DSC Unit 14: Guided Capstone

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?

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.

1. **Criteria for success:** Determine ticket price for Blue Mountain resort service.
2. **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.
3. **Constraints within solution space:** Premium amount that needs to be charged with every sold ticket to accomplish business goals (revenue & profitability).
4. **Stakeholders to provide key insight:** 1. Jimmy Blackburn – Director of Operations 2. Alesha Eisen – Database Manager
5. **Key data sources:** 1. Data comparing different resort facilities and price information (single CSV file)

`

Data Wrangling

1. **Data Overview**: The original data file contained information for 330 resorts, consisting of 27 different features.
2. **Identifying Missing Values**: Initial analysis involved identifying missing values present in various columns of the dataset. This step is crucial for understanding the completeness and quality of the data.
3. **Verification of Resort Uniqueness**: The uniqueness of resorts was verified to ensure each resort is correctly represented in the dataset. This step helps in avoiding data duplication or confusion.

4. **Resolving Resort Name Conflict**: It was discovered that two resorts shared the same name but were located in different states. Resolving such conflicts ensures accuracy in data analysis and prevents ambiguity.
5. **Studying Resort Distribution and Ticket Prices**: Analysis focused on studying the distribution of resorts along with ticket prices during weekdays and weekends, considering both state and region factors. This step helps in understanding pricing patterns and geographical variations.
6. **Handling Missing Ticket Price Data**: Rows where ticket price information was missing for both weekday and weekend columns were dropped. This ensures that the dataset used for further analysis is consistent and does not contain incomplete information.
7. **Resulting Dataset**: After dropping rows with missing ticket price information, a new data frame was formed, containing 277 observations and 25 features.
8. **Studying Feature Value Distributions**: The distribution of feature values across different resorts was studied. This analysis provides insights into the variability and characteristics of the dataset.
9. **Adding External Data**: Population and state area data were added to the existing dataset from another reliable source (Wikipedia). This enriches the dataset with additional contextual information for more comprehensive analysis.
10. **Correlation Analysis for Modeling Target Selection**: Correlation analysis was performed between weekday and weekend ticket prices to determine an appropriate target for modeling. Weekend ticket price was chosen as the modeling target due to its lower number of missing values, which enhances the reliability of subsequent analyses.
11. **Data Cleaning and Saving**: After completing data cleaning and preprocessing steps, the refined dataset was saved into a different CSV file. This ensures that the cleaned data is preserved for further analysis and modeling without affecting the original dataset.

Each step in this report contributes to the overall data wrangling process, aiming to prepare the dataset for subsequent analysis and modeling tasks while ensuring data accuracy, completeness, and consistency.

Exploratory Data Analysis

1. **Identification of Top States for Summary Statistics**:
    - Total state area, total state population, resorts per state, total skiable area, total night skiing area, and total days open were computed for each state.

- Top states were identified based on these summary statistics.

2. **Definition of Resort Density Ratios**:
   - Resort density ratios, such as resorts_per_100kcapita and resorts_per_100ksq_mile, were defined to normalize resort data across different states for fair comparison.
3. **Identification of Top States for Resort Density Parameters**:
   - Top states were identified based on the defined resort density parameters.
4. **Principal Component Analysis (PCA)**:
   - PCA was performed to disentangle the relationships between different features in the dataset.
   - The first two components were observed to account for over 75% of the variance, while the first four components accounted for over 95% of the variance.
5. **Visualization of PCA Results**:
   - The Seaborn package was used to visualize the relationship between the first two principal components (PC1 and PC2) for different states via a scatterplot.
   - This visualization helps in understanding the underlying structure and patterns in the data.
6. **Generation of Feature Correlation Heatmap**:
   - A feature correlation heatmap was generated to understand the relationships between different variables in the dataset.
   - This heatmap provides insights into the strength and direction of correlations between pairs of features.
7. **Generation of Scatterplots**:
   - Scatterplots were generated to understand the relationship of numeric features with ticket price.
   - Strong positive correlations of ticket price with vertical_drop, fastQuads, and total_chairs were observed in these scatterplots.

