Guided Capstone Project Report

1. Introduction

Big Mountain Resort is a ski resort located in Montana. It offers spectacular views of Glacier National Park and Flacthead National Forest, with access to 105 trails. Every year about 350,000 skiers and riders of all levels and abilities ski or snowboard at Big Mountain.

The resort's pricing strategy has been to charge a premium above the average price of resorts in its segment. But pricing on just the market average does not provide the business with a good sense of how important some facilities are compared to others. And there is a suspicion that Big Mountain is not capitalizing on its facilities as much as it could. They recently install an additional chair lift to help increase the distribution of visitors across the mountain which increases their operating costs by $1,540,000 this season. And they also consider a number of other changes. But the business wants some guidance on changing which factors will either cut the costs without undermining the ticket price or will support an even higher ticket price and how to select a better value for their ticket price.

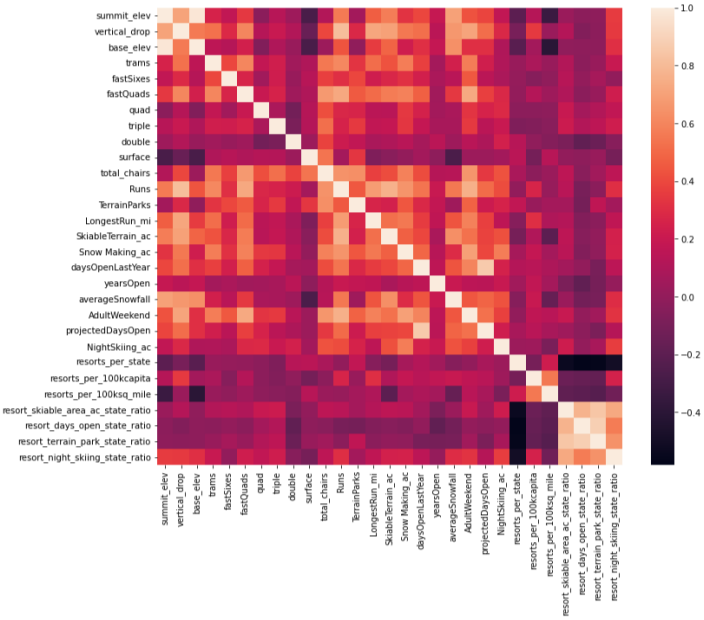
2. Dataset and Features

The data is from the database manager. The dataset consists of 24 features from 330 resorts in the US and those from Big Mountain Resort. In this dataset, we train and test our model with other resorts and then use the best model to predict the ticket price for Big Mountain Resort. Over 82% of resorts have no missing ticket price, 3% are missing one value, and 14% are missing both. We remove the 14% of rows missing both price values because price is our target, and these rows are of no use. Compared the data missing one value, we find the weekend prices have the least missing values, so we drop the weekday prices and then keep just the rows that have weekend price. Then, we merge the "state summary" features (the state population, the total skiable area of the state, the total open days, the total number of terrain parks, the total night skiing area, the resorts number per 100k capitals, and the resorts number per 100k square mile) with the dataset, and add the "state resort competition" features (the ratio of resort skiable area to total state skiable area, the ratio of resort days open to total state days open, the ratio of resort terrain part count to total state terrain park count, and the ratio of resort night skiing area to total state night skiing area).

