1. Summary On Data Wrangling

1.1 Data Wrangling on Original Ski Data

The original dataset has 330 rows of records and with 27 columns/features. Among 50 states, New York accounts for the majority of resorts (33 out of 330). Big Mountain Resort is in Montana, which comes in at 13th place and has 12 resorts in total. See Figure 1 in Appendix A for distribution of resorts by states. Aside from some relatively expensive ticket prices in California, Colorado, and Utah, most prices appear to lie in a broad band from around 25 to over 100 dollars. See Figure 2 in Appendix A for average price by states. Not all states are homogeneous. Some States show more variability than others. For example, Montana and South Dakota, show fairly small variability. Nevada and Utah, on the other hand, show the most range in prices. See Figure 3 in Appendix A for the side-by-side boxplots of the ticket price distribution by state

There are missing values. Some resorts have more missing than others, and some features of the resorts are missing more recorded information than others. The most important missing worthy of mentioning is in the two target quantities, the adult weekday ticket price and the adult weekend ticket price. They each has about 15-16% of values missing. The missing in adult weekend ticket price (~15.45%) is less serious that in the adult weekday ticket price (~16.36%), and 47 resorts are missing both prices. Just over 82% of resorts have no missing ticket price. Another heavy missing is on the number of fast eight person chairs ("FastEight"), missing more than half (166 out of 330 total). There are erroneous recordings as well, for example, with Pine Knob Ski Resort on opening years, Heavenly Mountain Resort on area snow making area, and Silverton Mountain Resort on skiable terrain area.

Table 1 in Appendix B summarizes missing information in the original dataset and actions we took in data wrangling. i). We dropped feature "Fast Eight-person chairlift" due to its over 50% missing; ii). We dropped the adult weekday ticket price and kept the adult weekday ticket price for two reasons. On one hand, missing is more serious in weekday prices than that in the weekend price. On the other hand, not only all 12 resorts in Montana but also more than half of all resorts (106 in total) have the two types of ticket prices matching. See Figure 4 for the matching feature. iii). We dropped two resorts, Pine Knob Ski Resort and Heavenly Mountain Resort, for erroneous recording, iv). We deleted 51 resort records that are missing adult weekend price, including Heavenly Mountain Resort. v). We kept Silverton Mountain Resort, and imputed its error input on skiable terrain area.

We are now left with a dataset of 277 resorts with 25 features/quantities for analysis.

1.2. Calculate State Summary and Pull in State Population and Sizes

The remaining 277 resorts are found to scattered in 35 states. On a state level, we then calculated state summary statistics, such as i). state total numbers of resorts, state total number of days of resort openings in last year, state total skiable area, state total number of terrain parks and state total night skiing area, ii) two state resort density measures: the ratio of state total number of resorts per 100k people, and the ratio of state total number of resorts per 100k square miles.

Because of the nature of the resort business and the heterogeneity of demographic features across 50 states of the United Sates, On these summary statistics, we also pulled in state population and area information from wikipedia.

2. Summary of EDA

2.1 State Level

Principal Component Analysis of State Level Summary Statistics vs Ticket Price, Shows Random Pattern

At this preliminary step, we did two layers of analysis, on the state level and then down to the resort specific level. On the state level, we applied principal component analysis (PCA) on state summary statistics, including 7 summaries on features and 2 calculated state resort density measures. It was found that there is no clear pattern in the distribution of ticket prices across states, as reflected in Figure 5 where the first two PCA scores were plotted. In Figure 5 ranges of ticket prices are reflected by different sizes and colors of the bubble. As we can see, ticket prices in States with similar features can range quite bit. This finding offers justification for building a pricing model that leaves out state as a predictor.

2.2 Resort Level

Onto the resort specific layer, we examined the binary correlation between the ticket price and various resort specific features. Figure 6 in Appendix A presents these binary scatterplots. We found that there's a strong positive correlation of a resort's ticket price with features such as 1) vertical drop from summit to base, 2) the total area covered snow making machines, 3) the number of fast four person chairlifts and the total number all chairlifts, 4) the number of runs.

For features such as resort skiable area, days of resort open in last year, the number of terrain parks at the resort, and the resort night skiing area, we also examined whether their resort-to-state shares matters the resort's ticket price. The measures we used for analysis are ratios of the resort specific values to the state summaries (total value). Among the ratios, the one ratio of a resort night skiing area to its state total stands out and appeared to be most correlated with a resort's ticket price.

