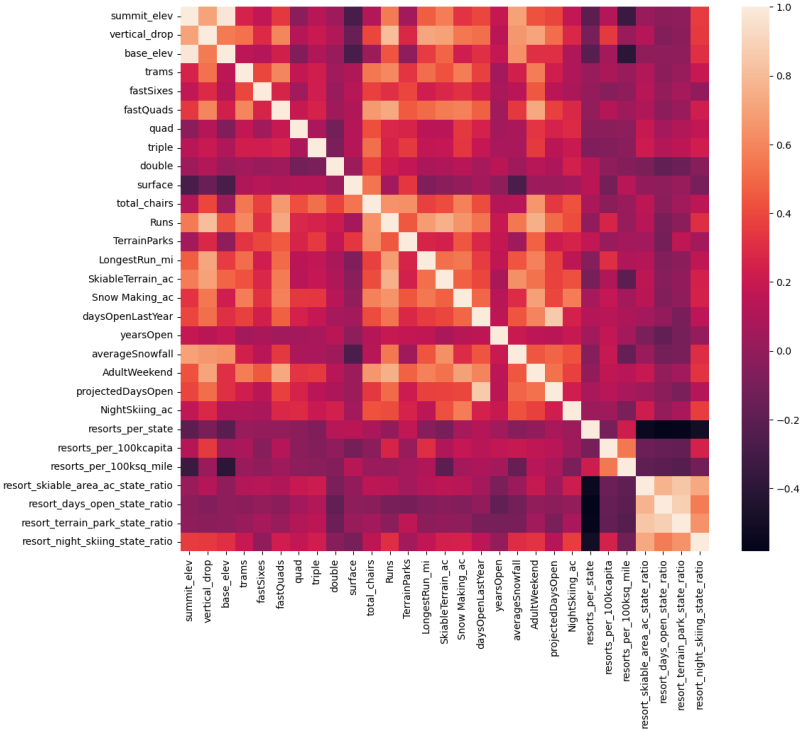
Big Mountain Resort Pricing Model

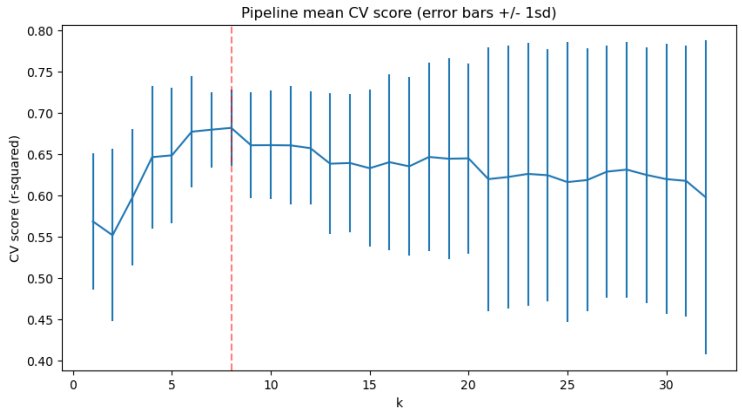
This report will summarize our findings on a new pricing model for Big Mountain Resort, as well as the basis for those findings.

Using the dataset provided, we aimed to build a predictive model for competitive pricing based on pricing trends in comparable ski resorts across the US. After evaluating the provided data, which contained some empty/irrelevant features and missing/implausible values, we narrowed our focus to a target value (weekend ticket prices, which had fewer missing values than weekday ticket prices, and we confirmed that both values were correlated strongly enough to drop one) and built a dataframe better suited to our purpose by turning certain features in the data into more comprehensible values (ex. merged external data on state areas and populations in order to better understand the relevance of impacted features). Since our goal was to create a predictive model specific to only comparable resorts, it was important to ascertain whether there were external factors that might require us to subset the data for our model, i.e. whether there were significant differences in pricing patterns between different states, which might suggest differences in demand/market size and would skew our findings. We will note here that an analysis of null values seemed to suggest that some information had been artificially/systematically removed from the set for unknown reasons.

