Guided Capstone Project Report

**Introduction:**

Big Mountain needs a new predictive model for ticket price based on the number of facilities and properties of their competitors in the market. The project provides guidance for Big Mountain’s ticket pricing and investment plans for other facilities.

**Data Wrangling**

Original number of rows in the data is 330, and the number of columns is 27. Firstly, the distribution of each variable was examined to detect any anomaly in the data. By checking the information on the web, the value of "SkiableTerrain\_ac" for "Silverton Mountain" which created a huge outlier for that variable has been fixed. The column "fastEight'' has been removed since half of the values are missing and most of them are 0s. One entry had the value of "2019" for "yearsOpen" which must have been a mistake and it has been removed as well under the assumption that such information about a new resort cannot be determined whether it is actual or mere projections. This left the data set with 328 rows and 26 columns.

Then, the data was grouped by the state to examine the state-wide summary statistics. Only the relevant columns, such as "TerrainParks," SkiableTerrain\_ac," "daysOpenLastYear," and "NightSkiing\_ac'' were selected to look at the sum of those for each state. The rows with missing values of both AdultWeekend and Adultweekday prices were dropped, and after examining the number of missing values, **AdultWeekend** has been selected as it had less missing values. The data ended up to have 277 rows and 25 columns, and they are in *Table 1*.

**Exploratory Data Analysis**

. After looking at some general trends for different states through state summary data, principal component analysis (PCA) was conducted. First of all, the data was scaled and made sure that the mean of each column is near 0 and standard deviation is 1. Then, the scaled data was transformed and plotted. Seeing that the first and second component of the PCA explains the 77.2% of the variance, the first two derived features were plotted and each point with state names were labeled (*Figure 1*). Then the average price of tickets for each state was calculated and appended to the PCA data frame to further examine its relationship to the ticket price for each state. A "Quartile" was created, which is a categorical variable that categorizes the values into different categories of values to plot with better details. Through the analysis of state level data, it was concluded that all states can be treated equally.

The data in individual resort level was examined by appending state level columns into the individual resort level data. Looking at correlation heatmap and scatter plots, we tried to find some significant correlations that are noteworthy (Figure 2). There seems to be some relationship between state and ticket prices on features such as vertical\_drop, fastQuads, runs, total\_chairs, and resorts\_per\_100kcapita. Number of chairs that the resort can accommodate determines the price of the ticket - the more chairs, the less resorts can charge for each entry.

**Preprocessing and Training**

The machine learning model is built, first by splitting the data into training splits (30%) and testing splits (70%). The missing values are imputed with the mean and the data is scaled. The linear regression is conducted and different k values were selected. Performance metrics such as mean absolute error, r squared, and mean squared error, are evaluated. Cross validation was performed to prevent overfitting. Hyperparameter was tuned using GridSearchCV. The best k that outputs the best CV score was 8 (Figure 3). The coefficient for the linear regression model (Table 2) suggests that vertical\_drop seems to have the greatest influence on the price.

Another machine learning model, random forest regressor model was performed and compared to the linear regression model. The CV score has been improved from the linear regression model. The feature importance (Figure 4) was plotted. The dominant top four features based on linear model and random forest regressor model are: fastQuads, Runs, Snow Making\_ac, and vertial\_drop.

Based on comparison of two models by calculating the mean absolute error using cross validation, the random forest model has lower cross-validation mean absolute error by almost $1, showing less variability.

**Modeling**

The model suggests that Big Mountain Resort should be charging $95.87, where the actual price that the resort is charging is $81.00. The expected mean absolute error is $10.39. Even with the expected mean absolute error, there is some room to increase the ticket price.

Different scenarios of either cutting costs or increasing revenues through increasing ticket price were tested out. Closing 1 run does not make any change in revenue. Closing 2 and 3 runs successively reduces the revenue, however, closing 3, 4, and 5 runs do not make any difference in ticket price, and therefore the revenue stays the same, while closing further runs leads to larger drop in ticket price and revenue.

Increasing the vertical drop by 150 feet and installing additional chair lift increases support for ticket price by $1.99. And this effect is the same as adding 2 acres of snow making. Increasing the longest run by 0.2 miles and adding 4 acres of snow making capability does not have any effect in model predicted ticket price. Therefore, the business should consider all these factors, and make decisions accordingly.

