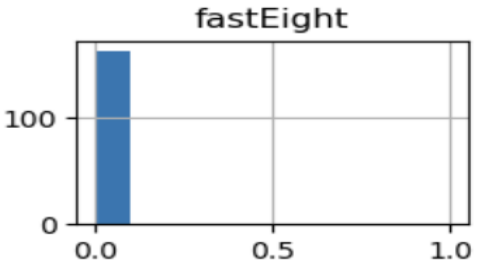
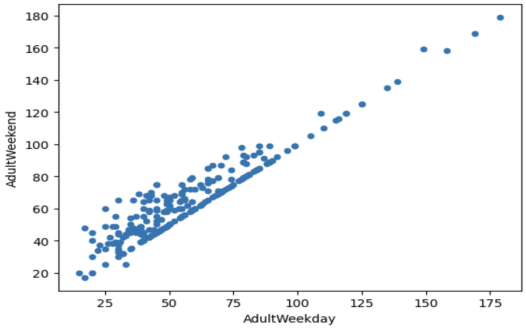
PROBLEM STATEMENT

Big Mountain Ski Resort had a good business model but weren't really sure how it performed when compared to other ski-resorts in the country. They were popular in their services enough to proceed with the installation of a chair lift to accommodate their growing needs which would add an additional operational cost of $1,540,000. **However, they wanted to know if this was a reasonable investment and were open to suggestions on what could be alternate strategies to increase their revenue, which was measured by the ticket price.** They had a csv file that contained market data on 330 ski-resorts in the United States which was the only data source available to make reasonable suggestions.

DATA WRANGLING

With the challenge to conduct a study on the ski-resort prices across the country, we started the data wrangling process to understand if the csv file captured the information needed to assist in addressing our question of - Have we priced Big Mountain Resort ticket price accurately. Some of the questions we wanted to answer were - **Do we have the target value ? Do we have all the relevant data collected to support our decision of the target value ? Do we have any issues with the data ?** With csv file starting at 330 rows and 27 columns, after the data wrangling process of dropping columns after verifying distributions of features that didn't influence the target variable AdultWeekend i.e *fastEight* and *AdultWeekday.* Fig on Left shows the number of fastEights across resorts with is 0 and Figure on right indicates how AdultWeekend and AdultWeekday columns are almost similar after $100. and cleaning up rows with multiple missing values we ended the Data Wrangling process by storing the cleaned dataset with 227 rows and 25 columns. One key step in this process involved collecting state-wise population and area statistics from Wikipedia and appending it to state-wise aggregate content from the csv file before some rows were dropped. In the end, we had 2 tables to use as a starting point for Exploratory data analysis (EDA).

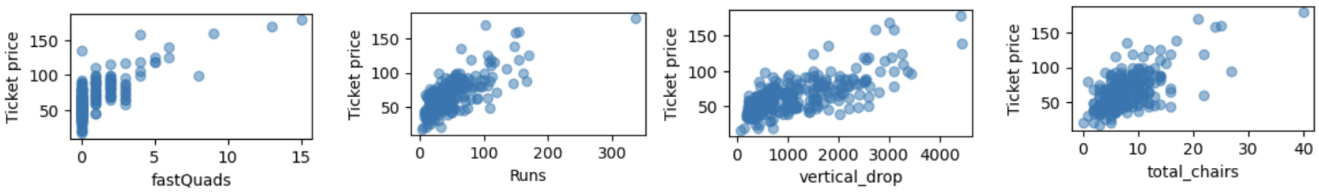




EXPLORATORY DATA ANALYSIS

The purpose of the EDA performed in the notebook is to evaluate relationships between the attributes we thought could potentially influence the prices of the ticket. These evaluation/detective tasks were chosen to guide us to making appropriate assumptions and conclusions on variables. We carried out state-wise ranking for features representing resort facilities and state statistics and found there was no direct relationship between the characteristics of what made a certain resort popular. To establish a linear relationship between the most influential determinants of the ticket price, we conducted Principle Component Analysis (PCA) and found that that although there wasn't any grouping between states, we found that **resorts_per_100kcapita, resorts_per_100ksq_mile** affected the ticket prices so it is more appropriate to calculate some

more resort-wise statistics. After including resort-wise columns and constructing a heat map we concluded that the most influential features are fastQuads, Runs, vertical_drop and total_chairs. Here's Figures from Left to Right as fastQuads, Runs, vertical_drop, total_chairs



MODEL PREPROCESSING WITH FEATURE ENGINEERING

The primary purpose of pre-processing is to make sure that the values are scaled i.e the effect of each attribute is captured accurately and is not skewed solely based on its numerical values. We also take care of imputing missing values with appropriate statistical measures in this step because ML models don't like missing values. First, we divided the dataset into a Train/Test set with 0.7/0/3 ratio. Second, we impute the missing values with median and mean sequentially and observed that imputing with median yielded a lower Mean Absolute Error by a slight margin.

ALGORITHMS USED TO BUILD THE MODEL WITH EVALUATION METRIC

We build a model by sequentially applying the process of imputing missing values, scaling the features using StandardScaler(), training a model using Linear Regression and Random Forest Model, and calculating model performance by checking the Mean Test Score and the Standard Test Score for both the models.