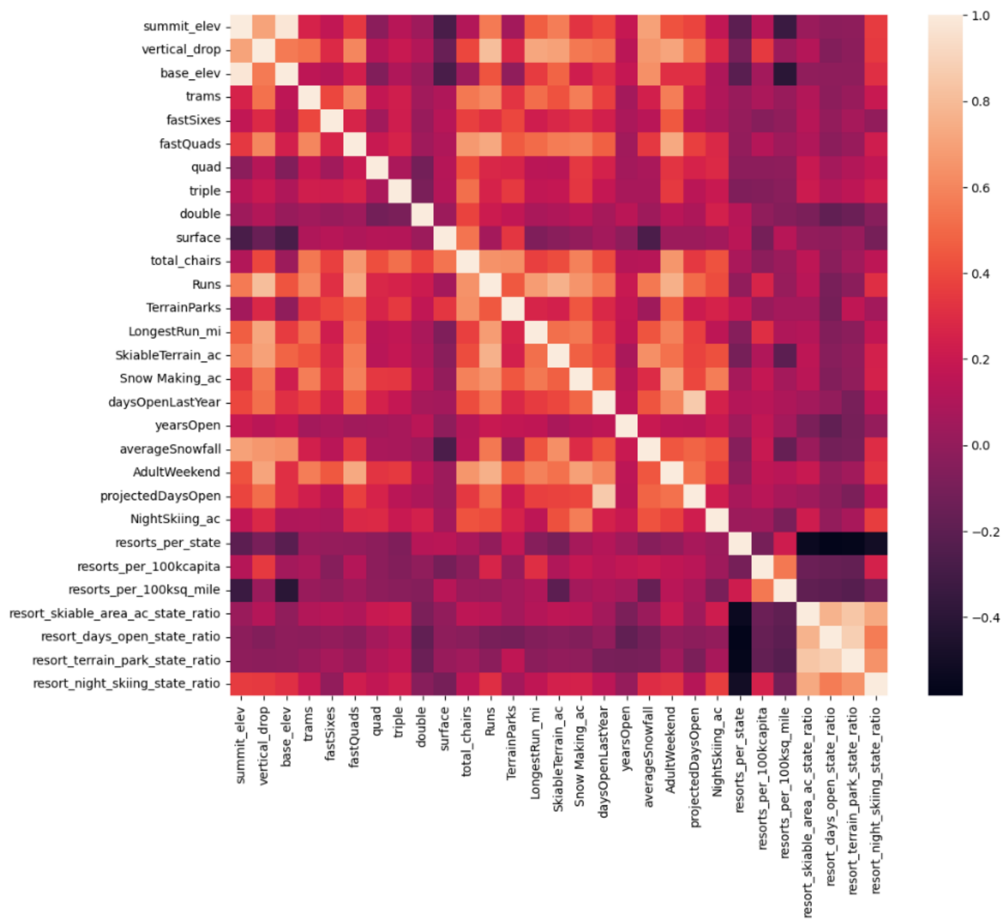
Each step in this report contributes to the exploratory data analysis process, aiming to understand the underlying patterns and relationships within the dataset. These analyses provide valuable insights for further interpretation and modeling tasks.

