**Guided Capstone Project Report: Big Mountain Resort Business Problem**

**Data Science Problem**

Big Mountain Resort suspects it may not be maximizing its returns relative to its market position. It also does not have a strong sense of what facilities matter most to visitors, particularly which ones they're most likely to pay more. This project aims to build a predictive model for ticket prices based on resorts feature facilities. This model will be used to guide Big Mountain's pricing and future facility investment plans.

**Data Analysis, Modeling, and Results**

The ski resorts data was obtained from Springboard. We performed preliminary assessments of data quality and refined the question to address our business problem. We determined that predicting the adult weekend ticket price was our primary aim. The target feature to predict ticket price is the resorts features that matter most to visitors, and they are most likely to pay more for them. We assessed data quality by examining the raw data observations and features distributions to identify any issues such as outliers and missing feature values. We dropped records with missing price data. We look for any patterns between the states, but since we did not observe any apparent relationship between state and ticket price, we decided to treat all states the same when building the price modeling.

After cleaning the raw data and including states' population and area information as part of the ski cleaned data, we examined relations among features. We observed some positive and negative correlations between ticket prices and some features, and other less significant correlations among other features. We decided not to remove any features at this stage of the analysis.

Then, we employed machine learning to build ski price models. First, the data was split into training and testing sets. We then created a baseline model using the training set's average ticket price to predict ticket prices. This model's performance was evaluated on the test set by computing the R squared (R2), mean absolute error (MAE), and mean squared error values. The MAE result showed that, on average, we might expect to be off by $19.00 if we guessed ticket prices based on the average of the known values.

Next, we proceed to build a linear regression model and addressed the missing values imputing them with the median and mean values. The MAE from the two imputing cases were very similar. These results showed that the linear model could predict prices within $9.00 of the accurate average prices. The preliminary results of the linear model outperformed the baseline model without removing any features. Then we performed 5-fold cross-validation on features space exploration to identify the predictors that contribute the most to the ticket prices. The results showed that out of 32 features, 8 are the most significant contributors (vertical\_drop, snow making\_ac, total chairs, fastQuads, Runs, LongestRun\_mi, trams, and SkiableTerrain\_ac) to make price predictions in the linear model. We also observed that the performance results' variance increases for the number of features higher than 8 (Fig. 1).

