GUIDED CAPSTONE PROJECT REPORT FOR BIG MOUNTAIN RESORT

BACKGROUND

Big Mountain Resort, a ski resort located in Montana offers spectacular views of Glacier National Park and Flathead National Forest, with access to 105 trails. Every year about 350,000 people ski or snowboard at Big Mountain, it can accommodate skiers and riders of all levels and abilities with the service of 11 lifts, 2 T-bars, and 1 magic carpet. The longest run Hellfire is 3.3 miles in length, with a base elevation of 4,464ft, a summit of 6,817ft, and a vertical drop of 2,353ft. Big Mountain Resort has recently installed an additional lift chair to help increase the distribution of visitors, however, the chair increases their operating cost by $1,5400,000 this season. In response to the cost change, the resort's pricing team has decided to charge a premium price for the tickets. The price increase criteria will be analyzed from collected data of other resorts across the country to maximize the profit margin.

PROBLEM STATEMENT HYPOTHESIS

What recommendations exist to cover up the additional operating costs of Big Mountain Resort by $1,540,000/- over the season? Also, to forecast current years' revenue if the recommendations are implemented.

METHODOLOGY FOR THE SOLUTION

The Data Science Method (DSM) steps as shown below adopted to develop and analyze the ticket-pricing model for the Big Mountain resort compared to other resorts in their market:

RECOMMENDATIONS

The ski data and state summary data were explored and resort density per capita and area were calculated. Then Principale component analysis (PCA) was performed to find linear combinations of the original features that are uncorrelated with one another and order them by the amount of variance. The categorical features were names of the resort, its region, and state, the numerical features are numbers of chairs, vertical_drop height, number of runs, etc.

Based on seaborn heatmap of correlations, fastQuad, Runs and Snow Making_ac were observed to have correlations with the ticket price; based on scatterplots of price with other numeric features, the vertical drop has a strong positive correlation with the ticket price. These features that were observed should be used for Big Mountain Resort ticket price modeling.

The first model is a baseline performance comparator for any subsequent model; a 70/30 train/test split is performed on the ski data. Three different metrics were used: R-squared, mean absolute error and mean squared error. The R-squared on the test set is - 0.00312, it is expected to be negative since performance on the test set generally is slightly worse than on the training set (in this case 0). The mean absolute error is higher on the test set than on the training set, but the test set's mean squared error performs better than the training set. The initial model started with imputing missing feature values, where both median and mean methods were used. The imputation was applied to both train and test splits and the data were scaled for them to be on a consistent scale. Those data were used to train the linear regression model to make predictions. The R-squared performance was around 0.8 for both median and mean impute on the training set and around 0.7 for both test sets. MAE and MSE were similar for both the mean and median filling methods for the training set and test set. This means that the result would be similar no matter whether using mean or median to fill missing values. To verify if the model was overfitting, the SeletKBest function used to select k best features and the score function will be f_regression. After using the cross-validation technique to estimate model performance, GridSearchCV was used for hyperparameter finding that the best k is 8. After displaying these eight features, we can observe that vertical drop has the biggest positive value, and the area covered by snow-making equipment is also strongly positive but trams and skiable terrain are negatively associated with the ticket price. A random forest model was performed to test if it fits in this case, cross-validation will use the default setting so we can investigate different hyperparameters. In this model, the dominant features are fastQuads, runs, the area covered by snow-making machines, and vertical drops. By comparing its performance with the linear regression model, the random forest model has lower cross-validation mean absolute error by almost $1 and exhibits less variability so it would be the model I have decided to use going forward.

I want to draw attention to Big Mountain Resort's current ticket price of $81.00. My model suggested that to Big Mountain Resort should be charging ~$95.87. Now, the model has a variety of errors equal to $10.39; meaning that Big Mountain Resort certainly has room to increase its price, but the $95.87 is just the model's prediction of the ticket price and the price should be within +/- $10.39 of the predicted price. Overall, as facilities currently stand, without factoring in final changes, an increase in the ticket price is already justified.