Big Mountain Resort Ticket Price Analysis

Big Mountain Resort is a ski resort located in Montana that is interested on if their lift ticket price is reflective of the benefits that their facilities have to offer. Historically, the pricing strategy was to charge a premium based on the average ticket price of resorts in their market segment. There is rising suspicion that the resort is not taking complete advantage of their offering.

While exploring data, we redefine the question to, “What should weekend ticket pricing be?” due to the linearity in weekday and weekend pricing, see scatter plot below.

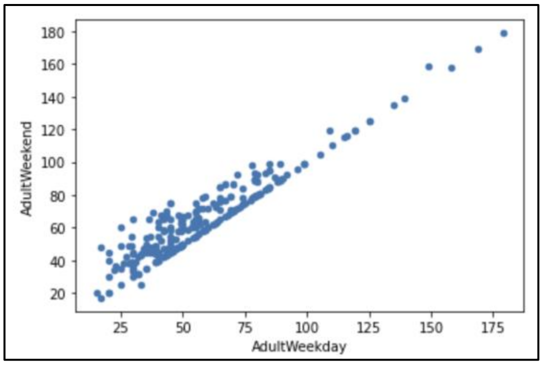

Fig 1. Correlation between weekday and weekend price

Exploratory data analysis was able to unbury some key components of the data. When constructing a correlation heat map, fastQuads stands out, along with Runs and Snow Making_ac. Visitors would seem to value more guaranteed snow by manual snow making. Of new features added, resort_night_skiing_state_ratio seems the most correlated with ticket price.

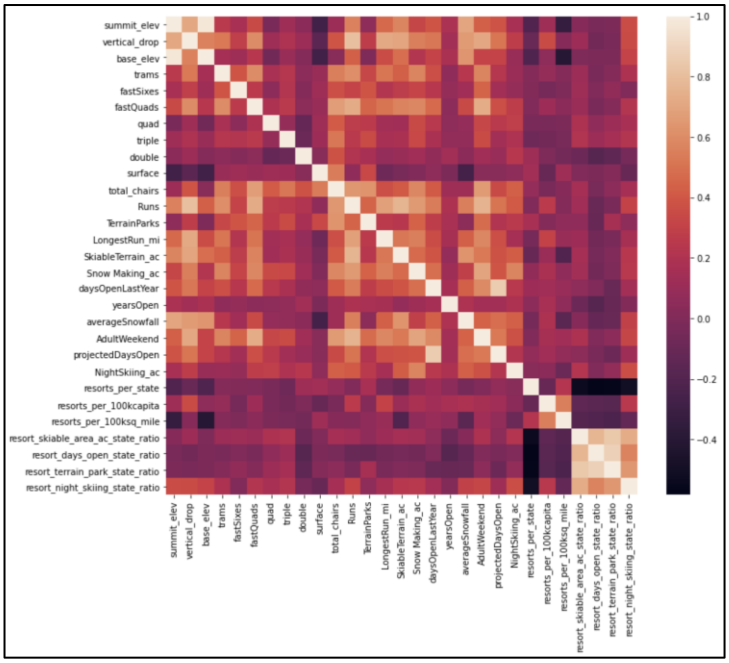

Fig 2. Heatmap correlation of all features

Completion of a principal component analysis revealed that the two top components accounted for 77% of the variance in the dataset. In the end, there exists justification for treating all states equally, and working towards building a pricing model that considers all states together, without treating any one particularly specially. At this point, there is not any clear grouping between features, but we have captured potentially relevant state data in features most likely to be relevant to the business use case.

When starting to train the data, the data is split into a train and test set to evaluate performance of the model. To start, we wanted to see how good the mean is as a predictor. In other words, if we simply say our best guess is the average price. The main metric that we will use to summarize the difference between predicted and actual values is the mean absolute error or the average of the absolute errors, see equation below. Our mean absolute error in this case told us that we might expect to be off by around $19 if we guessed ticket price based on an average of known values.

$$MAE = \frac{1}{n}\sum_{i}^{n} |y_i - \hat{y}|$$

Yi are the individual values of the dependent variable
ŷ are the predicted values for the depended variable

Next we decided to impute missing values with median. We trained the model on the train split. This was a linear regression model on scaled data to account for the various magnitudes of the measured data. The mean absolute error values suggested that on average you'd expect to estimate a ticket price within $9 or so of the real price, much better than the $19 in the previous model.

We then moved on to refining the Linear Model as we suspected the model was overfitting. We implemented a feature selection function that defaulted to using 10 best features. What we saw though, was that 10 features is worse than using all features as the mean absolute error was over $11. Instead of selecting different k values and tuning the model to the arbitrary test set, we used a technique called cross-validation to pick the value of k that gives the best performance. We found that a good value for k is 8. These results suggest 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 in this particular model.

A random forest regressor was then tried, where we go straight from defining the pipeline to assessing performance using cross-validation. The dominant top four features are in common with our linear model: fastQuads, Runs, Snow Making_ac, vertical_drop.

We built a best linear model and a best random forest model and need to choose between them. We decided to calculate the mean absolute error using cross-validation. The linear model came in at 11.79 and the random forest model came in at 9.54. The random forest model has a lower cross-validation mean absolute error and has less variability so we decide to move forward with that model for production.

We also see that we have sufficient data sizes and the client does need to do anymore collecting and investing of further resources. There's an initial rapid improvement in model scores as one would expect, but it's essentially levelled off by around a sample size of 40-50.

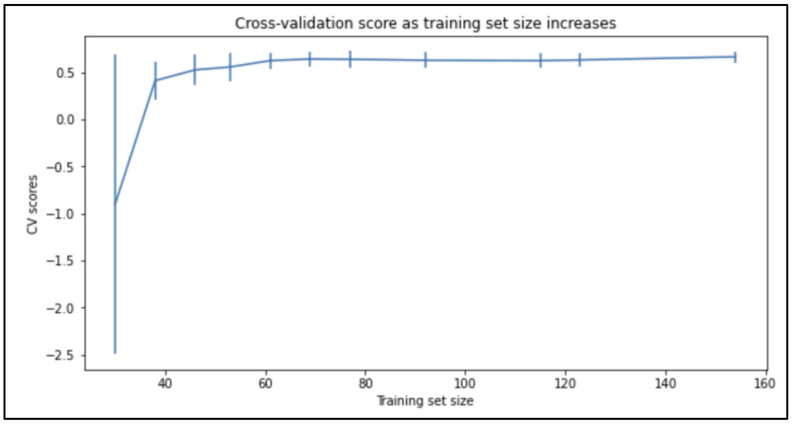

Fig 3. Performance with differing data set sizes

In our analysis, we follow the assumption that all other resorts are largely setting prices based on how much people value certain facilities. It might be the case that some of these resorts are overpriced and some are underpriced. We know nothing about operating costs, and that could help us understand how raising ticket prices could offset some of the costs, or the costs associated with making change in facilities.

In conclusion, Big Mountain currently charges $81 for an Adult Weekend ticket, the pricing of interest in this study. Our random forest regression model suggests that with the facilities that Big Mountain resort offers, the resort could support a ticket price of $95.87. Even with the expected mean absolute error of 10.39 dollars, this suggests there is room for an increase. Big Mountain ranks high in major categories like Vertical Drop, Snow Making Area, Total Number of Chairs and Fast Quads even further supporting a ticket price increase.