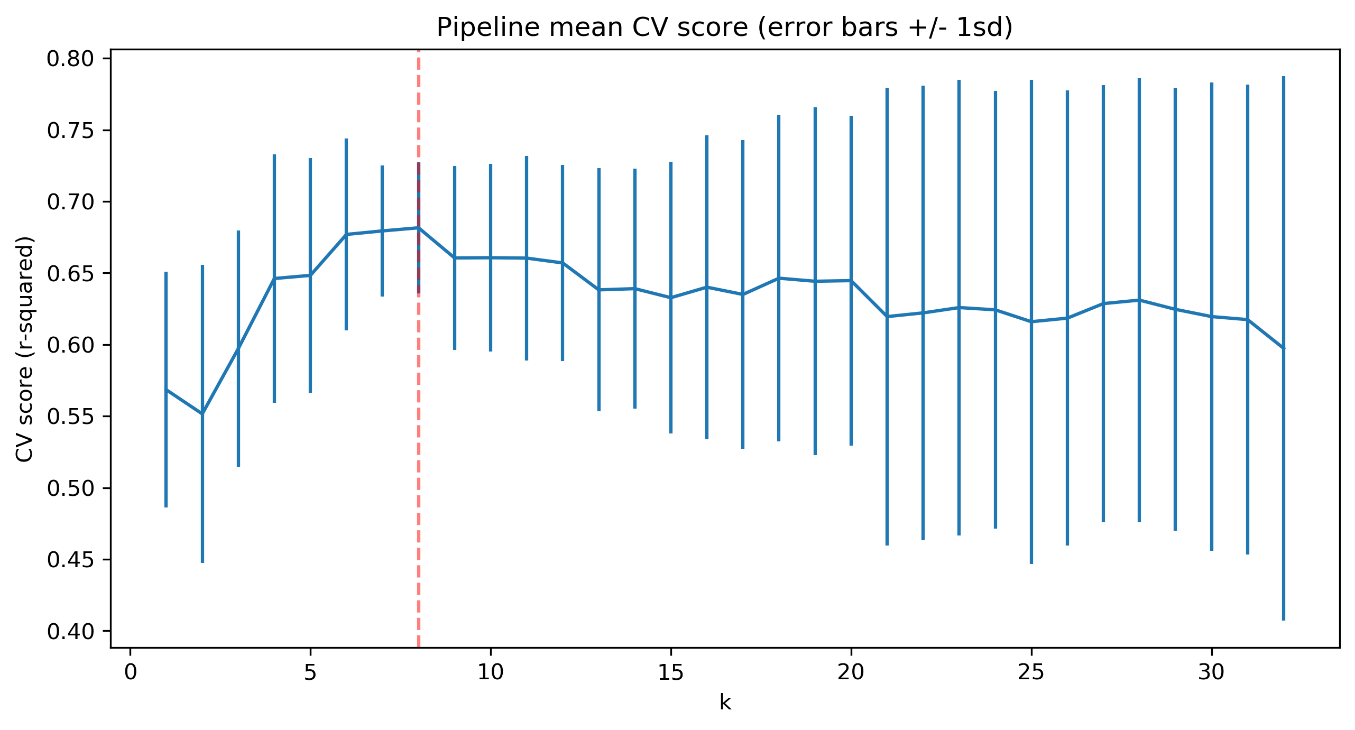


Figure : Linear model features space mean cross-validation (CV) score results shown in blue and the optimum number of feature k=8 shown in a vertical red dashed line. The k = 1-32 features are shown on the x-axis, and the r-square values are shown on the y-axis.

Lastly, we build a random forest model. The random forest model has several parameters to explore. We focused on with and without scaling, imputing missing values using the mean and median values, and employing 5-fold cross-validation to evaluate this performance on the training set. The results showed that a random forest model without scaling and using the median for imputing missing values gives the best performance. We also obtained the contribution scores of each of the 32 features to the model (Fig. 2). The top contributors to the selected random forest model are fastQuads, Runs, Snow Making\_ac, and vetical\_drop, which are similar top contributors in the linear model.

