Data Analysis and Investment Strategy Report

PREPARED FOR: Big Mountain Resort
PREPARED BY: Andrew Chen

Introduction

Located in the state of Montana, Big Mountain Resort is a 105-trail ski resort with an annual average attendance rate of 350,000 people. On-site operational equipment consists of 11 lifts, 2 T-bars, and 1 magic carpet for novice skiers. Additionally, Big Mountain Resort has recently installed an additional chair lift which helps clear visitor congestion but increases seasonal operating costs by $1,540,000. Although the current pricing strategy for the resort is to charge a premium above the average price of resorts in its market segment, this does not accurately optimize the investment strategy across separate facilities. To increase the resort's revenue in the upcoming season, ticket pricings need to be improved across different facilities to reflect by minimizing costs and/or developing the resort to support a higher ticket price. This report aims to build a predictive, machine-learning model for ticket price by exploring the relationships between multiple different facilities and their respective geographical and operational characteristics within the given data.

Method

The method used to develop this model is broken down in this order: problem identification, data wrangling, exploratory data analysis, pre-processing/training data, and modelling. Data wrangling consists of acquiring raw data, defining variable data, transforming dataframes, and

identifying any missing or non-sensical values. Diving deeper into the wrangled data, exploratory data analysis consists of understanding data, examining relationships between different variables, and replacing missing or non-sensical values. Next, pre-processing/training data consists of establishing a train/test split distribution, making pipelines, testing trained data, estimating variability/correlation, cross-validating trained data, and deciding on the best trained model to use. Modelling consists of calculating the expected target value, establishing important features that have heavy correlation towards the target feature, observing the graphical relationship of the target feature with each important feature, computing the results of modelling different scenarios to attain the most optimal solution to the problem statement.

Data Analysis

The final cleaned ski data consists of 277 facility entries and 36 variable columns with 11 of those variables created from. The goal was search for patterns and relationships that have a strong correlation to the target ticket price feature, *AdultWeekend*. As the main resort of interest, the data subset of Big Mountain Resort is shown below in (Fig. 1).

Name	**Big Mountain Resort**		
Region	Montana		
state	Montana	daysOpenLastYear	123
summit_elev	6817	yearsOpen	72
vertical_drop	2353	averageSnowfall	333
base_elev	4464	AdultWeekend	81
trams	0	projectedDaysOpen	123
fastSixes	0	NightSkiing_ac	600
fastQuads	3	resorts_per_state	12
quad	2	resorts_per_100kcapita	1.1227776
triple	6	resorts_per_100ksq_mile	8.1610446
double	0	resort_skiable_area_ac_state_ratio	0.1401214
surface	3	resort_days_open_state_ratio	0.1293375
total_chairs	14	resort_terrain_park_state_ratio	0.1481481
Runs	105	resort_night_skiing_state_ratio	0.8450704
TerrainParks	4	total_chairs_runs_ratio	0.1333333
LongestRun_mi	3.3	total_chairs_skiable_ratio	0.0046667
SkiableTerrain_ac	3000	fastQuads_runs_ratio	0.0285714
Snow Making_ac	600	fastQuads_skiable_ratio	0.001

Fig. 1. Based on the final ski_data for the Big Mountain Resort subset consisting of its variable features and its respective categorical or numerical values. The variable features are highlighted in blue, and the feature values are highlighted in yellow. The target feature *AdultWeekend* and its corresponding value are denoted in red.

The target feature *AdultWeekend* was chosen through an evaluation of the relationship between weekday and weekend price to decide which target feature to use when modelling. The following plot examines the variability and distribution of the average ticket price for each resort based on its corresponding state (Fig. 2).

