Adam Lang

4/2/2024

Springboard Data Science Fellow

Guided Capstone Project Report

**Problem Statement**

Big Mountain Ski Resort is a popular winter destination in Montana for skiers and snowboarders. A new state of the art lift system recently brought their lift count from 11 to 12. However, the new lift is going to increase operating expenses by $1.5 million over the next winter season. Previously Big Mountain had set lift ticket prices at a premium above average price of resorts within its market segment across the U.S. Management realizes this strategy is not practical as the market ticket prices do not account for Big Mountain’s individual operational facility costs. **How can Big Mountain Ski Resort implement a data-driven business strategy to select a better valued lift ticket price for the upcoming ski season that will offset the sudden increase in operating expenses while aligning with the market, customers, and resort facility costs?**

**Data Wrangling**

A raw csv file of 330 ski resorts in the same market as Big Mountain Resort was available for analysis. The file had 330 rows and 27 columns. **AdultWeekday** and **AdultWeekend** were the target variables for lift ticket prices. Big Mountain Resort did not have any missing values. Data quality issues included missing values. The variable **fastEight** had about 50% missing values and the variable **NightSkiing\_ac** was a close second with 43% missing values. The 2 target variables **AdultWeekday** and **AdultWeekend** were missing around 15-16% of their values. **AdultWeekend** prices were higher than **AdultWeekday** prices for all states. The 3 most expensive states were California, Colorado and Utah. For the 2 target variables **AdultWeekday** and **AdultWeekend** there were 7 and 4 missing values respectively. Based on the notebook instructions I dropped **AdultWeekday,** and kept **AdultWeekend**. This changed the dataframe shape to 277 rows and 25 columns. The reason for this was based on a scatter plot constructed with **AdultWeekday** (x-axis) and **AdultWeekend** (y-axis) **(Appendix - Fig 1).** A strong correlation was noted between the 2 variables at $100 and below, and the state we are concerned with (Montana) aligned with having equal weekday and weekend prices. In addition, for weekday prices above $150, the weekend prices were $160 and above and these were basically outliers compared to the rest of the dataset as there were only about 7 total. Two new datasets were created for analysis: *Ski\_data\_cleaned.csv* and *state\_summary.csv.*

**Exploratory Data Analysis**

EDA began with creating 2 variables: **resorts\_per\_100k\_capita** and **resorts\_per\_100ksq\_mile**. Vermont had the most resorts per 100k capita. Montana, home of Big Mountain Resort, ranked 4th. The top 5 states in **resorts\_per\_100ksq\_mile** were 5 of 6 states in the New England Region. These states all have lower square mileage than western states like Colorado or Montana so it makes sense that they are the top 5 in terms of number of resorts per 100k square miles. For weekend prices, Colorado had the highest average price with Arizona, California, Nevada, and New Jersey rounding out the top 5. Montana was 15th. Principal Component Analysis (PCA) was performed to investigate states vs. ticket prices. A strong positive correlation was seen with **vertical\_drop, fastQuads, Runs** and **total\_chairs (Appendix - Fig. 2).** This correlated with the most expensive states that were also most popular for skiing: Utah, Colorado, Vermont, California. For **resorts\_per\_100kcapita**, when the value is low, there is significant variability in ticket price, although it can go much higher. Prices climb with the number of resorts serving a population because it indicates a popular area for skiing with plenty of demand - this is easily understood when we see the top 3 resorts in this aggregation are: Utah, Colorado and Vermont **(Appendix - Fig 3).** Lower ticket prices were seen when a smaller number of resorts served a population because it was a less popular state for skiing, these were all midwest states: Iowa, Illinois, Ohio, Missouri **(Appendix - Fig 4).** This may be “an exclusive vs. mass market resort effect”; if you don't have a lot of chairs, you can charge more for your tickets, although with fewer chairs you're going to be able to serve fewer skiers. Your price per visitor is high but your number of visitors may be low.

