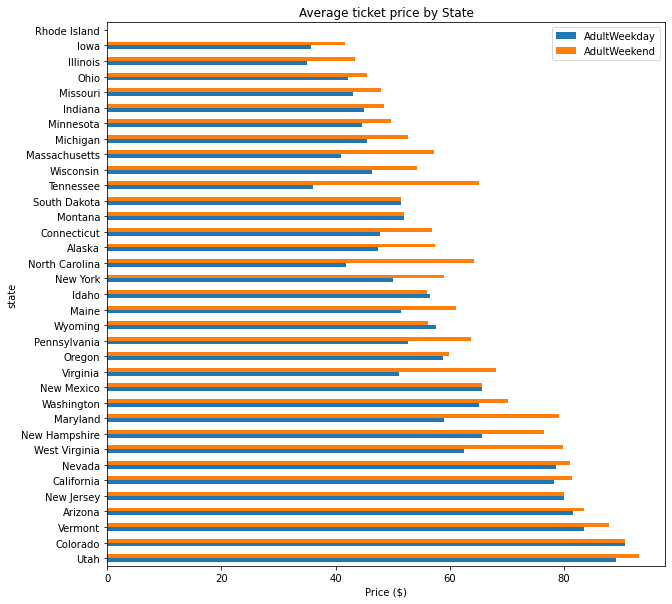
We start with the problem statement that captures the problem that we are trying to solve for Big Mountain: “How can Big Mountain Resort obtain a better value for its ticket price by optimizing its facilities to increase its customer spending volume by at least 15% over the next year?” This problem statement puts a framework around what is to be accomplished (increasing ticket price) by the data science project by giving a specific set of actions (through optimizing the resort’s features), to be accomplished in a specific time frame (one year).

After “framing” the project we began to wrangle the data. At this stage of the project we proceed to take steps to ‘clean up’ the data by making sure all the data types are correct,

getting rid of rows that have lots of null values (assuming those rows make up a small percentage of the dataset), correct any typos and data that has been verified to be false. In addition to doing cleanup, this is the time to explore the data, inspecting the different columns for missing values (which may be treated in the pre-processing stage) and counts of different values in the columns.

For this project, we see that ‘Region’ and ‘State’ are different columns but they both use state names as values. Therefore, it was necessary to verify that no two observations were the same. We also investigated ticket prices, which was necessary since the point of the project is to determine a feasible ticket price but there turns out to be two prices. The following is a horizontal bar graph that depicts both the Weekday price and the Weekend price for each state.



Based on the bar graph, the Adult Weekend prices are greater than the Adult Weekday prices for nearly all the states, and the two prices are very close for most of the states. Another set of data worth investigating was the distribution of the features among the states. This was helpful because it gave insight into which features were common to the different resorts, thereby giving an indication of the importance of those features.

For example, it is apparent that ‘fastEight’ lifts are not very necessary, but it is a good idea to have vertical drops of 2000 meters. The main information gathered from this leg of the project is that it is probably best to work with the AdultWeekend prices since there were fewer missing values than there were for AdultWeekday prices.

After wrangling the data, it is time to perform exploratory data analysis. In this segment of the project, we scale the data and then we reduce the dimensionality of the data. The scaling is necessary because the data in the columns are given in differing units, but any model will require the data to be on the same scale. We use the ‘scale’ function to scale the data. Since the ‘scale’ function returns a numpy array, the previous column names from the corresponding DataFrame are lost; for that reason we saved the column names before the scaling in order to use them later.

The reduction of dimensionality makes the data easier to graph and comprehend while simultaneously keeping the most important features for building the model. Using the previously scaled data, the PCA function takes that data and restructures it into principal components. A chart of the Cumulative ratio variance explained by PCA shows which of the principal components makes up the largest part of the data.

A blue line on a white background

Description automatically generated

In our case, the first 2 components account for a little less than 80% of the variance.

A powerful piece of insight can be gained by using a heat map, which shows the correlation of each feature with every other feature. Below is an example from this project using all the features from the original data except the ‘AdultWeekday’ feature.

A screenshot of a computer screen

Description automatically generated

Each intersection of a row and file is a square of a certain color, where the lighter the color the stronger the correlation. Heatmaps are especially useful in determining which features should not be included in a model since having highly-correlated features in the model can negatively affect the results of the model.

Finally, we were able to plot each feature versus the AdultWeekend ticket price, with the idea of determining which features have a linear relationship with ticket price.

A group of blue dots

Description automatically generated

After EDA comes preprocessing and training of the model. Typical train/test splits can be 80/20, 70/30 etc., but for this project we did a 70/30 split. Thus 70% of the data is used to train the model and then the other 30% of the data is given to the model to see how well the model does. Before actually building a model, we constructed what amounts to what a ‘good guess’ at a solution would be by simply taking the average of the AdultWeekend Ticket prices.

To measure the effectiveness of the model, we use the ‘coefficient of determination’ metric, which measures the amount of variance of the ticket price that is explained by the model. Before building the model, it is important to make sure that there are no empty cells in the DataFrame so we applied imputation using median values. The mean was a better choice than the median because many of the predictor feature distributions were skewed and the mean is more sensitive to skewness than is the median.

Creating the model requires three steps: Imputation, scaling, and choosing a model. SKlearn has a ‘pipeline’ function which can combine all of these into one line of code. In this project, we used the pipeline to try two separate models: the Linear regression model and the random forest model. The model that yielded the best results was the random forest model because it gave less variability. As to the question of the current data at our disposal being sufficient, the answer is yes, according to a chart of Cross-validation scores vs. Training set size.

The final step in the project is modeling, where the model is finalized with features. We start with fitting the model to the data but excluding the data for Big Mountain so that Big Mountain’s features do not bias the data. We were then able to compare Big Mountain’s features to those of the other resorts.

In general, the features of Big Mountain were competitive with those of the other resorts, but the AdultWeekend price of Big Mountain is rather on the high end when compared to the other resorts in Montana. Finally, we consider 4 scenarios where features are manipulated to see the effect on ticket prices. Based on the results, the scenario of adding a run and increasing the vertical drop by 150 ft and installing an additional chair lift tied with the scenario of adding a run, increasing the vertical drop by 150 ft, adding an additional chair lift and adding 2 acres of snow making in terms of the ticket price increase that could be supported. The implication here is that additional snow making would not add ticket value.