Data Science Report

Prediction of ticket pricing for Big Mountain Ski Resort

Denzel Williams

Springboard Data Science Career Track, September 2020 Cohort

Data Science Guided Capstone

October 2, 2020

Abstract

Big Mountain Resort a ski resort in Montana, United States is looking to improve the way they price their tickets. Their current ticket model takes the average price of all the tickets in the market and charges a premium above that average. With this strategy they aren't able get insight on what facilities and features support a higher ticket price and which can be safely removed. These insights are essential components on choosing what investment strategy to follow. Using the data provided along with state data sourced from the web, the research used a Random Forest machine learning technique to predict, based on the current features and facilities of the resort, what the price of their ticket should be. The results of the model suggest that Big Mountain Resort has facilities and features that justify a higher ticket price of $95 compared to their current ticket price of $81. Of the four investment strategies that the resort proposed to increase revenue or cut costs it is recommended that they improve their vertical drop feature by adding a new run to a point 150ft. lower down and install a chair lift to bring skiers back up.

---

1 - Introduction

To price tickets for their resort, Big Mountain Resort uses a pricing model where they charge a premium above the average price of all the resorts in its market segment. Their current pricing model doesn't give them insight on how to allocate funds for their investment strategy. By only utilizing average price it is unclear if the facilities they have or want to invest in support a higher ticket price. Additionally, it is unclear in which directions the business can costs such that the ticket price doesn't become overvalued. Understanding the impact of different facilities and features of the resort give Big Mountain Resort a keener sense on how to allocate their money to drive revenue and shed weight.

2 - Exploratory Data Analysis

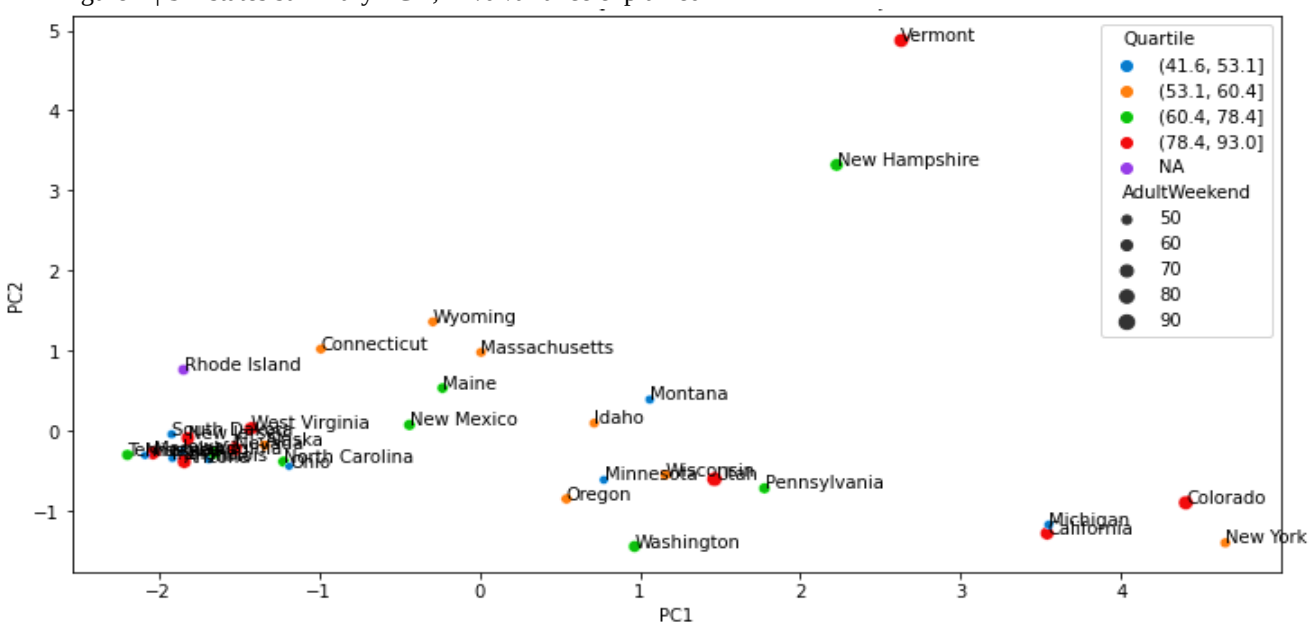
The Database Manager at the resort provided data that contained different properties of 330 resorts across the country. After inspecting and cleaning the data for various issues the number of resorts in the data set was reduced to 277. The target feature that was in focus was the *AdultWeekend* property which describes the cost of an adult weekend chairlift ticket.

