Data Report: Big Mountain Ski Resort Pricing

Allen Wang

8/31/2020

Springboard

**Introduction**: Big Mountain Resort in Montana recently just installed a new ski lift that increases their seasonal expenditures by 1.5 million dollars and they are trying to implement different business strategies. They are trying to either cut costs to accommodate or they are to charge a premium on the regular price of the tickets.

**Data Wrangling**: The dataset is a single csv provided by Springboard. The first step in the process is data wrangling in which we uncover the dataset and change it into an appropriate format through data manipulation and cleaning, then see what each of the variables are and how they correlate with others and find a general trend in the dataset. In the data wrangling method, we first checked for missing values. Then we checked on the categorical features and did a quick graph on the number of resorts by region and state with Montana coming in at 13th place. These graphs raise questions and are important details to keep note of. The next check is on feature values. A histogram plot was used on the numeric feature values. These plots revealed outliers on the skiable terrain column in which it was revealed there was a data entry error in which it was corrected. The fast\_eight column was dropped since it had over 50% of missing valuables which proves invaluable. The next column that needed cleaning was years\_open and it was revealed that the column had outliers as well which resulted in the entire row with the outlier dropped. We then proceeded to drop the rows with no pricing data and then we found that there were state names with square brackets in the entries in which we fixed.

**Exploratory Data Analysis**: Further analysis of the dataset requires exploratory data analysis which is a procedure in which each of the variables are examined and looked at in correlation with each other. First we examine the numeric columns such as the resorts per state, state population, state area, and days available. From our examination of each column, it seems as if New York is the most popular state for ski resorts. However, New York does not have the most available ski area which raises the question. Are the ski resorts popular because they are most convenient for wealthy New Yorkers? Looking at resort density, the charts show the Vermont and New Hampshire dominate in terms of resorts per capita and per state area. We decided to run a PCA process which will show the best linear combinations of variables and show the variance that these variables account for. We found that two variables account for over 75% of the variance and we graphed the PCA with component 1 on the x axis and component 2 on the y axis and from matplotlib and seaborn graphs there doesn’t seem to be any correlation with price. Two datasets were merged for feature engineering and these columns were derived from the merging: ratio of resort skiable area to total state skiable area, ratio of resort days open to total state days open, ratio of resort terrain park count to total state terrain park count, ratio of resort night skiing area to total state night skiing area. A seaborn heatmap was used and it indicated that total night skiing area correlates with ticket price as well as number of chairs. Then a scatterplot was used to identify more positive correlating features with ticket price such as runs, vertical drop, fast quads, and total chairs. A few columns with ratios of the other columns were created. We added a ratio of chairs to runs and a ratio of chairs to total skiable area. At first these relationships are quite counterintuitive. It seems that the more chairs a resort has to move people around, relative to the number of runs, ticket price rapidly plummets and stays low. What we may be seeing here is an exclusive vs. mass market resort effect; if you don't have so many chairs, you can charge more for your tickets, although with fewer chairs you're inevitably going to be able to serve fewer visitors. Your price per visitor is high but your number of visitors may be low. Something very useful that's missing from the data is the number of visitors per year.

It also appears that having no fast quads may limit the ticket price, but if your resort covers a wide area then getting a small number of fast quads may be beneficial to ticket price.

**Data Preprocessing and Training Data**: In this section, we used the data to create machine learning models. Using sklearn’s train\_test\_split function, we split 70% of the data into training data and 30% of the data into testing data. We then defined 3 different metrics as functions: R^2 value, Mean Absolute Error, and Mean Squared Error. These metrics are used to describe different aspects of efficiency in a test. R^2 value accounts for the total percentage of variance of the data explained by the model. Mean Absolute Error gives a numeric value to the deviation of price from the mean value. Mean square error calculates a numeric value of the sum of the squares of the error. We imputed median and mean into the missing values of the model and both seem to be generating linear regression models around similar efficiencies. The median imputed model produced an R^2 of (0.8177988515690604, 0.7209725843435146) and the mean imputed model produced (0.8170154093990025, 0.716381471695996). We created a pipeline that compiled all the steps and streamlined the process. We continued to refine the model by implementing cross validation. We found that the best number of parameters to use in our equation is 8 and after that the variance spikes. We then used a different algorithm which is the Random Forest algorithm which is a group of decision trees because this algorithm generally produces good results for a variety of problems. We found that the dominating four features fastQuads, Runs, Snow Making\_ac, vertical\_drop. The random forest model has a lower cross-validation mean absolute error by almost $1. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.

**Modeling:** In this section we want to refit all the data into our new model that we created in the previous step. We want to create the model without the data from Big Mountain and insert the Big Mountain data into the model as test data. These are the achieved results through repeating the steps with a pipeline: a modelled price of $94.22, the actual price being $81.00 there is a room for increase in the pricing. In the random forest model, there were 8 other covariates that we should consider for the pricing: vertical\_drop, Snow Making\_ac, total\_chairs, fastQuads, Runs, LongestRun\_mi, trams, SkiableTerrain\_ac. We then modeled 4 different scenarios for adjusting for these covariates. In scenario one, we can reduce the number of least used runs. The chart that we plotted shows that decreasing runs by 1 makes no different in pricing but reducing 2 sharply reduces the price therefore revenue. Reducing 3 is the same as reducing 4 or 5 so might as well reduce 5 runs leading to no further loss in ticket price. In scenario two, the number of runs is increased by one, another chair lift will be installed, and the vertical drop will be increased to 150 feet. This scenario increases support for ticket price by $1.99. Over the season, this could be expected to amount to $3474638. In scenario three, the same goes as scenario two but two acres of snow making is added. The results are the same. In scenario four, the snow making is increased by four acres and the longest run is extended by 0.2 miles. It results in no change. Therefore, the most ideal scenarios are two and three in which support for the ticket price is an increase by $1.99.

**Summary**: Big Mountain currently charges a mean price of $81.00 and our model suggests that we should be charging $94.22. Therefore, Big Mountain has a room for increasing their prices. However, this doesn’t necessarily mean that we should as the market for skiing might change and some resorts are underpriced while others are overpriced. We can consider several options. We can close down the one of the least used runs or we can close down 2. If we close 3, then we might as well close 4 or even 5. We can also add a run, increase vertical drop, and install a chair lift which will lead to a support of $1.99 leading to an increase in revenue of $3474638.