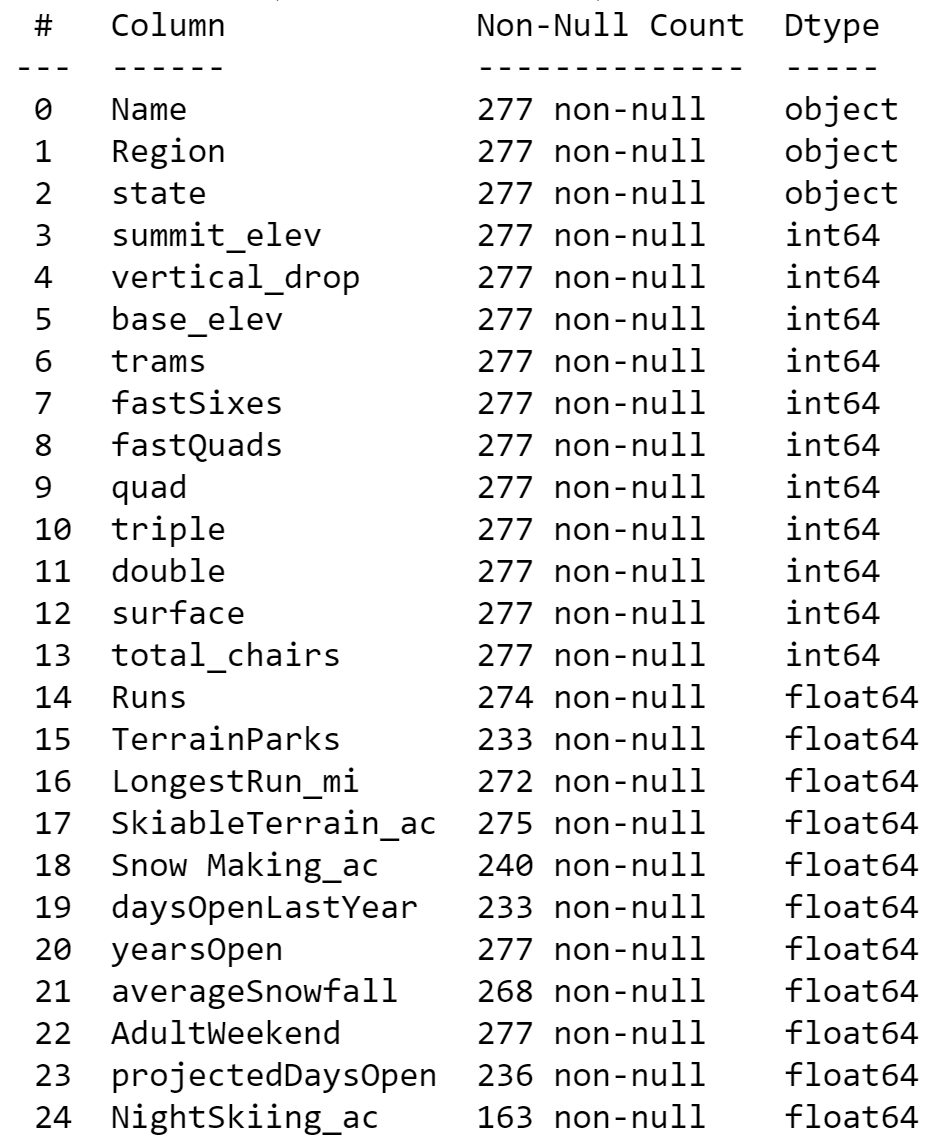


Table 1. Features after