The first major influence that drove the analysis was that the resorts were in different states, each with their own properties. And those properties could potentially change the dynamics of how resorts, in that state, price their tickets. For example, if is lacking in skiable terrain area then a resort may increase their ticket pricing if they are able to capture much of that area. With this understanding we had to determine how we were going to handle the state data. Using the data set, the state's properties were derived from the individual properties of the resorts along with some outsourced data on the population and area of each state. With the state data in hand, the question now became: *Are all states going to be treated equally or will the pricing model treat certain states in a special manner.*

2.1 - Principal Component Analysis

The derived state data set had 7 features to choose from. A principal component analysis (PCA) was used to visualize the data in a lower dimension and see if it can give us insight on how to proceed with handling the state data. The PCA gave us justification for treating all states equally and

Figure 1 | Ski states summary PCA, 77% variance explained



Notes: PCA on state data showing no clear way to group states

to continue by building the model with all states being equal *(Figure 1)*.

2.2 – Feature Engineering

Outside of the PCA the state data isn't generally useful. However, when put in relation to the individual resorts, the state data allows us to determine the ratio that each resort captures of specific state properties. The four new features that were engineered answered the following questions:

- How much of the state's total skiable area does the resort use *(resort_skiable_area_ac_state_ratio)*?
- How much of the state's total terrain parks does the resort own *(resort_terrain_park_state_ratio)*?
- How much of the state's total night skiing area does the resort use *(resort_night_skiing_state_ratio)*?
- How much of the state's total days open is the resort open for *(resort_days_open_state_ratio)*?

With these engineered features made, a correlation map was made between every feature in the dataset. The graph on the following page isolates the AdultWeekend column of the heatmap *(Figure 2)* revealing some interesting correlations. Of all the engineered features the night_skiing_ratio seems to be the most correlated to ticket price. If this is true, then perhaps seizing a greater share of night skiing capacity is positive for the price a resort can charge. It seems that the features that are more positively correlated with ticket price are the features that are causally related to a customer either getting on the snow or being on the snow.

3 – Training the Model

After splitting the data in train and test sets, three models were selected to be fitted to the data: A Dummy Regression Model for the baseline test, a Linear Regression Model, and a Random Forest Model. In both the linear regression and random forest models a hyper-parameter search combined with a 5-fold cross-validation was used to find the parameters that produced the best metrics on the training set. After the best parameters were found another cross-validation was run using those specific parameters to assess the models by using the average of the 5 resulting mean absolute errors. The results of the training are as follows:

- Dummy Regressor *(Baseline)*: simply using the mean as the predictor, resulted in a mean absolute error of $19 when used on the test set. This means that this model can predict a ticket price within $19 of the real price.
- Linear Regression: after optimizing the model and running a cross-validation the average of all the mean absolute errors was $10.50 with a standard deviation of $1.62. The performance of this model is significantly better than the Dummy Regressor.

When used to predict the price on the test set this model outputted a ticket price of $11.80.

- **Random Forest**: completed the same steps as in the linear regression model with an average mean absolute error of $9.64 and a standard deviation of $1.35. With a smaller average mean absolute error and a tighter standard deviation the random forest model performed the best and was chosen as our model going forward. When used to predict the price on the test set this model outputted a ticket price of $9.53.

Table 1 | Performance assessment of models

Model	Mean	Std. Dev.	Test Prediction
L. Regression	$10.50	$1.62	$11.80
R. Forest	$9.64	$1.35	$9.53