Further exploration required us to apply PCA to determine next steps in our data cleaning. This revealed a positive relationship between ticket prices and the ratio of resorts vs. state area and population. No data were removed in this step but it is relevant to include here should the client want the model altered further on. Analysis of relationships between features did allow us to draw out a number of factors most correlated with ticket price, namely vertical drop, number of fast quads, number of runs, and chairs:

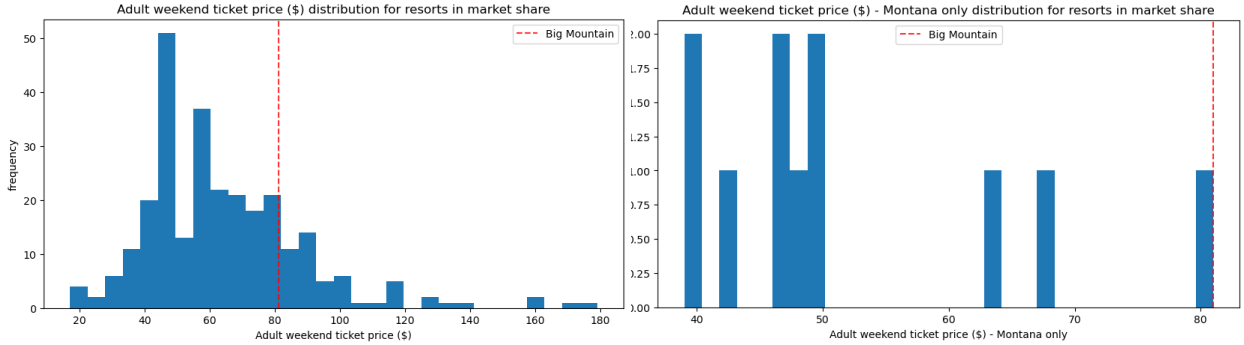


After producing a usable dataset, we moved on to building and training our final model. Using a 70/30 split, we first tested the effectiveness of simply using the mean: compared to a linear regression on the raw dataset (not corrected for feature relevance or multicollinearity), performance of the mean as a predictor was poor. We moved forward and built a pipeline to scale the data, impute missing values, and perform a linear regression, and used k best feature selection to refine the dataset by correcting for feature relevance and multicollinearity. To avoid over- or under-fitting to the training set, the model uses a grid search cross validation to determine a value for k (number of features used):



Our most positive features were consistent with earlier exploration: vertical drop was best, followed by snowmaking area. To further assess the accuracy of our model, we ran our dataset through a random forest regression as well, which again aligned with earlier exploration and our linear model: the best features were: number of fast quads, number of runs, snowmaking area, and vertical drop. Comparison of our linear and random forest models actually revealed that the latter produced a lower MAE and exhibited less variability. Assessing this model in relation to dataset size also confirmed that the size of the given dataset was sufficient and increasing our sample would not significantly improve model accuracy.

Applying the final model to the data given for Big Mountain revealed that, even taking MAE into account, the ticket price is currently undervalued according to the given market data. We evaluated the accuracy of this output by assessing all the features we determined earlier (in exploration/preprocessing) were most relevant to pricing:



Evaluated against (not pictured): vertical drop, snowmaking area, number of chairs/runs/trams/fast quads, longest run, skiable area.

Big Mountain's values for these features validated its relatively high (especially high for its state) ticket price and led us to move forward with the current model.

We were asked to compare a few options for the resort, i.e. closing runs, extending runs, and/or increasing snowmaking coverage, against estimated values for ticket sales. From the options given, our model found that either the first or second solution would be best, though the first must be implemented incrementally to assess for changes in visitor numbers with each implementation, and the second would benefit from being evaluated alongside concurrent operating costs, since in the absence of any data on operating costs, these solutions were only evaluated in terms of revenue, not profit.

It is relevant to note that the model could be improved if we had information on visitor numbers (not clear on how/whether pricing affected or was affected by visitor numbers), operating costs (same reason), and additional pricing data (not clear on how/whether weekly/season passes are accounted for in the provided price data). In the absence of this information the current model makes several assumptions, especially on the accuracy (over- or under-pricing) of other resorts' ticket prices, and also imputes some missing information.