Guided Capstone

Project Report

Svyatoslav Andriyishen

# Introduction

Big Mountain Ski Resort is located in Montana, has yearly customer base of around 350,000 people that come to enjoy its 105 trails, 11 lifts, 2 T-bars, and 1 magic carpet for novices. It has recently installed a new ski lift that increases the operational costs by $1.54 million this season. As a result, the business is evaluating its ticket pricing strategy. Up to now, the strategy was to charge a premium above the average price in the market segment, which makes it difficult to understand the value of various facilities and create a sound investment strategy. There is a need to come up with a more data-driven approach to ticket pricing.

# Methods and data exploration

The dataset obtained from Big Mountain Ski Resort’s stakeholders has information on 330 resorts, Big Mount among them. However, 52 resorts were removed due to missing target variable of ticket price and one additional resort was removed due to a typo in the years open variable that could not be made sense of within the data.[[1]](#footnote-1)

Weekend ticket price was selected as the main target variable because it had less missing data than weekday ticket price and there was no evident other benefit of selecting weekend vs weekday ticket pricing. One had to be selected as some resorts had set price regardless of the day of week. Furthermore, Montana, where our resort is located, had no difference in weekend and weekday ticket pricing.

fastEight column, representing presence of a specific type of lift at a resort, was removed because it was not informative, with only one positive value and having over 50% of its data missing.[[2]](#footnote-2) Region and state variables have too much overlap and were flagged for further examination along with numerous column data based on examination of their distributions. For the purposes of this project, state variable was preferred over region. Silverton Mountain's skiable terrain variable was flagged as an outlier and upon examination was determined to be a typo that was corrected.[[3]](#footnote-3) Finally, the data on skiable terrain, night skiing, days open per year, and number of parks within the resorts were summarized by state and added back to the data as variables. State population data was scraped from Wikipedia and added to the dataset.

The heatmap of correlations among variables provided initial insight, and along with PCA, pointed to importance of variables in the data. It showed that excluding some variables would be beneficial. State summaries’ data was recalculated as ratios to add some granularity to the data, and number of runs, snow-making, total chairs, fast Quads, longest run, skiable terrain, and days open or projected days open were identified as most correlated with ticket price, along with night skiing in some circumstances.[[4]](#footnote-4)

# Modeling

The data was split into a standard 70%/30% train/test buckets. The train data was then fitted into a naive means prediction, linear model, and random forest, in that order. Missing values were imputed at each step with a mean and median estimator, of which the median provided better performance. Standardizing around mean=0 also yielded better results, except in the random forest prediction.

The mean naive prediction was used as a benchmark to compare and the other models performed much better. In order not to overfit, 5-fold cross-validation was used for linear and random forest models as well as K-best features selection method to find the optimal number of features to include and the random forest was also optimized on the number of trees trained. Data quantity was also assessed using a learning curve method to see if more data is needed to yield better cross-validation results. It was found that there is more than enough data (60 rows would be sufficient for good results) and random forest yielded the best prediction with 69 trees and mean absolute error of roughly 9.5 on both the training and test data, with standard deviation of less than 1. Linear model was rejected because of its weak performance as well as test data result being on the high end of standard deviation from prediction, suggesting that it would not perform as well on new data.

# Result of Scenario Application

Big Mountain Resort currently charges 81 dollars for weekday and weekend tickets. According to the random forest model we trained, there is evidence to increase the price to up to 95.87 dollars. The mean absolute error was 10.39 dollars in the model, but even accounting for it, we can still increase the price safely by at least 4 dollars. Moreover, the current ticket pricing for Big Mountain Resort is the highest in the state of Montana and is justified to be so due to the Resort's leading numbers in the major predictors of ticket price identified by the model, such as the vertical drop, snow making area, total number of chairlifts and fast quad lifts, number of runs and the length of the longest run, and skiable area when compared to the national ski resort offerings.

On the other hand, of the four suggested cost-cutting scenarios:

1. Increasing the longest run should not lead to higher profits and is not recommended.
2. Closing the least popular run should have no effect on ticket pricing, but will surely save money on maintenance and is therefore recommended. Closing additional runs will have to be reviewed on the basis of profit to maintenance trade-off as per the chart below as the ticket pricing will have to be reduced.

Chart, line chart

Description automatically generated

1. The remaining two scenarios involve adding a run and the associated chairlift and thereby extending the vertical drop by 150ft, with the difference being adding or failing to add the associated snow making equipment to cover 2 acres of the run's skiable area. Scenario without the snow making equipment is wholeheartedly recommended, as it supports an increase of 8.61 dollars on the ticket price, and based on the expected seasonal patronage of 350,000 people with an average of 5 days spent at the resort, would yield an additional profit of 15,065,471 dollars. That should be more than enough to cover the estimated 1.54 million dollars of expenditure for operating an additional chairlift.
2. By adding the snow making equipment to the previous scenario, we can justify an increase of 9.9 dollars to the ticket price and a profit of 17,322,717 dollars. Bearing in mind that these are estimated numbers, if the snow making equipment is significantly cheaper to operate than the ~2 million dollar difference in profit between the scenarios, then this option is recommended instead.

# Potential Problems and Future Directions

The ticket pricing strategy outlined here was hampered by a number of factors:

a) Competitiveness was not properly addressed from popularity perspective: there was no indication of the number of seasonal visitors in the data, which is likely to be a big factor in ticket pricing.

b) Competitiveness was not properly addressed from geographical perspective: in addition to seasonal patronage, examining number of out-of-state vs in-state visitors would be beneficial. The resorts in the data could be further categorized by their specialty and geography.

c) In conjunction with a) and b) above, those insights could be better informed through temporal variables that would give us a clearer picture of historical trends and patterns of visitation. Examining historical data would also help us confirm whether the facilities historically impacted ticket pricing, and if ticket pricing fluctuations impacted visitation, as all of these factors influence each other.

d) We only had operational costs for the new suggested chairlift, but the other operational costs could help us to determine true projected profits from modifying the facilities.

e) Finally, marketing and ad spend should definitely be included as they are important for a travel-based seasonal destination like the Big Mountain Resort.

f) All of the above is supported by a modest amount of variance (about 30%) and a relatively high absolute error we have remaining after doing the modeling. We certainly have enough rows of the data but could benefit from a few more targeted columns based on the above.

# Conclusion

The above being the case, there is enough evidence to use the trained model to make business decisions (and in any case, it has been demonstrated that it would be better to use this model than randomly guess or approximate effects with means or medians). We could provide the business stakeholders with a comprehensive report based on achievable scenarios or build a dashboard for analyzing those scenarios on their own based on the selected final model as next steps and while we address the issues outlined in the previous section. However, for the latter option, it would be useful to include the operational costs to determine true profit increases or decreases before attempting it.

1. This should be followed up on with the client and the data provider. [↑](#footnote-ref-1)
2. Although fastSix is more informative in its distribution, it also has many missing values and may be removed. [↑](#footnote-ref-2)
3. This should be followed up on with the data provider. [↑](#footnote-ref-3)
4. These are prone to autocorrelation with each other and should not be all included in the model. [↑](#footnote-ref-4)