Elinor Velasquez
November 29, 2020

Guided Capstone Project Report

The Big Mountain Ski Resort in Montana sells tickets for $81.00. The goal of this report was to compute whether or not the resort was under-charging (or over-charging) skiers. Big Mountain Resort in Montana has a vertical drop of 2353 feet and a summit elevation of 6817 feet, a total of 14 chairs and 105 runs. It has 3000 acres of skiable terrain and 600 acres of snow making, with an average snowfall of 333 inches. It has been open for 72 years and was open 123 days last year. Montana has 12 resorts, while some of the more populous states such as California have 21 resorts.

The average ticket price in Montana is $52, while states such as California have $81 as their average ticket price. The distribution of skiable area in acres for all states is heavily skewed to the right. The feature fastEights has half of its entries missing and 163 resorts have zero fastEights. So, the feature 'fastEights' was dropped from the data set. One resort had as years open 2019 years. This resort was dropped from the data set because this value appeared erroneous. Resorts with no information for ticket price were also dropped. Many of the histograms for the distribution of the features showed a skewed distribution, while a few features appeared as normally distributed. Ticket prices for weekday and weekend are mostly similar, although a significant portion showed higher prices for the weekend ticket. Four principal component analysis (P.C.A.) components were seen to be significant in capturing the variance. Two P.C.A. components were seen to capture 77% of the variance. A heat map of correlations was constructed for all the features. Adult weekend ticket price showed a correlation between fastQuads, Runs, and Snow Making (acres), Total Chairs, and Resort Night Skiing Per State (Ratio) , and Vertical Drop in the heatmap (Figure 1). Ticket price could be correlated with the number of resorts because the number of resorts serving a popular area allows the ticket price to increase due to demand. From scatterplots, there seemed to be a linear relationship between Ticket Prices and Night Skiing, Days Open Last Year, Projected Days Open, Skiable Terrain,

Resort Night Skiing Per State (Ratio) and Total Chairs. The target feature was Weekend Ticket Price.

A Dummy Regressor modeled the Mean as the single feature to predict Weekend Ticket Price. The mean was $64, which was less than Big Mountain's Weekend Ticket Price of $81. The baseline model, which predicts the ticket price to be the Mean, gives $R^2 = 0$. Models that are worse than the Mean give negative R^2 and models that are better belong to the interval [0, 1]. Testing the model of the Mean to predict Ticket Price gave a negative value when tried on test data, -0.003 (training data gave 0.0). The mean absolute error gave $18 for training data and $19 for test data, meaning if ticket price is based on the mean, then the variation in price is $19. The root mean square error was $25 for training data, and $24 for test data. The missing values with the training data were replaced by medians (medians because the histograms of most features were skewed).

The ticket price was then predicted with a linear regression model. The R^2 score, which assessed model performance was 0.82 for variance between the training data for ticket price and the predicted value for training data, and 0.72 for the test data and the predicted value for the test data. The Median-Mean Absolute Error was 8.5 and 9.4, respectively.

The make_pipeline function used the SimpleImputer (which replaced missing values for each feature with the median of that feature's values), the StandardScaler which normalized the data (mean equals zero, and the standard deviation equals one), and the LinearRegression, which computed a linear model with the training data. The make_pipeline function had two methods, "fit" and "predict," used to fit a model to the training data, and predict ticket price based on the model. Use of SelectKBest selected the k best features for training the model. The obvious question posed was what was the best number of features for the model? The best features to use could be found by inspecting the heatmap. (Figure 1) The Linear Regression R^2 score for the training data was 0.77, and for the test data was 0.63. Cross validation with the number of folds equalling 5 and 15 as the best number of features to use gave a 95% confidence interval of [0.44, 0.82] for R^2. Cross-validation was then used to choose the value of k that gave the best features for the best performance of the model: To find the k best features to achieve the model's best performance, the function GridSearchCV was used with the make_pipeline function. The graph (Figure 3) showed that k=8 gave the optimized value for achieving the best model performance. The eight best features were Vertical Drop, Snow Making (acres), Total Chairs, Fast Quads,

Runs, Longest Run, Trams (negative coefficient) and Skiable Terrain (negative coefficient). The negative coefficient implied that increasing Trams and Skiable Terrain caused negative effects on the resort, since keeping the number of chairs constant caused the chairs to be heavily used (not optimized).