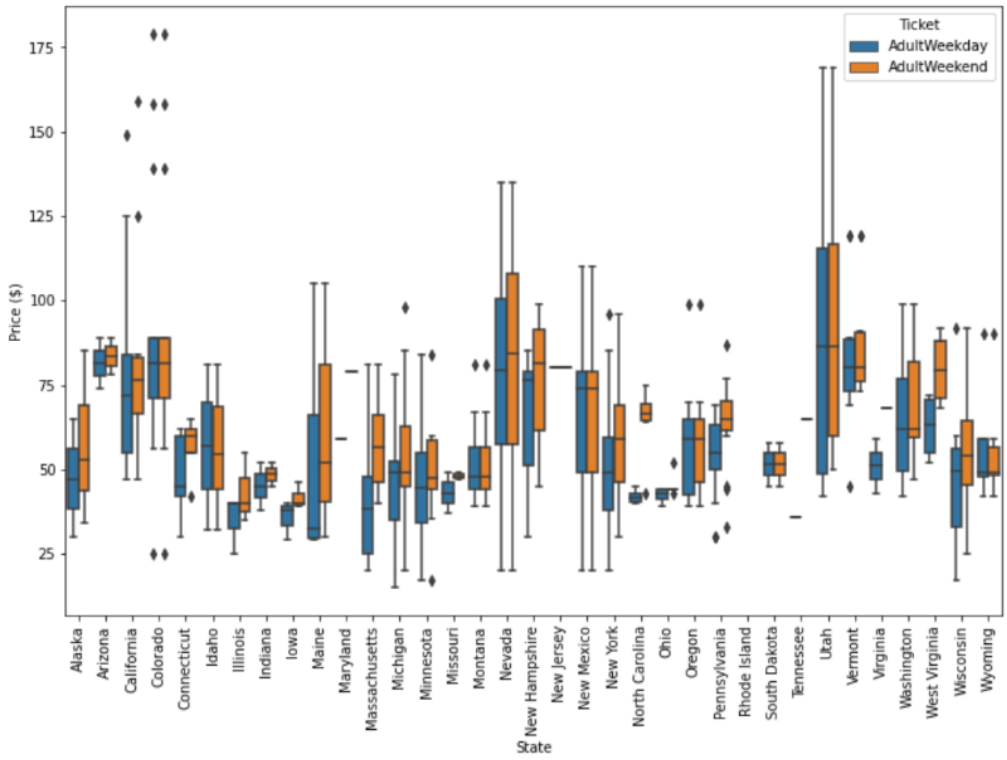

Fig. 2. Horizontal bar plot of the distribution of average ticket Price vs State in which average ticket Price consists of the two ticket price variables: *AdultWeekday* and *AdultWeekend*. Note that most state weekend prices that are higher than its corresponding state weekday prices seem relatively constrained to resorts that have ticket prices less than $95.

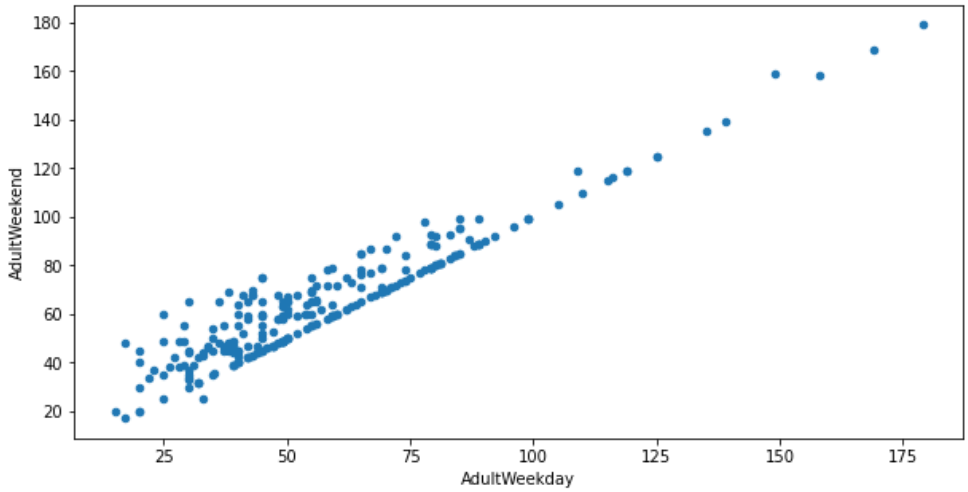

Fig. 3. A scatter plot with a linear relationship of *AdultWeekend* vs *AdultWeekday* in which both its feature variables consist of ticket price values.

Not only is there a visible best fit line in which weekend and weekday prices are equal, but more specifically the distribution for weekday and weekend prices in Montana seemed equal (Fig. 3). Therefore, *AdultWeekday* was removed over *AdultWeekend* because it has a greater number of missing values. After deciding on the target feature, the distribution of each individual feature is explored.