**Model Preprocessing with feature engineering**

Data pre-processing involved creating a 70/30 train/test split, performing an initial price prediction using a ***DummyRegressor*** model on the dataset mean, and then building 2 machine learning models, a linear regression and a random forest classifier. I started by using the mean as a price predictor. This was implemented using the ***sklearn DummyRegressor***. Baseline predictions revealed I would be off by about $19 if I predicted the ticket prices based strictly on the mean. I then pre-processed the training set by imputing the median for missing values and built a linear regression model which estimated a price of $9 within the real price, which was an improvement. The mean was imputed instead of the median for missing values. However, the mean model performed worse than the median imputed model. Cross validation was performed and showed a k-value of 8 was sufficient for model training; past that, the variance significantly increased. Vertical drop was the most important feature followed by snow making acreage and total number of chair lifts, fastQuads, and Runs. The most surprising finding was that Trams and amount of skiable terrain were negatively associated with ticket price **suggesting that larger resorts can lower their ticket price by accommodating more skiers (Appendix - Fig. 5).** A random forest regression model was built by imputing the median which was based on the better results of imputing the median with the linear regression model. It was found that imputing the median worked well, but scaling features did not. **The top 5 most important features found via random forest were: fastQuads, Runs, snow making acreage, and vertical drop (Appendix – Fig. 6).** Trams were the least important of all the features. The random forest model had a lower cross-validation mean absolute error by almost $1 and had less variability. I verified performance on the test set which produced performance consistent with the cross-validation results.

Based on these results, I decided to **move forward with the random forest model as it had less variability in the output features than the linear regression model.** The random forest would likely give more accurate price predictions, better interpretation and comparison of the independent features/variable weighted importance for predicting ticket price. Lastly, using cross-validation, a data quantity assessment was performed, and showed a sample size of 40 to 50 is sufficient before model performance levels off **(Appendix - Fig. 7).**

**Algorithms used to build the model with evaluation metric**

As mentioned briefly above, Linear Regression and Random Forest Models were built. Metrics used for evaluation included the R-squared (coefficient of determination), Mean absolute error (MAE), and Mean squared error (MSE). MAE was the “most intuitive” of the metrics as it is an average of the absolute difference between actual and predicted values. Cross-validation was used to compare models and predict how they would perform on unseen data without having to use the test set. GridSearchCV was used to find the optimal hyperparameters for each model. The sklearn pipeline was used to implement this **(Appendix – Fig. 8).** A k value of 8 was noted. Final results were then compared using these metrics for both the Linear and Random Forest models. **Random Forest had the lower cross-validation mean absolute error by almost $1 with a result of 9.53 vs. an MAE of 10.49 for the linear model.** The test set was used to validate these results and again the MAE for the random forest was lower at 1.37 vs. an MAE of 1.62 for the linear model. The conclusion was the Random Forest regression model as the best model to move forward with.

**Winning model and scenario modeling**

The Random Forest model was saved and used as the “winning model” as it had the best evaluation metrics. The model was re-loaded and fit on all data except Big Mountain Resort. The results showed that Big Mountain Resort currently charges $81 for an adult weekend ticket. In the context of the entire U.S. ski resort market, Big Mountain sits at the upper end of “standard ticket prices”, which ranges from just above $20 to just above $100 **(Appendix - Fig. 9)**. In Big Mountain’s home state Montana, they own the highest price as most of the resorts charge between $40 and $50 for a weekend ticket. Big Mountain is 1 of 3 resorts that charges above that, the 2 resorts below them charge between $60 and $70.

Modeling suggested the ticket price for Big Mountain Resort could be $96.32. The models also suggested we could increase prices by about $10.41. Big Mountain sits in the “upper tier” for 7 operational dependencies of all ski resorts in the market. This means Big Mountain could potentially leverage various scenarios to manage their ticket price based on these operational expenses. We looked at 4 different scenarios:

1. **Closing ski runs:** closing 1 run makes no difference in price. Closing 2 and 3 runs would reduce support for ticket price and revenue. By closing 3 runs it seems they may as well close down 4 or 5 as there's no further loss in ticket price. Increasing the number of runs closed down to 6 or more leads to even more drop in revenue. Overall it seems that reducing the ticket price by $1 would mean closing 6 runs and the trade off in lost revenue would be about $2.50. That could add up over time per customer so they may want to re-think this strategy.
2. **Add a new ski run and chairlift:** If Big Mountain added a new ski run, they would increase vertical drop by 150 feet and add a new chairlift. This would increase ticket price support by almost $2, and could increase revenue by almost $3.5 million dollars. In theory this could offset the $1.5 million increase in operational expenses by the new lift they installed. The question though is whether or not the new run and chairlift would mean closing any existing lifts or keep them running? We would need to see the full operational expenses for doing this as this model is based only off of runs, vertical drop and total number of chairs of each resort in the market.
3. **Adding 2 acres of snowmaking:** This really has no effect.
4. **Increase longest run by 0.2 miles and add 4 acres of snow making:** This again amounts to no difference!

