1. Introduction

Big Mountain Resort (BMR) is a ski-resort in the Montana state with hundreds of thousands of skiers and snowboarders every year. They have recently installed new chair lift to better distribute the ski-resort attendants around the trails and slopes, which brought additional operational costs to over $1.5M. In order to accommodate to the new costs, BMR needs to identify where better options for ticket prices can be adjusted and where costs can be cut down. The prices of tickets and costs can leverage how the market is favoring some facilities more than others and use that information to appropriately value the tickets for the benefit of the profits of BMR by up to the additional costs of BMR from the new chair lifts. In summary, we have to know:
How much can cost be cut down from snow and lift operations and which tickets can be priced appropriately based on the market to bring higher revenues and reduce costs for a profit margin of at least $1.5M by the end of the year?

2. Data Cleaning, Wrangling, and Exploratory Analysis

To apply a predictive model to find the optimized price for ski resort tickets of Big Mountain Ski Resort (BMSR), we used 2 sets of data: ski resort data for the country provided by BMSR and demographic/state statistics obtained from Wikipedia.
Originally we had the ski resort data set with 330 supposed ski resort records, including BMSR, with 25 potentially useful fields for the model along with an adult ticket price for weekends and weekdays, separately. With 13 fields missing up to 50% of records, as seen in the Fig. below, there was a need to clean the data.
First and foremost, the following were addressed by removing or correcting data:

- **SkiableTerrain_ac** needed corrected values because they clustered down the low end
- **SnowMaking_ac** for the same reason
- trams also may get an amber flag for the same reason
- **fastEight** column was dropped because all but one value is 0 so it has very little variance, and half the values are missing
- **yearsOpen** records above 1000 years were dropped because most values are low and anything over that strongly suggests someone recorded year rather than number of years

Lastly, the population and state statistics data were cleaned to have a one to one relationship with the ski resort data, according to the state/region they were associated with to join them.

In terms of the usefulness of these parameters for ticket price, it was clear to see the disparity between the prices of the different parts of the country, as shown on the next figure on the following page. Overall, all of the data missing both the weekend and the weekday prices were dropped. Additionally we kept all records except the ones with missing weekday price data because there were almost twice as many missing weekday price data then weekend ones. Hence, the weekday price data is the most useful as the dependent variable for our model.

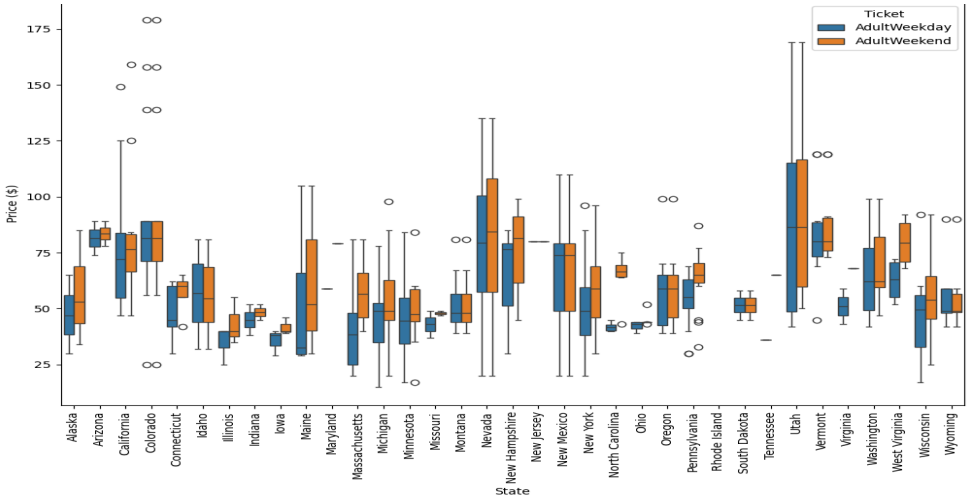

Fig 1. Distribution of all ski resort prices by state, indicating a range roughly between $25 to $100 for most resorts.

Exploring the parameters for the population and state statistics (PSS), we found applying a PCA to optimize and reduce the dimensionality of the dataset. Only 2 of all the features, resorts_per_100kcapita and resorts_per_100ksq_mile, in the PSS dataset are needed to explain over 3/4th of the variance. Including those features to the original ski resort dataset, along with creating ratios with their respective related features (s.a. Ski resort area and total state skiing area), there were key correlations at our disposal. In short, Weekend ticket prices had a R-squared correlation with fastQuads, Runs, Snow Making_ac, total_chairs, vertical_drop, resort_night_skiing_state_ratio, fastQuads, and resorts_per_100kcapita.