A resort's chairlift resource availability was measured by four ratios and their correlation with the ticket price was examined. The four ratios are the ratio of total number of chairs to runs, the ratio of total number of chairs to skiable area, the ratio of total number of fast four person chairlifts to runs and the ratio of total number of chairs to skiable area. Among the four, the first two actually showed negative correlation with a resort's ticket price. The less available our chairlifts is, the less people are willing to pay? It seems odd? To further investigate, we would need data on the number of visitors at the resort in last year.

3. Data Pre-processing

3.1 Taking Average Approach

Taking Average Approach, the one we are currently adopting. Conclusion: under this approach, we would expect to be off by around $19 on average.

3.2 Linear Regression modeling

Implemented python GridSearchcv capability. Cross-validation (with 70/30 split between train/test data) was utilized to compare various models with different combinations of some or all of features. Grid search results are presented in Figure 7 in Appendix A, where performance score (r-square) is plotted for various models. According to the plot, the optimal linear model was found to be with eight

best predictors, namely, a resort's vertical drop, areas covered by snow making equipment, total number of chairlifts, total number of four-person chairlifts, length of the longest run, the number of trams and the area of skiable terrain, in a descending order of significance. The middle column of Table 2 in Appendix 2 lists the estimated coefficients. The larger the absolute value of a coefficient is; the more power the feature has to affect the ticket price. A resort's vertical drop is found to be the biggest positive feature to bring up its price, Also, the area covered by snow making equipment is a strong positive as well. It seems that resorts with bigger vertical drops, larger area covered by snow making equipment, more chairlifts (especially fast four person ones), longer and/or more runs, less skiable terrain area and fewer number of trams have more leverage to set a higher price; Under this modeling approach, on average we would expect to estimate a ticket price about $10 or so off the real price (see Table 3). This is much, much better than the $19 from just guessing using the average.

3.3 Random Forest Regression modeling

Implemented python GridSearchcv. Cross-validation (with 70/30 split between train/test data) was utilized to compare models/forests of different sizes. A random forest with 69 decision trees were found to be optimal. Importance of all features were ranked, and the result is plotted in Figure 8 in Appendix A. Although the order of importance is different, this model does confirm the four most important features we found in linear regression, namely the number of fast four person chairlifts, the number of runs, area covered by snow making equipment and vertical drops. The importance of all features has a big drop/decline after the fourth one (vertical drop). So we suspect that a random forest model including only top four features might be sufficient enough, if a simpler version is preferred. Compared to the linear model, this random forest model provided better and more reliable/consistent predictions.
See below for three model comparison.

3.4 Model Performance Comparison

Compared to the linear model, this random forest model provided better and more reliable/consistent predictions with less variability. The stats are shown in Table 3 (model comparison) In Appendix B. Besides, it does not assume linearity between the ticket price with other resort features. The linearity assumption is rarely true in reality.

3.5 Data Quantity Assessment

Do we have enough data to support our finding? The answer is yes, as shown in Figure 9 (learning curve plot) in Appendix A, where we plotted Python learning curve, training data size vs performance score. The plot shows an initial rapid improvement in model scores as training data size increases, but it's essentially levelled off by around a sample size of 40-50. With our 70/30 train-test data split, we ended up with maximum of 193 data records in the training dataset, far more than enough.

4. Summary of Modeling

4.1 Big Mountain Resort is undercharging

With all resort data but Big Mountain's, the estimated mean prediction error is found to be $10.41 with a marginal error of $1.46. Our model predicts $96.32 as Big Mountain's fair market price, about

$15 more than our actual price of $81.00. Even with the expected error of $10.41, this suggests there is room for an increase. If we assume 350,000 expected visitors over the season and 5 tickets per visitor, then a conservative $5 increase per ticket would mean about $8,750,000 increase in revenue.

4.2 Explore Business Scenarios

Should we add more runs? Expand existing skiable area? or add more chairlifts? Knowing our standing in the league tables of these features helps us answer these questions. Table 4 in Appendix B summarizes information of our standings, together with potential feature modifications we proposed to make and our expectations. Table 5 In Appendix B summarizes business scenarios that we experimented and their findings.

