Documentation

1. Problem statement

In this project we’re investigating how to select a better value to ticket price for the ski resort. Big Mountain Resort is a ski resort located in Montana. It has recently installed an additional chair lift to help increase the distribution of visitors across the mountain. This additional chair increases their operating costs by $1,540,000 this season. The resort's pricing strategy has been to charge a premium above the average price of resorts in its market segment. There's a suspicion that Big Mountain is not capitalizing on its facilities as much as it could. The business wants some guidance on how to select a better value for their ticket price.

The key criteria that will deem this work useful is the growth of capitalization.

The focus of this business initiative is to cut costs without undermining the ticket price or will support an even higher ticket price.

The constraint that may prevent this business initiative from succeeding is that resort's pricing strategy has been to charge a premium above the average price of resorts in its market segment.

The key stakeholders that need to be involved in this project are Director of Operations, Jimmy Blackburn, and Alesha Eisen, the Database Manager.

The key places of data to use to solve the problem are SQL database or an S3 bucket.

In this example, I’ll work with a single CSV file that I got from the database manager.

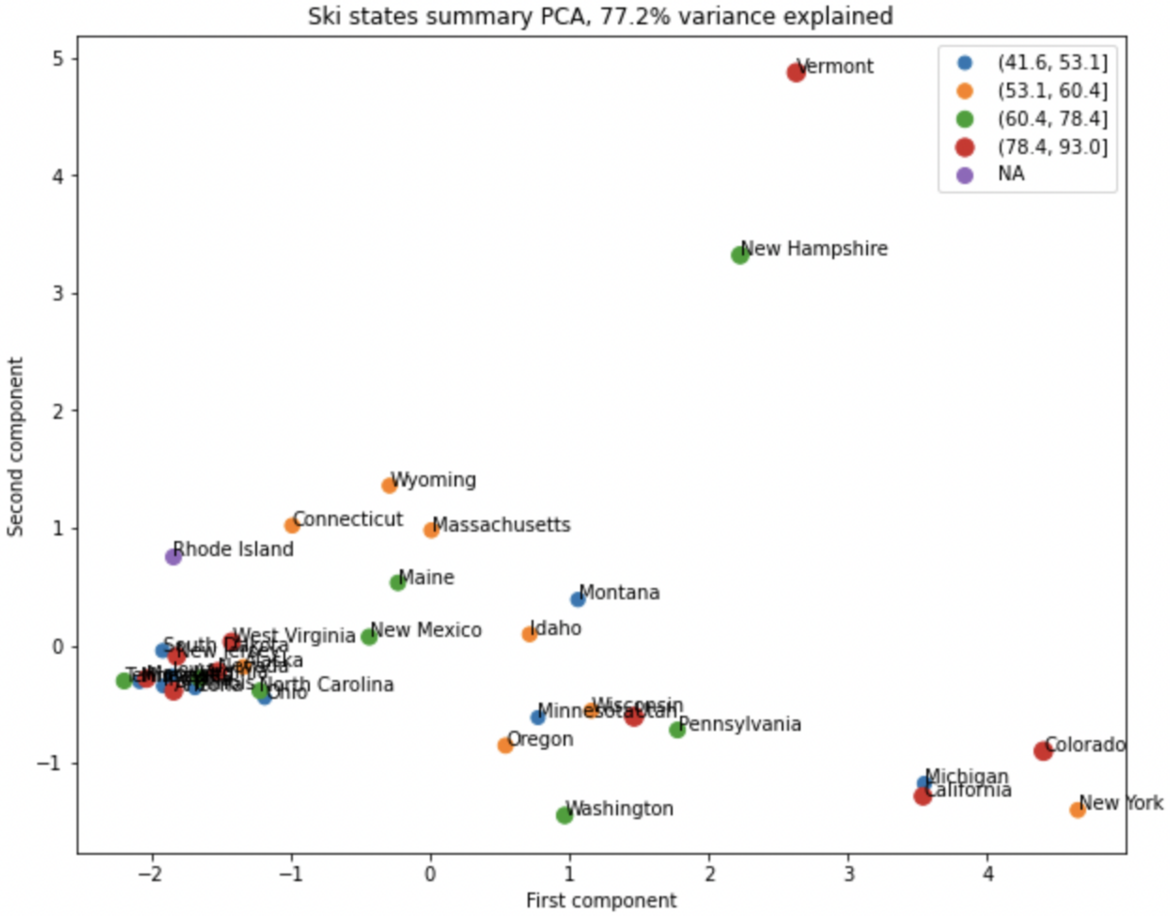
2. Data wrangling

Original data has 330 rows, Big Mountain Resort is present in the data. Rows with no price data were dropped because of the absence of the target variable. The fastEight column was dropped too because half of the values are missing and the other half is equal to zero. Final data has 277 rows and 25 columns. All rows are unique, no duplicates. There are 2 resorts with the same name, but they are in different states. There is a distribution of ticket price by state. Also it is different on weekdays and weekends. Also found some outliers in values. One resort (Silverton Mountain) has an incredibly large skiable terrain area. Found a mistake in the yearsOpen column Population and area data for the US states can be obtained from wikipedia. Some state names were fixed to match the desired format.

3. Exploratory data analysis

We had Name, Region and state - categorical features in the data, and all other - numerical. Was considered the ratio of resorts serving a given population or a given area. When the histograms of density distribution were created, was discovered some structure. To explore this more we used principal component analysis. We paid attention to the first two components of PCA. To discover correlation between features we used heatmap. And created scatterplots of numeric features against ticket price.

We considered the number of resorts per 100k population and per 100k square miles. Distributions of them showed some structure there. We scaled the data and made the PCA transformation. It helped explain the variance of the data. We created a cumulative variance ratio explained by PCA components for the state/resort summary statistics plot. And it showed that the first two components seem to account for over 75% of the variance, and the first four for over 95%. Was created the plot of 2 first components which showed distribution of states with quartiles of prices where they belong.



