Detailed information with the topics below

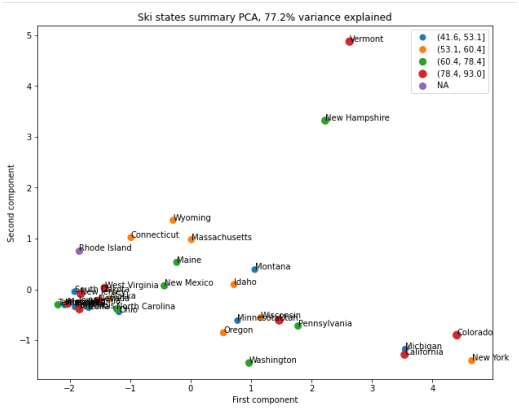
1. Introduce topic and Problem statement
2. Explain your approach to solving problem
3. Data Wrangling (introduce dataset and features)
4. Exploratory data Analysis
5. Model Preprocessing with feature engineering & Algorithms used to build the model with evaluation metric
6. Winning model and scenario modeling & Pricing recommendation
7. Conclusion & Future scope of work

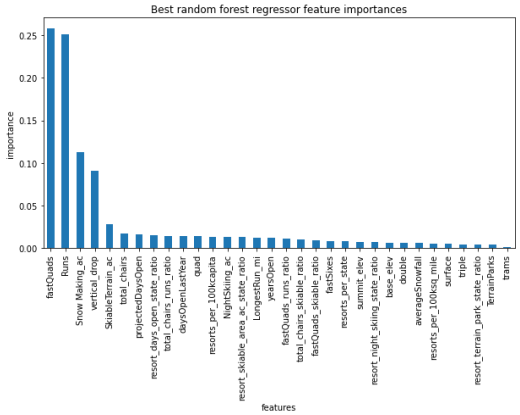
*documentation includes graphics such as charts and plots
*charts/maps/plots show data in most intuitive way
*model/metrics are clearly explained
*trends are drawn from data and recommendations are provided
*valid conclusions and future scope of work is defined
*scenario modeling completed
*concise report with proper flow, format, and story

Big Mountain Resort decided to install an additional chair lift, which increased their operating costs and decreased their incoming revenue. Executives believe the resort isn't capitalizing on its facilities as much as it could and wants to formulate new ideas on how to select better ticket prices or cut down costs to increase revenue. Therefore, the problem statement at hand is how can Big Mountain Resort increase annual revenue by 10% within the next year by either selecting a better value for their ticket prices or cutting down costs without undermining ticket prices based on data (chairlift speeds, total number of chairlifts and amenities, and average ticket prices) of other ski resorts compared to Big Mountain Resort's? In order to find a solution, we must analyze the data of other ski resorts' and produce models to compare them to Big Mountain Resorts' through data wrangling, exploratory data analysis, algorithms, and modeling.

I was given a CSV data file containing 330 different ski resorts along with 27 features for each. In order to find out which rows, columns, or specific values I needed to remove to clean the data, I analyzed numeric and categorical features first. Since our main focus is our Big Mountain Resort, our primary target response feature should be either adult weekday ticket prices or adult weekend ticket prices. I removed any rows that have missing/null values in these particular columns and plotted the column distributions to discover any features with incorrect values that need to be replaced. Following these similar steps, I continued to clean small details in the data files. I can confirm the target feature for my desire to predict ticket price is the Adult Weekend Ticket price. After all important steps were taken to wrangle the data, I saved the new ski data with 277 rows and 25 columns.

Next, I loaded in state summary statistics to explore and visualize the data. I added a new data type, a categorical variable, called 'Quartile' to create a list of quartiles with an informative legend to see points colored by quartile and sized by ticket price. Through plotting the ski states summary PCA, I found that there wasn't an obvious relationship between state and ticket price. This led me to look at all other numeric data to compare with ticket price. There seemed to be a lot of reasonable correlations between ticket price and fastQuads, ticket price and Runs, ticket price and Snow Making_ac, ticket price and resort_night_skiing_state_ratio. I believe almost all numerical columns of the data should be taken into account when performing feature selection for modeling, and missing useful data should also be taken into account.





After removing my resort from the data and splitting the rest of the data into training and testing data subsets, I began the preprocessing step by simply taking the average price to see how good it was as a predictor and comparing it to different metrics. I built a linear model and trained it on the training data. Through the linear model that used the mean, I found over 80% of the variance on the training set and over 70% of variance on the test set. The linear model that used the median produced the same results as the mean model, so I used sklearn's pipeline, which is a better model for linear regression, to confirm my findings. Cross-validating my results displayed that model

performance is always open to variability. Additionally, I tried a random forest regressor, which had an improved estimated performance. After comparing my linear regression model and my random forest regression model, I found that my random forest model had a lower cross-validation mean absolute error than the linear model by almost $1. Therefore, I chose to go with the random forest model because it displays less variability and the test set was consistent with the cross-validation results.

Big Mountain currently charges $81 dollars for its tickets, but my model sets ticket prices at a predicted $95.87. Based on features, it seems our resort is underpriced compared to other resorts and the features they offer. By exploring and plotting different features that seem impactful to deciding ticket price with respect to Big Mountain Resort, I was able to come up with 4 different scenarios that could potentially solve the resort's original problem. In scenario 1, I explored what would happen to the total revenue when up to 10 of the least used runs were closed down. The model showed that closing one run doesn't make a difference, but closing 2 or more would just lead to a loss in ticket price and revenue. Scenario 4 showed that increasing the longest run and its snow coverage made no difference to the ticket price. In scenario 2, I found that adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift would support increasing the ticket price by 8.61 dollars. This would lead to an increase in revenue by 15,065,471 dollars over the season. Therefore, I would suggest increasing the vertical drop by 150 feet, adding a run and chair lift, and increasing the ticket price by $8.91 in order to increase revenue for Big Mountain Resort. Additionally for future improvements, I would recommend considering closing down 2 of the least used runs and 3-5 of the least used runs and see how it affects the revenue.

It would have been useful if I was given the cost of adding each feature. For example, I don't know what the cost is to increase the vertical drop by 150 feet, which would contribute to what features the resort should add in relation to how much it would cost them. Big Mountain's modeled price is so much higher than its current price because it has so many different facilities offered, but executives weren't charging the prices they deserved. I think this mismatch would come to a surprise to executives because they are aware of all the many facilities the resort offers, but they were being generous with their ticket prices. After presenting my models and findings, if executives feel that my model is useful, the business can use the same model to further test other ideas they have to cut costs or increase revenue through altering many different features. After giving them my model, I wouldn't expect them to come to me every time they wanted to test a new combination of parameters because inputting the parameters in the model should give them the same results that I would give them. Business analysts would be able to freely test their ideas using this model without my help, and consistently improve the resort's incoming revenue.