**Mark Kincaid**

**UMGC Data Science Bootcamp**

**Guided Capstone Step 6**

**Background.** Big Mountain Resort, a Montana ski resort recently installed an additional chair lift to increase distribution of visitors across the mountain. This investment increased operating costs by ~$1.5 million, and the resort has asked a data science team to examine the pricing structure to optimize ticket revenue and operating costs.

**Problem Statement.** The team will recommend a pricing strategy to Big Mountain Resort aimed at recouping $1.54 million in expansion-related operating costs this year. The recommended strategy will aim to replace the current ticket-pricing strategy, which uses market average ticket price as basis. A successful strategy will maintain current profit margins, highlight market facility/amenity advantages, and forecast anticipated annual revenue impacts for future seasons.

**Methodology.** This project followed the standard progression of a data-informed examination, beginning with “Data Wrangling,” or examining and preparing available data for further analysis, followed by “Exploratory Data Analysis,” or an initial stage of analysis. Insights gained from this phase made it possible for the team to conduct “Model Preprocessing and Feature Engineering,” to build predictive toolsets for the data to become useful in solving the stated problem. The team developed and tested varied algorithms, resulting in a determination of the most accurate model. The selected model successfully generated an accurate pricing recommendation, and identified remaining information gaps to generate recommendations for future scope of work.

**Data Wrangling.** The original dataset comprised of 330 rows, in which the subject resort was included. The dataset was modified to remove data values in that column found to be of no use with either missing, irrelevant, and null values, with a focus on keeping as much available data related to ticket pricing as possible. Duplicate values were cleaned and cross-checked using online research, ensuring that each resort in the available data was a unique entity. Cleaned in this manner, the dataset is composed of 277 rows. This phase initially revealed a ticket-pricing structure higher than the current levels, and identified relevant amenities as possible discriminators for ticket pricing during further examination.

**Exploratory Data Analysis.** The examined dataset cross-referenced features of all U.S.-based resorts, such as geographic location, demographic data of that locale, average ticket prices, average annual snow fall, total skiable terrain, etc., to form a heat-map highlighting correlation of these features with ticket price. This heatmap suggested a strong positive correlation between the variable features of ticket price, vertical drop, snow-making acreage, number of chairs and total number of runs. This processA screenshot of a computer screen

Description automatically generated revealed a negative correlation between state population and density of resorts, suggesting that low geographic density in the number of resorts can push ticket prices higher. These correlations led the analysis to model the relationship between ticket price, and the following ratio-integrated variables: total number of chairs and the total number of runs; total number of chairs and the skiable acreage; total number of fastQuad lifts and the total number of runs; the total number of fastQuad lifts and the skiable acreage.

**Model Preprocessing with Feature Engineering.** Baseline understanding of average price was established by authoring two model sets; a “train” set, and a “test set. These models were examined by determining the mean price of the train set, and comparing this simple calculation against sklearn DummyRegressor. The result was an unacceptable margin for error, requiring the use of a more accurate (linear) model. To do this, the median value is used to predict and input missing values into the test set, scale the data, and select the best features to focus on for further examination.

**Algorithms and Evaluation Metrics.** The examination proceeded by cross-validating a linear regression and a random forest regression for accuracy, with mean absolute error as the discriminating variable. The linear model demonstrated a higher degree of variability, yielding differing results depending on the quirks of the data points in each fold. Testing for mean absolute error determined that sklearn RandomForestRegressor had a lower cross-validation mean absolute error than did the linear regression model by ~$1 while exhibiting less variability. This examination took a further step to determine how the study would be impacted by increased data, resulting in the determination that any gains from increased sample size would be marginal, at best.

**Chosen Model, Conclusion and Pricing Recommendation.** The random forest regressor was chosen for its more accurate and precise output, as determined by the lower cross-validation mean absolute error. The model revealed that there were no differences in pricing models resulting from increases to the longest run or snow coverage area, and it may even increase overall operating costs. Adding additional run however, increasing vertical drop, and adding additional chairlift would allow for an increased price by $1.33 dollars, as well as the annual revenue by over $2.3 million.

A graph with blue and white text

Description automatically generatedThe data examination additionally made a compelling case against the option of closing runs to reduce operating costs. The model shows that closing one run had no impact on revenue and ticket price, only marginal improvements when closing between two and four runs, with significant drop-off of revenue thereafter.

The data suggests that, with an assumed annual visitation of 350k customers, purchasing on average, 5-day passes, the additional investment of another chairlift will increase competitive advantage of Big Mountain Ski Resort, and the additional revenue (~$2.3 million) will reflect it. Such an investment does not come without the associated operating costs related to the maintenance of such an infrastructure improvement.

**Future scope of work.** Operating costs associated with each recommended capital improvement were not available in the examined data set. Were access to such data possible, the business could examine the impact of ski run closure on the total operating costs of the mountain (presumably including everything from ski-lift operators maintainers, and groomers, increased Ski Patrol, etc.). This analyst presumes that the problem in this knowledge gap, is the potentially proprietary nature of operating costs, which separates it from the data under current examination. Such data, which is subject to the efficacy of management practices, may not be publicly available.

The modeling of this data provides a prediction of increased revenue after the execution of infrastructure improvements on the resort. A business analyst will likely find this information useful in the conduct of a feasibility study on whether or not the cost of those infrastructure improvements sufficiently bost the return on investment (ROI), given additional operating costs associated with the improvements, to justify the injection of capital into those improvements.