5. What is Next

5.1. Management

1. Consider increasing both types of ticket price immediately; and If time allows,
2. consider adopt the business proposals given in Table 6 in Appendix B, where we present our recommendations for executives to consider. Of course, the feasibility of some will need additional information, as stated in Table 6, to further investigate.

5.2. Data Science Team

For that our data science team will need to immediately get on the following tasks:

- Gather information on number of visitors from the past skiable seasons;
- Identify runs with least average number of visitors and the operational cost of each;
- Estimate potential educed increase in operational cost of remaining operating runs;
- Cost and time to add vertical drop by adding a run of 150-200 feet deeper
- Cost and time to add a new fast four-person chairlift

APPENDIX A

Figure 1. Resort Distribution by State

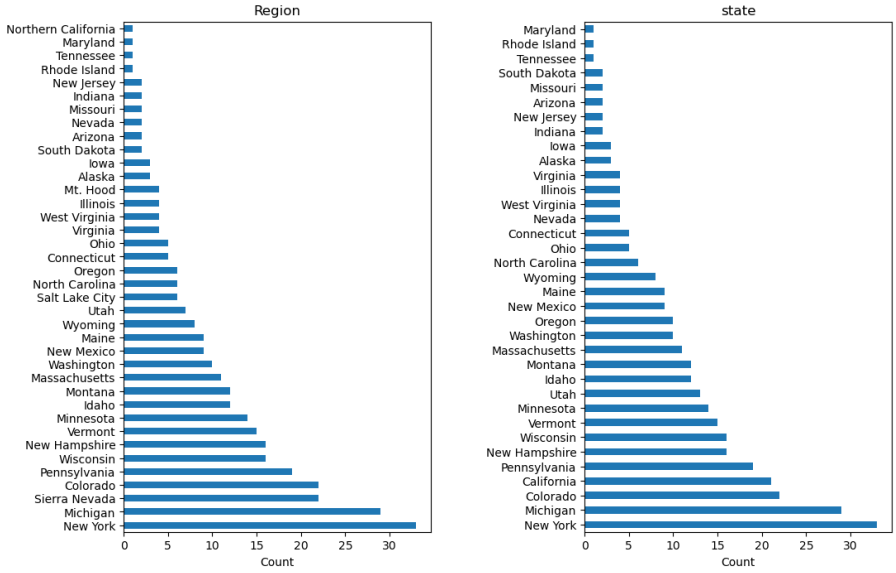


Figure 2. Avderage Ticket Prices by State

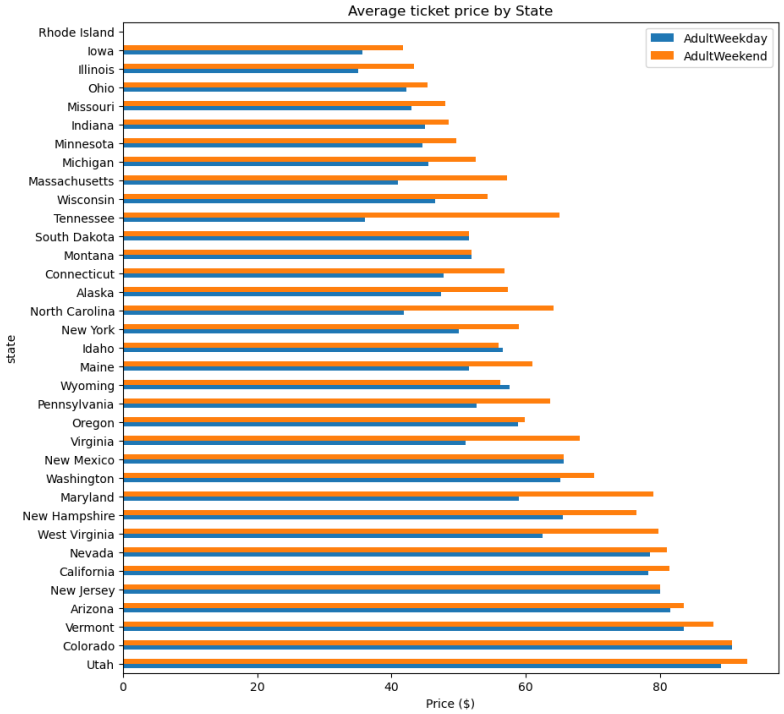


Figure 3. Boxplots of Ticket Price by State

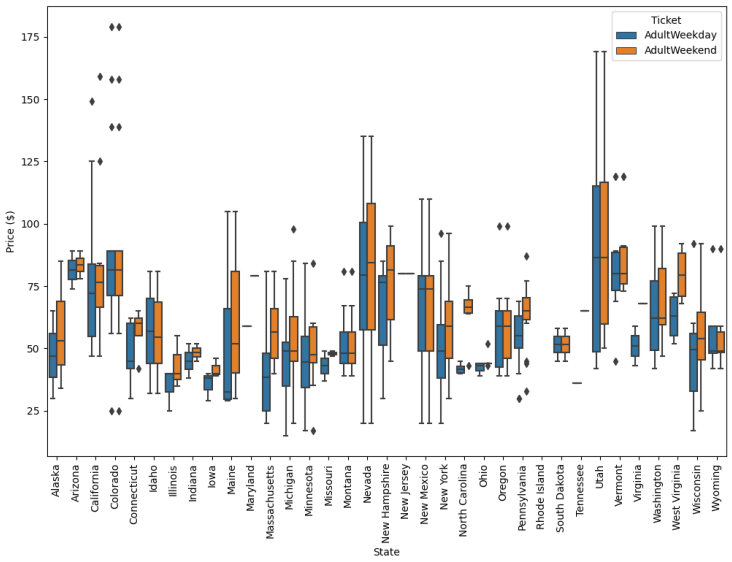


Figure 4. Weekday Price vs Weekend Price by Resorts

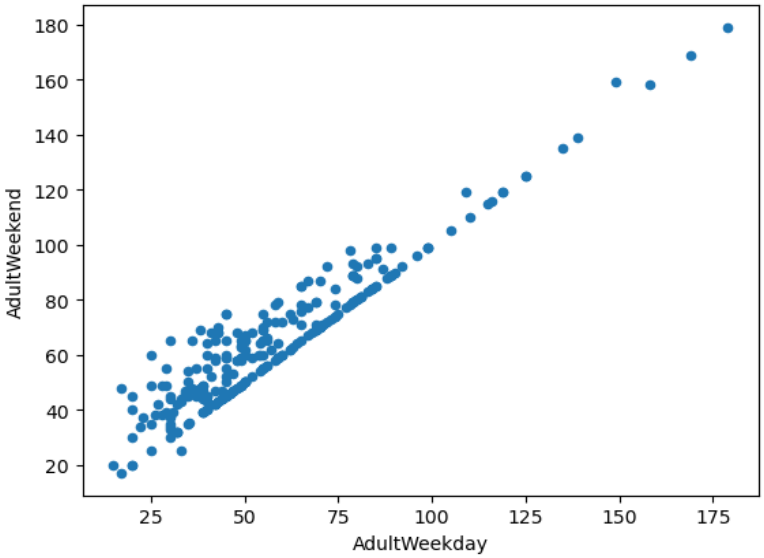


Figure 5. Principal Component Analysis of
State Level Summary Statistics vs Ticket Price:
Shows Random Pattern