We extracted some summary features: resorts per state, state total skiable area ac, state total days open, state total terrain parks, state total nightskiing ac, resorts per 100k capita, resorts per 100 ksq mile. Also were created: 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. AdultWeekend ticket prices showed high correlation with fastQuads, Runs, Snow Making_ac, total_chairs, resort_night_skiing_state_ratio.

4. Model Preprocessing with feature engineering

At first we tried to take the mean price as a predictor. To measure the quality of the Dummy Regressor model were considered R-squared (got 0 as the result on the training set and negative value on the test set). Also we used Mean Squared Error and Mean Absolute Error. To create initial models, we tried to impute missing predictor values with median and with mean and performed the next 4 steps for each situation: impute missing values, scale the features, train a model, calculate model performance. Which we later added to the pipeline. The results obtained with different fold (or different quantity of folds) in cross-validation can vary. We used GridSearchCV to find best 8 features: vertical_drop, Snow Making_ac, total_chairs, fastQuads, Runs, LongestRun_mi, trams, SkiableTerrain_ac.

5. Algorithms used to build the model with evaluation metric.

A Dummy Regressor model and RandomForestRegressor were tried. After hyperparameter search with GridSearchCV, imputing missing values was done with the median, data wasn’t scaled, and the best n estimators equal to 69. The best random forest regressor feature importances for random forest regressor were fastQuads, Runs, Snow Making_ac, vertical_drop.

6. Winning model and scenario modeling.

After comparing performance of the models, the random forest model showed a lower cross-validation mean absolute error by almost $1. It also exhibits less variability. About the size of the data sets, CV validation score becomes the same after 40-50 samples in the set.

7. Pricing recommendation

Current Big Mountain ticket price is 81. Model suggests to increase the price to 95.87. Considered business options with modeling:

Permanently closing down up to 10 of the least used runs. This doesn't impact any other resort statistics. - Closing one run makes no difference. Closing 2 and 3 successively reduces support for ticket price and so revenue. If Big Mountain closes down 3 runs, it seems they may as well close down 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a large drop.

Increase the vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up, without additional snow making coverage - This scenario increases support for ticket price by 1.99. Over the season, this could be expected to amount to 3474638

Same as number 2, but adding 2 acres of snow making cover - This change doesn’t make a difference.

Increase the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres - Also no difference.

8. Conclusion

The price could be increased to 95.87 to grow the capitalization with the additional chair installed.

9. Future scope of work

Snow making areas have a lot of NA. Skiable terrain area, runs and longest run also have some missing values. This could limit the work. Store availability for buying or renting the equipment also could be useful.