**Pricing Recommendation**

Based on modeling scenarios above, I would recommend Big Mountain Resort cover their increased operational expenses by adding a new run and chairlift. This would involve increasing current price by ~$2 to $83 for an adult weekend day ticket. Overall this is a win-win for Big Mountain because this could potentially increase revenue by $3.5 million for the ski season. Assuming no additional expenses arise from maintaining any additional lifts and channeling resources to the new terrain and lift, this results in positive gains all around.

**Conclusion**

Big Mountain is the largest ski resort in Montana with the highest ticket price. Within the U.S. ski resort market, they reside in the “upper tier” of resorts for many operational expenses. However, based on exploratory data analysis and predictive modeling, we can see that Big Mountain is able to increase their weekend ticket price by ~$2 to $83 a day, and potentially increase revenue by $3.5 million for the ski season, which would more than offset the price of a new trail and new lift. There are still other factors to analyze such as Big Mountain’s individual operational expenses and their individual customer data as there may be outliers compared to the rest of the ski market, but this was a good start for our stakeholders to have evidence based information to begin A/B testing our results on customers and our infrastructure and potentially proceed with implementation.

**Future scope of work**

The biggest issue in these predictive modeling scenarios is that we are predicting ticket price on: vertical drop, snow making acreage, total chairs, fast quads, runs, longest run, trams and skiable terrain. While some of these variables account for operational expenses of the resort, they do not account for all operational expenses. We absolutely need more data to formulate a more concrete plan. I suggested in my original problem statement that we really need to know the full operational costs of Big Mountain Resort, not just lifts but also other facilities such as hotels, restaurants, electric power and more. What power sources are used and how much do they cost? Would investing in solar power be more efficient?

* Weekday ski prices should be analyzed since they were dropped from the original dataset. Weekday prices could offset weekend prices and could change the price predictions.
* Seasonal, holiday, and multi-day lift ticket prices need to be examined. It is well known in the ski industry that prices change as the season goes on, snow amounts, and whether or not you purchase a ticket online vs. at the window, or through a 3rd party. There are also skiers that use season passes or multi-resort passes and it would be worth knowing how many skiers at Big Mountain and other resorts buy these passes as this can also offset a ticket price prediction and operational expenses.
* Customer data is needed. Who are the customers buying lift tickets? What are their demographics? Are they mostly young families with kids, college kids, and/or older adults? Where do most customers travel from? Are they local vs. regional vs. national vs. international? We need to consider the costs that the consumer spends to get to the resort. Are they staying in a hotel or just coming for the day? These factors can weigh our price predictions differently. We should consider if people are driving, what is the cost of gas and how does this affect them getting here? At the same time we need to consider the cost of a plane ticket and does the consumer have direct access from the airport to our mountain or are they paying for a plane ticket, transportation and a hotel and then buying a lift ticket? All important factors that are not directly related to our operational expenses, but they factor into the customer's reason(s) for buying a ticket in the first place.

The modeled price of $96.32 is well above the actual price of $81. This is validated with the mean absolute error of $10.41. This absolutely would come as a surprise to the business executives, they would want to know what is causing the over prediction?

* I would like to assess the predictive model and specifically the variables and error metrics. The random forest regressor was saved as the best model and was used for price predictions. GridSearchCV was used to validate the best parameters and cross validation was used to compare the linear model vs. random forest model. However, the model was overfitting the training data and over-predicting on the test set.
* We will need to go back to the previous notebook in step 4 and look at MSE differences between train and test sets. We should utilize a “Gap” plot and this equation: ***Gap = Test MSE - Train MSE*** on the y axis and the ***n\_estimators*** on the x axis. This would reveal how much the model is overfitting. We should take a closer look at grid search parameters used to build the random forest and parameters selected. It would be helpful to re-eval this model by looking at other hyperparameters such as max\_depth vs. number of estimators. The max depth of each tree that is built by the random forest model would allow us to see the optimal parameters for the model (Molas, 2022).
* One more issue with the random forest model is in the grid search algorithm. Only n\_estimators was used as a hyperparameter and other important hyperparameters were not considered such as: max\_features, max\_depth, max\_samples and bootstrapping or the oob\_score=True setting which splits the data into in-bag and out-of-bag (OOB) datasets (Ram, 2020). Lastly, we need to consider the concept that shallow depth trees in a RF model tend to reduce overfitting when there is less noise in the training data (Zhou et al. 2021). Certainly during our PCA we were able to reduce some of the noise in the data but perhaps further exploratory analysis to reduce any further noise would be helpful as well.

