Guided Capstone Project Report

Introduction

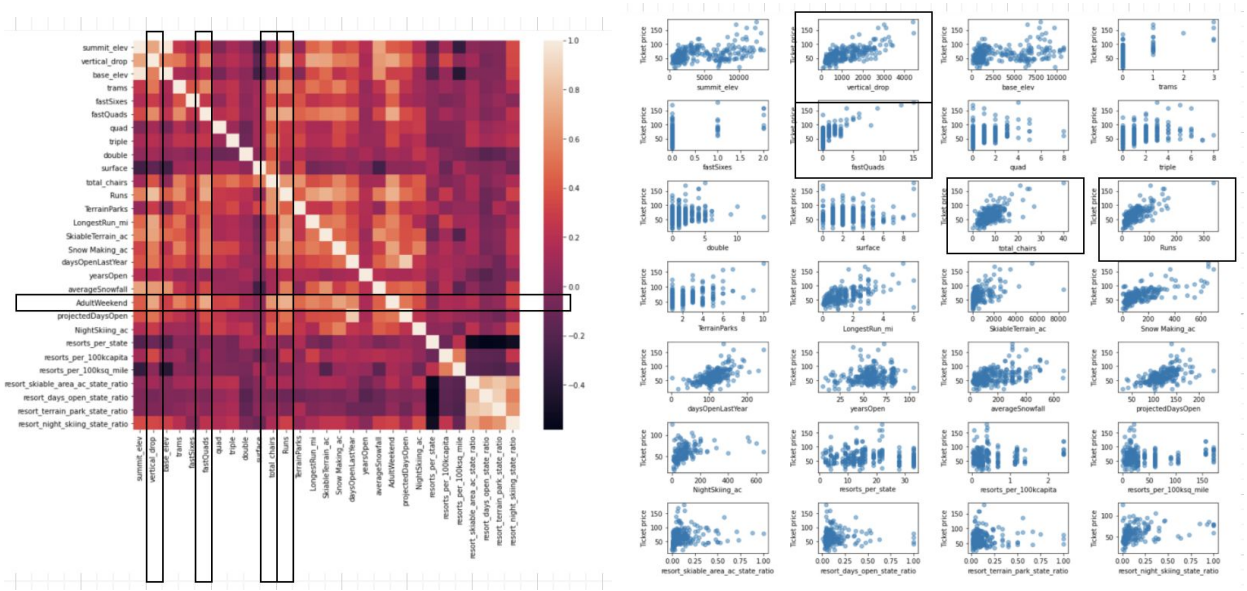
A ski resort located in Montana, Big Mountain, promises stunning views of Glacier National Park and Flathead National Forest and serves about 350,000 people a year. The resort has access to 105 trails, the 3.3 miles longest run, 4,464 ft base elevation, and 6.817 ft on the summit. It provides 11 lifts, 2 T-bars, one magic carpet for skiers. Big Mountain resort invested an additional chair lift, which increases its operation cost by **$1,540,000** this season. So the resort plans to charge the price above the average price in the market, but still not sure if this will benefit the business. Besides that, the resort may not capitalize on its facilities, so it changes in business strategy to improve its profit margin. Using a collected dataset from other resorts across the country to identify competitive ticket prices and opportunities in the next year to increase the profit margin.

Data Wrangling

The dataset has many resorts in different states and regions of the United States, including Big Mountain Resort in Montana. The data provides the facilities' information and services with their ticket prices: adult weekday and adult weekend prices. Big Mountain resort doesn't appear to have any missing values. However, the other resorts have missing features. Columns with a large amount of missing value and incorrect data were dropped. Both ticket prices are similar to each other and almost identical in Montana, so the AdultWeekend feature is kept because the weekend prices have the least missing value from the two. Besides, suspicious data such as 26819 acres in skiable terrain had been replaced with an accurate number from a trustable source. There are 227 rows left from 330 rows initially.

Exploratory Data Analysis

The categorical features are Name, Region, and state. The rest are numerical features: summit_evel, vertical_drop, chairs, runs, and so on. Grouping numerical characteristics for each state corresponding to its state population and space distance is the way to discover and detect any new pattern. After analyzing all of these traits and noticing different trends, I used principal component analysis (PCA) to reduce the dataset's dimensionality into smaller sets that can be easily visualized and analyzed. Once the six numeric features are scaled and fit the PCA transformation, the first two components (resort_per_state, state_total_skiale_area_ac) 75% account for over 75% while the first four accounts for over 95%. I applied the average ticket price to the scatter plot from the first two PCA elements, but the pattern wasn't clear. I created a heatmap and scattered plots to see the relationship and correlation amongst the features.



Based on the heatmap and scatterplot, there are positive correlations between the ticket price with vetical_drop, fastQuads, total_chairs, and runs for each state.

Preprocessing and Training

In the preprocessing process, a baseline model is an essential solution to see how good the mean value is, which is also a metric compared to other models. I split the dataset into a 70/30 train/test split that only contains numerical data. I use three metrics: R-squared, mean absolute error (MAE), and mean squared error (MSE). On the test set, the R2 is -0.00312, which is also lower than on the training set. The MAE on the test set is worse than the training set, but the MSE did better on the test set. Therefore, I check the two models: the linear regression model and the random forest model.

For the initial models, I impute missing features with the median for one model and with the mean for the other model on both the train and test splits. Then, I scale the data to have a consistent measurement. I train the linear regression model on the training set and make a prediction. The R-squared performance is over 80% on both models on the train set and over 70% on the test set models. The MAE and MSE train test scores are similar, whether using median or mean to replace missing data. The model may be overfitting. Therefore, I use a cross-validation pipeline to find the best k, which divides the training set into k folds and trains the model on k-1 folds. The best k is 8. Vertical_drop has the most impact, and other essential features are Snow Making_ac, total_chairs, fastQuads, and Runs. With a random forest model, I impute the missing value with mean and median without feature scaling and split the train set into 5-fold cross-validation. On this model, the four dominant features are fastQuads, Runs, Snow Making_ac, and vertical_drop. The mean absolute error using cross-validation is 11.79 and 9.537 for linear regression and random forest, respectively. With a smaller error to almost $1 different, the random forest model is better.

Modeling

The model suggests $95.87 for a Big Mountain ticket price, which is higher than its current price of $81.00 because the model assumes that other resorts set their prices on a market-based basis.

Big Mountain can either cut costs or increase revenue and evaluate the estimated result for each scenario. The first scenario is close down to up 10 of the least used runs. Statistically, closing one run makes no difference, but closing from 2-3 runs reduces ticket prices and decreases revenue. Closing 4 or 5 runs have the same result as closing three runs, but closing six runs or more leads to a large revenue drop. The second scenario is increasing the vertical drop by 150 feet and installing an additional chair lift, while the third one is adding 2 acres of snowmaking. In both cases, the business can increase the revenue up to **$3,474,638** over the season. The last scenario is increasing the longest run by 0.3 miles and adding 4 acres of snowmaking capacity, which shows no revenue difference.