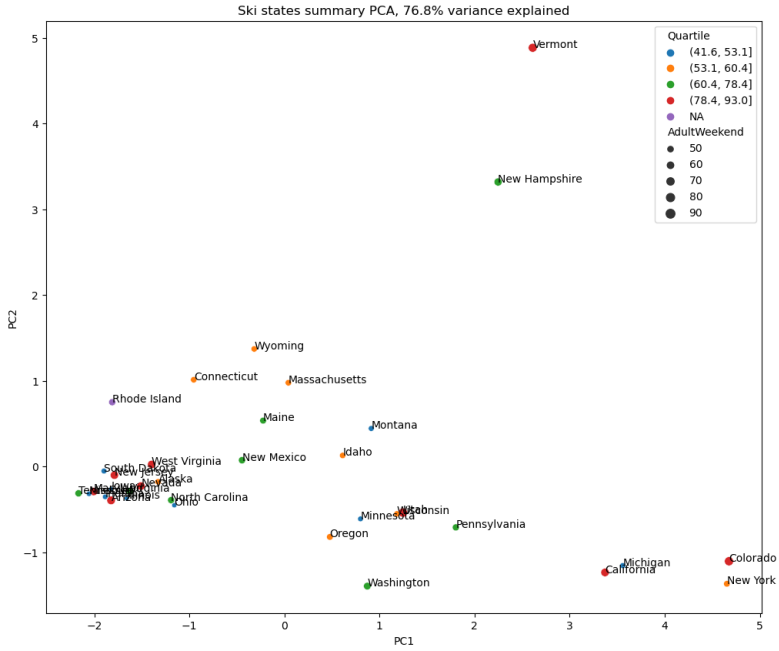


Figure 6. Binary Scatterplots of Ticket Price vs Resort Features

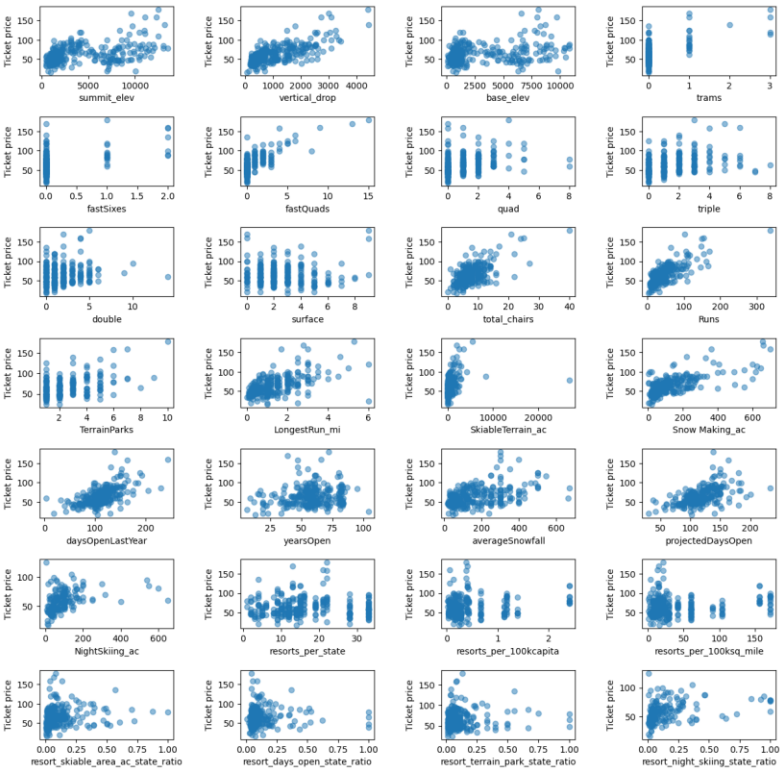


Figure 7. Linear Regression GridsearchCV Result

