Guided Capstone Project Report

Introduction

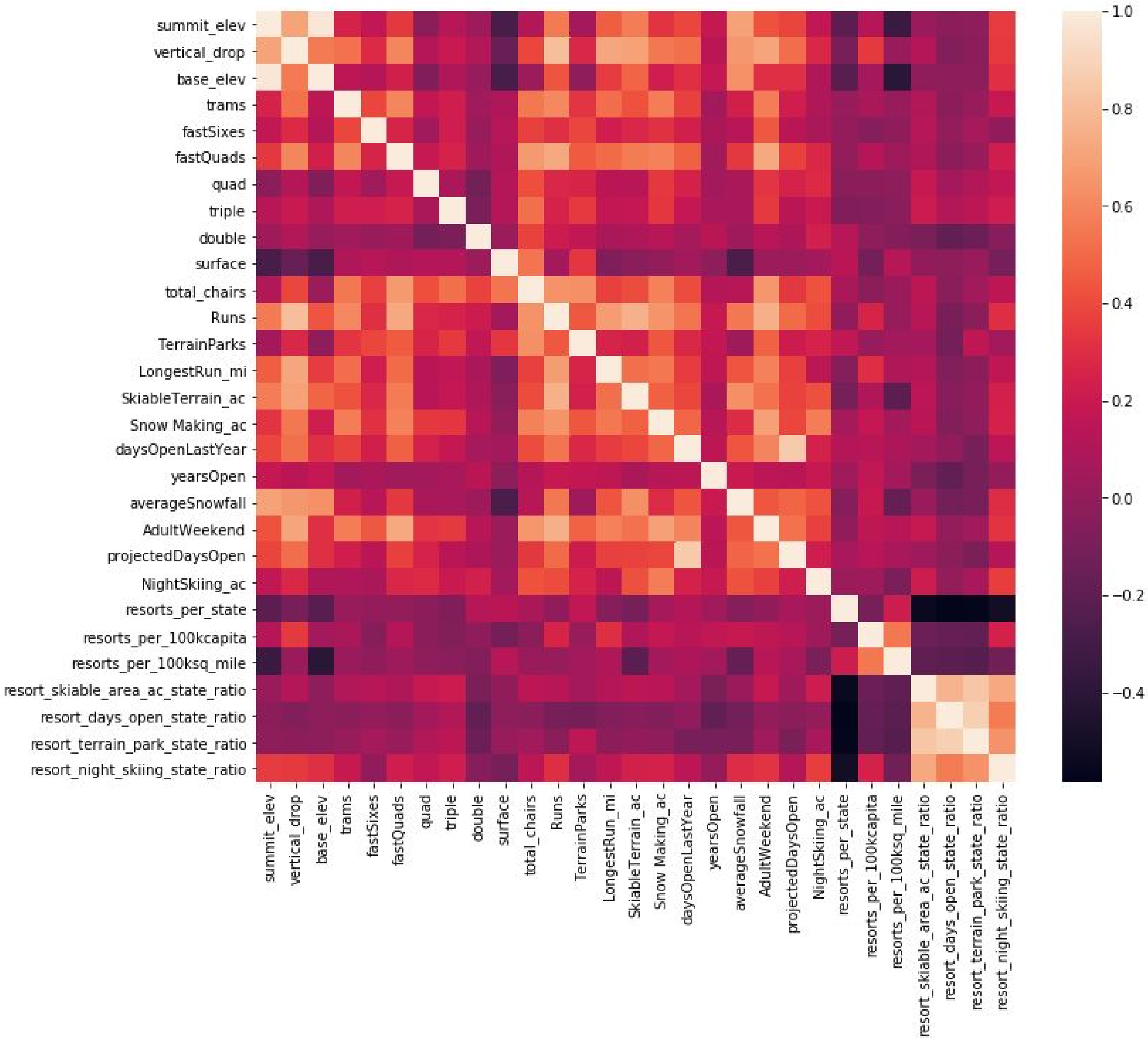
Big Mountain Resort is a ski resort located in Montana that includes beautiful views of the Glacier National Park and the Flathead National Forest. The services Big Mountain Resort provides are 11 lifts, 2 T-bars, and 1 magic carpet. Furthermore, the longest run is 3.3 miles, the base elevation is 4,464 ft, the summit is 6,817 ft, and has a vertical drop of 2,353 ft. In this upcoming season, they have installed an additional chair lift to increase the distribution of visitors on the resort. However, with the additional chair, this increases the operation cost of $1.54 million for the upcoming season. With the assets and services provided by Big Mountain Resort, their ticket prices are compared and analyzed to other resort ticket prices across the United States in order to find a ticket price model for Big Mountain Resort.