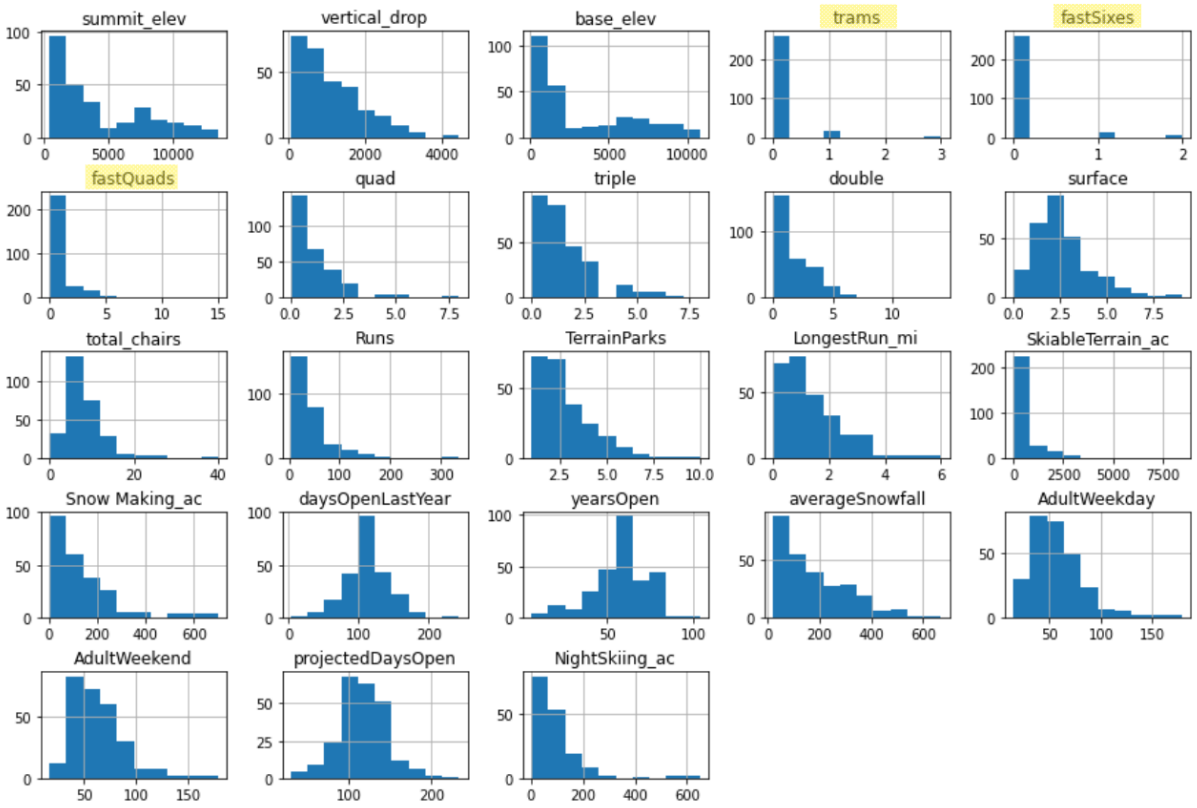

Fig. 4. Distribution of each individual feature values in which y-axis for all subplots is the count of its respective feature value. The highlighted features are plots with the largest leftward skew distribution.

After replacing missing values and nonsensical outliers, a distribution of the feature values was created to visualize, identify, and fix any odd distributions for each feature (Fig. 4). Because the distributions of *fastQuads*, *fastSixes*, and *trams* are skewed towards the left, these feature values have a good number of values that are close to zero if not, equal to zero. Therefore, there is a probability that these features are categorical and/or not weighed heavily towards the target feature of *AdultWeekend* ticket prices.