data cleaning

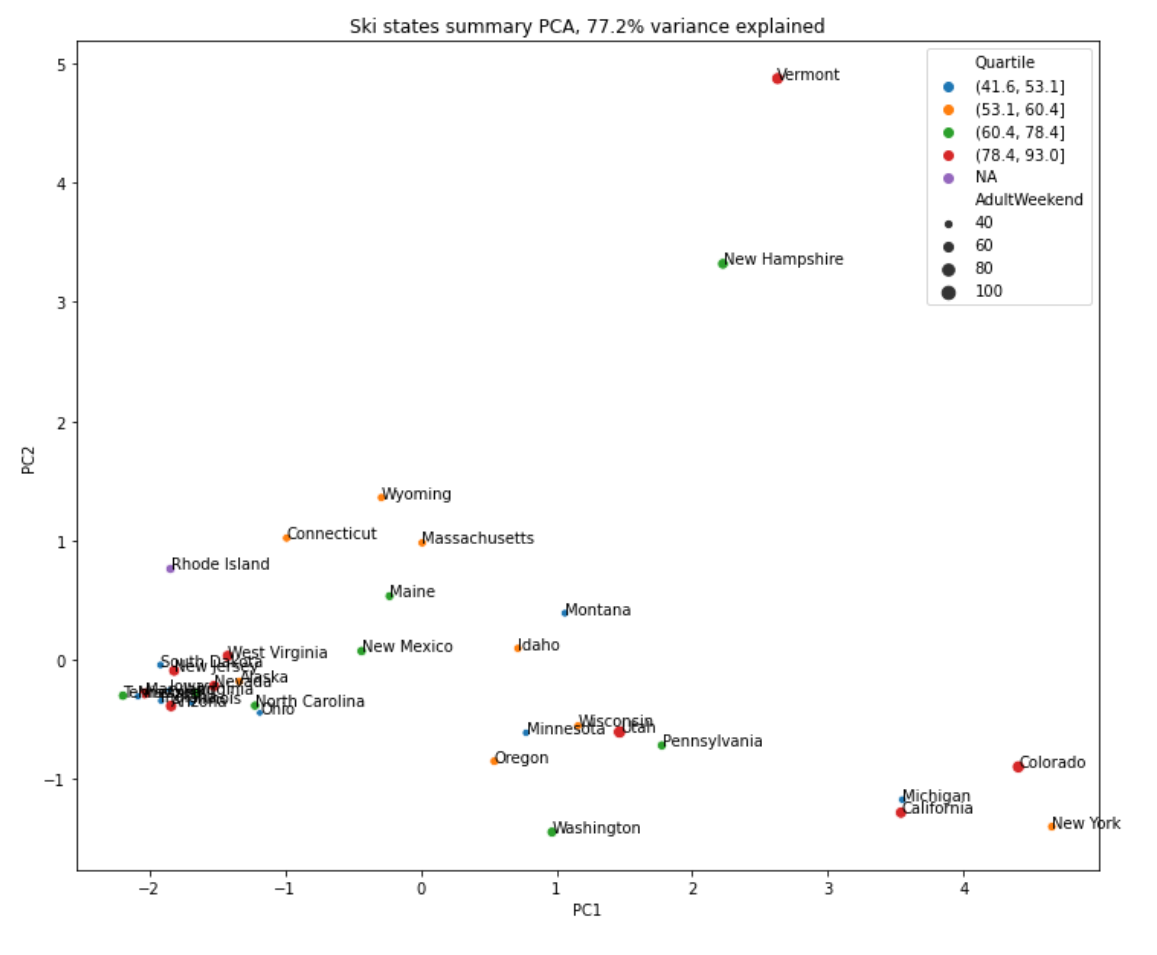