Assuming the model is deemed useful for Big Mountain Resort, I would recommend hypothesis or A/B testing with customers prior to implementation. This would allow us to perform “real-world” validation and mitigate any risks that may occur during implementation. Furthermore, this would allow us to compare the machine learning model predictions vs. other strategies so we don’t lose time and money when we do implement the model recommendations (Natarajan, 2023).

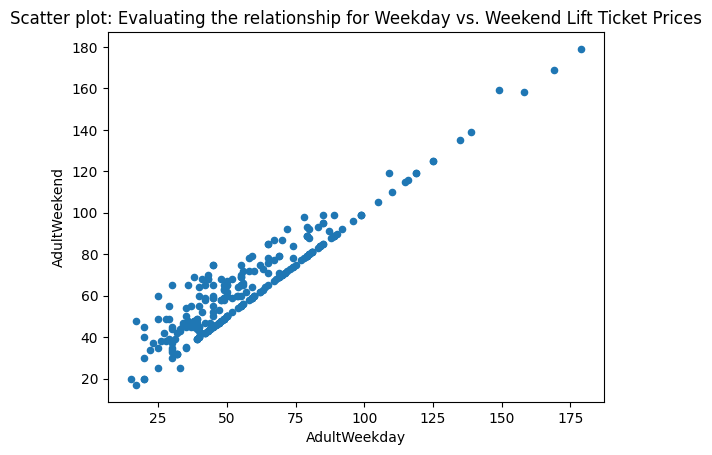
* Source code should be stored in a git repository so that everyone has access to. We can also make the models automated in a machine learning pipeline such as Azure DevOps that can easily be tested when new parameters and data become available.
* We need to be wary of “model drift” and the idea that the data for this model will become “stale”. We will need to consider appropriate timeframes for data and model updates and reassess our plan of action and predictions. Model drift can be detected using machine learning based approaches and statistical tests (All, 2023).
* Lastly, we may want to consider sharing our results with regional and national stakeholders to notify them if we are going to implement these changes such as raising our ticket prices.

**References**

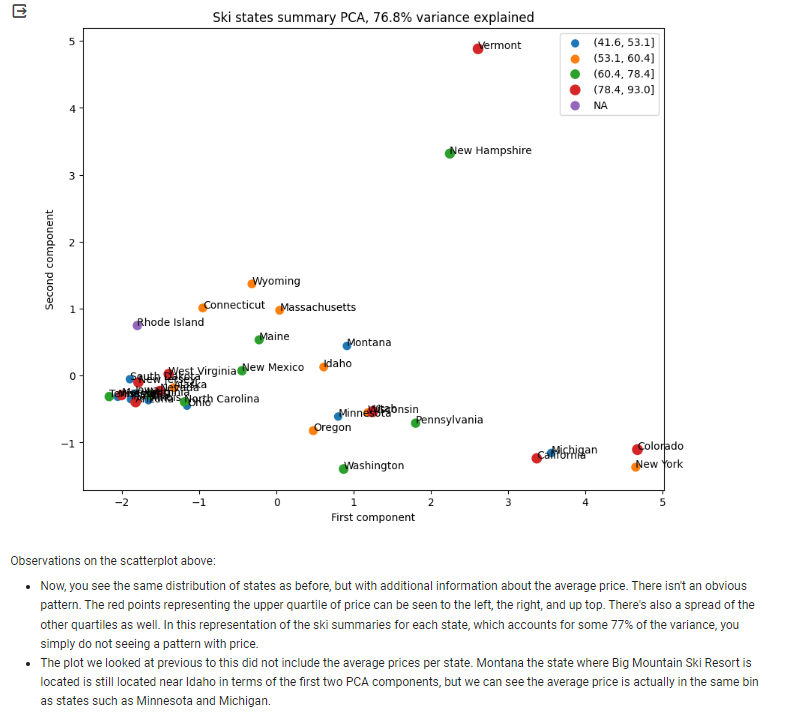
1. All, M. 2023. "Understanding Data Drift and Model Drift: Drift Detection in Python." Retrieved from: [https://www.datacamp.com/tutorial/understanding-data-drift-model-drift](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fwww.datacamp.com%2Ftutorial%2Funderstanding-data-drift-model-drift)
2. Molas, A. 2022. “Can Random Forests Overfit?” Retrieved from: [https://medium.com/@alexmolasmartin/can-random-forests-overfit-a743755251b4](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fmedium.com%2F%40alexmolasmartin%2Fcan-random-forests-overfit-a743755251b4)
3. Natarajan, M. 2023. "A/B Testing in Machine Learning: Unraveling the Experiment." Retrieved from: [https://medium.com/@megha.natarajan/a-b-testing-in-machine-learning-unraveling-the-experiment-10780228cfb4#:~:text=Understanding%20A%2FB%20Testing%20in%20Machine%20Learning&text=Here's%20where%20A%2FB%20testing,predictions%20from%20the%20new%20model](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Fmedium.com%2F%40megha.natarajan%2Fa-b-testing-in-machine-learning-unraveling-the-experiment-10780228cfb4%23%3A%7E%3Atext%3DUnderstanding%2520A%252FB%2520Testing%2520in%2520Machine%2520Learning%26text%3DHere%27s%2520where%2520A%252FB%2520testing%2Cpredictions%2520from%2520the%2520new%2520model).
4. Ram, S. 2020. “Mastering Random Forests: A comprehensive guide." Retrieved from: [https://towardsdatascience.com/mastering-random-forests-a-comprehensive-guide-51307c129cb1](https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Ftowardsdatascience.com%2Fmastering-random-forests-a-comprehensive-guide-51307c129cb1)
5. Zhou et al, 2021. “Trees, Forests, Chickens, and Eggs: When and Why to Prune Trees in a Random Forest.” arXiv:2103.16700v1 [stat.ML] 30 Mar 2021