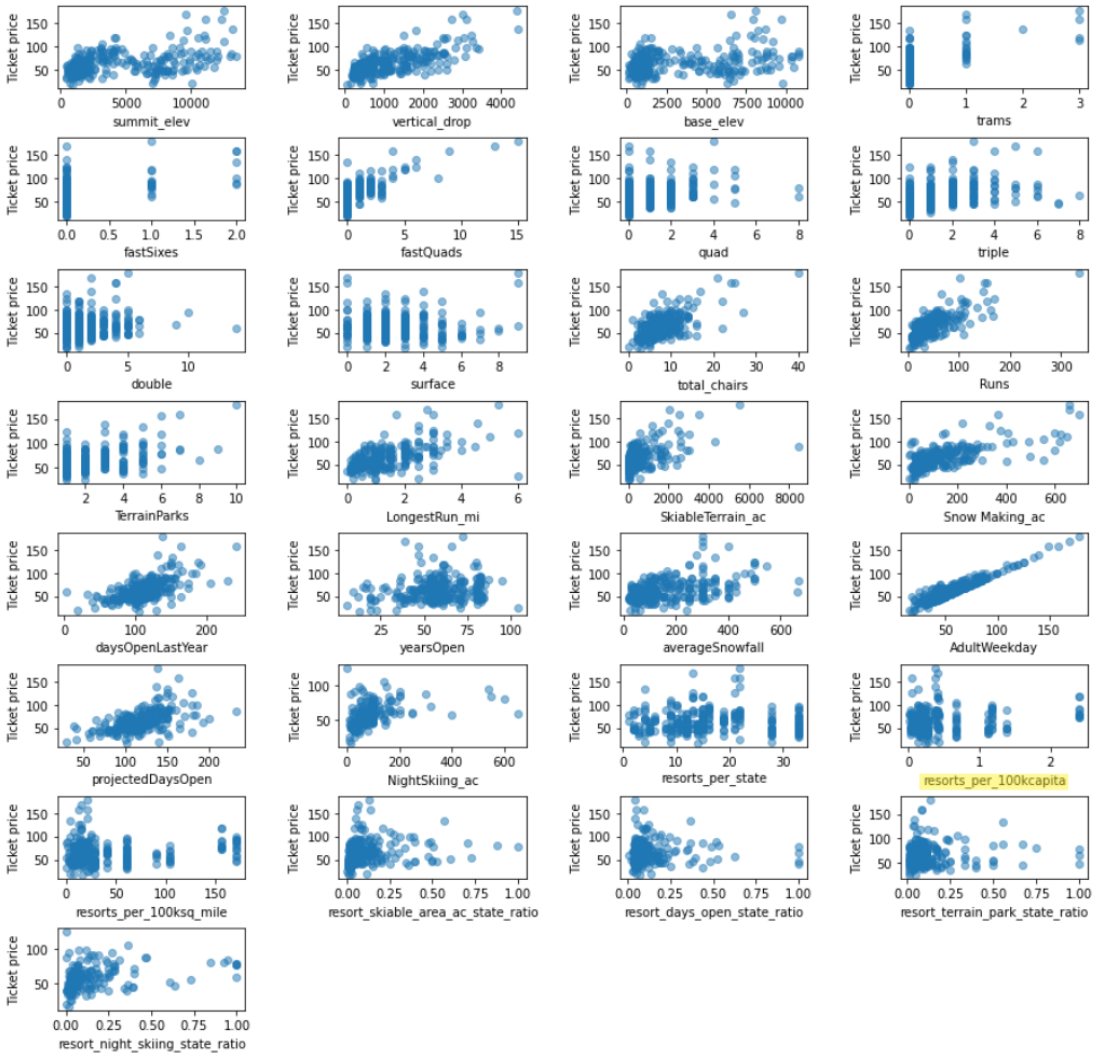

Fig. 5. Series of scatter subplots with the relationship of Ticket Price vs each individual numeric feature in the wrangled dataset. The numerical feature pairs that have a strong positive correlation are *vertical_drop* with *fastQuads* and *Runs* with *total_chairs*. The highlighted subplot of Ticket Price vs *resorts_per_100kcapita* has a unique distribution that can cater towards a non-linear relationship.

By using scatter plots to visualize ticket price variability with each numeric feature, we can more efficiently and accurately examine correlation patterns and relationships of the numeric feature values (Fig. 5). There is a substantial amount of ticket price variability on the lower end of the feature value spectrum as you approach zero in which prices range from a minimum of $17 to a maximum of $179. Because there is a moderate possibility that ticket price will dip before it raises due to the increasing number of resorts per capita, a non-linear relationship is formed. Whether ticket prices follow an uptrend or downtrend is dependent on the specified location's demand for ski tickets. States that have resorts that are uncommon, isolated and/or small, may be able to instill an expensive ticket price monopoly due to a lower population density.