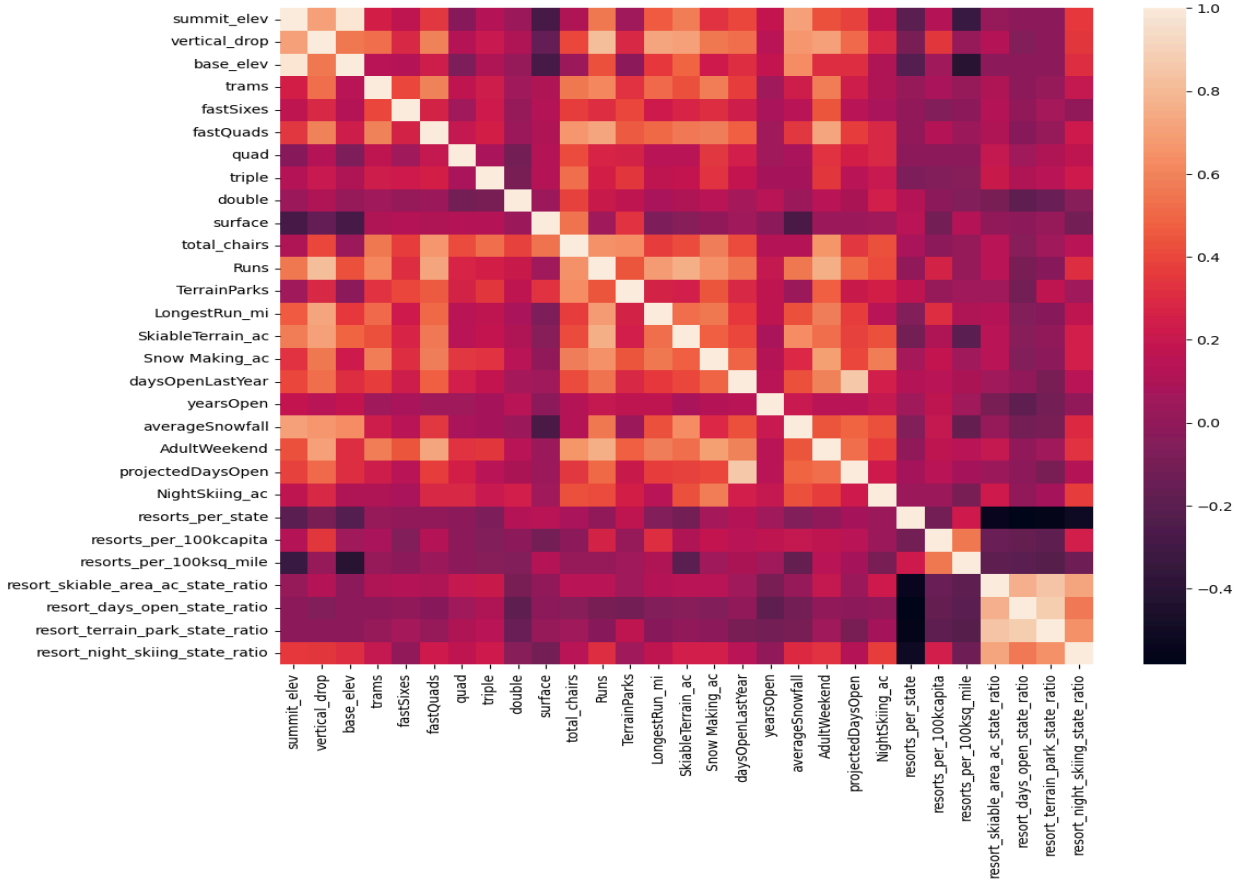

Fig 2. Heat map of the correlation of all features in both data sets, where Adult weekend is shown to correlate with fast Quads, Runs, Snow making capabilities, ratio of resort days open during the year, and resorts per state.

3. Modeling Results and Summary

The average of the y_values we used for training was identical to the constant value of the fitted model. And after fitting the linear model to the average price of the rest of the dataset, the R-squared value on the test array was -.00312! By using the mean-squared error method and mean-absolute error method, the result was just over 19 and 24 dollars, respectively, for the average weekday adult price predictions. Then, we fitted the model to predict values, using the median and mean that were missing in the dataset. And after using those in the trial set and rechecked the mae and mse-root values. There was less than an 11 dollar difference between the predicted amount and the actual price of weekday tickets. But this was the result assuming the entire dataset was useful and not over-fitting. Most importantly, cross-validation made way to see vertical_drop, Snow Making_a, total_chairs, and skiable area as the most correlated for predicting price of tickets. And both cross-validations point to this as well, especially when a random forest regressor was applied to the data set. Thus, with the lowest dollar variability in ticket cost predictive capabilities, as shown below, the random forest regressor is the best model with only a 9.50 dollar mean absolute error.