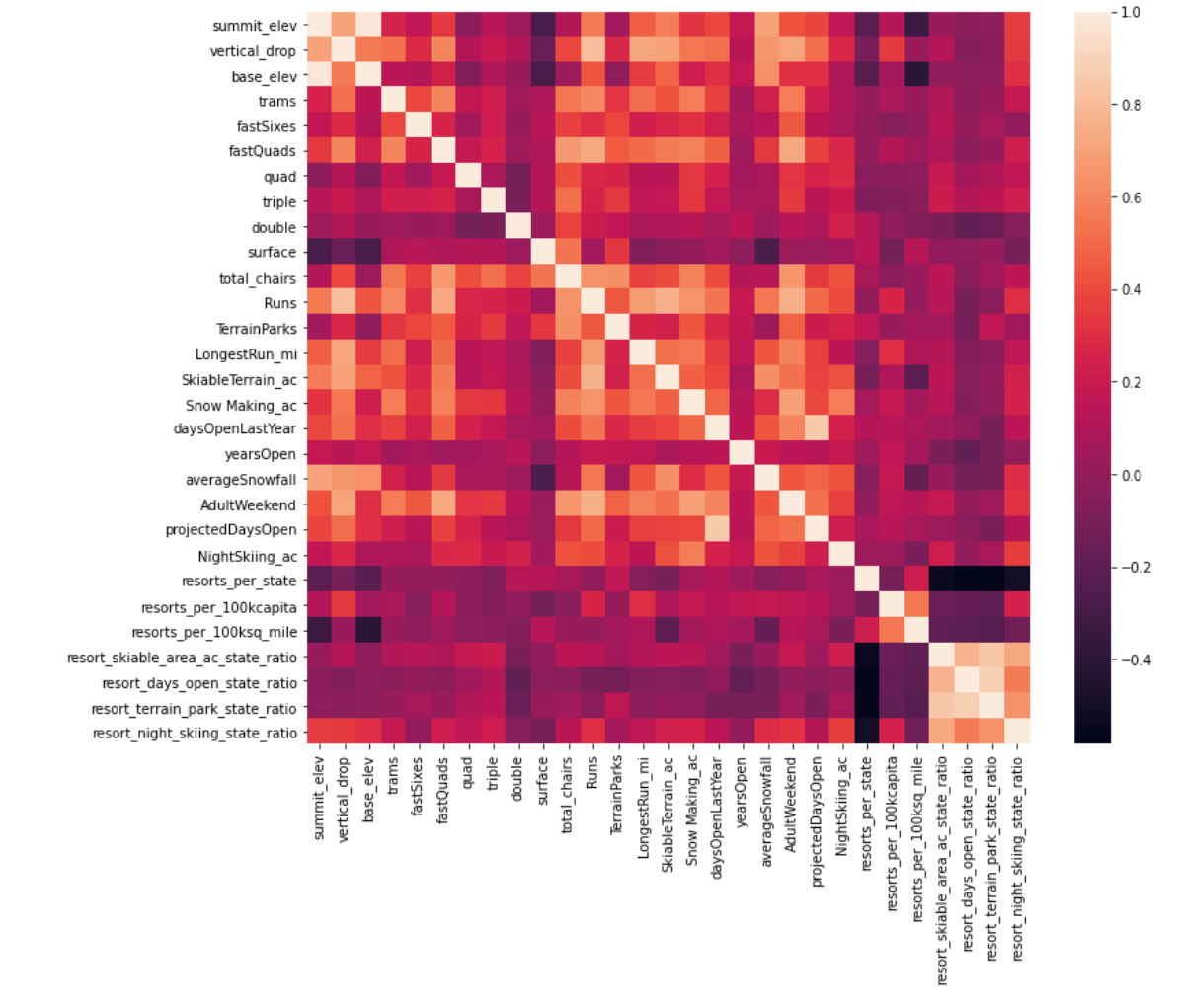
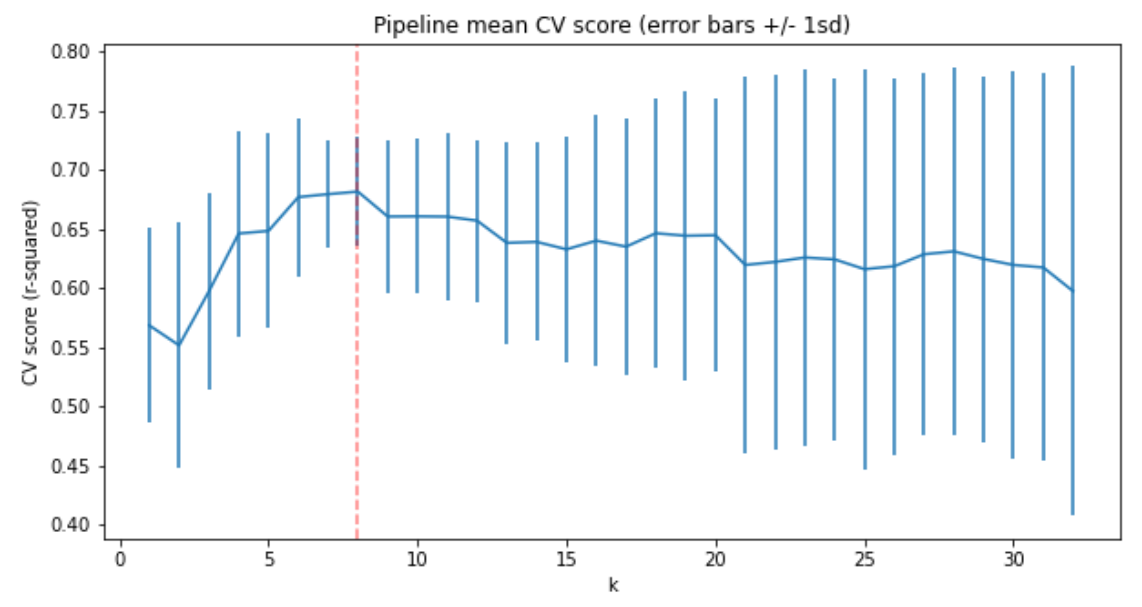


