Data-driven strategies for ticketing values and operation cost for Big Mountain

Kiman Park

**Overview**: Statistical analysis and machine-learning modeling are performed to identify strategies to increase ticketing values by 10% and reduce the operation cost by 10%. We have identified important features contributing to an increase in the ticket price as well as possibly reducing the operation cost, such as fastQuads, Runs, Snow Making\_ac, vertical\_drop, and SkiableTerrain\_ac. Using the Random Forest model, we were able to calculate an expected price of $95.87, which is an 18.36% increase compared to the actual ticket price of $81.00. Furthermore, a reduction of the longest run, snow coverage, and snowmaking might decrease the operation cost, which will increase revenue as well. However, additional data collection and analysis are needed to make a more realistic and accurate prediction of the business model. Several features are needed such as the operation cost of individual parts in the resort, the one-time cost of setting up the equipment/land, and other costs that might be important. With this, we will be able to create realistic scenarios to simulate the options for facilitating the decisions and bettering the business.

**Methods:** The methods included data wrangling, exploration data analysis, model preprocessing with features engineering, building models and evaluations, and creating scenarios to evaluate the strategies.

**Results:**

**Data Wrangling** – First, we loaded the data and looked to see if any data was missing. The numerical data had several variables that were missing. The fastEight variable was missing 50%, which was the most, followed by the NightSkiing\_ac with 43%. The ticket price variables were also missing 16% and 15% missing also (AdultWeekday and AdultWeekend, respectively). Surprisingly, categorical data were not missing or duplicated. Then, we explored the relationship between the average ticket price by the state, which did not yield any significant relationships. We also explored distributions of numerical features and identified variables with anomalies, which were corrected in this process. Furthermore, we collected population data from Wikipedia, which we cleaned and saved with the ski resort data.

**Exploratory Data Analysis** – Using the data from above, we calculated the resort density per 100k population and 100k square miles, which gave us some ideas about the competition in each state. Then, we scaled the numerical features and verified the scaling by observing the mean and standard deviation values. The mean value was correct, but the standard deviation had to be corrected with ddof=0 since the default value was 1, making the standard deviation value drift a little. Principle component analysis was used to reduce the dimensionality of data, which revealed that the smallest k over the threshold of 95% is 4. We displayed the first two components to verify 77% variance using matplotlib and seaborn. Then we merged ski resort data with state summary data and calculated several ratios. To gain insights into the correlation among variables, we graphed a correlation heatmap, which revealed various relationships with the ticket price. Then, we graphed a scatterplot of the ticket price as a function of each variable and verified the correlations in the heat map.

Preprocessing and Training Data – The training and test data set were split 70% and 30%, respectively. First, we calculated the mean value for a simple check with the dummy regressor. Then, we calculated various metrics, such as R-squared, mean absolute error, and mean squared error. We also tested these metrics with sklearn package as well. For the initial model, we used the median to fill in the missing data, scaled the data, fitted the data with a linear regression model, and predicted the ticket price. Upon the assessment, R-squared values for the train and test sets were 81.78% and 72.09%, respectively, indicating that the model was overfitted. Also, the mean absolute error revealed that the estimation can be within $9, which is better than guessing with the average value. We performed the same process with the mean value for filling in the missing data. We achieved the same results more or less. The pipeline was created and verified the whole process, which made the prediction worse by looking at the metrics. We redefined the pipeline using a different number of features, cross-validation, and hyperparameter search using GridSearchCV. We have identified the best k value as 8. Furthermore, we used a random forest as a model with the same process, which displayed better metrics. We also identified important variables in relation to the ticket price, which were fastQuads, Runs, Snow Making\_ac, Vertical drop, and Skiable Terrain. A graph with blue and white text

Description automatically generated

We selected a final models for the linear regression and the random forest model, which performed well with mean absolute error of almost $1. Furthermore, we verified that the data quantity is plenty by using learning curve as a function of training size.

**Modeling:** Using the models above, we refitted the data without Big Mountain and calculated the expected ticket price, which displayed $95.87. This prediction is an 18.36% increase from the actual price of $81. Then, we graphed to verify that claim, which intuitively seems right as Big Mountain is among the top for several important variables. Finally, four selected scenarios were examined, which revealed several important aspects. Closing the runs decreases the ticket price, also the revenue. However, the operation cost for the run is not included, which should be reviewed again once the operation cost for the run is collected. The installation of chairs increases the ticket value. However, the cost of the installation and the operation cost are not considered. The long-term view is needed here with time-series modeling. An increase in snow making does not seem to affect the ticket price. However, decreasing the area of snow making should be explored with the operation cost, which might increase the revenue. Finally, increasing the longest run does not seem to affect the price either. Decreasing the run might be a viable option as well in order to reduce the operation cost as long as the ticket value is not affected.

**Future Work:** The additional data collection is needed to achieve more realistic and accurate model of the business overall. Further calculations of accumulation of revenues are needed as well. As mentioned above, operation cost and one-time installation cost should be included to take account of the revenues and strategizing with short-term vs. long-term views.