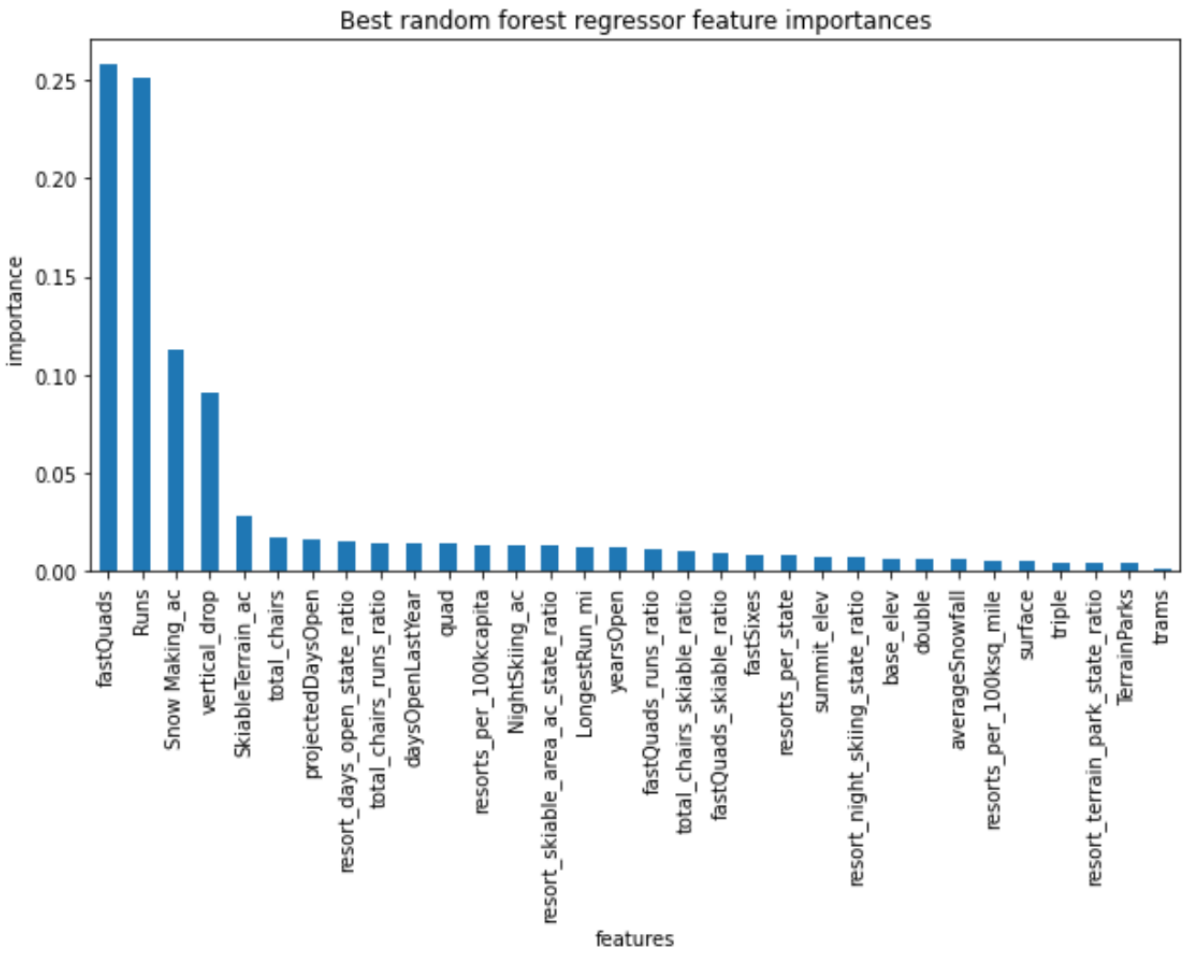

Fig. 6. A bar plot with the relationship of importance vs each numeric feature created with a random forest regression pipeline, cross-validation, and a hyperparameter search through *GridSearchCV*

The random forest regression model was created & fitted by making a pipeline with *SimpleImputer()*, *StandardScaler()*, and *RandomForestRegressor()*. The results were cross validated for the pipeline, *X_train*, and *y_train* with 5 folds followed by a grid search that was created and fitted through *randomforestregressor__n_estimators*. The important features from the random forest regression model were found to be *fastQuads*, *Runs*, *Snow Making_ac*, and *vertical_drop*, respectively (Fig. 6). The cross-validation estimates for mean and std are 0.698 and 0.071, respectively. On the other hand, the test set estimate for mean and std are 0.709 and 0.065, respectively. The performance on the test split was mainly consistent with this estimate because there is a minimal mean increase of 1.6% and std decrease of 8.5%. Therefore, the model best fit to continue is the random forest regression model because it has a lower cross-validation mean absolute error by 0.85, exhibits less variability, and produces consistent performance between the test set and the cross-validation results.