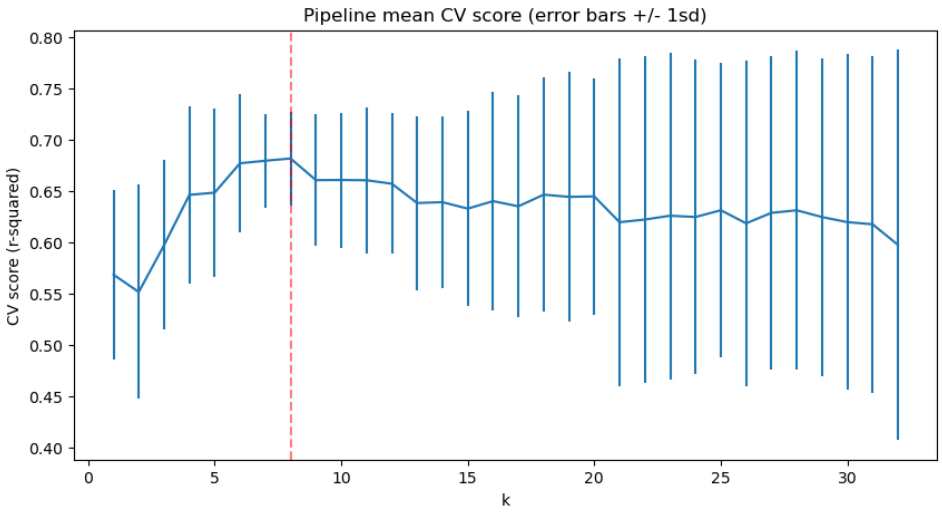


Figure 8. Feature Importance Rank Plot

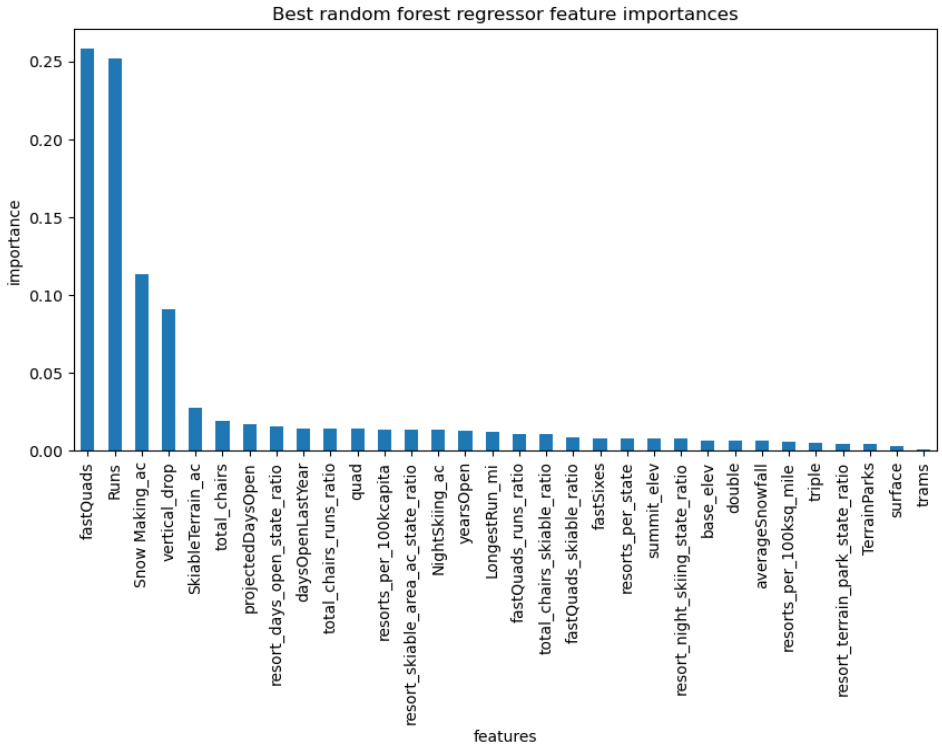
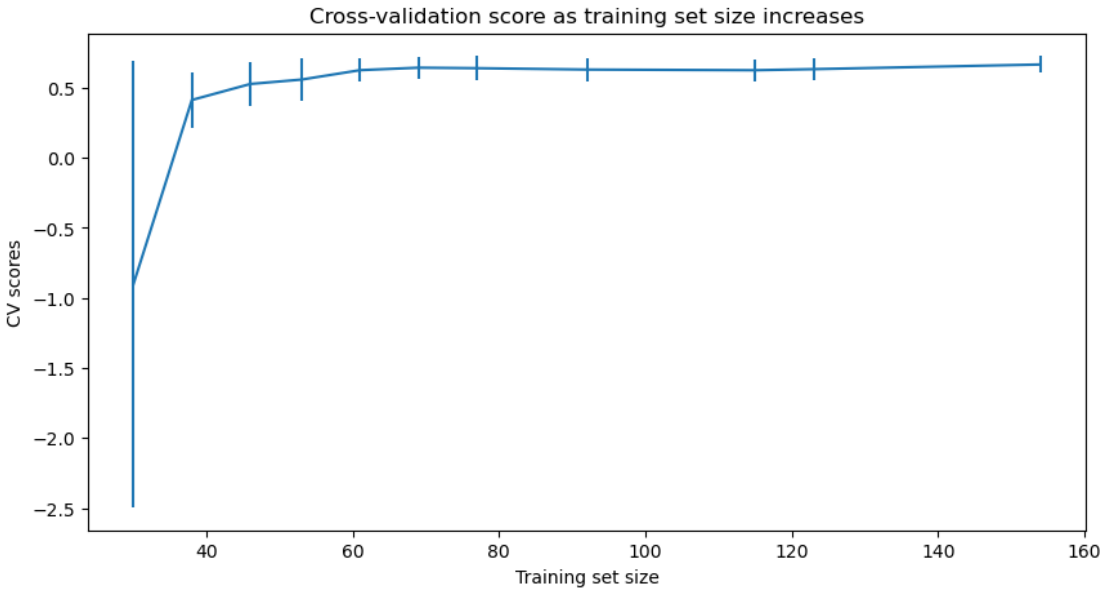


