**Guided Capstone Project Report**

1. **Problem statement**  
   What opportunities exist to increase the revenue by $1.54 million for the Big Mountain Resort through selecting a higher ticket price or cutting costs with keeping the current price? Big Mountain Resort, a ski resort, has installed an additional chairlift to help increase the distribution of visitors across the mountain, which increases their operating costs by $1.54 million for this reason. But basing their pricing on just the market average does not provide the business with a good sense of how important some facilities are compared to others. So, they decided to set a new ticket price for the coming season. Set the ticket price (higher price or keeping the current price with cutting costs) before the coming ski season.
2. **Data Wrangling**  
     
   A graph of different colored columns

   Description automatically generated with medium confidence A screenshot of a graph

   Description automatically generated  
    Figure 1. Overview boxplot of raw data. Figure 2. All features’ data distribution.

Figure 1 shows some states show more variability than others. Montana and South Dakota, for example, both show small variability as well as matching weekend and weekday ticket prices. Nevada and Utah, on the other hand, show the most range in prices. Some States, notably North Carolina and Virginia, have weekend prices far higher than weekday prices. Then we remove missing values and correct outlier if we can. Figure 2 shows us that some feature’s data distribution has missing values or outliers, so we remove missing values and correct outlier. FastQuads, fastSixes, and trams features still have issue of value distribution, considering machine learning model, we need to focus on them if we apply those features to training our model or avoid overfitting. So, we may need to think about non-linear transformers of them.  
we added population and area values to our data. but the state's name of addition data needs to be corrected before merging.

AdultWeekday and AdultWeekend price are the same in Big Mountain Resort, thus we want to keep one of them. Finally, we found AdultWeekday has more missing values than AdultWeekend and we dropped AdultWeekday feature. So AdultWeekend is our target feature. 277 rows are left in the final data.

1. **Exploratory data Analysis**

For the state population and resort number, each feature cannot explain why some state with more resorts still has fewer resorts per capital, from this aspect, we need to introduce new density data to describe feature of each state like density in physics. The number of resorts per 100k population and number of resorts per 100k square miles are better for explanation. Next, we scaled the data which normalizes the data for PCA.

A graph of states with numbers and names

Description automatically generated with medium confidence

Figure 3. The plot of the two most components after PCA.

After PCA, we picked the two most important components those seem to account for over 75% of the variance, and added the average ticket prices per state data, to better understanding the relations between them. By scatter plot (figure 3), we simply do not see a clear pattern with price. In brief, we cannot cluster or group states only by two principal components, but we find potentially relevant state data in features most likely to be relevant to our business. So, we finally added states summary data to our ski data to dig more and added "state resort competition" features (ratios).

A screenshot of a computer screen

Description automatically generated A group of blue dots

Description automatically generated  
 Figure 4. The heatmap and scatter plots of all features.

A heat map was created to look for correlations in features (column variable). We find that the price is highly correlated with fastQuads, along with Runs and Snow Making\_ac. But Resort\_night\_skiing\_state\_ratio seems to be the most correlated with ticket price; runs and total\_chairs is also quite well correlated with ticket price.

After plotting price with all other features, we find there exists some correlations between them, including snow making, runs, vertical\_drop, fastQuads, total\_chairs. Aslo, we added some further features that may be useful in that they relate to how easily a resort can transport people around, including total\_chairs\_runs\_ratio, fastQuads\_runs\_ratio and so on. Finally, for better training our model, we choose those features those are highly correlated with price, which could include chairs, runs and having fastQuad lift or not.

1. **Model Preprocessing with feature engineering**  
   We regards price as dependent variable, others as independent variable, and use those independent variables (features) to predict price. First, we tried to take the average price as a predictive value for train and test data. We can use R-squared, mean absolute error, and mean squared error to estimate it. Then, we think median would be better for the imputing missing values, and we build a linear model to estimate it. We get a better result, also take R-squared (1 represents good prediction) for example, which value is close to 1 (train: 80%, test: 70%), much better, due to skewed data. We then used cross-validation to get our top 8 features (Figure 5) those are highly correlated with price, including vertical drop, Snow Making\_ac, total\_chairs, fast Quads, Runs, LongestRun\_mi, trams, SkiableTerrain\_ac, wihch are explainable. And random shows top 4 features on figure 6.  
     
   A graph with blue lines

   Description automatically generated  
    Figure 5. Best K features to apply.

A graph with blue and white text

Description automatically generated   
 Figure 6. Top features of random forest method.

1. **Algorithms used to build the model with evaluation metric**  
   After evaluating they test scores of the two models by cross-validation, including linear regression and random forest. Although we improved score on linear model, but the test split performance was not consistent with train set ((train: 80%, test: 70%)). Then tried random forest to avoid overfitting. Obviously using hyperparameter is better way to build model. Then, we got better train split score, which is close to test split score. The random forest model has a lower cross-validation mean absolute error by almost $1. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.
2. **Winning model and scenario modelling**

Cause the train split score of the random forest model is close to test split, and it also exhibits less variability. So, we choose the random forest model to go forwards.  
Look at where Big Mountain sits overall amongst all resorts for price and for just other resorts in Montana (figure 7).  
A graph of a number of tickets

Description automatically generated with medium confidence A graph with blue lines

Description automatically generated  
  
A graph of a vertical drop

Description automatically generated A graph of a graph

Description automatically generated with medium confidence  
  
A graph of chairs distribution

Description automatically generated A graph with numbers and a number of quads

Description automatically generated  
  
A graph with numbers and lines

Description automatically generated A graph of a running graph

Description automatically generated  
A graph with numbers and lines

Description automatically generated A graph of a mountain area

Description automatically generated with medium confidence  
 Figure 7. Main features comparison between Big Mountain and other resorts

The expected number of visitors over the season is 350,000 and, on average, visitors ski for five days. Assume the provided data includes the additional lift that Big Mountain recently installed.

The business has shortlisted some options:

1. Permanently closing to 10 of the least used runs. This doesn't impact any other resort statistics.   
     
   A graph of a price change

   Description automatically generated with medium confidence

Figure 8. Revenue change on Runs closed.

1. Increase the vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up, without additional snow making coverage. Increasing price by *$1.99* can increase **$3474638** on revenue.
2. **Pricing recommendation and Conclusion**  
   We found that the predicted price **$95.87** is **$14.87** more than actual price **$81.00**, which suggests there is room for we found that Big Mountain resort key features are all on the top left among the resorts. So, it's reasonable to increase the price of Big Mountain.

I suggest option 1 and option 2:  
Option 1: may close 4 or 5 runs to make well on cutting cost without hurting revenue. But we don't know the additional operating cost data, it's hard to say how much the revenue could increase. For this reason, we need to get operating cost data.  
Option 2: increasing price by *$1.99* can increase **$3474638** on revenue, but it needs additional installation of chair lift. We need to know this additional cost to calculate how many vertical we need to drop.

1. **Future scope of work**  
   First of all, we missed three important data: Ticket Price in weekdays, visitors of weekday and visitors of weekend, operating cost. We should consider if increase on weekday price will hurt the number of visitors. Or decrease on this price will attract more visitors. And without visitors of weekday and visitors of weekend data, we don't know revenue distribution which can help us to make better strategies on price change. The final one is operating cost. Without operating cost for each chair lift, we cannot calculate the accuracy revenue increase and how many runs to close.

And why Big Mountain's current price is much lower than the predicted model. Maybe some superior facilities have lower operating cost. There exists a room for price adjustment. We could use this model to build a whole application to let us and other user to set their own parameters to get better predictive price.