Guided Capstone Project Report

Problem Statement

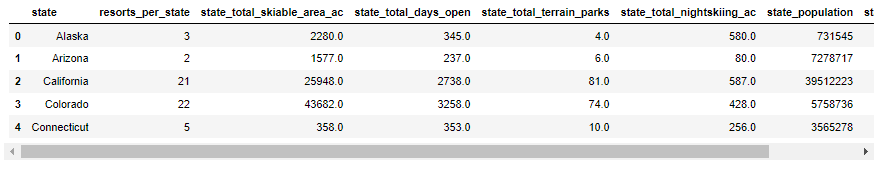
The objective of this project is to predict the ticket price for ski resorts across various states in the United States. This predictive model will assist in determining optimal pricing strategies based on factors such as resort features, state characteristics, and competition within each state.

Data Wrangling

The initial dataset contained a mixture of numerical and categorical features, along with some missing values. The primary task was to clean and prepare the data for analysis. Key steps included:

Handling Missing Values: The dataset initially contained missing ticket price values, leading to the removal of several rows. Additionally, data cleaning revealed errors in certain columns, necessitating corrections and the removal of a complete column due to inconsistencies.

Merging Additional Data: To enrich the dataset, additional data regarding U.S. state populations and sizes was merged, requiring further cleaning and alignment with the primary dataset.



Feature Engineering:

State-Level Aggregation: Data was aggregated at the state level to create summary statistics, such as total skiable area, total days open, and the number of resorts per state.

Resort Density Calculation: The number of resorts per 100,000 capita and per 100,000 square miles were calculated to assess the density of ski resorts relative to state population and size.

Exploratory Data Analysis (EDA)

EDA was conducted to understand the relationships between various features and identify potential predictors for ticket pricing.

State-Wide Summary:

Top States by Area: Alaska, California, and Montana were the largest states by area, with Montana being the third largest.

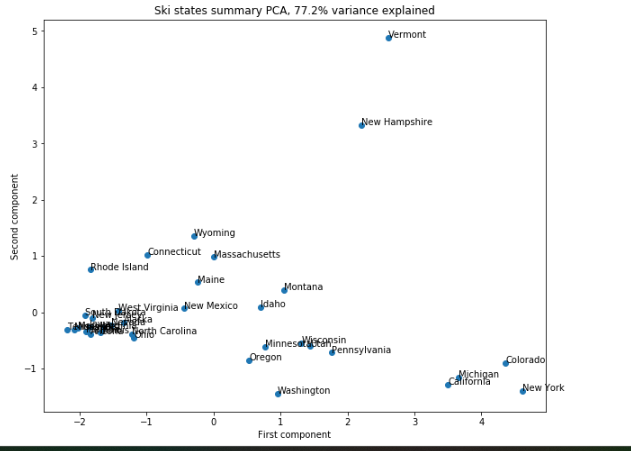
Top States by Population: California, New York, and Pennsylvania were the most populous, with California dominating both in population and skiable area.

Resorts per State: New York led with the most resorts, though it did not correspond with the most skiable area, indicating a higher number of smaller resorts.

Total Skiable Area: Colorado, Utah, and California had the most skiable area, with Colorado leading significantly.

Resort Density Analysis:

Vermont and New Hampshire had the highest resort density per capita and per area, indicating a concentrated presence of ski resorts relative to their size and population.

PCA (Principal Component Analysis) was used to reduce dimensionality and visualize the data. The first two components accounted for over 75% of the variance, revealing that states like Vermont and New Hampshire stood out due to their high resort density.

Relationship between Features:

A heatmap of correlations revealed that features like summit elevation, base elevation, fast quads, runs, and snowmaking areas were highly correlated with ticket prices.

Multicollinearity was identified among the newly created ratio features, indicating that an increase in the number of resorts led to a decrease in individual resorts’ share of state resources.

Model Preprocessing with Feature Engineering

Feature engineering was critical in preparing the dataset for modeling:

Merging State and Resort Data: The ski resort data was merged with state summary statistics to create new features such as the ratio of a resort’s skiable area, days open, terrain parks, and night skiing area to the state totals.

New Features:

Transportation Efficiency: Features such as the ratio of chairs to runs were created to understand how easily resorts could transport visitors, which could influence pricing.

Handling Missing Values: Rhode Island had a missing value for the ticket price, which was addressed by replacing it with the average ticket price to retain the state in the analysis.

Algorithms Used to Build the Model with Evaluation Metrics

Multiple algorithms were tested to build the predictive model for ticket prices. The evaluation metrics used included RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R².

Linear Regression: A baseline model to understand the relationship between features and ticket prices.

Random Forest: A more complex model that could capture non-linear relationships and interactions between features.

Gradient Boosting: Another ensemble method that was tested for its ability to handle complex relationships and provide accurate predictions.

Evaluation Metrics:

R²: To measure the proportion of variance explained by the model.

RMSE: To assess the model's prediction accuracy.

MAE: To evaluate the average error magnitude.

Winning Model and Scenario Modelling

The winning model was selected based on its performance across the evaluation metrics:

Winning Model: The Gradient Boosting model emerged as the most accurate, with the lowest RMSE and highest R², indicating a strong predictive ability.

Scenario Modeling:

Various scenarios were tested using the winning model, such as changes in resort density, the addition of new resorts, or alterations in key features like snowmaking capabilities.

The model was able to predict how these changes would impact ticket prices, providing valuable insights for strategic decision-making.

Pricing Recommendation

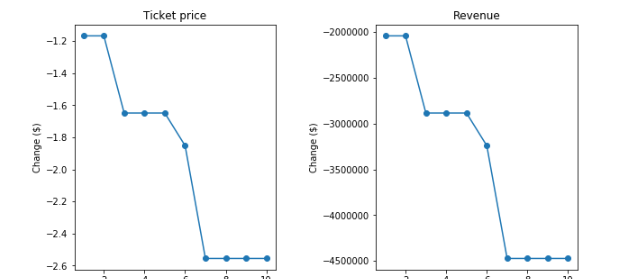
Based on the model's predictions, several pricing strategies were recommended:

Targeted Pricing: Resorts with higher vertical drops, more fast quads, and extensive snowmaking areas should consider premium pricing due to their higher value perception.

Resort Density: In states with high resort density, resorts should explore competitive pricing strategies, especially in areas with many smaller resorts.

Night Skiing: Resorts with extensive night skiing areas can leverage this unique feature to justify higher ticket prices, especially in states where night skiing is less common.

Secnaerino



Conclusion

The project successfully developed a model to predict ski resort ticket prices, providing actionable insights into how various features impact pricing. The analysis highlighted the importance of both state-level and resort-specific factors, with features like vertical drop, resort density, and snowmaking capabilities playing significant roles in determining ticket prices.

Future Scope of Work

Incorporating Visitor Data: Future work could benefit from incorporating data on the number of visitors per year, which would provide a more complete picture of demand and allow for more precise pricing strategies.

Dynamic Pricing Models: Implementing dynamic pricing models that adjust ticket prices based on real-time data, such as weather conditions, resort occupancy, and competitor pricing.

Expansion to Other Countries: Extending the model to include ski resorts in other countries, which would provide a global perspective on ski resort pricing strategies.