Guided Capstone Project Report: Big Mountain Resort

By Fatih Demiroz

Problem Statement, Success Criteria, and Scope of Solution

Big Mountain Resort (BMR) is a ski resort located in Montana with 350,000 visitors annually. This season, the company installed new equipment that cost $1,540,000. BMR thinks they don't capitalize enough on the facilities and need a data-informed strategy for determining the optimum ticket price.

The two options that BMR has are cost-cutting and increasing ticket prices. The company must cut costs or increase revenue by $1,540,000 effective immediately (i.e., success criteria).

The scope of solution for this problem is facilities. The company needs to make changes to the facilities that will either cut costs or support increased ticket prices.

Data Source, Data Wrangling, and Exploratory Data Analysis

The data for this project come from the company's database manager, Alesha Eisen. The data is a single CSV file with 330 rows (i.e., resorts across the country) and 27 columns (i.e., features).

Target Feature

'AdultWeekday' and 'AdultWeekend' features show ticket prices, while other features are about facilities. There were no significant differences between these two features, and 'AdultWeekday' had more missing values; therefore, I used 'AdultWeekend' as my only target variable.

Missing Values

I dropped the 'fastEight' column because it had 50% missing values. Interestingly, the percentage of missing values in some rows were multiples of 4 (e.g., 4%, 8%, 12%, 16%, 20%), as if values were deleted intentionally.

Outliers, Duplicate Values, and State Data

Visualization of each column showed that some columns had outliers, which skewed the distributions. I excluded such outliers from the dataset. Also, I checked and corrected suspicious-looking data points using information from resorts' websites. Distribution of features after cleaning outliers were presented in Figure 1 below. No duplicate values were found.