Data Wrangling

When loading the dataset and identifying what data is provided, many resorts are listed with their services as well as their ticket prices. Within the ticket prices that are listed, there are two types of tickets, one for the weekday and one for the weekend. When reviewing the correlation between Weekday and Weekend prices, there were enough similarities that the Weekday prices were dropped. Furthermore, any rows with missing data were concluded to be deleted since filling in without correct data will skew the model price. Specifically, there were 14% of rows that contained missing values and were dropped. I also imported another data set of state populations to gain a better insight to see if there are any patterns with ticket prices and population. Small adjustments were corrected such as the skiable terrain was corrected to 1819 to be more accurate. The Fast Eight column was dropped due to not providing good information.

Exploratory Data Analysis

To begin the analysis, I began to do general statistical summaries of each state to see if there were any patterns. I found that there are a variety of resorts among many states and that it would be best to know the number of resorts per 100,000 people and 100,000 square miles per state. Once these numbers have been included, the data was scaled and validated through mean and standard deviation with a pass in ddoff=0. I then found the Principal Cumulative Analysis (PCA) and found the first two components to account for over 75% of the variance as well as the first four for over 95% of the variance. Using the first two components, I added the ticking pricing to the scatter plot of the PCA. This allowed us to have a visual of the average ticket pricing. Although we had a visual, it was difficult to spot a pattern. Even though this portion of the data analysis was not enough to form a prediction model, it provided a sense of direction and gave an idea of each resort in each state. With more thorough data analysis of the resorts, I made a heatmap with each resort's asset to find a relationship between each asset. When focusing on the ticket price, the specific assets that seem to have a correlation with ticket price are the fastQuads, Runs, SnowMaking_Ac, and night skiing state ratio. This is a good indication to focus on these columns when making the predictor models.



Pre-processing and Training

4 categories were identified that played a crucial role to select the price of tickets. The first step after loading the data and filtering for usable data. The mean was calculated to identify if it was a good predictor on its own. The baseline model produced a R2 of 0 and in the test model, R2 is -0.003.

This was a good indicator that the mean was not a good model for this data set. In addition, the mean absolute error produced an error of around $18 in the ticket price and around $19 in the testing set. In creating a better model, I imputed missing data values with median on one and mean on the other. Using StandardScaler, the data was scaled so that the data will be consistent. Then linear regression was found to make predictions. When using the median for missing data, the training R2 was 0.82 and testing R2 was 0.72. On the other hand of using mean for missing data, the training R2 was 0.82 and testing R2 was 0.72. There were no differences and it did not matter, which value was used to impute for missing values. Furthermore, this model set identified an error of around $9 of ticket price. Pipeline was done to efficiently produce the identical results. To identify the best categories that affected ticket prices, I used SelectKBest and f_regression to refine the linear model. In order to this, the pipeline was redefined and the K value was tested to see which k value was best fitted. using GridSearchCV, the K value of 8 was found. With this value, we were able to see that Vertical Drop had the most impact on ticket prices and the least impact on ticket prices was Skiable Terrain. To use a more efficient model, we used a Random Forest Model. This helped identify that imputing the median value helps and the scaling feature does not help. We similarly found the vertical drop to be important in ticket pricing but was also able to visualize 3 other categories that were important. I compared the linear regression model and the Random forest regression model and found that the random forest model was much better. There was lower cross-validation mean absolute error by almost $1 and it had less variability.

Modeling

Big Mountain currently charges at $81.00 and the model price is at $94.22. The difference between the model price and the actual price is $13.22; however, the mean absolute error is $10.39. Even with the mean absolute error, there is room to increase the ticket price. Furthermore, the business has given options to increase revenue through increasing ticket price or to lower operation cost. I would approach the business with the model going through each scenario to show the changes. In the first scenario to close down up to 10 of the least used runs. If one run was to be closed, there will be no difference. If 2-3 runs are closed then this reduces support for ticket price, which affects revenue. If 4-5 runs are closed, then that is similar to the effects of 3 runs being closed. Anything from 6 or more closed runs, then increasingly drops ticket price and revenue overall.The 2nd scenario is to increase the vertical drop by lowering a run 150 feet lower, but this would need to install an additional chair lift to bring skiers back up without additional snow making coverage. The 3rd scenario is similar but adds 2 acres of snow making cover. The 2nd and 3rd scenario produce similar results in supporting a higher ticket price by $1.99, which overall increases the revenue to $3474638. In the final scenario to increase the longest run by .2 miles and guaranteeing snow coverage by 4 acres of snow making capability. This scenario has no difference on ticket price and revenue.