Use of Random Forest Regression instead of Linear Regression improved the model. The graph (Figure 2) showed the best features to use with Random Forest Regression. Mean Absolute Error for the Linear Regression was $11.8. Mean Absolute Error for the Random Forest was $9.5. The partition size with a 70/30 train/test split were 193 resorts for the training data and 83 resorts for the test data. The total number of features was 35. For the Big Mountain Resort data, the Random Forest model predicted $95.87 for a Weekend Ticket. The actual price was $81. The mean absolute error was $10.39. The graph showing the Adult Weekend Ticket Price showed the variation of ticket prices was normal with a moderate skew to the right. Closing 4 - 5 runs would have the same effect as closing 3 runs. For another scenario, Big Mountain Resort added a run, increased the vertical drop by 150 feet and installed another chair lift. This scenario suggested the ticket price be increased by $8.61. Over the season the total amount from this new revenue would be $15,065,471. Scenario 3: This scenario was the same as Scenario 2 with the addition of Snow Making (acres) equalling 2. The ticket price increase was $9.90 and the total amount from this new revenue is $17,322,717. Scenario 4: In this scenario, 0.2 miles were added to the Longest Run (miles) and 4 to Snow Making (acres). The increase in ticket price was $0.0 and the total amount from this new revenue is $0.