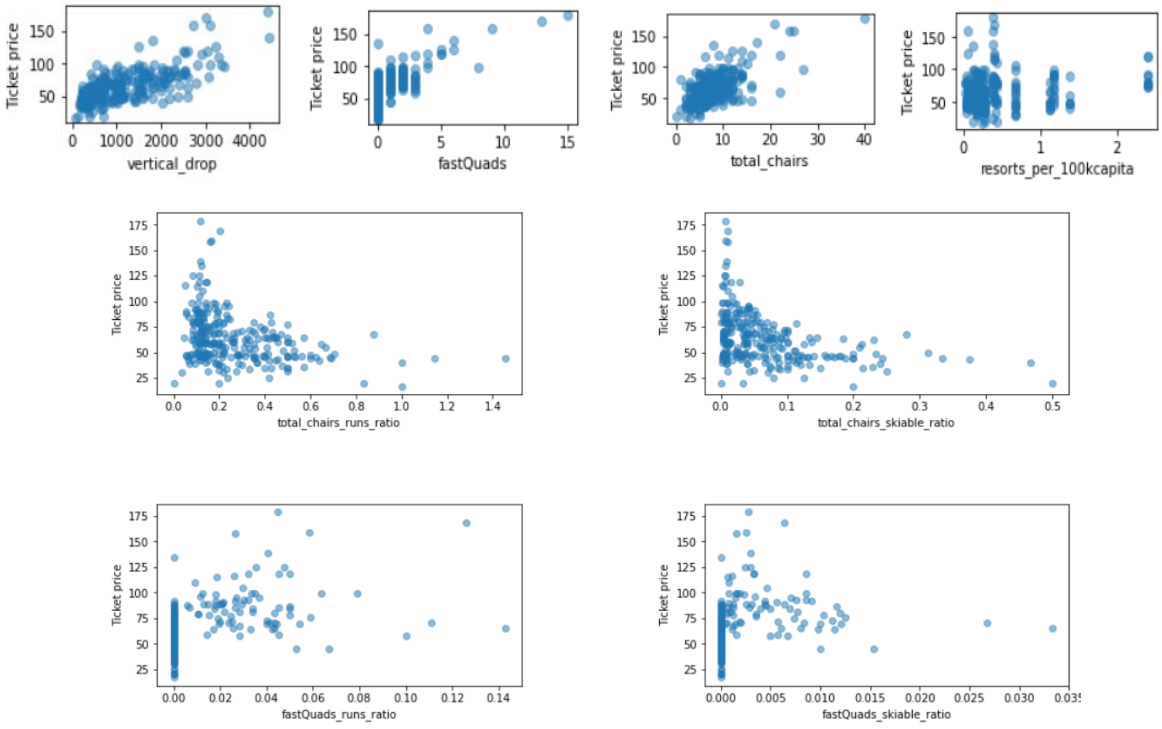
We also calculate the correlations between the ticket and its strong positive correlation features: the vertical_drop, fastQuads, Runs, and total_chairs as shown in Figure 1(a). To explore the details of the correlation between the ticket price and each feature, we explore the scatterplots (Figure 1(b)).

We can see that there's a strong positive correlation between ticket price and vertical_drop. As for the fastQuads, when the value is low, there is a variability in ticket price, and then the price climb with the fastQuads number. This means that the resort with more fastQuads has higher ticket price. To get it clear, we check the ratio of fastQuads and runs and the ratio of fastQuads and resort area. We find that having no fast quads my limit the ticket price, but the price may significantly increase when the resort covers a wide area. We also find the chairs number may be useful. At the beginning, the ticket price is low when a resort has more chairs. This means that you can charge more if you don't have so many chairs. This is quite counterintuitive. But here we noticed that more chairs don't mean that there are more visitors. Your price per visitor is high but your number of visitors may be low. Our dataset missing the very useful feature is the number of visitors per year. This is also important to explain the number of fastQuads feature. The resorts_per_100kcapita also show the unclear picture individually. When the value is low, there is quite a variability in the ticket price, and we cannot see the trend of the ticket price change with the number of resorts per capita. This may be also due to the number of visitors per year. The ticket price increases may

arise from more demands for skiing in some popular area, and the lower ticket price may be due to the small number of resorts or the less popular states.



(a)



(b)

Figure 1 Feature correlation heatmap and the scatterplots

3. Preprocessing and Model Selection

Here, we use cross-validation to set the train and test datasets in order to avoid overfitting. In this work, we use cv=5. And then, we design a pipeline to impute the missing value, scale the data, select the k best features, select the best model. When we use linear regression model, we find using 8 features gives the highest R^2 and the smallest error range. The 8 features are: vertical drop, Snow Making_ac, total_chairs, fastQuads, Runs, LongestRan_mi, trams, and SkiableTerrain_ac (see Table 1). The vertical drop is the biggest positive feature to increase price, and the area covered by snow making equipment is also strong positive, because people like guaranteed skiing. However, the skiable terrain area is negatively associated with ticket price. This is odd because usually people would like to pay less for smaller resorts. It could be an effect whereby larger resorts can host more visitors at any one time and so can charge less per ticket, but the data are missing information about visitor numbers. To improve our fitting, we also try Random Forest Regression method. Here, we explore the trees numbers through picking up 20 numbers from 10 to 1000 and try both mean and median as strategies for imputing missing values. We also try to standardize our features. The results show that the best feature number is 69, we need to use median to fill the missing value, and we don't need to standardize our features. Among the 69 features, the top 4 dominant features are in common with our linear regression model: fastQuads, Runs, Snow Making_ac, and vertical_drop (see Figure 2). Other features' effects are relatively low.

