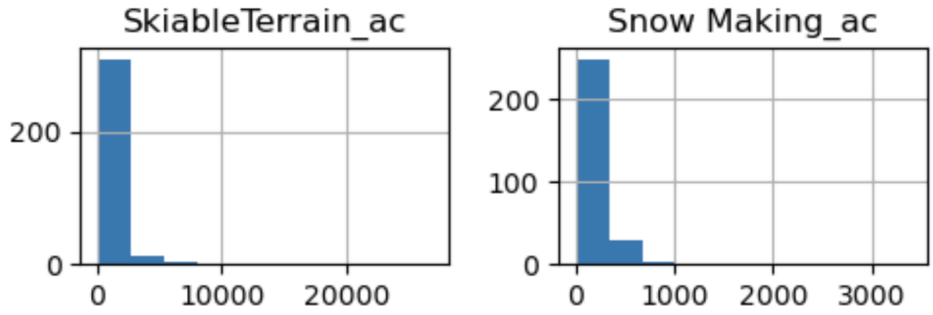
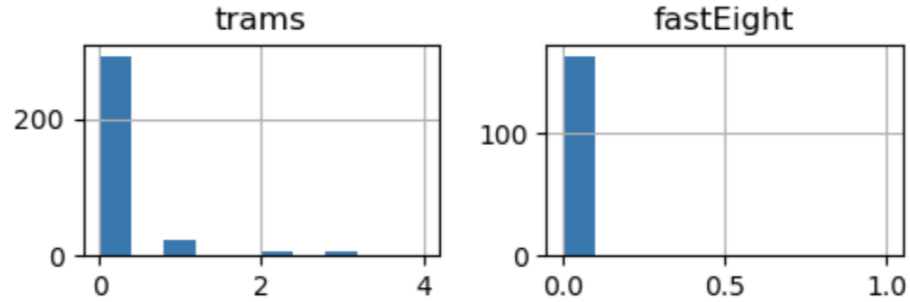
Big Mountain Resort is a ski resort in Montana. This resort has a wide variety of attractions, trails to accommodate for any skier and rider level, and approximately 350,000 visitors each year. Big Mountain Resort recently installed a new chair lift that increased their operational costs by $1,540,000 this year and wants to increase ticket prices in order to make a profit. The current ticket price is based on a market average, but could be higher if the price was based on examining key resort features such as amount of runs, vertical drop, number of chairs, longest run, and weekend vs weekday prices. So then the big question was: what opportunities exist for Big Mountain Resort to increase revenue by $1.54M through a new ticket price that considers amount of runs, vertical drop, number of chairs, longest run, and an increased weekend price?

In order to begin to answer this question, the first step was to clean the data provided. While cleaning the data it appeared that there were two entries for 'Crystal Mountain'. In order to confirm or deny that these were duplicates, I took into account state and region. By doing so, I concluded that the two entries were not duplicates because one of the resorts was in Michigan and the other was in Washington. When examining the distributions of feature values trams, fastEight, fastSixes, SkiableTerrain_ac, Snow Making_ac, and yearsOpen stood out. SkiableTerrain_ac and Snow Making_ac stood out for their clustered distribution on a low end. This clustered distribution implies that most resorts had the very similar, low values for skiable terrain and snow making machines.



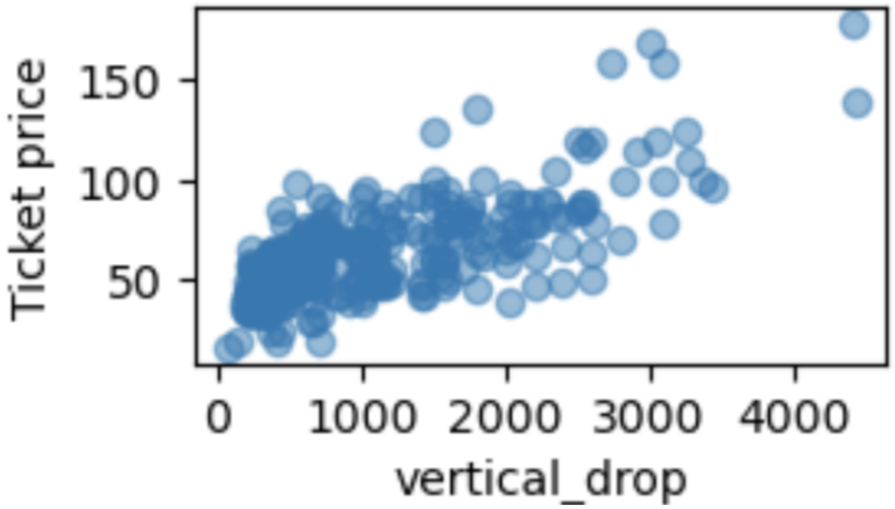
The fastEight, fastSixes, and trams distributions all stood out due to their little to no variability.