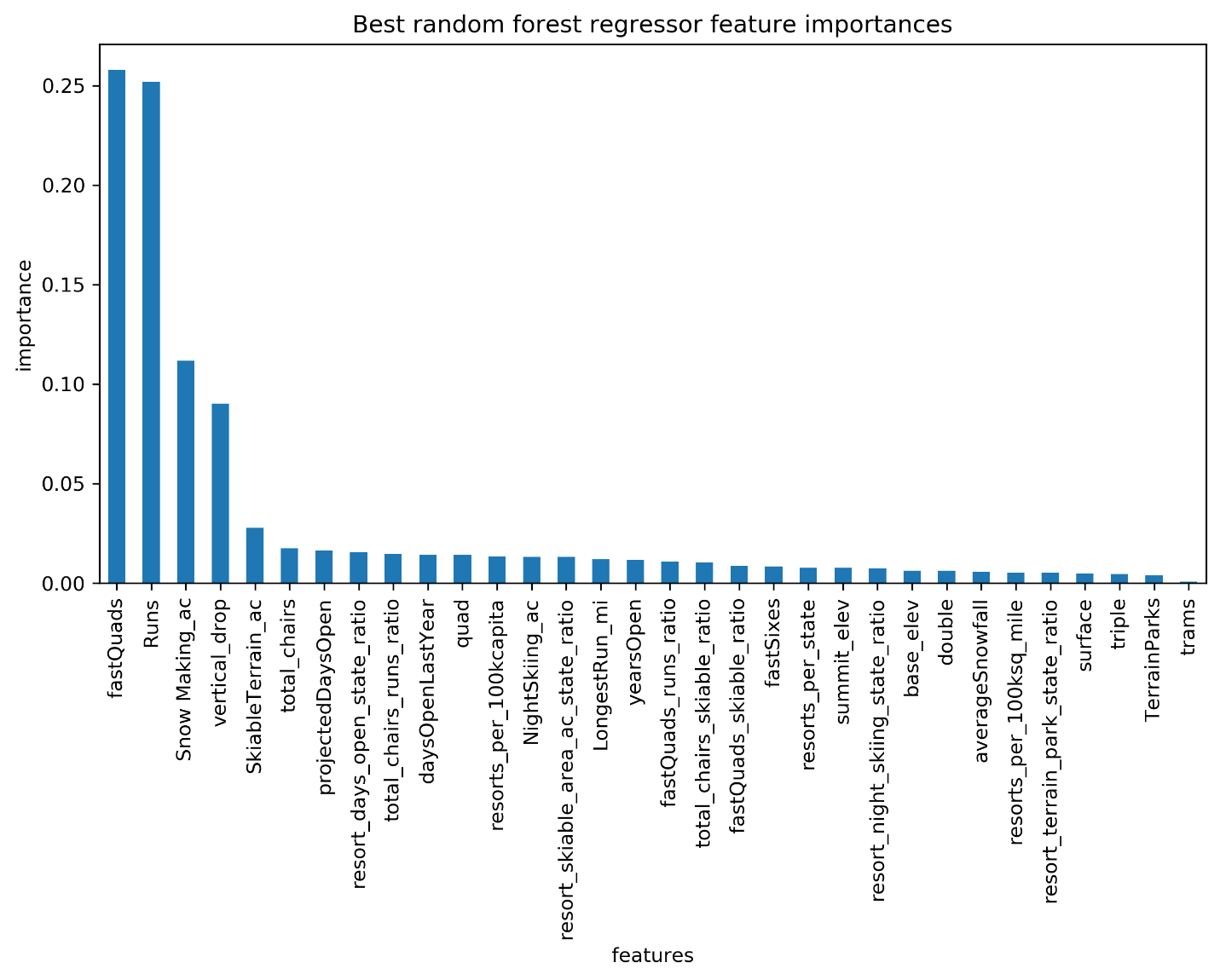


Figure : Random forest model feature importance distribution.

When comparing the linear model's cross-validation MAE values to the random forest model, we found that the random forest model outperforms the linear model by almost $1.00. It also exhibits less variability than the linear model.

Lastly, we applied the random forest model to predict Big Mountain resort prices. The random forest model prediction price for Big Mountain Resort is $94.22. With an expected mean absolute error of $10.39, this result gives Big Mountain Resort room to increase its current price of $81.00. Then, we explored four potential scenarios for Big Mountain resort for either cutting costs or increasing revenue from ticket prices. 1) Permanently closing down up to 10 of the least used runs. 2) Increase the vertical drop by adding a run to a point 150 feet lower down but requiring installing an additional chair lift to bring skiers back up, without additional snowmaking coverage. 3) Same as number 2, but adding 2 acres of snowmaking cover. 4) Increase the longest run by 0.2 miles to boast 3.5 miles length, requiring additional snowmaking coverage of 4 acres. Considering an expected number of visitors over the season of 350,000 who ski an overage of five days, our results from scenarios 1-4 suggested going with procedure 2 to set up a new ticket price. As we recall, by adding a chair lift, Big Mountain's increases their operating costs by $1,540,000. Therefore, by increasing the price by $1.99 provides an increase of $3,474,638 in revenue, which covers the installed chair's operational cost.

In summary, we have shown that Big Mountain Resort should not base their ticket pricing on just the market average. This approach does not provide the business with a good sense of how valued some feature facilities are than others. The linear model has shown eight features as the most significant contributors to ticket prices. In contrast, the random forest price model retaining all features provides a smaller MAE than the baseline and linear models. We selected the random forest as the best ticket price model and delivered it to the Big Mountain Resort business decision team to make future price explorations and adjustments.