Guided Capstone Project Report

**1.Problem statement：**

Big Mountain Resort, a ski resort located in Montana,with access to 105 trails, offers 11 lifts, 2 T-bars, and 1 magic carpet for its 350,000 annual visitors. The longest run is 3.3 miles in length. The base elevation is 4,464 ft, and the summit is 6,817 ft with a vertical drop of 2,353 ft. Now Big Mountain Resort needs to make a decision on whether to make changes for capitalizing its facilities or raise up tickets price so that the resort can compensate for the extra $1,540,000 operating costs caused by installing an additional chair lift this season. As Big Mountain Resort’s data science team, we are providing our recommendation on the price strategy and further investment strategy here by conducting a series of analyses on a nation-wide ski resorts data set.

**2.Data Wrangling**

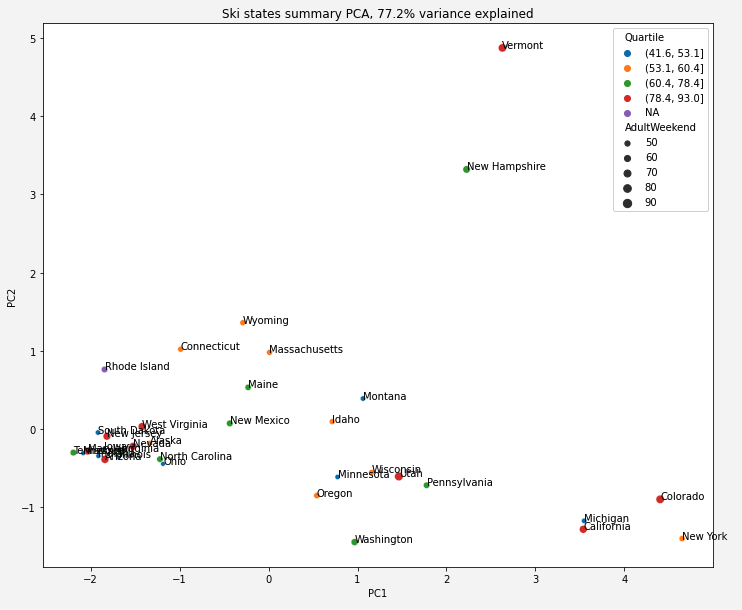
What does our data set include? Do we have all the information ready for further analyses? To answer these questions, we did data wrangling as first step of this project.

There were 330 resorts (including Big Mountain Resort) and 27 features of a resort (including name, state, number of chair lifts, days open, ticket price, etc) in the original data set. After examining the data set, we dropped the ‘fastEight’ column, for half the values are missing, the existing values are 0 excpet for one value is 1; ‘AdultWeekday’ column, for according to the price box plot, weekday and weekend prices in Montana seems equal and 'AdultWeekday’ column has more missing value than 'AdultWeekend’; rows that missing ‘AdultWeekend’ value, for they missed the ‘AdultWeekend’ prices to compare with our own resort; and the row which ‘yearsOpen’ column value is 2019, for the obvious incorrect value. We also amended Silverton Mountain resort’s ‘SkiableTerrain\_ac’ column value to 1819 and Heavenly Mountain Resort ‘Snow Making\_ac’ column value to 2880. Besides, we decided to target on ‘AdultWeekend’ value to predict ticket price. After data wrangling step, there were 277 resorts and 25 features of a resort left in our data set. Additionally, we pulled the population and area information of each state from online open resource.

1. **Exploratory Data Analysis**

Was there any pattern suggested of a relationship between state and ticket price? We used principal component analysis approach on the state summary data set to find the answer. It turned out the first two components explained 77.2% of price variance. From the graph “Ski states summary PCA,77.2% variance explained” below, there is no clear pattern for state and ticket price. This led us to use the components\_ attribute of the fitted PCA object. The component value of the features tell us how important (and in what direction) each feature contributes to each score (or coordinate on the plot).

Which features are relevant for ticket pricing? We conducted a heatmap on the combination of resorts data set and state summary data set. Ticket price is highly related to “fastQuads”, “Runs”, “Snow Making\_ac”, “vertical\_drop” features in a positive way.



1. **Model Preprocessing with feature engineering & Algorithms used to build the model with evaluation metric**

How should we wisely use our data set when training a model? What’s the metrics for evaluating a model?

Firstly, we applied a linear model which includes the following four steps: 1. replace missing values with the median for each feature; 2. scale the data to zero mean and unit variance; 3. train a linear regression model; 4.cross-validation technique for estimating model performance. It turned out linear model has a much higher R squared score (mean:0.63 standard deviation: 0.10) and a lower mean absolute/squared error(9.21/10.49). Because in cross-validation, each of the five blocks has been used as a test set, its performance on the test split was consistent with this estimate. Additionally, we used GridSearchCV function `best\_params\_` attribute to select the k best features and got a result of eight. We then used get\_support() method and `coef\_` attribute to see how exactly the best eight features is related to the predicted price. After sorting the coefficient values, the top five important features are: “vertical\_drop”, “Snow Making\_ac”, “total\_chairs”, “fastQuads” , and “Runs”.

Then we applied a random forest model, the preprocessing steps are the same as linear model except we trained a random forest model on step three. For the default data set, the mean R squared score is 0.70 and the standard deviation is 0.07, for the best parameter, the mean R squared score is 0.71 and the standard deviation is 0.06. This model’s accuracy is even better than linear model, so we decided to going forward with random forest model. Encouragingly, the dominant top four features are highly in common with your linear model: “fastQuads”, “Runs”, “Snow Making\_ac” and “vertical\_drop”.

1. **Winning model and scenario modeling**

Are we charging a fair price? Now we can calculate the expected Big Mountain ticket price with our winning model- a random forest model with 69 trees. Big Mountain is currently charging $81.00 per ticket while our model is suggesting $95.87. It look like we have enough room to raise up the ticket price! In fact, a $0.88 raise is enough to cover the the new lift’s operating cost. Among the four potential scenarios, the first three worth further consideration, for the first scenario, we can definitely close one run since it won’t affect the revenue. The second and third scenarios will work too as long as the related extra operating costs are less than 15.07 million and 17.32 million.

**6.Future scope of work**

The validity of our model lies in the assumption that other resorts accurately set their prices according to what the market (the ticket-buying public) supports which we need to further verify. Runs and snow making operating cost would be useful for the first three scenarios analysis. We don't know yet if our current low price is our business executives' strategy to attract more visitors. We can test how sensitive visitors react to the price change later. The business leaders or business analysts should be able to reuse this model by change different parameters or update the original data set.