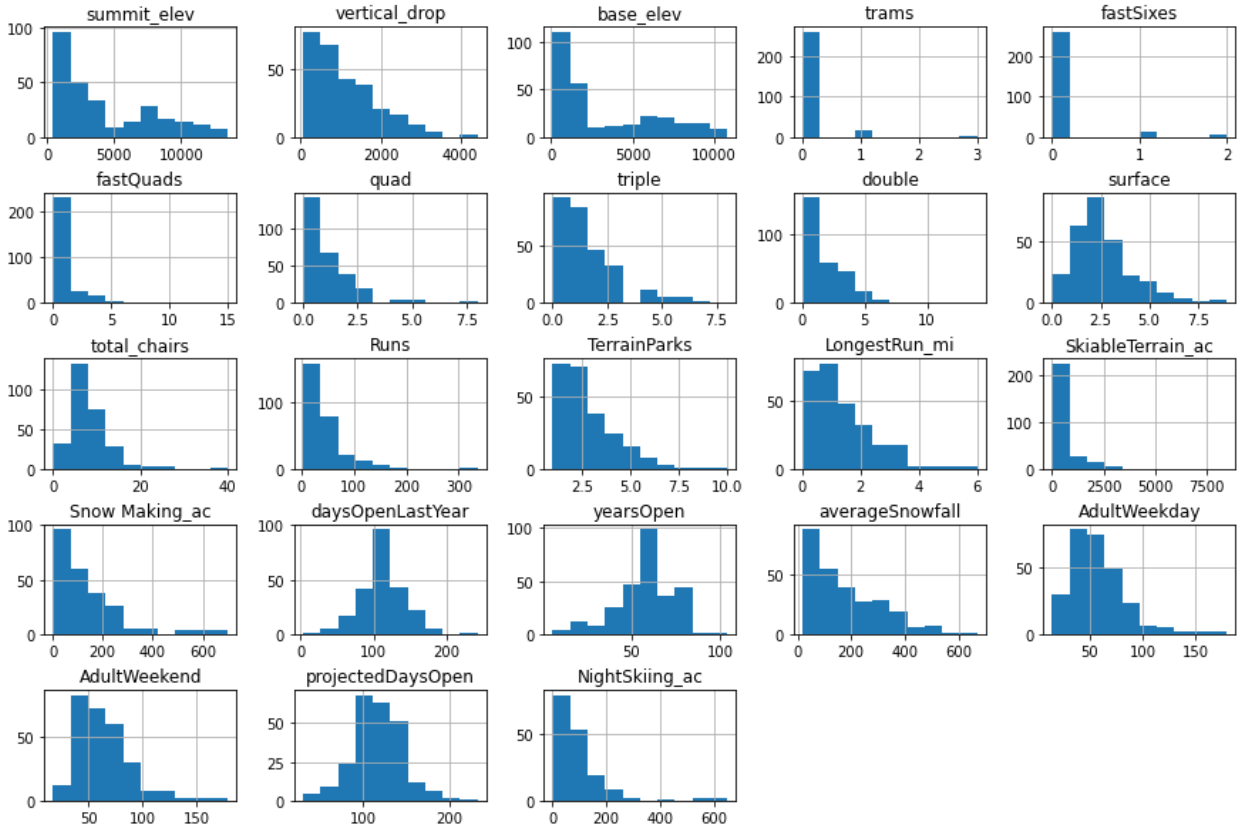

Figure 1 Distribution of Features After Outliers Excluded

Exploratory Data Analysis

To have a better understanding of the characteristics of each state, I created new variables/columns by calculating the ratio of (i) resorts per 100,000 people (variable name resorts_per_100kcapita) and (ii) resorts per 100,000 square miles (variable name resorts_per_100ksq_mile).

I ran a principle component analysis (PCA) and identified two components that explain more than 90% of the variability in 'state_summary' data (see Figure 2).

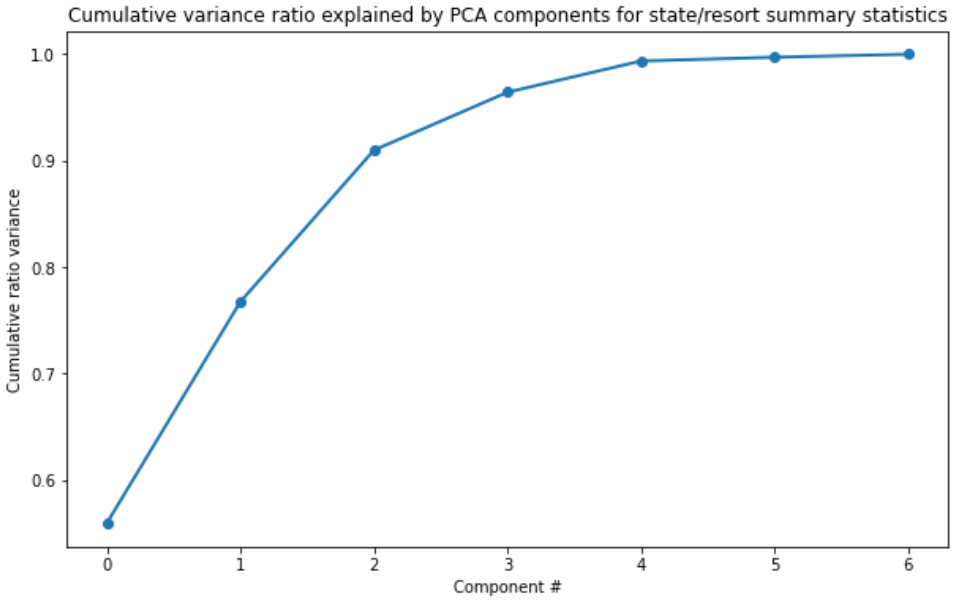

Figure 2 Result of PCA analysis. The first two components explain over 90% of variability for state/resort summary statistics.

Model Preprocessing with feature engineering

I developed multiple machine learning (ML) models with different hyperparameters.
I imputed missing values with median and mean values and tested which imputation method generates better metrics (Mean Absolute Error, R2, and Mean Squared Error).

I first built a multiple regression model and tried to identify the Best-K-features for maximum parsimony and best model accuracy. I cross-validated my models on five different sections of the data using sklearn's built-in functions. My analysis identified the following eight features (see Figure 3 below).

