**Guided Capstone Project Report**

**Overview**

Big mountain Ski resort, which is in Montana gets 350,000 visitors for every ski season. They recently installed additional chair lifts to increase the visitors across the mountain. The additional chair lifts increased the operational cost by $1,540,000. Big mountain resort must implement a new pricing strategy to increase the revenue and reduce the operational cost.

**Objective**

To develop a machine learning model with the data from all the other resorts across nation to determine the ticket price which could increase the revenue and decrease the operational cost.

**Data Preparation**

The Ski resort data set had 330 rows and 27 columns. The columns were analyzed to find the percentage of missing values. The fastEight column had the most missing values, at just over 50%. Almost 82% of the ticket price, which was our column of interest was without missing values. Only 14% of ticket price values were missing. The rows with missing values in these ‘AdultWeekday’ and ‘Adultweekend’ columns were removed. These these columns TerrainParks, SkiableTerrain\_ac, daysOpenLastYear, NightSkiing\_ac were aggregated.

**Exploratory Data Analysis**

In terms of total area, Montana was the third largest state. When looking at the population, California was on top and seems to be densely populated. However, Montana was not on the top thus it was not densely populated. For the number of Resorts per state, New York was on the top with 33 resorts. Even though, New York had highest number of resorts it was not on the top of the list for Total Skiable Area. Montana was 4th on the list for the Total skiable Area. Colorado was on the top for the total number of days open. Then Resorts per capita and Resort per square mile was calculated. The state population and the area of the state columns were dropped. While looking at the Top states by resort density, Vermont was particularly high in terms of resorts per capita, and both New Hampshire and Vermont top the chart for resorts per area. New York did not appear in either!

A heatmap was created to show the correlation between the features. When looking at target feature, AdultWeekend ticket price, quite a few reasonable correlations were found. fastQuads stands out, along with Runs and Snow Making\_ac. Of the new features, resort\_night\_skiing\_state\_ratio seems most correlated with ticket price. If this is true, then perhaps seizing a greater share of night skiing capacity is positive for the price a resort can charge. As well as Runs, total\_chairs were quite well correlated with ticket price. Created scatterplot for each feature with the Ticket Price. There was strong positive correlation with vertical\_drop. fastQuads, Runs and total\_chairs appear quite similar and useful.

Chart

Description automatically generated

Figure 1

All the features were reduced using PCA to 2 and explained 72% of variance in the ticket price is visualized in below Figure.

Chart

Description automatically generated

**Data Preprocessing**

The data set was partitioned to 70% training and 30% testing set. The average price was a good place to start in predicting the ticket price. When we use the mean price for prediction, then ticket price might be off by $19. Then, The Linear regression model was created which explains over 80% variance on training set and 70% variance on testing set. he Mean Absolute Error and Mean Squared Error was around 9. When we use this model, the ticket price would be $9 off. When, Linear Regression model was developed using different subset of parameters model performance was reduced to 76% for training set and 62% for testing test. Finally, used developed a RandomForst model was developed, and which has the advantage of find the important feature.

Chart, histogram

Description automatically generated

The important features were displayed in a bar chart. The top 4 features were fastQuads, Runs, Snow Making\_ac, vertical\_drop. The absolute mean error for this model was $1. The Randomforest Regressor was selected was further modeling.

**Modeling**

The final data set was trained with the model the was developed before and it was evaluated by Cross\_validated with 5 folds. The average of Mean absolute error was around 10. Then, Big mountain data was used to predict the ticket price and it was $97.96. Even with an error of 10, the model suggested that there was room to increase price.

Overall, Big Mountain was on the higher end of ticket prices, within Montana they charged the highest ticket price. Then, developed the modeling scenarios and calculated the price based on different scenarios. First scenario was to close down up to 10 ‘Runs’. There was no difference in the price when we close one run. If Big mountain might have closed 4 or 5 runs then no loss in ticket price. However, closing 6 or more leads to drop in price. Second scenario was to add one more ‘Runs’, increasing ‘vertical drop’ by 150 feet and adding one additional ‘chairs’. This scenario supported increasing the ticket price by $2.22. Third scenario was to repeat the second with an additional 2 acres of snow making. The was no difference from the second scenario. Fourth scenario was to increase the longest run by .2 miles and adding 4 acres of snow making. There was no difference in the price. It appears that closing to 5 of the least used runs would be the best way to go as we charge less for ticket prices which does provide less revenue. However, it is not incurring costs to install a new lift, new snow making capabilities and additional costs of providing the new terrain which certainly would add to the operating costs and lower the revenues. The business could test the scenario of closing runs by ranking the runs based on usage and eliminating the runs one at a time to determine if ticket sales start to slide.

**Conclusion**

In this project, only ticket price was considered for modeling. Other costs should be considered for deciding the ticket price. The price of new chairlift, the price of adding skiable acres, cost of additional snow making should be considered. This facility seems to be rated high on most of the data, hence we could see that the model predicted higher price than the current price. It would be beneficial to develop a framework for where the business could input different combination of features to generate the ticket price. This could potentially help the business to find out how could they increase the revenue by combination of different feature.