WINNING MODEL AND SCENARIO MODELLING

After modelling, Linear Regression model suggested 8 parameters as influential and Random Forest Model deemed 4 as influential. However, we need to find common metrics to evaluate them since these two models are different in the outputs they have produced. To compare them we rely on mean_absolute_error of both the models for the data row in the question i.e Big Mountain Resort. Linear Regression Model had (Mean, STD) as (10.499, 1.622) and a mean standard error(mae) of (11.793). The random forest model had (Mean, STD) as (9.644, 1.352) and a mean standard error(mae) of (9.537). The random forest model has a lower cross-validation mean absolute error by almost $1. It also exhibits less variability. Our winning model is RANDOM FOREST MODEL and therefore the features of importance are - **Runs**, **fastQuads**, **Snow Making_ac** and **vertical_drop.** Fig below shows the most important features in a bar plot.

Once we finalised the model i.e determined the key features influencing the prices of the ticket, we created 4 scenarios to map the differences in revenue i.e incremental, decremental or no change measured in these scenarios.

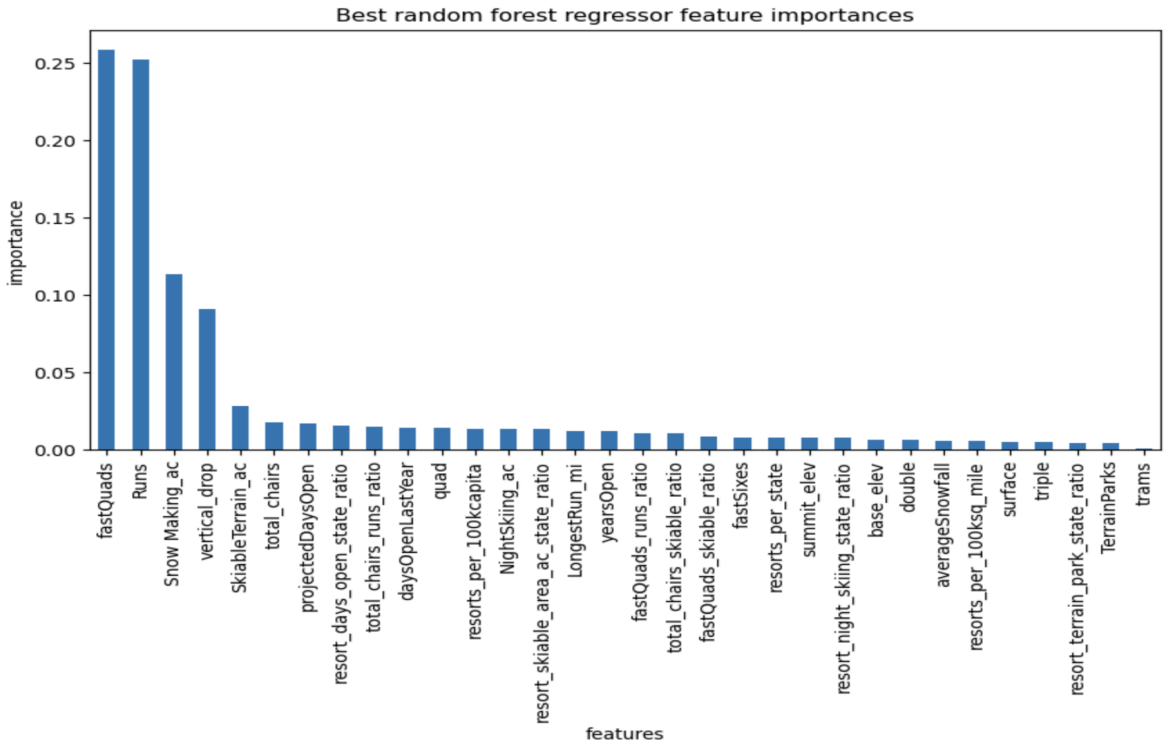
Scenario 1: Measured Revenue changes when Runs were decreased

Result 1: Closing a run didn't affect the revenue negatively, however the drop of closing 2 and 3 runs were similar and the trend in the revenue reduction was correlated with the number of runs reduced.

Scenario 2: 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

Result 2: We observed that we could charge an extra $1.99 per ticket and it would compensate for the installation of the extra chair lift of $1,540,000 by generating a total of $3,474,638

Scenario 3: Same as number 2, but adding 2 acres of snow making cover



Result 3:We saw the same increase in total revenue as scenario 2 , so we can say we don't necessarily need to add 2 acres of snow making cover.
Scenario 4: Increase the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres
Result 4: Case 4, yielded no improvement in revenue after increasing the longest run, we can say it is not an influential feature in determining the price of ticket, even if it was in the top seven feature list.

PRICING RECOMMENDATION

The recommendation is to execute Scenario 2, since we are able to make up for the expenses of installing an extra chairlift and also gain a profit by a margin of $1,934,638

CONCLUSION

Big Mountain Resort surely now has a few metrics to evaluate how certain business decisions will influence the revenue generated. The model however hasn't considered the operational costs to execute the scenarios, so the predicted ticket prices could be subject to slight change if/when we introduce these costs.

FUTURE SCOPE OF WORK

The future of this project is to understand and realign with the business leaders if there is value in putting the insights into action and to figure out additional tasks of maintenance of the model or updation to the model depending on changes in the data captured moving forward.