The yearsOpen and Skiable Terrain distributions also stood out due to extreme outliers, but they were corrected. Another big problem was missing values. The ski data dataframe started out with 330 rows and 27 columns. Our resort of interest, 'Big Mountain Resort' was present in the dataframe and had no missing values. There were a couple of columns with missing information. Some of these columns had to be removed due to their significant amount of null values. The

‘fastEight’column, for example, had approximately 50.3% missing entries. The ‘AdultWeekday’ column was also removed as it had approximately 16.4% missing values. Rows (resorts) with no price data were also dropped. Price is the target of this analysis, so rows without this information are not useful. There are now 277 rows and 25 columns left

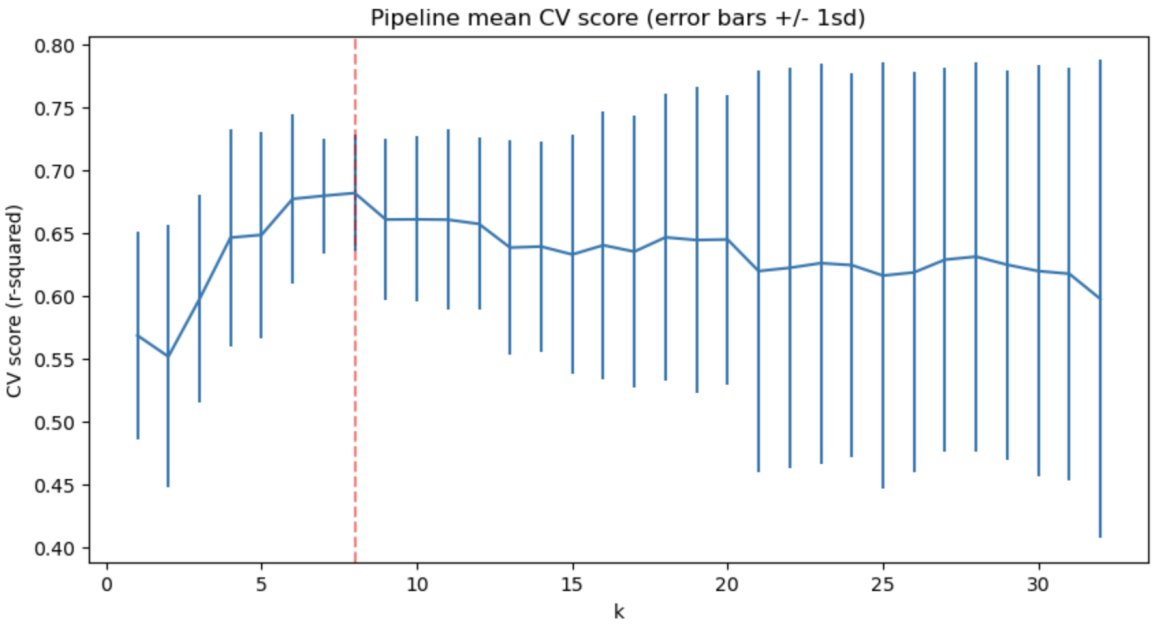
After data cleaning was complete, exploratory data analysis was done to see whether or not state data should be included in the ticket model. There were no clear patterns that suggested that states should be treated differently than one another, so states will not be included in the ticket model. It is important to note that there was a strong correlation between ticket price and vertical drop, which suggests that vertical drop should be included in the final ticket model.



