Springboard Data Science Guided Capstone Project Report

By: Bini Teklehaimanot

# Problem statement

What can Big Mountain Resort do to increase revenue by more than $1,540,000 over

the next three ski seasons by leveraging its facilities and adjusting prices where

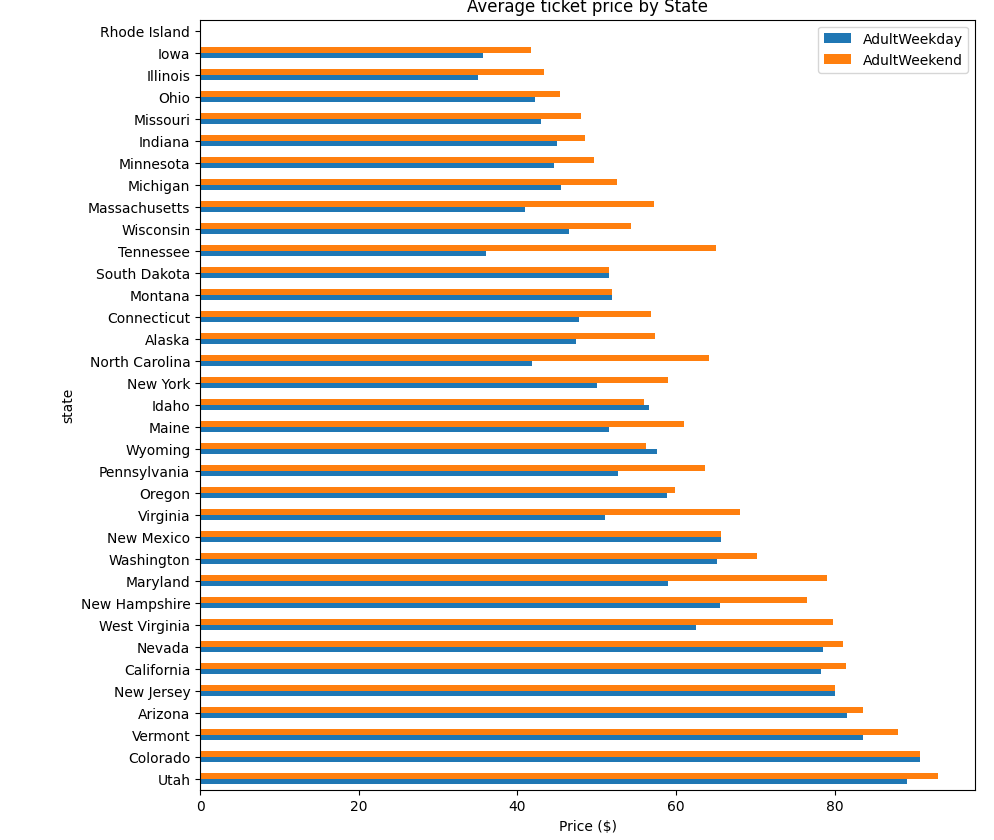
appropriate?

# Data Wrangling

The data set mostly contains potential numeric features. Resort name, region, and state are the only three categorical features. Since price is the driver of our problem, a decision was necessary to determine whether to use adult week day ticket price or adult weekend tick price. For our resort, both tickets have the same price. Therefore, I used the one with the least missing values as the target feature, adult weekend price.

For categorical features, I discovered the data set had one duplicate resort name, Crystal Mountain. The combination of name + region or name + state did not produce any duplicates, as region and state are the same for the majority of the resorts. California, Nevada, Oregon, and Utah use a region name different than the state name. I found that this information is not essential for our analysis.

The distribution of resorts by state shows New York as the leader. Montana comes in 11th place. This might be useful information when adjusting prices. The distribution of ticket prices by state paints a different position for Montana. We rank in the bottom third of the list, see chart below.



I focused on ticket price to impute missing values, as that is the target feature. Any row of data that is missing both weekday and weekend tick prices has no value to us. I dropped the rows from the data. I dropped the weekday price column since the weekend price is the most available.

# Exploratory Data Analysis

Exploring the data further, I found that Montana is the third largest state by square miles in the dataset. While it is a large state, its population is sixth from the bottom. Could this be a factor in pricing the tickets? The population does not seem to affect the resort count by state much. Monta has 12 resorts, about 1/3 of the New York count. New York’s high count is not surprising as the state boasts many wealthy residents.

By fitting the PCA transformation on scaled data, I saw that four components account for 95% of the variance and two components account for 75%, see chart below. After looking at the ski state summary, 77.2% variance, no trend or pattern was present. Most of the states are spread across the first component.

A graph with a line of cumulative variance ratio explained by PCA components for state/resort summary statistics



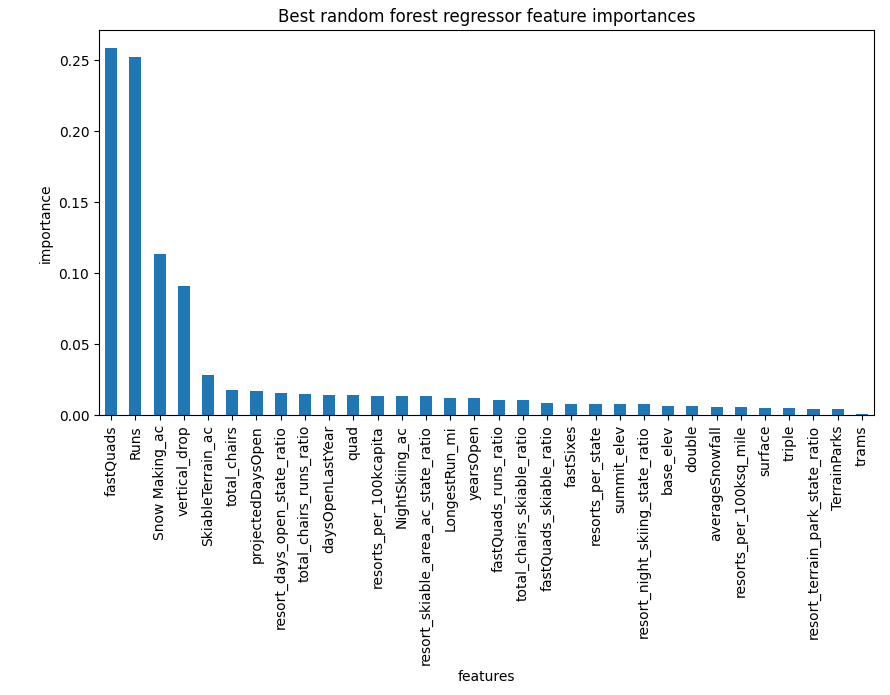
A correlation heatmap helped me uncover the highly correlated features. Summit and base elevation in our dataset are highly correlated. Some key observations include: negative correlation between number of resorts in each state and ratio features, positive correlation between the ratio of night skiing area and number of resorts per capita. In relation to our target feature adult weekend ticket price, a sensible correlation exists with fastQuads, runs, and snowmaking.

# Model Preprocessing with feature engineering

5

# Algorithms used to build the model with evaluation metric

Linear Regression and Random Forest algorithms were tried to build our model. The data was split 70/30 to train and test so we could get an unbiased dataset with the test. Three different metrics were incorporated. R-squared, Mean Absolute Error, and Mean Squared Error within sklearn\_metrics. To impute missing data, I used sklearn’s pipeline method SimpleImputer. The random forest model was selected because it has less variability and lower cross-validation mean absolute error. To check for feature importance, I plotted a bar plot. See chart below.



# Winning model and scenario modeling

5

# Pricing recommendation

5

# Conclusion

5

# Future scope of work

5