**Appendix**

**Fig. 1 - Data Wrangling: Evaluating the relationship of 2 target variables**

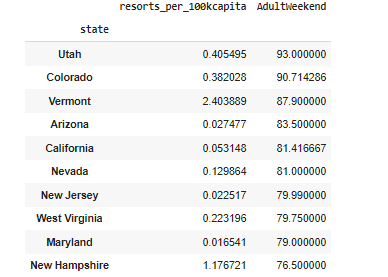
****

**Fig. 2 - EDA: Principal Component Analysis (PCA) with variance explained**

****

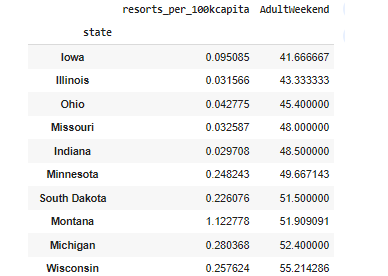
**Fig. 3 - EDA Chart: Resorts per 100k capita vs. Adult Weekend price - ranked in descending order**

* We can see the top 3 in this chart are: Utah, Colorado, Vermont.

****

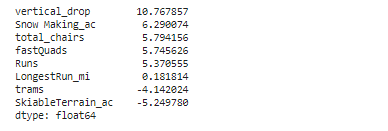
**Fig. 4 - EDA Chart: Resorts per 100k capita vs. Adult Weekend price - ranked in ascending order**

* We can see the bottom 10 in resorts\_per\_100k\_capital are mostly midwest states and this also includes Montana where Big Mountain Resort is located.



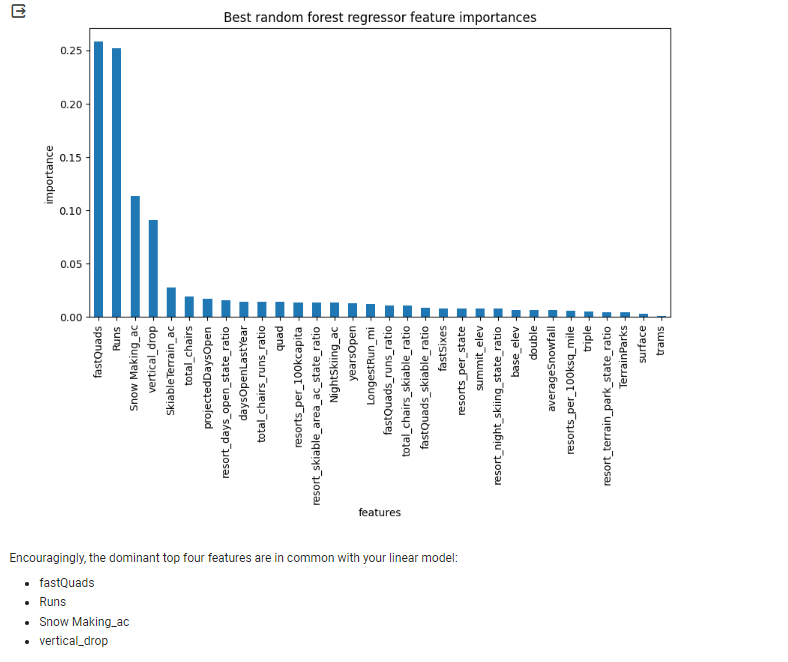
**Fig. 5 - Model pre-processing and feature engineering: Linear Regression Model Results**

* Below we can see the linear regression model coefficients.
* The most interesting finding was the negative correlation seen with SkiableTerrain\_ac suggestive that more terrain means that a larger ski resort can accommodate more skiers and thus support a lower ticket price.



