**Guided Capstone Project Report**

Problem statement

How can the Big Mountain Resort overcome additional cost of $1,540,000 incurred by additional lift chairs through ticket pricing strategy and tuning available resort services within next two seasons?

Data Wrangling

Data Wrangling is one of the most crucial steps in the Data Science. In this problem, Initially, there were 330 rows and 27 feature columns. The column (fastEight) with many missing values was dropped. Also, data values were checked to see if they are making sense, incorrect values were replaced or dropped. Around 14% of the rows has price values missing for both weekend and weekdays, so they were also dropped. At last, we had 277 rows out of 330 rows. The distribution of all features was visualized using histogram, seaborn's boxplot, which gave sense if the data are correct. The scatter plot was plotted to see the relationship of target features (weekend price Vs weekday price). The Big Mountain Resort has same price for weekend and weekdays.

Chart, scatter chart

Description automatically generated

Exploratory data Analysis

Most importantly, the principal component analysis was performed. The key steps behind it are to first scale the data, fit the PCA transformation, and apply the transformation to the data to create the derived features. I visualized the data using heatmap and scatterplot for individual features.

I think adult weekend price should be the target variable. I did not yet see the concrete pattern suggested by a relationship between state and ticket price. I guess further modeling of data is necessary.

Algorithms, model Preprocessing, feature engineering, and evaluation metric

The train\_test\_split method was used to split train and test set in 70:30 ratio. The missing values were imputed with median/mean for both training and test sets. For estimating model performance, cross-validation for multiple values of k was performed and the value of k that gives the best performance was picked. I compared the performance of two models: linear regression model and random forest regression model by calculating mean absolute error using cross validation. It came out that random forest model gave the mean absolute error lower by ~ $1. Finally, cross-validation score as a function of training set size was plotted to know if the data, we have is enough. The test confirmed that we have enough data required for this problem.

Chart, box and whisker chart

Description automatically generated

Winning model, scenario modelling, and Pricing recommendation

Random forest regression model was used for this purpose. The main idea is that we want to train a model to predict Big Mountain resort’s ticket price based on data from all the remaining resorts.

Previously, Big Mountain Resort used to charge $81 for the ticket price. **The expected ticket price obtained from the model is $98.85**. So, business leadership are strongly recommended to increase the ticket price.

It is seen that adding one run with vertical drop of 150 Feet and 1 additional lift chair increases support for ticket price by $1.36 and assuming each of 350,000 expected visitors buys 5 tickets for their expected five days of stay, this will increase the revenue by $2,386,364. This will alone cover the additional operating cost of $1,540,000.

Additionally, up to 8 least used runs can be closed as this doesn’t change much the revenue. Closing only 2-3 runs on the other hand can even increase the revenue.

Conclusion

The business leadership should consider increasing the ticket price at least 10 % from the current price. It is also strongly recommended to close few least used runs.

Future scope of work

Having complete data would enhance more balanced and systematic study and hence more mature proposition can be made regarding the new ticket price. key data, for example operating cost is missing in our dataset. The validity of our current model lies in the assumption that other resorts accurately set their prices according to what the market supports.

Also, apart from random forest regression model, we should also try gradient boosting regression and support vector machine regression models. One of the advantages of gradient boosting is that it can handle both categorical and continuous variables, whereas support vector regression model allows for non-linear relationships to be captured.