Austin Cody 10/13/23

Guided Capstone Project Report

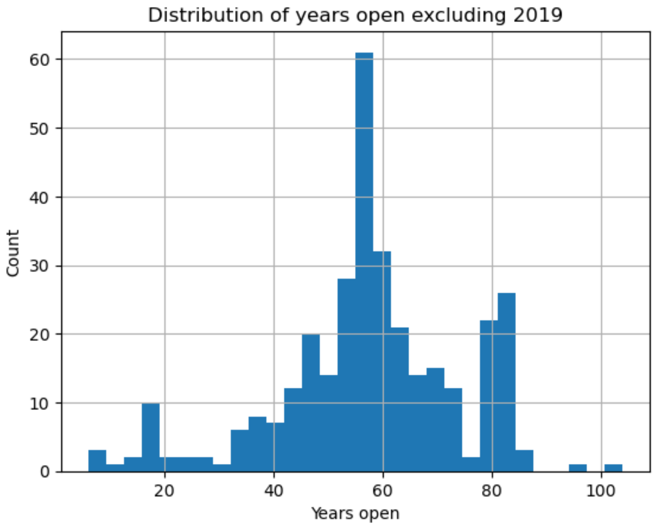
Overview

Big Mountain Resort (BMR) supports skiing and snowboarding for 350,000 customers in Montana. They charge a premium above the market price for chairlift tickets but want to use data to inform their pricing. Data on other resorts will determine if their premium is justified. If they charge 10% over market value for tickets, then through data analysis we can examine if they are providing a corresponding number more services in each category. Our plan is to build a model that will evaluate what their ticket price should be based on their facilities in comparison to other ski resorts in the market.

Data Wrangling

Our first step involved collecting, organizing, and cleaning our data. We were given a csv file by Big Mountain Resort's Database Manager. The originally generated table contained 330 rows of data and 27 columns. Our resort of interest, Big Mountain Resort, was present in the table and luckily was not missing any values.

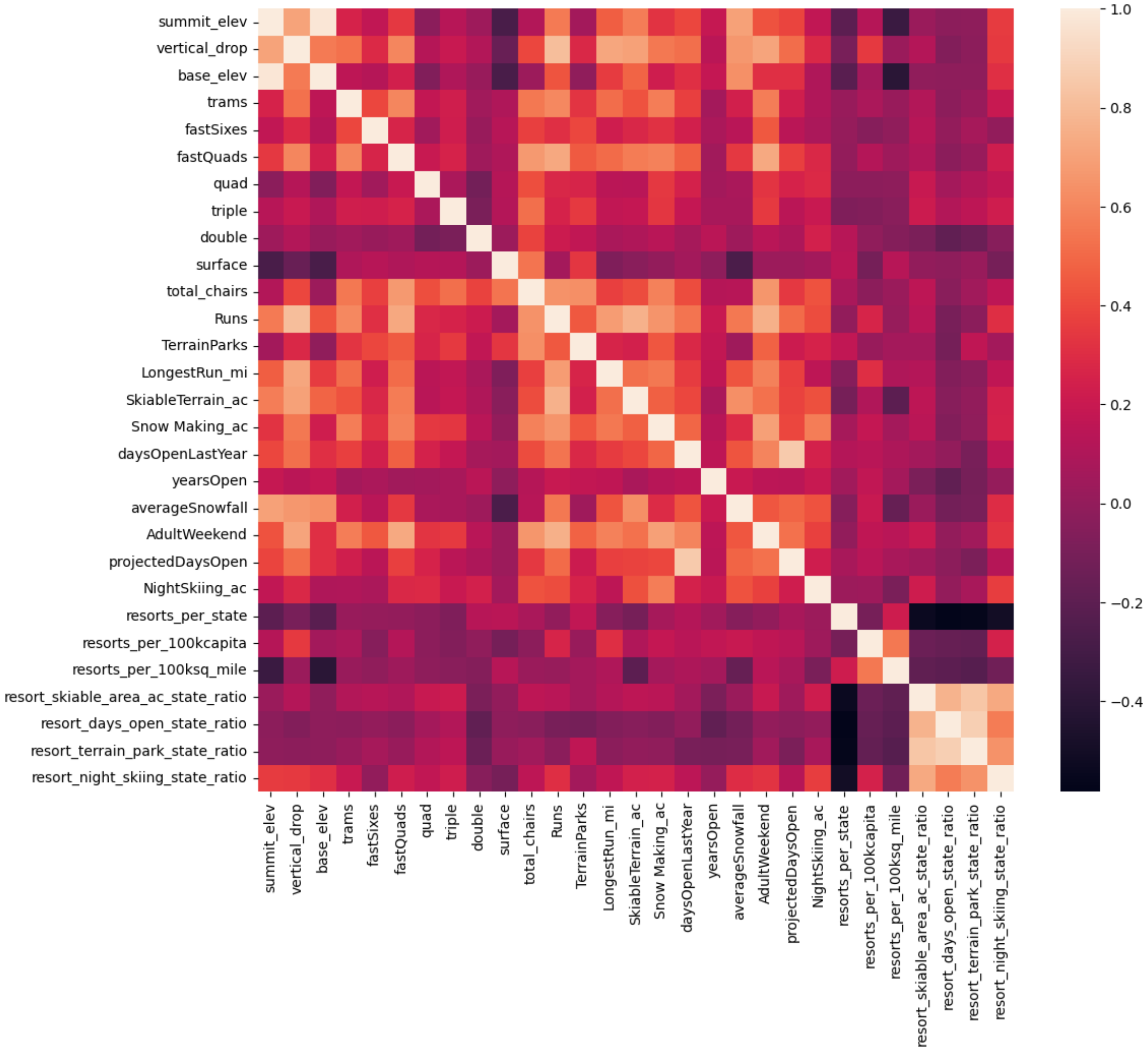
We cleaned the data by removing columns that were missing large percentages of their data, and we dropped rows that were missing values or that had obviously incorrect data, such as this column that, through visualization, we noticed had an abnormally high number of years open: 2019. It was severely skewed so we removed the value making it much more normal:



We manipulated our table to get a single price column and a categorical column for type of ticket. Of our many visualizations at this stage, perhaps our most useful one was our histograms of each feature value, which led us to observe outliers in a few categories and remove them. Eventually we obtained our final "clean" data table, ready for examination.

Exploratory data Analysis

Our next phase was visualizing the data such that we could get an idea for the relationship between ticket price and different variables, and a sense for the presence of resorts in different states and how that might be related to the population size, state area, and the skiable area of each resort. One of our most useful visuals was a heatmap displaying the correlation between different variables:



Through our heat map, we saw that our target variable, AdultWeekend (price), had correlations with a few variables: fastQuads, Runs, total_chairs, and Snow Making_ac. All of these are features that influence our price and thus will be useful in our predictive model.

Model Preprocessing with feature engineering

When selecting our model we started by splitting our data into a training set and a test set so that we had data to compare our model to later on.

Our first check for baseline performance was to check how well the average price did at predicting prices. As expected when we used the mean we got an R^2 value of 0.003 indicating that using the mean accounts for very little of the variability in our data. This was not good enough.

Next we replaced any missing values in our data set, scaled our features down so we could compare them against each other, and trained our data with a linear regression. This model accounted for much more variability in our data (higher R^2) than in our check using the mean price.

Algorithms used to build the model with evaluation metric

We refined our model using a pipeline and a feature selection metric called SelectKBest. For evaluation of our model, we used a technique called cross validation to run our model many times on our data constantly comparing them to a different test set. Consistent with our heatmap from the previous step, our cross validation techniques revealed that our four best features to use were vertical_drop, Snow Making_ac, total_chairs, and fast Quads. We then repeated these steps on a different model called a random forest regressor and found that it had a lower error and variability, and seems to be the best model to use on our data.

Winning model and scenario modeling

With our random forest regressor as our winning model, we used it to investigate the scenarios laid out by Big Mountain Resort and discovered that Scenario 2 was the best. This scenario involves adding a run, increasing the vertical drop by 150 feet, and installing a new chair lift. Our model predicts that this would support a $1.99 increase in ticket price.

Pricing recommendation

Our model also predicted what our price should be given our facilities offered compared to other resorts. Our model suggests $95.87 per ticket versus the actual price of $81.00.

Conclusion

Our model, based on the data provided, shows that Big Mountain may be undercharging on tickets based on their current facilities offered. Our model also shows that, of the scenarios provided by Big Mountain Resort, Scenario 2 would be best suited for increasing revenue and that the other scenarios cannot be recommended.

Future scope of work

In the future, with more data on other production costs as inputs we could provide more data to Big Mountain about how a scenario would directly affect its revenue. We only had the cost of adding a chairlift. But it would be useful to know the maintenance costs for each mile of each run, cost of adding snow, cost of adding trams.

Additionally in the future we could turn this entire model into a function that could be used by a business analyst at Big Mountain Resort so that they could look at how different features would affect ticket price and revenue on their own and run their own tests without having a data scientist make a new model every time they want to consider a new option.