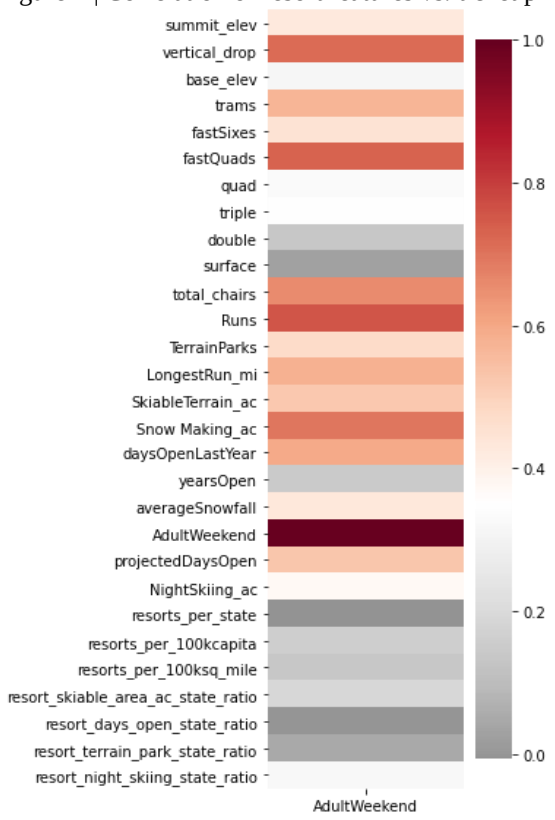
Notes: Mean absolute error metric after a 5-fold cross validation on hyperparameters

In both models there are common features that were the most dominate in determining ticket price. This observation suggests that there are in fact specific facilities and features of resorts that can drive up the ticket price. Some common dominant features in both models were the vertical drop, total amount of chair lifts, the number of runs, the amount of fast quad lifts, and the area of snow covered by snow making machines.

4 - Conclusion

When fit on all the data and used to predict the price of Big Mountain Resort the model predicted a ticket price of $95.87 as compared to the current price of $81. Even with the expected mean absolute error of $10.39, this suggests there is room for a price increase. Based on the features that are dominant in both models, Big Mountain Resort is in the upper echelon when it comes to most of those features and explains why the model outputted such a high-ticket price. Of the four courses of action the business is willing to take the second scenario is the best course of action: **Increase the vertical drop by adding a run to a point 150ft. lower, requiring an installation of a new chair lift to bring skiers back up.** This change in the resort would support an increase in ticket price by $1.99 and increase revenue by

Figure 2 | Correlation of resort features vs. ticket price



approximately $3.4M for the season. Assuming the new chairlift would have the same operating costs as the old one, $1.54M should be deducted from the $3.4M generated. Other suggestions that should be taken into consideration.

- Closing a single run does not affect the ticket price. However, closing 2 or more runs can cost the company anywhere from $700k to $2M+ in revenue.
- Adding a chair lift in isolation loses the business money. A new lift only supports a revenue increase of $507k *($0.29 per ticket)* but would cost $1.54M in operating costs.
- Close the least popular run, which our model suggest conserves revenue. Then chose another run from the least popular runs list that could easily be extended to increase the vertical drop metric by 150ft and install the required chairlift. Ideally, the most popular run from the least popular list would be chosen to potentially save marketing costs needed to drive interest into the new run.

4.1 – Limitations

The model that was built lays on the foundation of the price of the other resort's ticket price. Which is dependent on their individual pricing strategies. If Big Mountain Resort is mispricing its ticket price, then there could potentially be other resorts doing so as well. It is possible that the analysis was built on resorts that themselves are over/underpricing their ticket prices. Or if they aren't mispricing, then their pricing strategies are good which might mean that our model is lacking some key data that could be used.

4.2 – Things to Consider

One of the biggest things that the data is missing is operating costs per item. Revenue is a great metric, but profit is more important. As with an isolated chair lift installation, it brings in only a third of what it costs to operate. On the flip side, dropping two runs reduces revenue by $700K, but what if those unpopular runs cost $1M to operate. Without operating costs, the suggestions from the model are too theoretical and can serve only as a starting point for further exploration.

Another potentially useful piece of data would be the number of unique runs fast quad lifts give access to. A ratio could be created:

$$\frac{fastQuadAccess}{totalRuns}$$

Although not presented, adding a fastQuad in isolation suggests a large increase in revenue. However, is it worth it if the Quad only goes to one new unique run changing the ratio from $\frac{5}{105}$ to $\frac{6}{105}$. Increasing this ratio significantly would allow customers to get to more runs faster and then the business could try to use a separate "premium ticket model" that charges extra for access to fastQuads. A downfall of this premium ticket model is that we wouldn't know if the fastQuad would have the same effect on ticket price if it is only limited to certain customers, because our model is built on the fact that everyone has access.