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

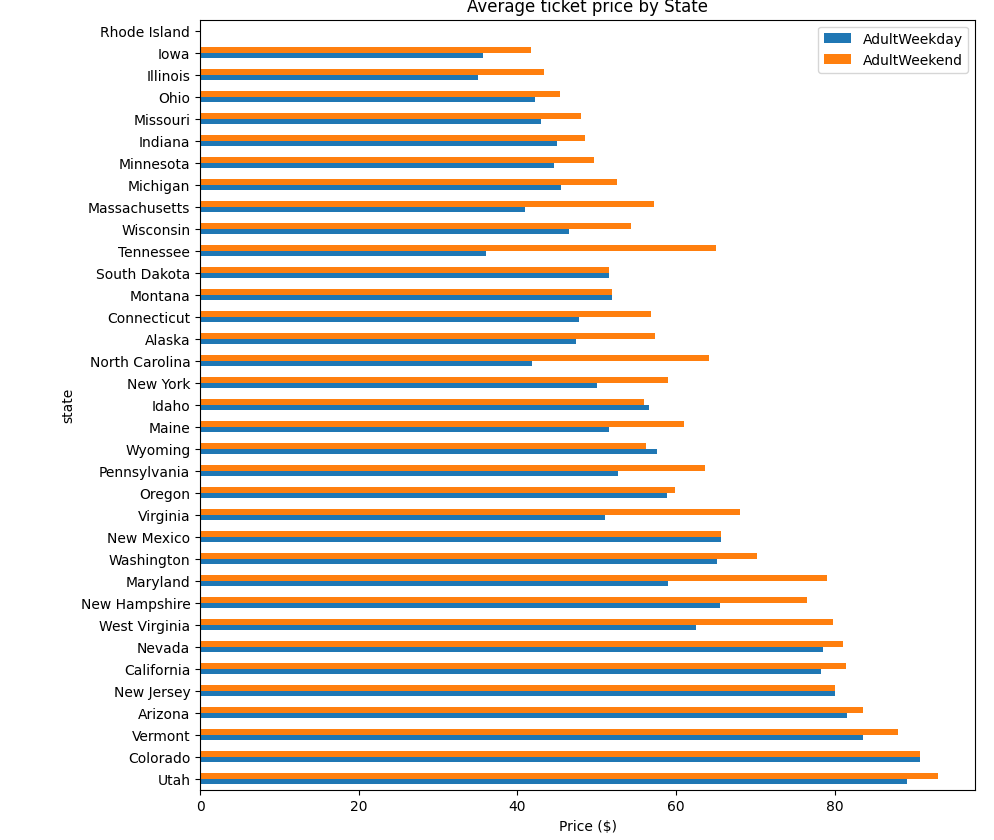
appropriate?

# Data Wrangling

The dataset mostly contains 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 the adult weekday ticket price or the 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, the adult weekend price.

For categorical features, I discovered the dataset 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 most 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 picture for Montana. We rank in the bottom third of the list; see the chart below.



I focused on the ticket price to impute missing values, which 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 has the most available values.

# 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. Montana has 12 resorts, about 1/3 of the New York count. New York’s high count is unsurprising 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.

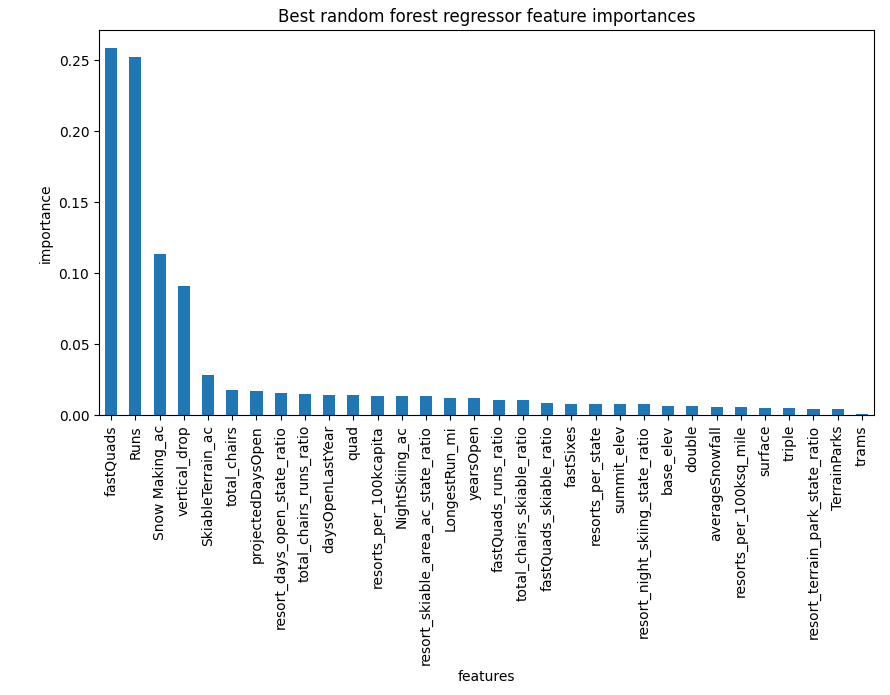
A graph with a line of cumulative variance ratio explained by PCA components for state/resort summary statistics



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

# Model Preprocessing with feature engineering

I followed a few steps to get the data ready for a model. I observed differences between the median and mean to impute missing feature values. The imputation is applied to both test and train splits. Then, the data is scaled so that all features are used on a consistent scale. This allowed me to make predictions on both splits and assess the model performance. The median and the mean presented similar results. For the final imputation of missing data, I used sklearn’s pipeline method SimpleImputer. To check for feature importance, I plotted a bar plot; see the chart below.



# 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 uncover the best model and scenario modeling. The random forest model was selected as the winning model because it has less variability and lower cross-validation mean absolute error. I verified that the dataset contained enough data and no further data acquisition was required.

# Pricing recommendation

I recommend increasing the price by at least $5 to be able to recover the additional cost of $1.54M to operate the additional chairlift. We should use a dynamic pricing structure and increase the price over time to the $9 - $10 max increase level.

# Conclusion

The Random Forest model selected for this work allows for increasing the ticket price by $9-$10. This translates to a revenue per season of $17.3M. Understanding the current pricing strategy is important to determining the best price for the new strategy.

# Future scope of work

There are a few areas for future scope of work. We should enhance the model complexity by implementing additional modeling techniques and analysis, such as time series analysis. Additionally, to help with decision-making, we must look at real-time demand forecasting, ethical considerations, and external factors that might impact our decisions. To make it easy for anyone to use the model, I will develop a user-friendly interface with adequate documentation, integrate it with existing systems and CRM, and conduct A/B testing.