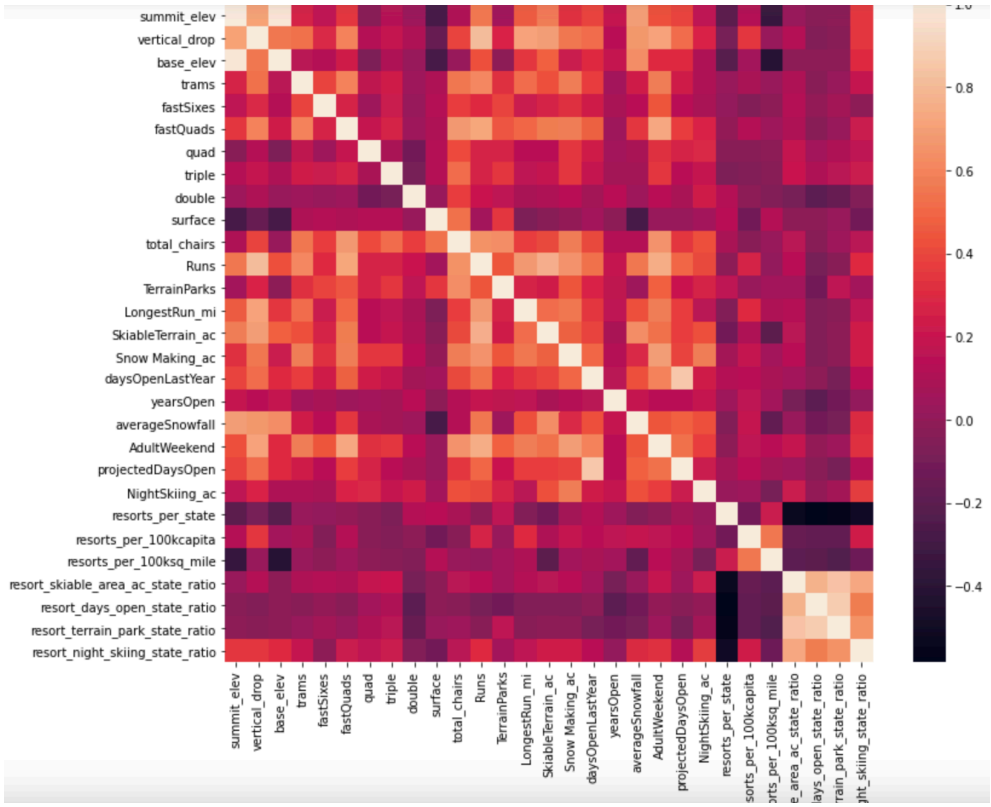

Figure 1. HeatMap

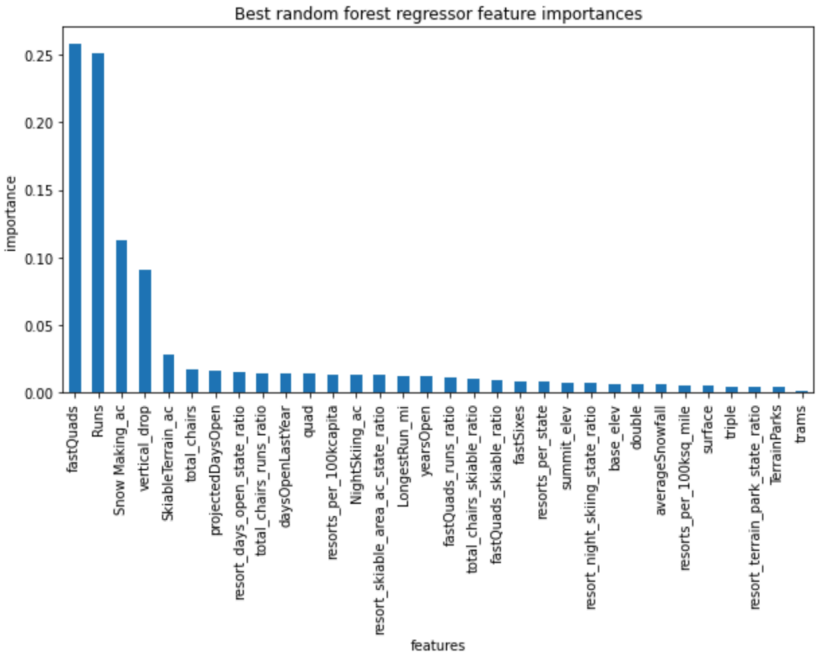

Figure 2. The Best Features to use with Random Forest Regression

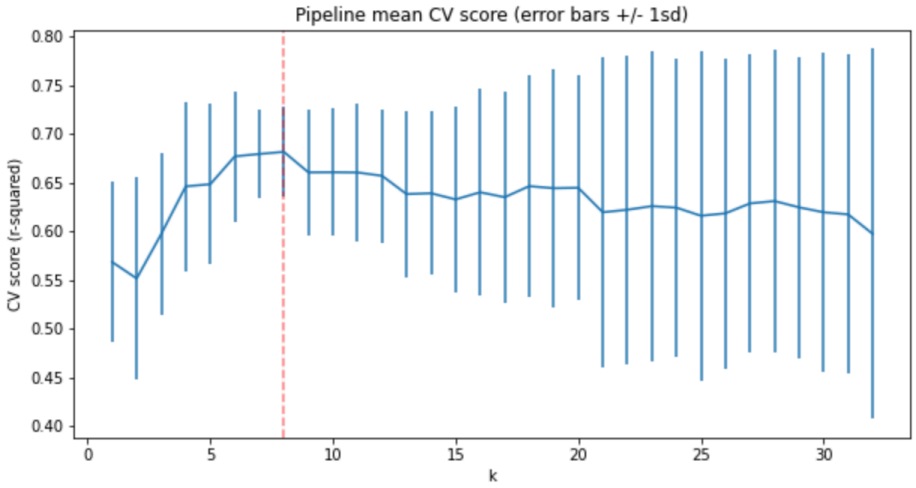

Figure 3. Graph of k Best Features