Big Mountain Ski Resort Preliminary Pricing Model

An Investigation Into a Data-Driven Approach for Optimizing Ticket Price Through Chair Lift Systems

Introduction

The client, Big Mountain Ski Resort, has installed a new chair lift system and wants to overhaul the pricing plan for their lifts, using a data-driven approach instead of simply charging a premium over the average price of resorts in the market segment. Big Mountain is aware this strategy is suboptimal and does not take into account facility importance. They also want to see if there is a way to cut costs or boost ticket price value further.

The goal of this project is to use market data from 330 other resorts to analyze and understand the intricacies of chair lift systems used by resorts in the market, and see which systems are optimal and which ones are not. Based on this analysis, changes can be made to Big Mountain in line with a more optimized strategy.

This project aims to build a predictive model for ticket price based on a number of facilities, or properties, at the resorts. This model will be used to provide guidance for Big Mountain's pricing and future facility investment plans.

Initial Data Observations and Cleaning

I want to make a model for ticket price based on physical features of a resort. This means the target will be either weekday or weekend ticket price, and predictors will be some combination of other features.

Because more rows are missing weekday values over weekend values, I drop the weekday column and remove any rows with na values for weekend. Therefore, the target variable in this model will be AdultWeekend.

As for predictors, the next step will be to decide what data to further look into. For example, Big Mountain is found in Montana, so I could only look into resorts from Montana to keep geographical and population factors unbiased, or I could include data from all states to generate more data points. To do so, I collect statewide statistics and enhance the original data set with

- state_population
- state_area_sq_miles

Looking into attributes of resorts in respect to population density and state size will give me better context for market insights. For example, resorts from a smaller state with a high population such as New York may have different accommodations than resorts from a larger state with lower population, such as Montana.

EDA

I start with examining different features from statewide summary statistics. These features are

- resorts_per_state
- state_total_skiable_area_ac
- state_total_days_open
- state_total_terrain_parks
- state_total_nightskiing_ac
- state_population
- state_area_sq_miles

While none of these features skew bias towards a single state, they do raise questions of how I will preprocess the data. Without any clear indicators, I need to find some kind of relationship between features. To do so, I use principal component analysis (PCA).

PCA flattens a dataset with high dimensionality down to a lower dimension for visualization. It does this by finding linear combinations of the original features and ordering them by the amount of variance explained.

The steps for this process are

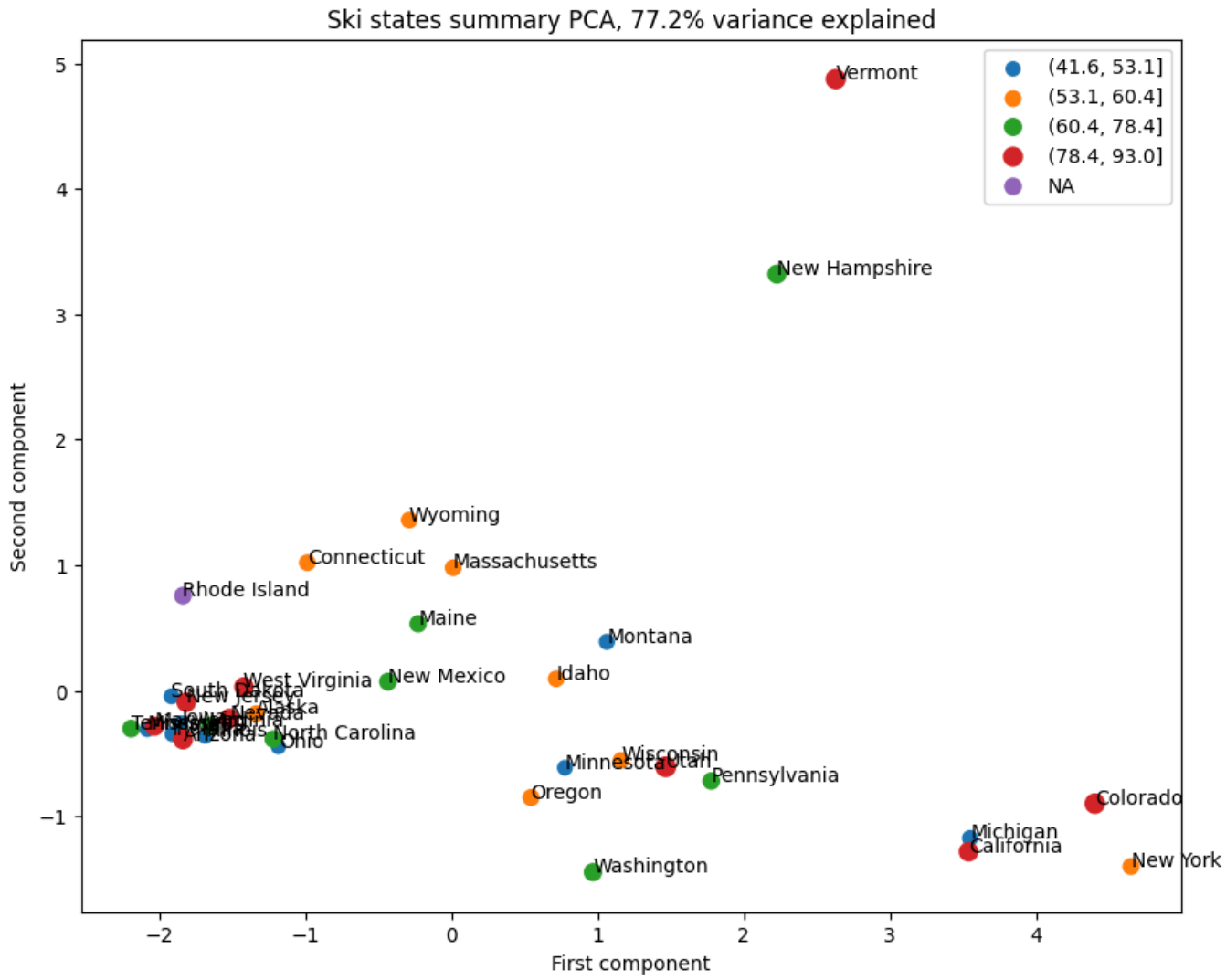
- scaling the data, as it is currently heterogenous
- fitting the PCA transformation
- apply the transformation to the the data to create the derived features
- use the derived features to look for patterns and explore coefficients

