# **DSC Unit 14: Guided Capstone**

## **Section 1: Problem Identification**

**Problem Definition:** What should be the ticket price for Blue Mountain resort service based on the following two factors? 1. Average price of other resorts 2. How do Blue Mountain’s facilities compare with other resorts?

1. **Context:** Big Mountain Resort, a ski resort located in Montana hosts 350K people every year for skiing or snowboarding. The resort's pricing strategy has been to charge a premium above the average price of resorts in its market segment. However, basing the pricing on just the market average does not provide the business with a good sense of how important some facilities are compared to others. The business seeks guidance on how to select a better value for their ticket price.
2. **Criteria for success:** Determine ticket price for Blue Mountain resort service.
3. **Scope of solution space:** 1. Price analysis for other resorts in the same market segment 2. Facilities benchmarking with other resorts to quantify value addition with current Blue Mountain services.
4. **Constraints within solution space:** Premium amount that needs to be charged with every sold ticket to accomplish business goals (revenue & profitability).
5. **Stakeholders to provide key insight:** 1. Jimmy Blackburn – Director of Operations 2. Alesha Eisen – Database Manager
6. **Key data sources:** 1. Data comparing different resort facilities and price information (single CSV file)

**`**

## **Section 2: Data Wrangling**

**Summary:** The original CSV file had information for 330 resorts, i.e., 330 rows and 27 columns. Blue Mountain resort information was available in row # 151. Missing value counts were identified in various columns. Uniqueness of different resorts were verified. Two resorts had the same name but were present in different states. Distribution of resorts and ticket price (both during weekday and weekend) with state and region were studied. Rows were dropped where the ticket price information was missing for both weekday and weekend columns. This resulted in a new data frame with 277 rows and 25 columns. Distribution of feature values were studied for different resorts. Population and state area data were added from another source (Wikipedia) to our existing data frame. Weekday and weekend ticket price correlation was examined to figure out appropriate modeling target. The weekend price was chosen for modeling purpose since it had a lower number of missing values. After data cleaning from the original data frame, new data were saved into a different CSV file.

## **Section 3: Exploratory Data Analysis**

**Summary:** Top states were identified for each of the summary statistics i.e., total state area, total state population, resorts per state, total skiable area, total night skiing area, total days open. Resort density ratios were defined i.e., resorts\_per\_100kcapita and resorts\_per\_100ksq\_mile to normalize resorts data for different states for a fair comparison. Top states were identified for these defined resort density parameters. PCA was performed to disentangle the interconnected web of relationships between different features. Based on this analysis, the first two components were observed to account for over 75% of the variance, and the first four for over 95%. Seaborn package was used to visualize the relationship between the first two components (PC1 and PC2) for different states via a scatterplot. Feature correlation heatmap was generated to understand the relationship between different variables. Scatterplots were generated to understand the relationship of numeric features vs. ticket price. Strong positive correlation of ticket price with vertical\_drop, fastQuads and total\_chairs were observed.

## **Section 4. Pre-processing and Training**

**Summary**: Baseline approach was implemented by simply taking the average price as the predictor. This resulted in a mean absolute error of $19 approximately if we guessed ticket price by using the average of known values. Next linear regression model was implemented. This showed a mean absolute error of $12 approximately. vertical\_drop feature showed the biggest correlation with price followed by Snow Making\_ac, total\_chairs, fastQuads and Runs. Next a random forest regressor model was tried. This showed a mean absolute error of $10 approximately. The top four features useful in predicting price were identified as fastQuads, Runs, Snow Making\_ac and vertical\_drop.Finally random forest model was chosen for modeling purpose since it had a lower cross-validation mean absolute error by almost $2. It also exhibited less variability.

## **Section 5. Modeling**

**Summary:** The modelled price for our client's resort was determined at ~ $96 approximately with a mean absolute error of ~ $10, suggesting room for price increase. The top four features that came up as important in modeling are: 1. vertical\_drop 2. Snow Making\_ac 3. total\_chairs 4. fastQuads. Below are the key observations for our client's resort facility regarding these important features. A. Big Mountain was found to be doing well for vertical\_drop, but there are still quite a few resorts with a greater drop. B. Big Mountain was found to be very high up the league table of Snow Making\_ac. C. Big Mountain has the highest number of total\_chairs. D. Big Mountain has 3 fastQuads, which puts it high up that league table.

1. **Resort’s feature correlation heatmap**

Chart

Description automatically generated

1. **Best random forest regressor feature importance**

Chart, histogram

Description automatically generated