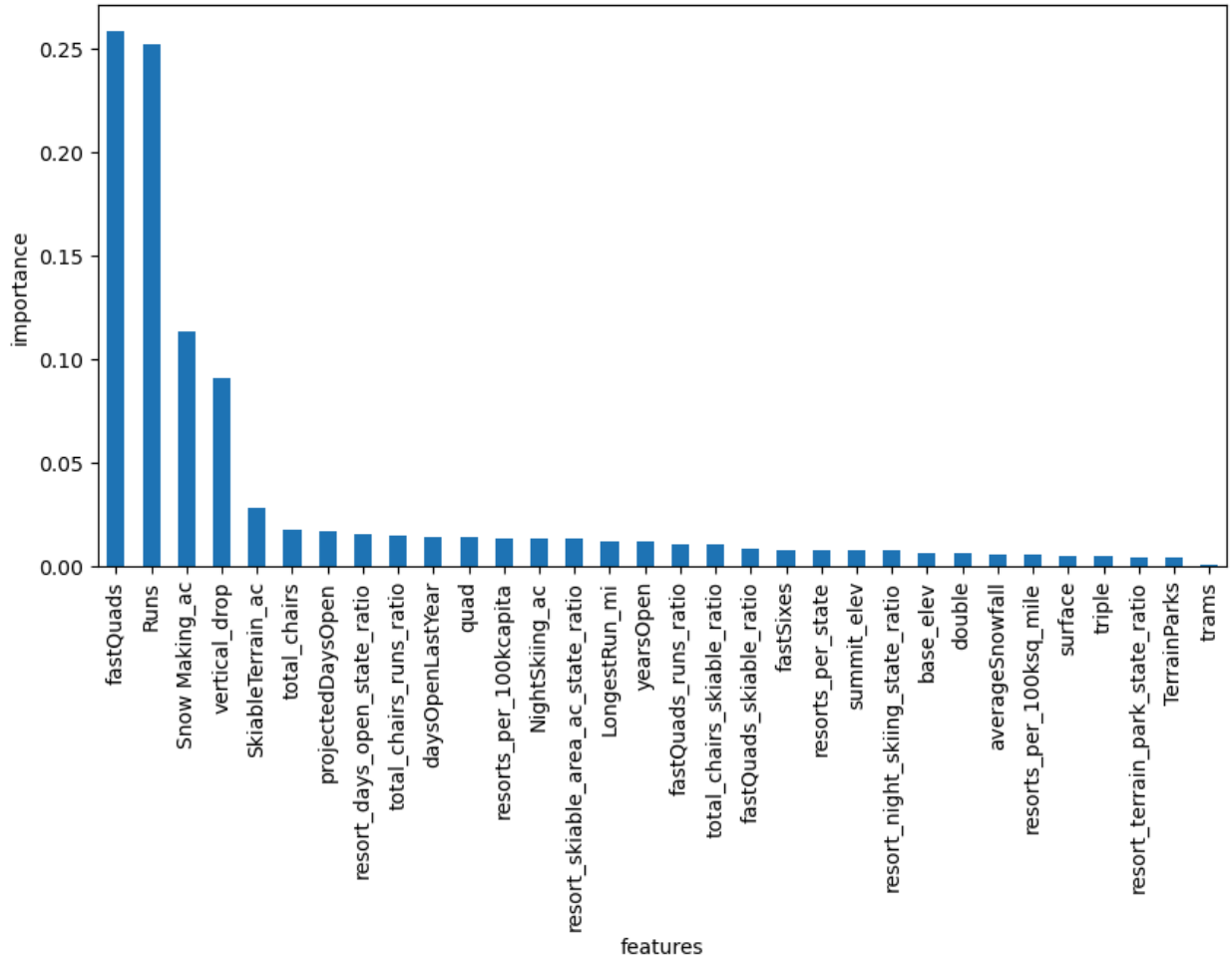

Fig 3. FastQuads, Runs, Snow_Making_ac, and vertical_drop were the most variable features which were best for predicting appropriate ticket cost for a ski resort.

4. Conclusion and Future Scope of Work

Therefore, I recommend adult tickets, on weekends and weekdays, to change from the 80 to 90 dollar range to the 95 dollar price, as indicated by the model. If all 350K guests of the resort, on average skied 5 times, in the coming season skied, this model predicts a revenue of 3.47M dollars while adding a run 150 ft lower down and a chair lift to support it.
On the downside, we have no means to capture any of the expenditures of the new lift and of any already on-going operations because this data has not been provided. In fact, we've had to assume that the additional chair lift would increase the operating cost by over $15M dollars. And surprisingly, closing down up to 5 slopes from 2 runs will not make any difference in possible revenue, as shown below.

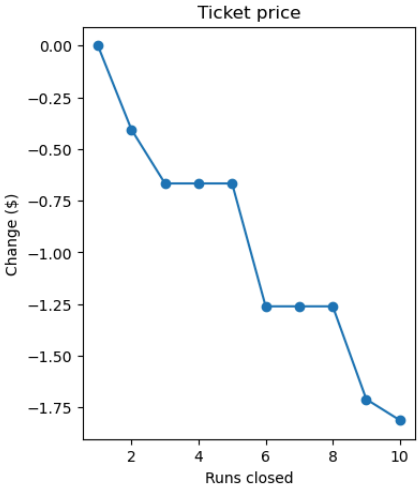

Fig 4. Closing down more slopes is only revenue deterring immediately up to 3 lifts being closed down.