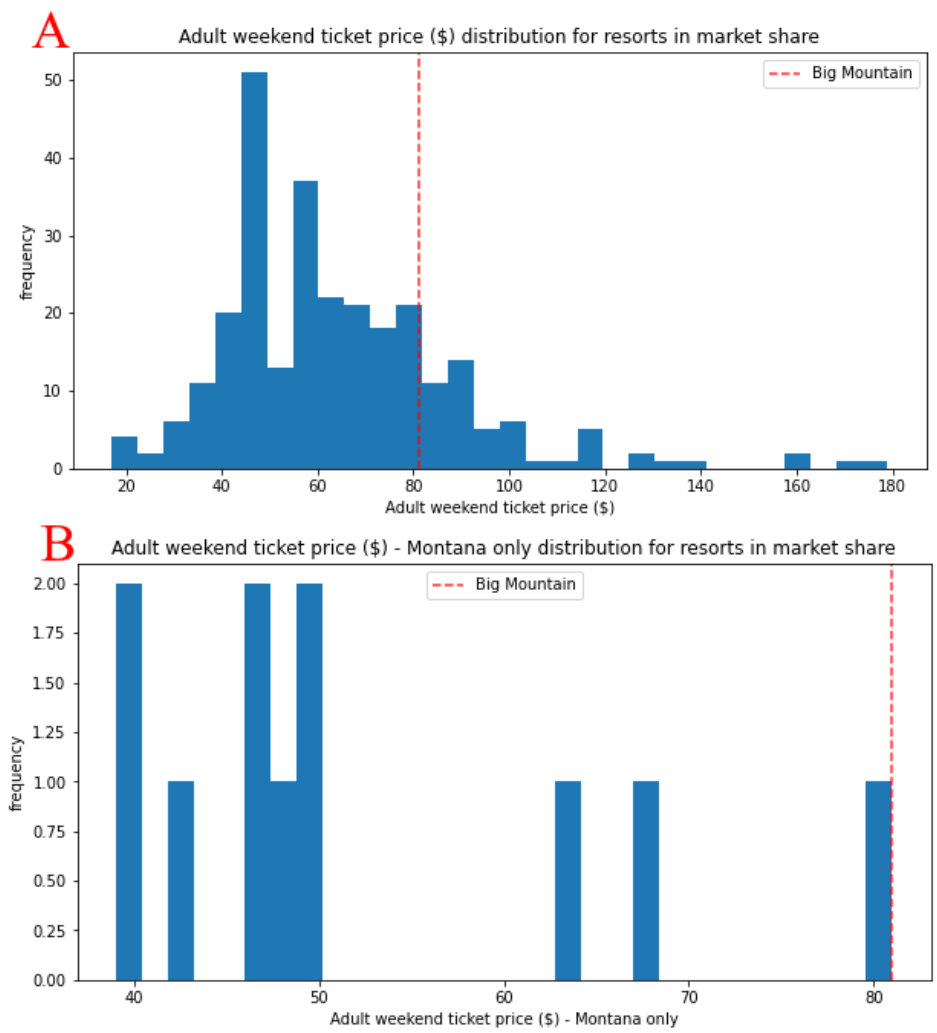

Fig. 7. Two histogram plots with the relationships of frequency vs *AdultWeekend* ticket prices in which (a) is the ticket price distribution for all resorts in market share, while (b) is the ticket price distribution for only Montana resorts in market share. The resort of interest, Big Mountain, is denoted with a vertical-dotted red line at its current ticket price of $81.

These two histograms help visualize the difference in variability between Montana resorts and all other resorts (Fig. 7). For plot (a), the current Big Mountain ticket price of $81 is 0.68 standard deviations above the mean *AdultWeekend* ticket prices for all resorts in the market share in which the calculated metrics of mean and standard deviation are 64.3 and 24.6, respectively. For plot (b), the current Big Mountain ticket price is 2.2 standard deviations above the mean *AdultWeekend* ticket prices for only Montana resorts in the market share in which the calculated metrics of mean and standard deviation are 51.9 and 13.1, respectively. The values calculated with the target feature for Big Mountain are all greater than or equal to 0.68 standard deviations above their respective mean feature values.