Figure 1. Ski states summary PCA, 77.2% variance explained.

Figure 2. Heatmap of correlations

Figure 3. Linear regression mean CV score for each k value

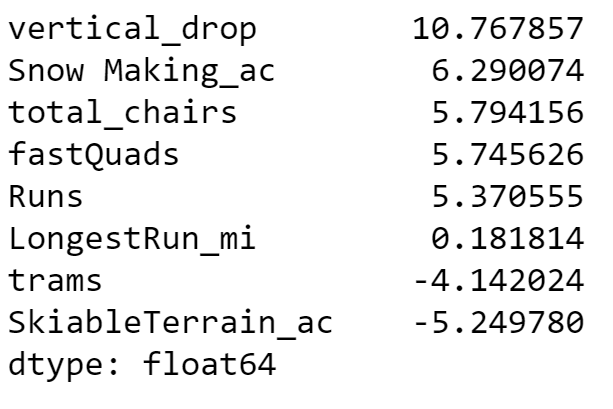


Table 2. Linear Model Coefficients

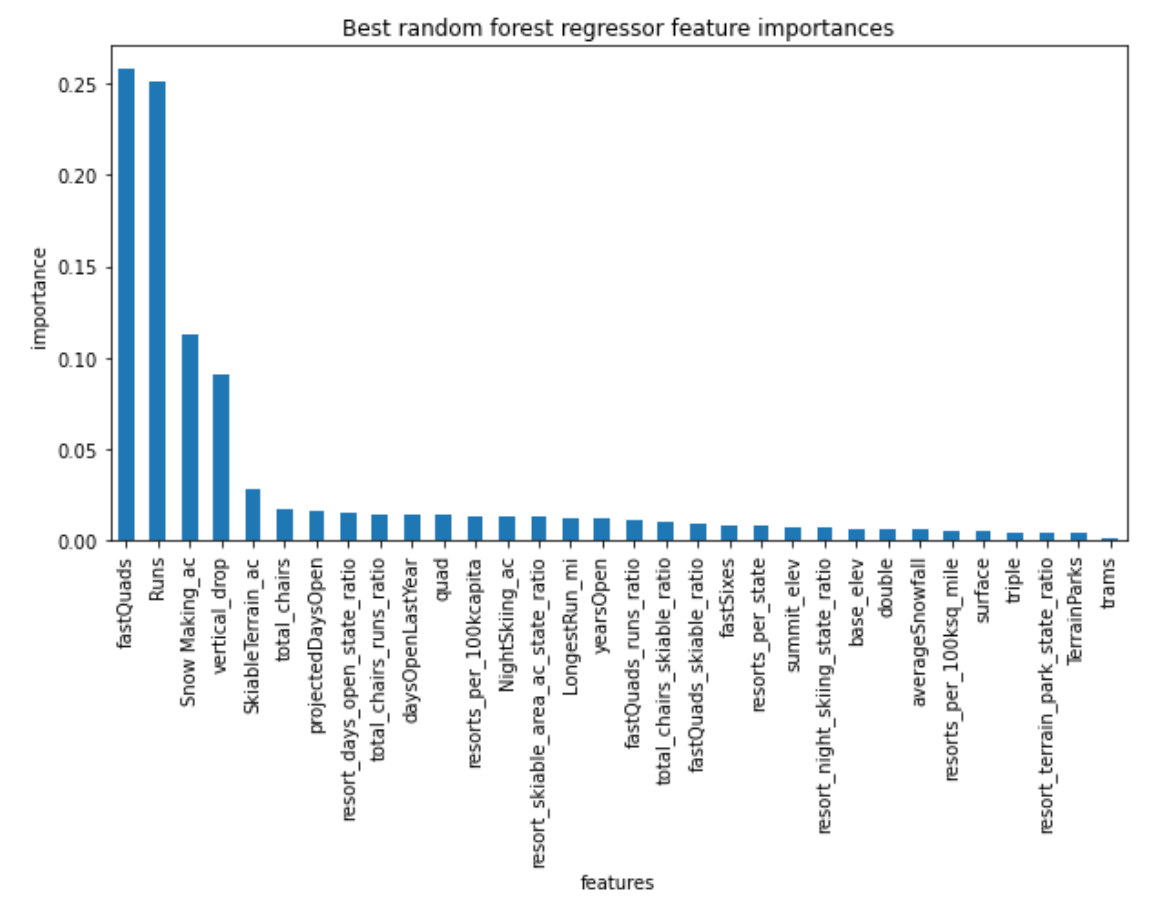


Figure 4. Random forest regressor feature importance

|  |  |  |
| --- | --- | --- |
|  | Linear Model | Random Forest Regression Model |
| Mean of Mean Absolute Error | 10.499 | 9.645 |
| Standard Deviation of Mean Absolute Error | 1.622 | 1.353 |
| Mean Absolute Error | 11.793 | 9.538 |

Table 3. Comparison of model performance through cross validation mean absolute error