Pre-processing and Modeling

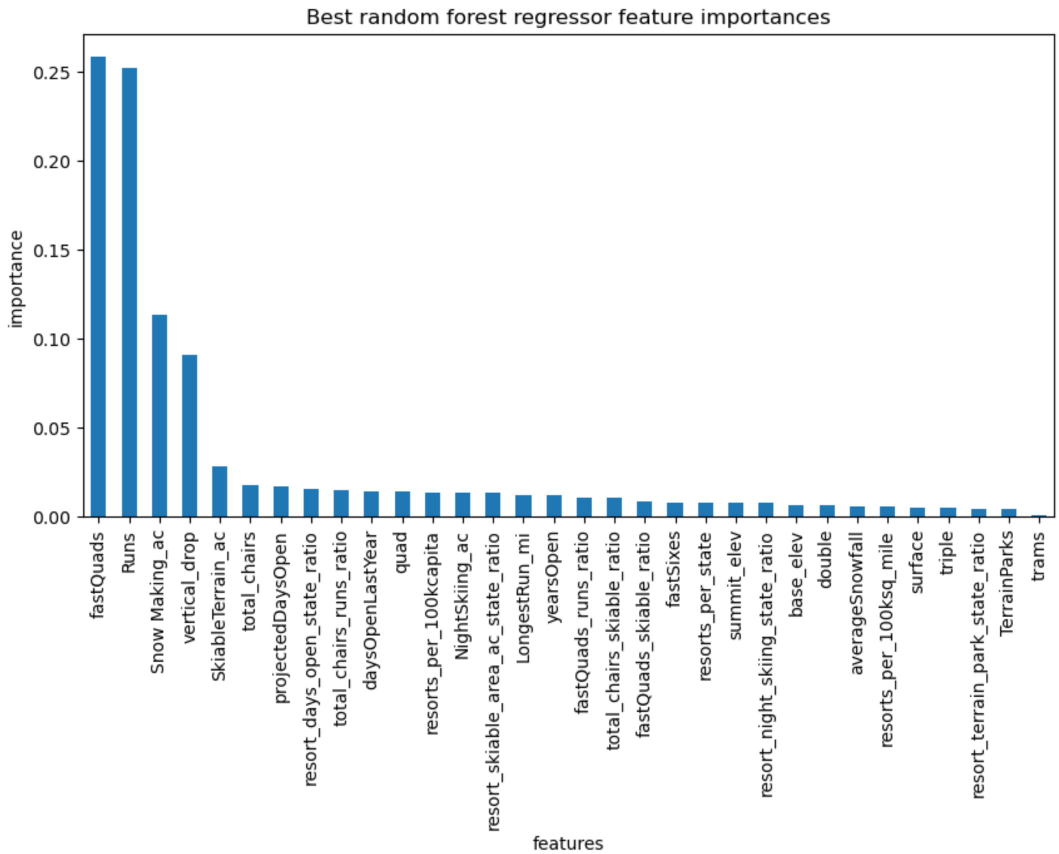
1. **Baseline Approach Implementation**:
    - The baseline approach involved simply taking the average price as the predictor.
    - This resulted in a mean absolute error of approximately $19 when guessing ticket prices by using the average of known values.
2. **Linear Regression Model Implementation**:
    - A linear regression model was implemented as the next step.
    - This model showed a mean absolute error of approximately $12, indicating an improvement over the baseline approach.
3. **Identification of Correlated Features**:
    - Analysis revealed that the vertical_drop feature exhibited the strongest correlation with price, followed by Snow Making_ac, total_chairs, fastQuads, and Runs.
    - Understanding these correlations helps in identifying key predictors for modeling ticket prices.
4. **Random Forest Regressor Model Implementation**:
    - A random forest regressor model was implemented as an alternative approach.
    - This model showed a mean absolute error of approximately $10, indicating further improvement compared to the linear regression model.
5. **Identification of Top Predictive Features**:
    - The top four features useful in predicting ticket prices were identified as fastQuads, Runs, Snow Making_ac, and vertical_drop.
    - Recognizing these influential features aids in building more accurate predictive models.
6. **Model Selection**:
    - Finally, the random forest model was chosen for modeling purposes.
    - This decision was based on its lower cross-validation mean absolute error, which was approximately $2 lower than that of the linear regression model.
    - The random forest model demonstrated superior performance compared to other models considered.

This step-by-step report outlines the iterative process of model development and selection, highlighting the progression from baseline to more sophisticated approaches based on their performance metrics. The identification of key features and the selection of the most suitable model contribute to improving the accuracy of predicting ticket prices for resorts.

1. Resort's feature correlation heatmap



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



Final Recommendation: The client should price the resort's ticket at ~ $96 suggesting room for price increase for the client.

1. **Vertical Drop**:
    - Big Mountain Resort is performing well in terms of vertical drop.
    - However, there are still several resorts with a greater vertical drop.
    - This suggests that while Big Mountain is competitive in this aspect, there is potential for marketing efforts to emphasize its vertical drop compared to other resorts.
2. **Snow Making_ac**:
    - Big Mountain Resort ranks very high in terms of snowmaking acreage.

- This indicates that the resort has invested significantly in snowmaking capabilities, which can be marketed as a competitive advantage, especially for ensuring consistent snow conditions throughout the ski season.

3. **Total Chairs**:

    - Big Mountain Resort has the highest number of total chairlifts among the resorts considered.
    - This suggests that the resort offers ample chairlift capacity, which can contribute to improved guest experience by reducing wait times and increasing accessibility to various parts of the mountain.

4. **FastQuads**:

    - Big Mountain Resort has three fastQuads, which positions it well in comparison to other resorts.
    - FastQuads are high-speed chairlifts that enhance the efficiency of transporting skiers and snowboarders around the mountain.
    - Having multiple fastQuads further enhances the resort's appeal by providing faster and more convenient access to different trails.

Based on these observations and the modeled price for Big Mountain Resort, which suggests room for a price increase, the resort may consider leveraging its strengths in these important features to justify and potentially implement a price adjustment. Additionally, marketing strategies can be tailored to highlight these strengths to attract more visitors and increase revenue.