Figure 9. Data Quantity Assessment



APPENDIX B

Table 1. Features/Records Dropped[2]

Feature/Record	Reasons		Action Taken
Fast Eight-person chairlift	Missing 166/330		Dropped
adult weekday ticket prices	Missing 54/330	47 overlap missing	Dropped
adult weekend ticket prices	Missing 51/330		Kept[2]
Silverton Mountain Resort	Missing Skiable Terrain Area		Imputed
Pine Knob Ski Resort	Error recording in years of open		Dropped
Little Switzerland	Missing years of open		Dropped
Heavenly Mountain Resort	Error recording in snow making area missing both ticket prices		Dropped[1]
51 Resorts	Missing the weekend price		Dropped
Silverton Mountain Resort	Missing Skiable Terrain Area		Kept and imputed

[1.] Heavenly Mountain Resort is included in the 51 resorts dropped due to missing ticket price

[2.] After this step, the data size is updated to 277 rows by 25 columns

Table 2. Feature importance

Features (standardized)	Linear Regression	Random Forest Regression
	Coefficients	Rank(top 8)
vertical_drop	10.767857	4
Snow Making_ac	6.290074	3
total_chairs	5.794156	6
fastQuads	5.745626	1
Runs	5.370555	2
LongestRun_mi	0.181814	NA*
trams	-4.142024	NA*
SkiableTerrain_ac	-5.249780	5

NA*: the feature did not make to the top 8.

Table 3. Model Performance Comparison

	Train Data		Test Data
Model	Mean Absolute Error	Standard Deviation of Error	Mean Absolute Error
Linear Regression	10.50	1.62	14.11
Random Forest Regression	9.58	1.37	9.54

Table 4. Big Mountain Resort's Standing in The League

Features	**standing**	**Potential change & Expectation**
runs	Compares well for the number of runs. There are some resorts with more, but not many	Closure of a few least used ones→cut down cost
vertical drop	Do well for vertical drop, but there are still quite a few resorts with a greater drop	adding vertical drop→ price increase→revenue increase
snow making area	very high up in the league table.	adding the snow making area→ make no much difference
Length of the longest run	Own one of the longest runs. Longer ones are rare.	Increasing the longest→ make no difference
Trams	The vast majority of resorts, such as Big Mountain, have no trams.	
Skiable terrain area	Amongst the resorts with the largest amount of skiable terrain	Adding skiable terrain → make no difference
Total # of chairlifts	Amongst the highest number of total chairs,	Only make changes when required by changes in other features
Fast quads	Most resorts have no fast quads. Big Mountain has 3, which puts it high up that league table.	

Table 5. Tested Scenario and Findings

Scenario Tested	Findings
• closing down up to 10 of the least used runs	• closing one run makes no difference • Closing 2 and 3 successively reduces support for ticket price and so revenue. • There is a plateau from closure of 3 runs till closure of 5 runs. • Ticket price drops rapidly when increasing the closure to 6 or more.
• Increase the vertical drop by adding a run to a point 150 feet lower down • install an additional chair lift to bring skiers back up.	• support a ticket price increase by $9.13. • increase revenue by $15,978,514
• The above, and • adding 2 acres of area covered with snow making equipment	• No difference from Scenario 2
• Increase the longest run by 0.2 mile • add an additional snow making coverage of 4 acres	• the ticket price stay unchanged.

Table 6. What is Next

suggestion	consideration
increasing both types of ticket price immediately	
Closure of 5 least used runs	Cut on cost would need to offset the loss in revenue. Need information such as • Operation costs of the 5 least used runs; • Projected number of visitors over the season; • added-on operational cost on the remaining runs and chairlifts.
Increase vertical drop & Install additional chairlifts	Increase in revenue (by ticket price) would need to offset the additional cost. Need information such as • Cost of adding vertical drop and installing new chairlifts • Projected number of visitors over the season; • Potential saving of operational cost on the other unchanged runs if visitors are more drawn to the new deeper run.