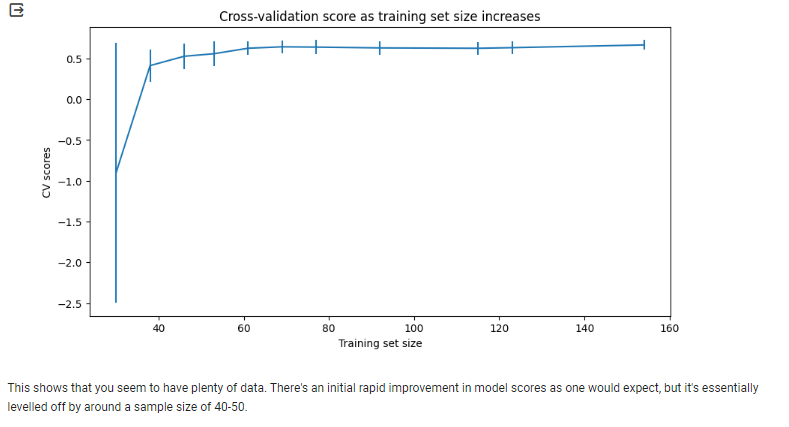
**Fig. 6 - Model pre-processing and feature engineering: Random Forest Model Results**

* We can see the most important features from the Random Forest Model plotted below.



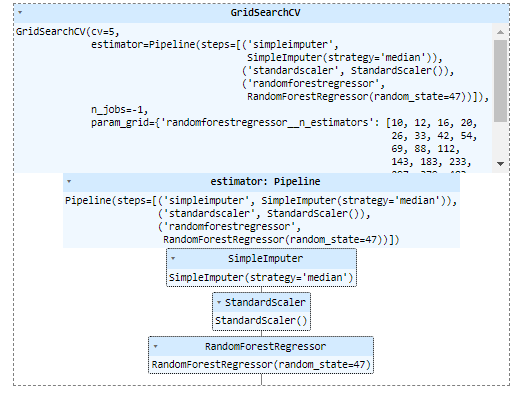
**Fig. 7 - Model pre-processing and feature engineering: Data Quantity Assessed using Cross-Validation**

* Results showed a sample size of 40 to 50 is sufficient before model performance levels off.



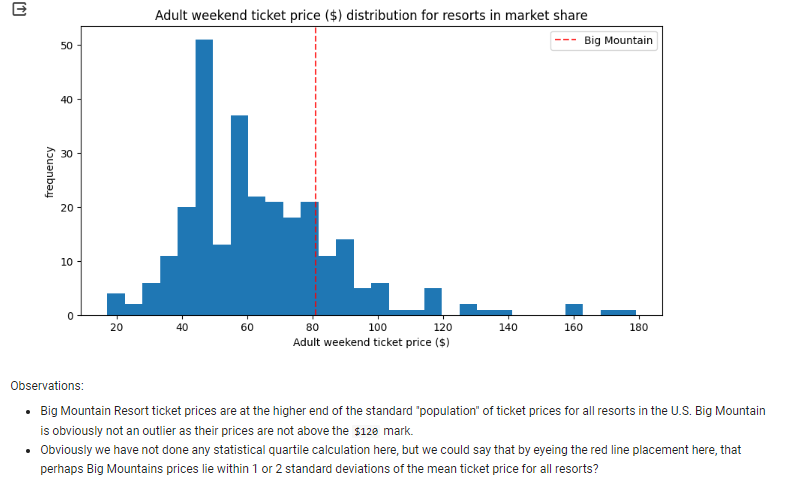
**Fig. 8 - Algorithms used: Sklearn pipeline with GridSearchCV for hyperparameter tuning**

* Example of the pipeline below with the GridSearchCV algorithm for hyperparameter search and fine tuning for the Random Forest Regressor model.

****

**Fig. 9 - Scenario Modeling - Big Mountain’s Ticket prices compared to ski resort market**

* Big Mountain is the **red line** below.

****