vertical_drop	10.767857
Snow Making_ac	6.290074
total_chairs	5.794156
fastQuads	5.745626
Runs	5.370555
LongestRun_mi	0.181814
trams	-4.142024
SkiableTerrain_ac	-5.249780

Table 1 The best features using linear regression model

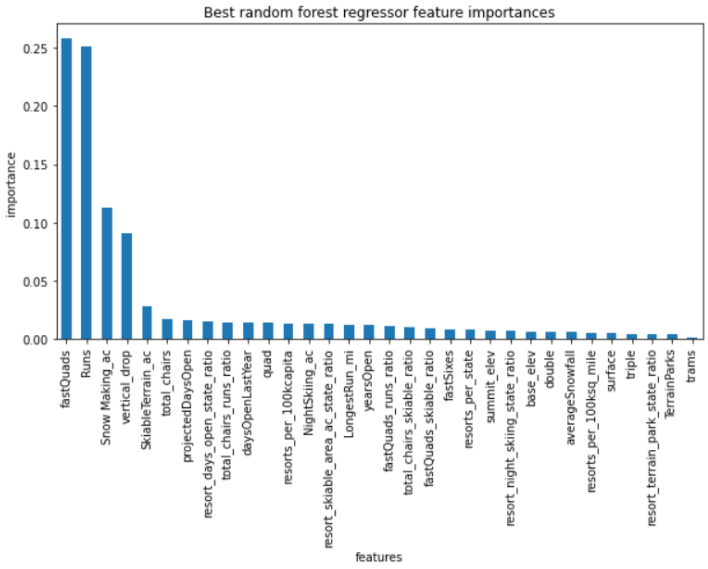


Figure 2 The best features using random forest model

4. Results Discussion

Here, we compare the value of Big Mountain and other resorts. Big Mountain is doing well for vertical drop, but there are still quite a few resorts with a greater drop. The snow making area and the number of chairs of Big Mountain are very high up the league table. Although Big Mountain has 3 fast quads, most resorts have no fast quads. Big Mountain also compares well for the number of runs. It also has one of the longest runs. Although it just over half the length of the longest, the longer ones are rare. Furthermore, it is also among the resort with the largest amount of skiable terrain. However, Big Mountain have no trams like most resorts. Here, we assume the expected number of visitors over the season is 350,000 and, on average, visitors ski for five days. When we close the runs, the ticket price and the revenue all reduce as shown in Figure 3. In this regard, we can add a run, increase the vertical drop by 150 feet, and install an additional chair lift. This can increase support for ticket price by $1.99 and over the season, this could be expected to amount to $3474638. However, adding the total area covered by snow making machines in the acres or the length of the longest run in the resort cannot help improving the price or revenue.

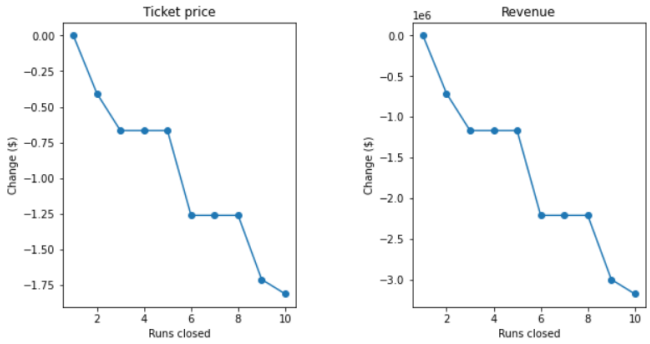


Figure 3 Relationship between the runs and Ticket Price/ Revenue

5. Outlook

In this model, we don't have the visitor number per year for other resorts, but this is a really important feature. It has strong relationship with other features and definitely impacts the ticket price. In addition, we only have the ticket price which is our revenue. But if we need to get the net revenue which is more useful, we need the operating cost of other facilities. For example, adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift can support for increasing ticket price. But we don't consider the operating cost of adding a run and increasing the vertical drop since we only have the operating cost of the new chair lift. The Big Mountain's current ticket price is $81. Our modeled price is $95.87, which is $10.39 higher than the current price. It could be that our model is lacking some key data, such as the operating costs and the visitors per year, or our model may be overtrained. To optimize our model, we can try to find the visitor number per year for other resorts and the operating costs. To find out the overtraining, we can try to reduce the R^2 and increase the mean square error for the train set. It the predicted value is closer to the current value, it is due to overtrain.