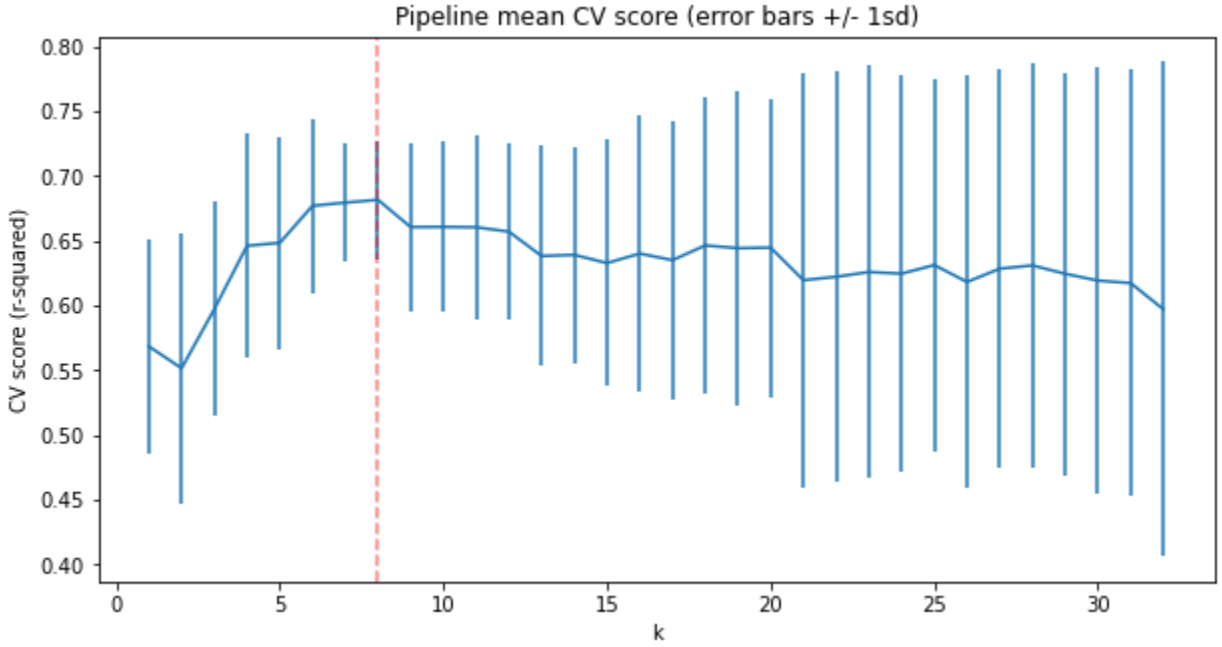

Figure 3 The Best K Features result is 8. Red line shows the cut off point for best line features.

I also developed a random forest model. The random forest model identified four features (see Figure 4).

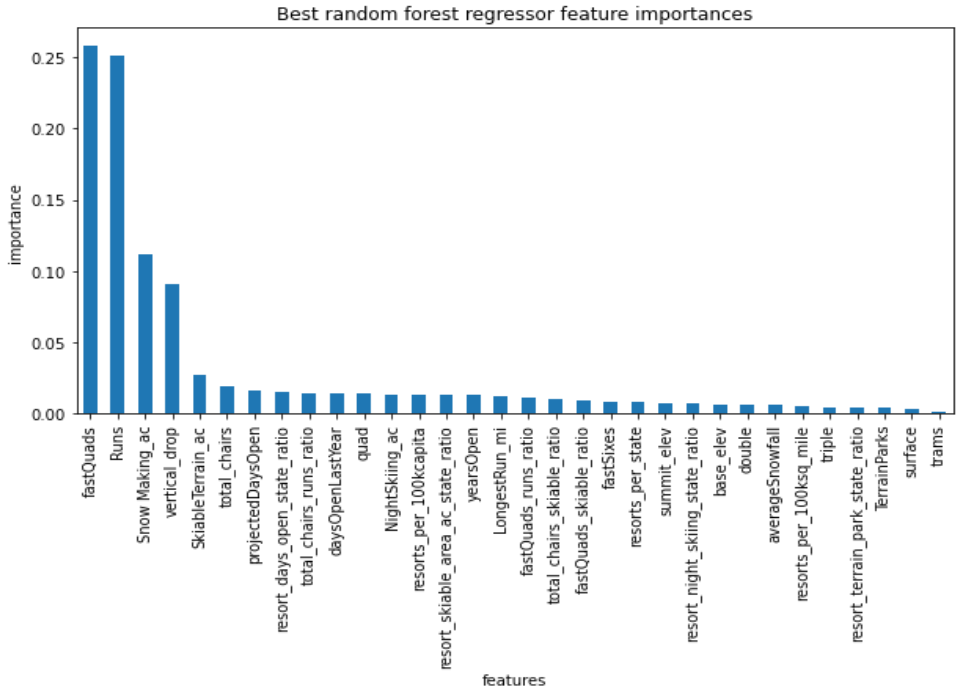

Figure 4 Most dominant features in the Random Forest Model

Model Selection and Scenario Testing

When the ML and random forest models were compared, the random forest model performed better in model metrics. Therefore, I continued my final analyses with random forest.

Using the random forest model, I predicted that the Big Mountain Resort could price its tickets at $97.33 with an expected mean absolute error of $10.36. This suggests that there is room for an increase in ticket prices. This was an optimistic result, but I wanted to do further analyses to ensure we were not missing anything.