I scale the data using Z - score normalization. After fitting the transformation, use a cumulative variance ratio plot to reveal the first two components being responsible for the majority of the variance, at 75%.

I then apply the transformation to the data and plot the states against the first two principal components.

Before examining how the states are ranked based on these two components, I add another layer of depth to the visualization through quartiles based on the mean weekend ticket price per state

- Q1($41.6, $53.1)
- Q2 ($53.1, $60.4)
- Q3 ($60.4, $78.4)
- Q4 ($78.4, $93.0).



The above PCA biplot shows the position of the states based on their relative similarity and relationships in the original high-dimensional space.

Ticket price distribution is varied across the plot, we see colors all over. For the first component, the majority of states are ranked close together, with Michigan, California, Colorado, and New York being the most positively correlated. For the second component, Vermont and New Hampshire jump out the most, scoring very positively compared to other more modest states.

To explain these variations from the group, I looked at coefficient scores for the two components.

For the first component, all of the coefficients are positive, with no clear difference in magnitudes. There are three coefficients which are stronger among the group, being

- resorts_per_state, 0.48
- state_total_days_open, 0.48
- state_total_terrain_parks, 0.48

New York and Michigan have higher values for resorts_per_state, being 2.99 and 2.35 std deviations from the mean. California and Colorado have higher values for state_total_days_open, being 2.19 and 2.81. California, Colorado, and New York have higher values for state_total_terrain_parks, being 2.61, 2.33, and 2.21 std deviations from the mean. None of these metrics within the group warrant special treatment.

For the second component, the only two positive coefficients stand out in the group, being

- resorts_per_100kcapita, 0.66
- resorts_per_100ksq_mile, 0.63

Both Vermont and New Hampshire have high values for resorts_per_100ksq_mile, being 3.11 and 3.48 std deviations away from the mean. Vermont also has a large value for resorts_per_100kcapita. This makes sense as these states are among the smallest by land size in the US, increasing resort to square mile density as well as per capita density.

Overall, the PCA biplot of differing characteristics of the states based on summary statistics does not reveal any particular trend or anomaly, even when investigating outliers. Therefore, I will treat all states equally and develop a pricing model using all data points available.