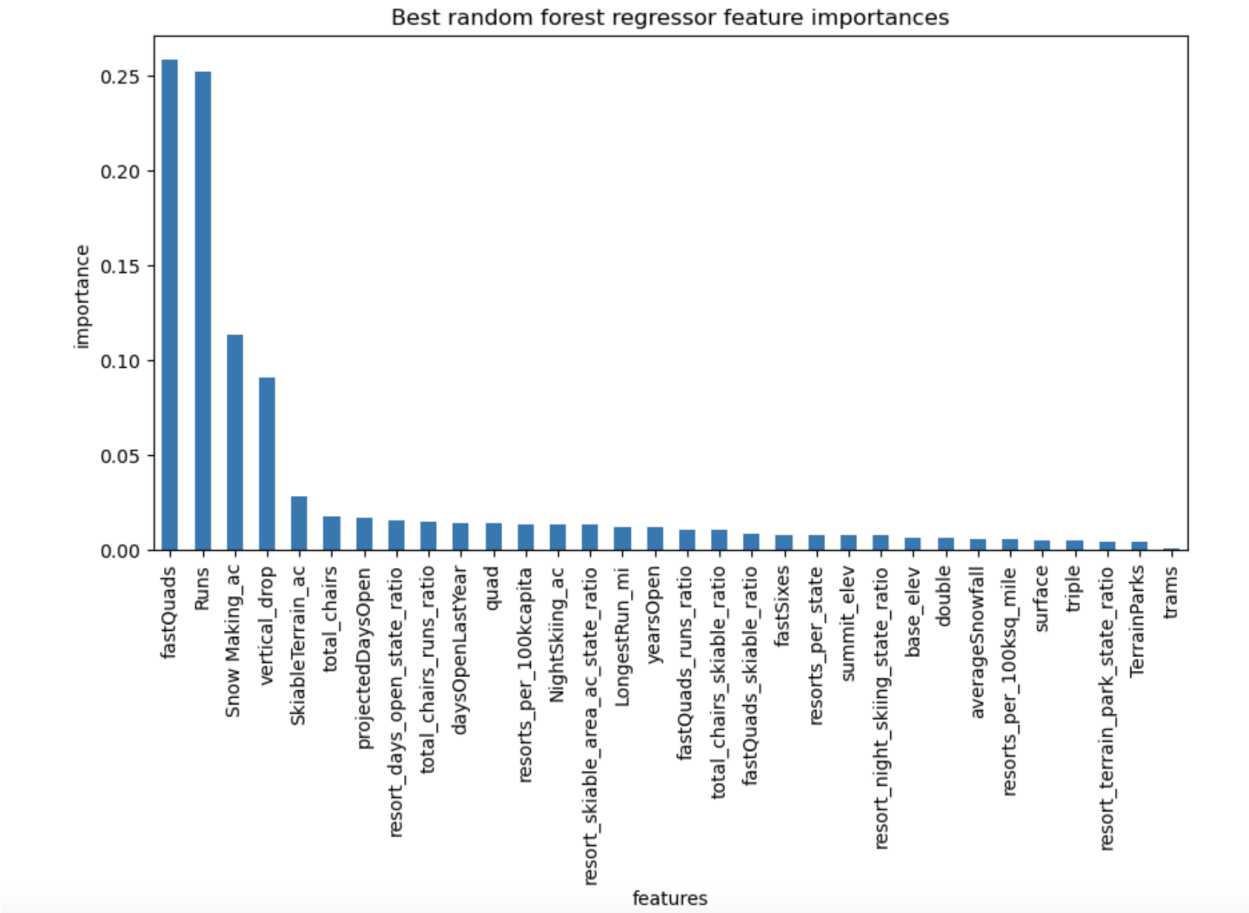
- Model Preprocessing with feature engineering

After the exploratory data analysis, preprocessing and training for the data began. We needed to know what to do with the missing values. There were a couple of options: guess, fill in with median, or fill in with mean. To figure which of the options was best, the data was first divided into two parts: 70% training and 30% for testing). Using a dummy regressor, missing values were filled using guesses based on averages. This model resulted in a mean absolute error of approximately $19.14, which means that, on average, it is expected that actual ticket prices are approximately 19.14 dollars away from these predicted ticket prices. The mean absolute error was approximately $9.41 for both the mean and median methods for missing values. Despite them being essentially the same, they were both still significantly better than just guessing using the average.

This simple linear regression model accounted for all of the remaining features of the resorts, but it doesn’t make sense to use all, since some features had no variability. In order to refine the linear regression model, a technique called cross validation was used. When using cross validation with CV=5 and default k, the mean r-squared was found to be ~0.633 +/- 0.095, which is consistent with the previous models. The cross validation model expects 95% of the r-squares are expected to be between 0.44 and 0.82. Further analysis showed that using a k=8 would ensure a higher r-squared with an even smaller error.



A random forest regressor was also tried. It found that the best parameter was median and that scaling the data did not help. The random forest regression model also showed that fastQuads, runs, snow making ac, and vertical drop had the biggest impact on ticket prices.



The random forest regressor model had a mean absolute error of 9.53, which was about $1 less than the linear regression mean absolute error. Due to its lower cross validation mean and smaller variability and consistent test results with the cross validation data, the random forest regression model was chosen.

In conclusion, Big Mountain should increase their ticket price to $85.48. Big Mountain currently charges $81 for their ticket price. The modeled ticket price is $95.87 +/- $10.39. This suggests that there is room for error because the lowest predicted price is $85.48, which is higher than their current price. This price was supported in the marketplace by Big Mountain's current facilities. A bigger profit could be made by adjusting some of the other features. If each visitor purchases 5 day tickets on average, a new chair lift is installed, the vertical drop increases by 150 ft, and one new run is opened, tickets could increase by $1.99. Another suggestion could be to close 5 runs. The model below suggested that closing one run would not make a difference in ticket price or revenue. Closing 2 runs makes a small difference. Closing 3, 4, or 5 runs makes the same amount of difference in both ticket price and revenue. Closing anything more than that would make a significant change in ticket price and revenue, so that wouldn't be a good idea. Closing 5 runs may optimize operational costs and revenue, so it would be something to consider. Further analysis could be done if we knew some of the operational costs for other features, such as runs or snowmaking costs.