Multiple Scenarios

Next, I wanted tested multiple scenarios that will either cut costs or support an increase in the ticket price. The four scenarios I tested were:

Scenario 1: Close up to 10 of the least used runs. The number of runs is the only parameter varying.
Prediction: Closing runs lead to drops in ticket prices (see Figure 5).

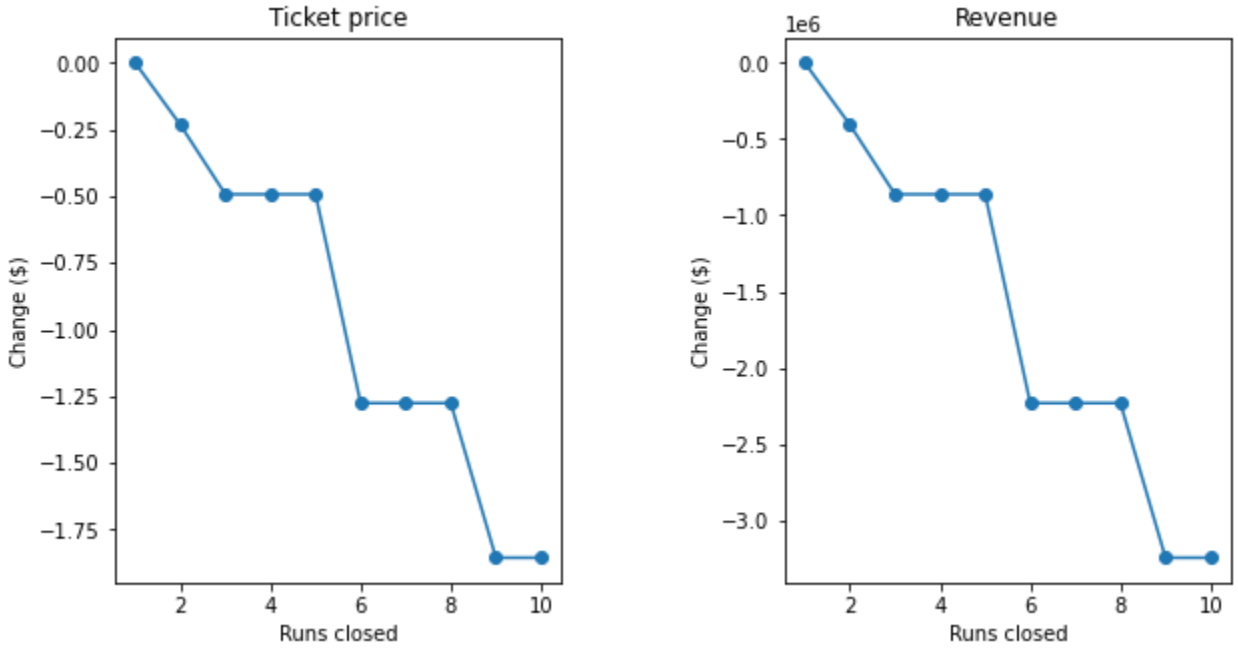

Figure 5 The impact of closing 'runs' on ticket prices. The more runs are closed the more ticket prices will drop.

Scenario 2: In this scenario, Big Mountain is adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift.
Prediction: This scenario increases support for ticket price by $1.99 (revenue increase of $3,474,638).

Scenario 3: Same as Scenario 2 but adding 2 acres of snowmaking.
Prediction: This scenario increases support for ticket price by $1.99 (revenue increase of $3,474,638).

Scenario 4: This scenario calls for increasing the longest run by .2 miles and guaranteeing its snow coverage by adding 4 acres of snow-making capability.
Prediction: No difference whatsoever.

Conclusion

The first scenario hurts ticket prices while the second and third scenarios support ticket price increase by $1.99 (predicted annual revenue increase by $3.47 million). The fourth scenario has no impact on ticket price.

Scenarios 2 and 3 support ticket prices, and both scenarios require adding a chair lift. Adding a chair lift costs $0.88 per ticket, which is less than the predicted $1.99 increase in ticket prices.

Furthermore, Scenario 3 is likely to add more to operational costs than Scenario 2 since an additional 2 acres of snowmaking is required in Scenario 3.

Limitations and Future Directions

Combining the current data with data regarding operational costs and where most of our visitors come from (e.g., out-of-state, in-state) would help improve the ML model.

I would be happy to help other business analysts should they use this model in the future.