I then perform feature engineering to create four ratios detailing the asset share within a state for each resort:

- resort_skiable_area_ratio = skiable_terrain_resort / skiable_terrain_state_total
- resort_days_open_ratio = days_open_resort / days_open_total
- resort_terrain_parks_ratio = terrainParks_resort / terrainParks_state_total
- resort_night_skiing_ratio = nightskiing_ac_resort / nightskiing_ac_state_total

With the ratios added to the main dataset, I visualize all relationships with a correlation heatmap, and pay attention for any that resonate strongly with the target of ticket price:

- fast_quads
- total_chairs
- snow_making_ac
- resort_night_skiing_state_ratio
- runs

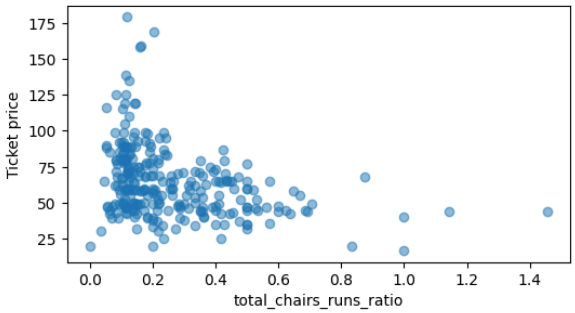
It makes sense for faster and a higher count of chairs to correlate strongly with ticket price, as customers are willing to pay for more efficient service. It is interesting to note how fast_quads specifically has the highest correlation to ticket price among all chair types, compared to fastSixes, or the regular quad, triple, and double. Perhaps fastQuads are the preferred chair type of the public.

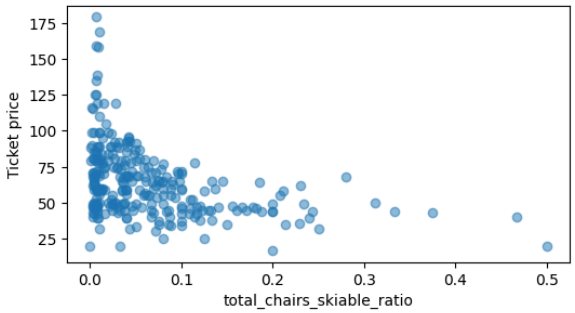
It also makes sense for more guaranteed snow (snow_making_ac) , greater skiing time availability (resort_night_skiing_state_ratio) , and trails (runs) to lead customers to accept higher ticket prices.

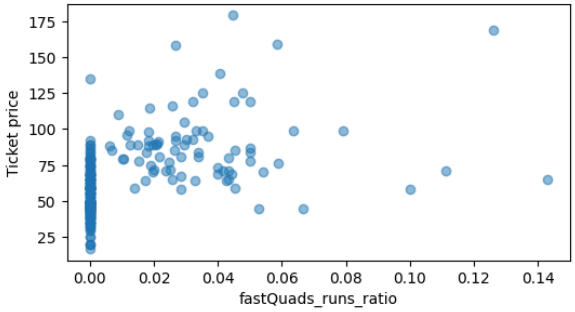
Looking into fast_quads and total_chairs specifically, I engineer four more ratios to get an idea of how a resort distributes its fast quad system over their entire property:

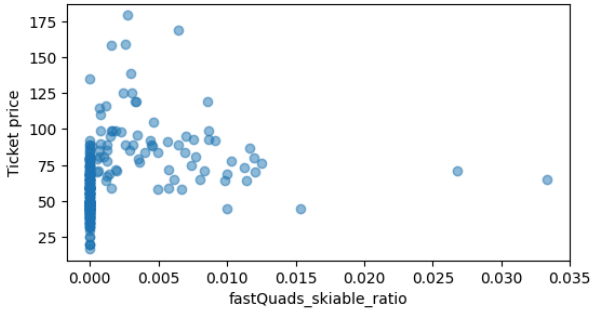
- total_chairs_run_ratio = total_chairs / runs
- total_chairs_skiable_ratio = total_chairs / skiable_terrain_ac
- fastQuads_runs_ratio = fastQuads / runs
- fastQuads_skiable_ratio = fastQuads / skiable_terrain_ac

Plotting these ratios against ticket price reveals a counterintuitive relationship between total chairs in regards to runs and skiable area.









It seems having more runs and skiable terrain, resulting in a more accessible skiing experience, does not correlate with a higher ticket price.

This observation hints at an intriguing dynamic in the ski resort industry. It suggests that there might be two distinct strategies at play:

- cater to an exclusive clientele
- targeting a mass market audience

Resorts with a smaller chair to skiable accessibility ratios but higher ticket prices may be looking to market their reserved seating as an exclusive skins experience, where perhaps their location has access to premier skiing attributes other resorts do not.

On the other hand, resorts with a high chair to skiable accessibility ratios may be looking to draw in crowds based on the constant availity of seating, where perhaps their location is not as desirable.

Without a metric for the number of visitors, there is no clear way to tell.

It is also apparent not having fastQuads severely limits ticket price, making these a key part of resort systems.

Base Model Creation

To start with the modeling process, I partition the data in a 70/30 train/test split.

I use the mean of ticket price as a dummy regressor. Doing so gives a MAE of $19 and a RMSE of 24 on test data.

