BIG MOUNTAIN RESORT – PRICING STUDY

The Big Mountain Resort is one of many beautiful skiing resorts in the United States. It is in Montana with a very large skiable terrain of about 3000 acres. They have multiple facilities such as lifts, T-bars, and a magic carpet. These facilities have high operating costs which expose the resort to the risk of making a loss. Furthermore, the resort is not sure about whether to get more facilities or which facilities they should procure if they do decide to get more facilities.

The pricing strategy currently used at Big Mountain Resort is a generic one that uses the average pricing of all other resorts in the United States. This pricing strategy does not consider any key facilities that may be unique to the resort, so even if there are highly demanded facilities in the resort, the pricing of tickets remains the same regardless.

The goal of this project is to produce a more ideal pricing strategy that not only provides a good price based on a set of high impact features for the resort, but also informs which facilities would be best to procure in order to increase profits optimally.

The project was done using python’s Jupyter Notebooks. The dataset was collected from a GitHub repository at <https://github.com/Thaps/DataScienceGuidedCapstone>. It was comprised of about 330 skii resorts across the United States with various features including the State, Region, and specific available facilities.

Another dataset for populations in states was also brought into the study and it was merged with the original dataset in order to get some insights that may not have been there otherwise. All the states that were in the original dataset were also in the additional dataset.

The data was not flawless, it had a few issues that needed to be addressed. A lot of missing numeric values were filled in using median and one specific column called ‘fastEight’ was completely removed from the dataset because half of its values were missing.

After cleaning the dataset, it was put through principal components analysis (PCA) to reduce its dimensionality. According to the pca, the top two components accounted for over 75% of the variance in the dataset. The two features that were identified to contribute largely to these two components were resorts per 100k people, and resorts per 100k sq. miles.

Next, some feature engineering was done to farther reduce the dimensionality of the data. As a result, the following features were produced:

* Resort skiable area to state skiable area ratio
* Days open last year to state total days open ratio
* Terrain parks to state terrain parks ratio
* Nightskiing to state total night skiing ratio

A heatmap was produced that showed a lot of strong correlations between these features.

The next stage of the project was to model the data using a simple linear regression model. As a baseline, the mean of all the resort ticket prices was used to predict all other ticket prices. As expected, the R^2 value was approximately zero. The mean absolute error was about 19 and the mean squared error was approximately 581.

The simple linear regression model used a pipeline that included imputing the data with the median, scaling the data, and using the optimal number of features which was 8 according to the grid search cv model. The other type of regression that was used was the random forest regression model. The results for the simple linear regression model were slightly inferior to the results of the random forest regression model.

Simple Linear Regression Model:

MAE = 11.79

Std = 1.62

Random Forest Regression Model:

MAE = 9.54

Std = 1.35

After modelling the data, a few scenarios were made to see how changing various features could affect the predicted price. As expected, the most significant features identified during the exploratory data analysis needed the lowest percentage change to make significant change on the predicted price.