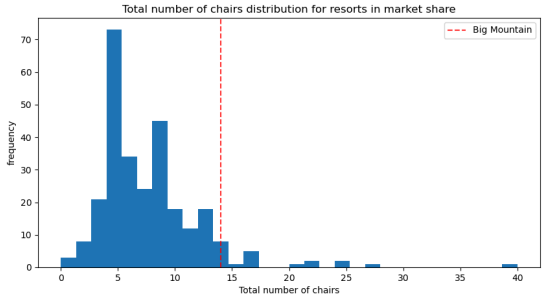


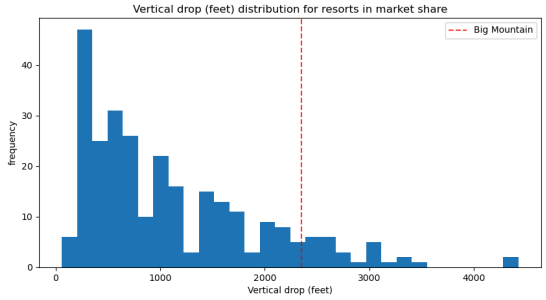


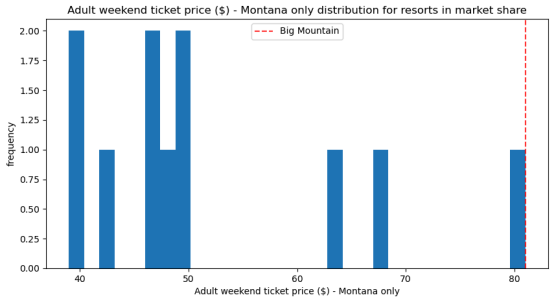












Big Mountain Resort has been reviewing potential scenarios for either cutting costs or increasing revenue (from ticket prices). Ticket price is not determined by any set of parameters; the resort is free to set whatever price it likes. However, the resort operates within a market where people pay more for certain facilities, and less for others. Being able to sense how facilities support a given ticket price is valuable business intelligence. This is where the utility of our model comes in.

The business has shortlisted some options:
1. Permanently closing down up to 10 of the least used runs. This doesn't impact any other resort statistics.
2. Increase 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
3. Same as number 2, but adding 2 acres of snow making cover
4. Increase the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres

The expected number of visitors over the season is 350,000 and, on average, visitors ski for five days.

I wrote a function, predict_increase, that takes features and deltas as arguments that would predict the ticket price given the difference between the scenario's prediction and the current prediction.
I used this function to predict prices based on changes in these features to identify which would be the most worthwhile choice. The model determined closing one run makes no difference. Closing 2 and 3 successively reduces support for ticket price and so revenue. If Big Mountain closes down 3 runs, it seems they may as well close down 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a large drop.

Increasing the vertical drop by 150 feet increases support for ticket price by $1.99. Over the season, this could be expected to amount to $3,474,638. By contrast, doing both this and adding 2 acres of snow making area makes no additional difference to the suggested ticket price. SImilarly, increasing the longest run would not have an impact on price.

Conclusion
Since operating the new chairlift will cost Big Mountain $1.5M, increasing the vertical drop by 150 feet to justify increasing the ticket price by $1.99 seems to be the best option to increase revenue to not only cover their new operating costs, but also have more profit.

Future Scope of Work
We did not have any data on the number of visitors per year for other resorts. Knowing this information could have allowed for further categorization of resorts based on ranges of visitors per year. It could have also provided more insight into the popularity of certain features.
We don't know how other resorts price their tickets. Our modeling assumed that prices were decided based on features and free-market principles.

Our modeling price might be so high because of the feature offerings Big Mountain has. That being said, Montana is not exactly close to any major cities and so traveling to the resort might be a barrier for increasing the number of annual visitors. Just a quick Google search shows that Mammoth in California gets over 1.3M visitors a year, nearly 5x as many as Big Mountain. However, Mammoth is much closer to big cities, including Los Angeles, San Francisco, and Reno. Executives may not be surprised about the suggested modeling price for this reason.

In the future, this model could be used to compare changes to existing features or adding new features and how that could impact the pricing for Adult Weekend tickets. They could also use it to see if changing other features could result in lower operating costs without drastically impacting revenue. The model could be turned into a simple application that would allow for selecting one or more features to run tests on with the specific changes to those features.