Using sklearn.pipeline, I fit a linear regression model which results in an improved MAE of $9 and a RSME of 12.6.

To further improve performance, I add in SelectKBest with GridSearchCV to find k = 8 as the optimized fold value.

Looking at the coefficients of the selected linear model reveals the following:

- 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

Vertical drop and snow made acres are the most correlated features with ticket price, as resorts with better elevation are naturally more desirable for skiing, and more guaranteed snow draws in more crowds.

To pair with the linear model, I implement a random forest model imputed with the ticket price median and tuned to 69 estimators.

When looking into feature importance for the random forest, the dominant top 4 features are in common with the linear model, being:

- fastQuads
- Runs
- Snow Making_ac
- vertical_drop

The relevance of these four features in both models indicates their importance to the dataset and should be explored in more detail in further analysis.

For now, comparing the best performing linear model against the best performing random forest model shows random forest exhibiting lower mae, variability, and mse, so random forest will the model of choice for scenario modeling.

Modeling

Running the model on Big Mountain Resort results in a model price of $95.97, compared to the actual price of $81.00. There is a mae of $10.39, suggesting a room for increase regardless of model error.

The validity of this result lies in the assumption of other resorts accurately setting their prices based on the public. It seems Big Mountain is underpricing their tickets, but it is possible other resorts could be under or even overpricing their prices as well. When comparing notable numerical features of Big Mountain to other resorts, Big Mountain seems to be on the higher end for all. This suggests Big Mountain can explore the possibility of raising ticket price, based on the fact they outperform other resorts based on physical attributes alone.

Continuing on with modeling scenarios, when adding an additional chair lift, ticket prices are raised by $1.99 to support operating costs. This can be paired with the dropping of a single unpopular run to further diminish operating costs without hindering revenue.

For future scenario exploration, the relationship between adding chair lifts and removing runs should be delved into, as up to 6 runs can be dropped before substantial drops in ticket prices are needed. By having more chair lifts to accommodate fewer runs, efficiency of the resort can increase, further supporting higher ticket prices. Of course, there will be a point where having too many chair lifts requires a heavy operating cost without any valuable increase to revenue.

Further Work

To expand on this data, I would recommend adding a numerical interpretation of customer volume for each resort. This can be done through past records, or expectations. Having the number of visitors as a metric would provide more avenues to explore within the data. For example, we can provide explanations for why resorts have the numerical features they do, thus providing more context for understanding the choice of ticket price. Adding in a customer volume metric to the data would add a "why" to aspects of the resort that we do not currently have access to.

Other information that would be useful includes:

- Operating expenses: Data on various operating expenses, such as labor costs, maintenance and repairs, utilities, insurance, marketing expenses, and administrative costs, would provide insights into the overall cost structure of the resort.
- Capital expenditures: Information on capital expenditures, such as investments in infrastructure, equipment, and facilities, would help assess the long-term investment and development strategy of the resort.
- Revenue streams: Data on revenue sources beyond ticket sales, such as revenue from food and beverage sales, equipment rentals, ski school programs, retail sales, and lodging accommodations, would provide a more comprehensive understanding of the resort's revenue streams.
- Cost of new features: In the case of Big Mountain, information on the cost of installing and operating the new chair lift would be crucial for evaluating the impact of this investment on ticket prices and overall profitability.

Additionally, it's important to investigate why the modeled price for Big Mountain was significantly higher than its current price. This discrepancy could be attributed to various factors, including the resort's positioning in the market, the perceived value of its facilities and amenities, competitive pricing strategies, and the willingness of customers to pay higher prices for premium experiences.

It's possible that business executives may be surprised by the model's findings, especially if they had not previously considered market comparisons or if they were unaware of Big Mountain's relative position compared to competitors.

To address these questions and concerns, business leaders may utilize the model as a strategic tool for decision-making. They may seek further analysis and insights to understand the rationale behind the model's predictions and to explore different scenarios and combinations of parameters. However, it may not be practical for business analysts to rely solely on data scientists to conduct every analysis or test new scenarios. Instead, the model could be made

available for business analysts to use and explore independently through user-friendly interfaces or dashboards.

One approach to making the model accessible to business analysts is to develop interactive tools or software applications that allow users to input different parameters and instantly visualize the predicted outcomes. These tools could provide intuitive interfaces for exploring various scenarios, adjusting parameters, and conducting sensitivity analyses. Additionally, user training and support resources could be provided to help business analysts effectively utilize the model and interpret the results.

Overall, making the model available for self-service exploration by business analysts would empower decision-makers to leverage data-driven insights for strategic planning, pricing optimization, and competitive positioning for Big Mountain Resort.