The purpose of this project is to come up with a pricing model for ski resort tickets in our market segment. The data are for resorts all belonging to the same market share. This suggests one might expect prices to be similar among them. There are two types of ticket prices in the data set: Ault weekday and adult weekend. We created a model for ski resort ticket prices to gain some insights into what price Big Mountain's facilities might actually support as well as explore the sensitivity of changes to various resort parameters. Big Mountain is doing well for the vertical drop, but there are still quite a few resorts with a greater drop. Big Mountain has the highest number of total chairs. Big Mountain compares well for the number of runs. There are some resorts with more, but not many. The vast majority of resorts, such as Big Mountain, have no trams. Big Mountain is amongst the resorts with the largest amount of skiable terrain. The expected number of visitors over the season is 350,000; on average, visitors ski for five days. We use our model to gain insight into Big Mountain's ideal ticket price and how that might change under various scenarios.

Scenario 1: Close up to 10 of the least used runs. The number of runs is the only parameter varying. The model says 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 the ticket price. Increasing the closures down to 6 or more leads to a large drop.

Scenario 2: Big Mountain is adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift. This scenario increases support for ticket price by $8.61. Over the season, this could be expected to amount to $15065471

Scenario 3: In this scenario, we are repeating scenario 2 but adding 2 acres of snowmaking. This scenario increases support for ticket price by $9.90. Over the season, this could be expected to amount to $17322717. Such a small increase in the snow-making area makes no difference.

Scenario 4: This scenario calls for increasing the longest run by .2 miles and guaranteeing its snow coverage by adding 4 acres of snow-making capability. No difference whatsoever. Although the longest run feature was used in the linear model, the random forest model (the one we chose because of its better performance) only has the longest runway down in the feature importance list.

Data Wrangling

There were missing values in the ticket price(15-16% missing values) About 14% of the rows have no price data, and we removed them. Weekend prices have the least missing values of the two, so we drop the weekday prices and then keep just the rows that have weekend prices. Silverton Mountain in Colorado has an incredibly large skiable terrain area. We replace the suspect value (26819) with the correct one (1819). We had a resort that has been open for 2019 years. It likely means the resort opened in 2019. We did not have any ticket pricing information at all for this resort, so we dropped the entire row. We Dropped the fastEight column in its entirety; half the values are missing, and all but the others are the value zero. There were some skewed distributions including fastQuads, fastSixes, and trams.

Exploratory Data Analysis

The average ticket price varies from state to state. New York accounts for the majority of resorts. Our target resort is in Montana comes in at 13th place. Some States show a marked difference between weekday and weekend ticket prices, so we let the model take into account not just State but also weekend vs. weekday. There are big states which are not necessarily the most populous. The states with the most total days skiing per season are not necessarily those with the most resorts. New York had the

most resorts but wasn't in the top five largest states, so the reason for it having the most resorts can't be simply having lots of space for them.

Summit and base elevation is quite highly correlated. If you increase the number of resorts in a state, the share of all the other state features will drop for each. There is some positive correlation between the ratio of night skiing areas with the number of resorts per capita. In other words, it seems that more night skiing is provided when resorts are more densely located with population. fastQuads, Runs, and Snow Making_ac are correlated with AdultWeekend ticket price.
resort_night_skiing_state_ratio seems the most correlated with the ticket price. As well as Runs, total_chairs is quite well correlated with the ticket price. the more runs, the more chairs you'd need to ferry people to them. They may count for more than the total skiable terrain area. For sure, the total skiable terrain area is not as useful as the area with snowmaking. People seem to put more value in guaranteed snow cover rather than more variable terrain areas.

Pre-processing

We sum up the features that we found interesting, including TerrainParks, SkiableTerrain_ac, daysOpenLastYear, and NightSkiing_ac. Weekend prices being higher than weekday prices seem restricted to sub $100 resorts. The distribution for weekday and weekend prices in Montana seemed equal. We used Principle component analysis(PCA) to find linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance they explain. After understanding what share of states' skiing "assets" is accounted for by each resort. We used SelectKBest in sklearn to select the k best features. We found the following "state resort competition" features: (a) the ratio of resort skiable area to total state skiable area; (b) the ratio of resort days open to total state days open; (c) the ratio of resort terrain park count to total state terrain park count; (d)ratio of resort night skiing area to total state night skiing area.
You impute missing values using scikit-learn. We created a pipeline that imputes missing values, scales the data, selects the k best features, and trains a linear regression model. We used cross-validation for multiple values of k and used cross-validation to pick the value of k that gives the best performance.

Modeling

We built a Random Forest Model with and without feature scaling and tried both the mean and median as strategies for imputing missing values. We train a model to predict Big Mountain's ticket price based on data from *all the other* resorts! We don't want Big Mountain's current price to bias this. We want to calculate a price based only on its competitors. Big Mountain Resort's modeled price is $95.87, the actual price is $81.00. Even with the expected mean absolute error of $10.39, this suggests there is room for an 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.

This result should be looked at optimistically and doubtfully. The validity of our model lies in the assumption that other resorts accurately set their prices according to what the market (the ticket-buying public) supports. The fact that our resort seems to be charging that much less than what's predicted suggests our resort might be undercharging. It's reasonable to expect that some resorts will be "overpriced" and some "underpriced." Or if resorts are pretty good at pricing strategies, it could be that our model is simply lacking some key data. Certainly we know nothing about operating costs, for example, and they would surely help.