**Guided Capstone Project Report**

**Introduction**

Big Mountain Resort (“BMR”) needs to determine if its current lift pricing strategy (given its perceived superior facilities due to both previous and recent $1.54 million in capital investments) is in line when compared to both regional and industry trends.  Analysis must be completed so any pricing changes can be reflected prior to upcoming seasonal market promotions. BMR would like to develop an algorithm that would quantitate the justification of making price moves instead of relying on gut feel.

**Data Wrangling**

Received a spreadsheet which listed 330 ski resorts from across the United States. Spreadsheet contained 27 columns. One column provided identity: Name. Two columns were categorical:Region and State. The remining 24 columns provided potential quantifiable dataConducted analysis on categorical variables and determine to utilize state as opposed to region (easier for state/state comparisons).

Reviewed numeric variables for completeness and reviewed distributions using histograms.

A picture containing crossword puzzle, window

Description automatically generated

Deleted columns for FastEight lifts due to lack of info. Deleted AdultWeekDay due to price similarity to AdultWeekend column. Deleted rows that lacked AdultWeekend values.

Chart, scatter chart

Description automatically generated

After deleting rows, 277 records remain. AdultWeekend will be the target feature.

**Exploratory Data Analysis**

After first reviewing summary statistics for: total state population and total state area, a clear pattern for state did not present itself. Created a couple of density metrics by state.

Combined the two list of variables into one dataframe (dropped total state population and Total state area). Scaled the column data using scikit-learn's ("SKL") scale function. Then performed Primary Component Analysis ("PCA") with SKL in order how much of the cumulative variance ratio could be ascribed to each of the seven variables. the first "elbow" appeared at the second element where 77.2% of the cumulative can be explained. Created a dataframe for these two PCA variables(PC1, PC2) by state and ticket price. Using a Seaborn scatterplot, compared PC1(x-axis) PC2(Y-axis) while using the hue to represent quartile, and size to represent ticket price. While there were some interesting data points, no discernable pattern appeared while using the state label in terms of pricing.

Merged ski\_data dataframe with state\_summary dataframe on the 'state' field and using a left join. This merge provided the fields necessary to add additional informative ratios.

Using Seaborn, created a heatmap to determine the correlation between variables. AdultWeekend appears to have a strong positive correlation with the following:

A picture containing text

Description automatically generated

Using scatterplots to compare the numerical columns with the AdulWeekend price confirmed the strong correlation for the below variables.

* vertical\_drop
* Runs
* total\_chairs
* FastQuads
* Snow Making\_ac
* total\_chairs\_runs\_ratio
* total\_chairs\_skiable\_ratio
* fastQuads\_runs\_ratio
* fastQuads\_skiable\_ratio

The total chair ratios tend to support the idea of too much of a good thing (mobility) may lead to the commoditization of the offering. Additionally, the ability to quickly move customers to the runs via fastQuads appears to be linked to higher prices.

**Preprocessing and Training**

Average Price Model

Prior to building a model, the mean ticket price is $63 with a Mean Absolute Error ("MAE") of $19. On average, the guessing the price based on the mean would produce an estimate that would be approximately $19 off.

Linear Price Model

First step in building a linear model was to build a pipeline: pipe with a strategy='median', SelectKBest(f\_regression), StandardScaler(), LinearRegression(). Initially, this actually performed worse than the Average Price Model because the SelectKBest has a parameter of K which defaulted to 10. Improved the training process by cross validation to create more training opportunities within the training data set.

Next, tuned the hyperparameter selectKBest\_k with GridSearchCV. Also found the coefficients for the linear model.

mean absolute error: train mean: 10.499032338015294 train stdev: 1.6220608976799664 test mean absolute error: 11.793465668669324

Random Forest Regression Model

Pipeline for random forest regressor was definedas:SimpleImputer(strategy='median'), StandardScaler(), RandomForestRegressor(random\_state=47). Performed cross validation and GridSearchCV on Random Forest Regression Model as well

mean absolute error: train mean: 9.644639167595688 train stdev: 1.3528565172191818 test mean absolute error: 9.537730050637332

Given that the test mean absolute error is smaller for the RFR Model as compared with the LR Model, the RFR Model is selected as the model of choice.

**Modeling**

Presently, BMR charges $81 for an adult weekend pass. The pricing model suggest a price of $95.87 with a Mean Absolute Error("MAE") of $10.39. The pricing model assumes that the free market has been allowed to set the current prices in the market for the other ski resorts used in the comparison with BMR.

As for covering the additional cost of the new chair lift, the cost per ticket would be approx. $0.88/ticket ($1,540,000 operating cost/ (350,000 forecasted skiers \* 5 tickets / forecasted skiers). The additional cost of the new chair lift would be covered by the ticket increase.

Business leadership should appreciate that BMR’s current price is $14.87 ($95.87 -$81) below the projected model pricing. Current pricing is also outside of the bound for the MAE by $4.48 (14.87 - $10.39). On average, there appears to be plenty of room to opportunity to both increase profitability as well as cover the additional cost of the new chair lift.

As for the above modeled improvement scenarios, scenario 2 is recommended because it is supported by a modeled ticket price increase of $1.99 and would generate a projected increase in seasonal ticket revenue of approx. $3,500,000. Scenario 3 forecast no increase in revenue over scenario 2, but would incur the cost of adding 2 acres of snow making. As for increasing the length of the longest run by 0.2 miles and adding 4 acres of snow making (scenario 4), this scenario did not support a modeled ticket increase, but would incur the cost of ski run modification and snow acreage creation.

Slope closure modeling (Scenario 1) as a whole, suggest that there are negative ticket price implications for the closure of slopes. This impact in per ticket price ranges from $0.00 for 1 slope closure to $1.75 for 10 slopes closures. There appear to be plateaus in the $/ ticket impact per slopes closed. For instance, the per $/ticket impact is the same in the 3-5 slopes closed range and another plateau is encountered when 6-8 slopes are closed. If the business wishes to test slope closures, it is recommended that they chose the midpoint of the range of the plateau to test impact on ticket pricing.

Chart, line chart

Description automatically generated

**Conclusion**

Given the information provided, an increase in price would appear to be warranted. I would recommend that more cost/expense data from BMR be secured and factored into the model where warranted.