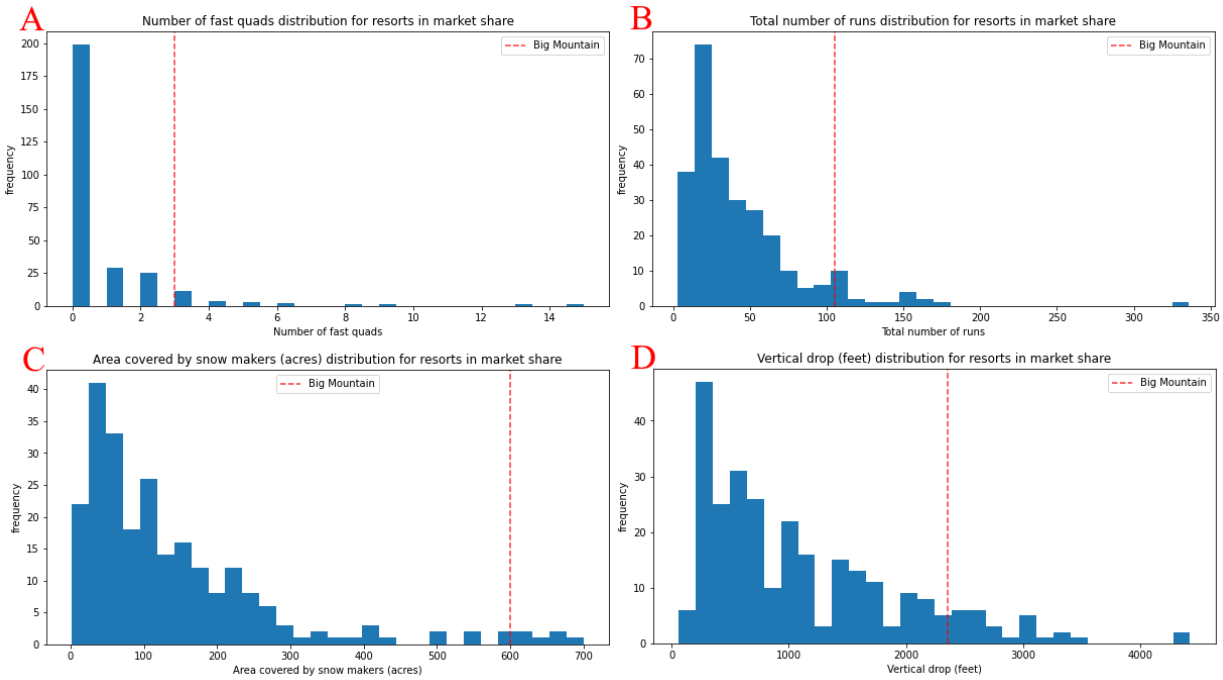

Fig. 8. Four histogram plots with the relationships of frequency vs each of the dominant top four features found using the cross-validation/grid search through a random forest pipeline in which (a) is the number of *fastQuads* distribution for resorts in market share, (b) is the *Runs* total distribution for resorts in market share, (c) is the *Snow Making_ac* covered area distribution in acres for resorts in market share, and (d) is the *vertical_drop* distribution in feet for resorts in market share. The resort of interest, Big Mountain, is denoted with a vertical-dotted red line at its current ticket price of $81.

Recall from Figure 6 in which the dominant features from the random forest regression model were found to be *fastQuads*, *Runs*, *Snow Making_ac*, and *vertical_drop*, respectively. Plotting these four dominant features as histograms yields a result in which all four histograms are skewed to the left not only due to outlying extremities, but also due to a decent portion of the respective dominant feature values being on the lower end of the spectrum (Fig. 8). For plot (a), the current Big Mountain *fastQuads* amount of 3 is 1.3 standard deviations above the mean *fastQuads* amount for all resorts in the market share in which the calculated metrics of mean and standard deviation are 0.7 and 1.7, respectively. For plot (b), the current total Big Mountain *Runs* of 105 is 1.6 standard deviations above the mean total *Runs* for all resorts in the market share in which the calculated metrics of mean and standard deviation are 43.6 and 37.6, respectively. For plot (c), the current area covered by *Snow Making_ac* for Big Mountain is 600 acres. This *Snow Making_ac* value is 3.3 standard deviations above the mean area covered by *Snow Making_ac* for all resorts in the market share in which the calculated metrics of mean and standard deviation are 140.0 and 138.6, respectively. Lastly for plot (d), the current Big Mountain *vertical_drop* of 2353 feet is 1.5 standard deviations above the mean *vertical_drop* for all resorts in the market share in which the calculated metrics of mean and standard deviation are 1106.5 and 843.7, respectively. The values calculated with the four important features for Big Mountain are all greater than or equal to 1.3 standard deviations above their respective mean feature values.

There are four potential scenarios that Big Mountain Resort has given for either cutting costs and/or increasing revenue through higher ticket prices. Scenario (3) and Scenario (4) feels erroneous because its effect on ticket prices is minimal to none. Therefore, Scenario (1) and Scenario (2) will be the forefront of model evaluation and analysis with the former being examined first. Scenario (1) dealt with permanently closing down up to 10 of the least used runs.

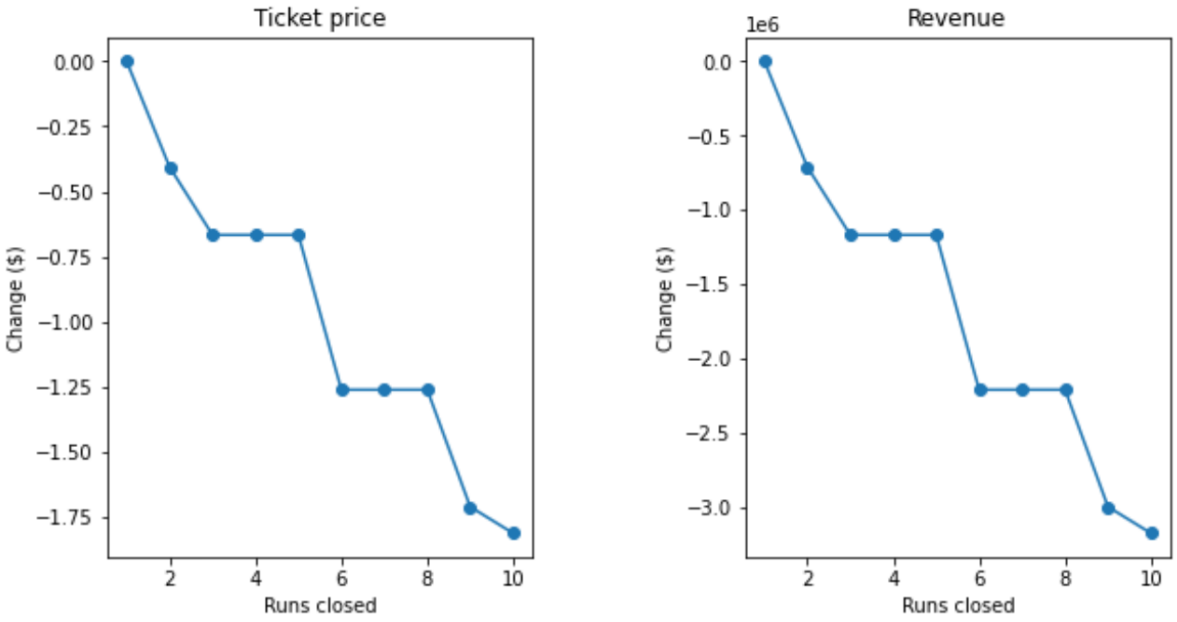

Fig. 9. Two separate plots for Scenario (1) with the relationship of the predicted ticket price change (*delta*) and the associated predicted revenue change vs number of *Runs* closed for each condition in the scenario. It was assumed that each visitor would purchase 5 tickets.

The Scenario (1) results indicate that closing 1 run does not change the ticket prices, closing 2-3 runs reduces ticket prices, closing 3-5 runs has no further obvious loss on the reduction of ticket prices, and closing 6+ runs greatly reduces the ticket prices. Based off the given data for Scenario (1), it seems keeping the same ticket price of $81 is favored. Subsequently if the cost of operating the least used runs supersedes the gain in revenue from the least used runs, then closing 5 runs may be appropriate. This would bring down the ticket prices to $80.33 but remove operational costs for the 5 closed runs.

Scenario (2) revolved around adding a run, increasing the vertical drop by 150 feet, and installing an additional chair lift. Based off the given data for Scenario (2), the ticket prices have increased by $1.99 totaling to $82.99. At the same time, it will not only increase revenue by ~$3.475 million, but also increase visitor appeal and improve visitor congestion. New runs and chair lifts are being added which drives up installation costs, so increasing ticket price is a highly viable option. If an additional chair lift costs about $1.54 million, increasing the seasonal ticket price by $0.88 to $83.87 can cover the lift's cost of installation.

Conclusion

As established before, adding another run with an increased vertical drop and an additional chair lift will increase construction/operational costs. This is a problem because there are many unknown variables when it comes to the costs of Big Mountain Resort. This data deficiency includes the lack of operational costs such as installation, staff, equipment/ground maintenance, electricity, and insurance. Another useful dataset can be the revenue and profit information could have been helpful when comparing the financial performance of the different facilities. Access to this data surrounding cost can narrow down variability and increase model training/performance efficiency.

In conclusion, Big Mountain Resort operates within a market where people pay varying ticket prices based on its respective facility; therefore, optimization of ticket price is crucial to stay competitive. Big Mountain Resort currently charges $81 for a ticket. On the other hand, the modeled price yielded a ticket price increase of $14.87 totaling to $95.87 for a single ticket. This 18% increase in ticket price may seem too high, but then because the expected mean absolute error is $10.39, the ticket price hypothetically still has space to grow. Out of all the facilities in Montana, not only does Big Mountain Resort already have the greatest ticket price, but also the resort's pricing strategy has always been to charge a premium above the average price of resorts in its market segment. Seeing as the next highest ticket price in Montana requires over a 20% increase to reach the value of Big Mountain’s current ticket price, an 18% markup on Big Mountain’s current ticket price is the optimal solution that can increase Big Mountain Resort’s seasonal revenue by $26,022,500. This holistic change in investment strategy is highly supported as not only are the new ski features enticing and marketable, but also the ticket price strikes the healthy balance between consumer affordability and increased revenue, while still being able to securely cover the short and long-term installation plus maintenance cost for the new operating equipment.