GUIDED CAPSTONE PROJECT REPORT

Big Mountain Resort

Ticket Price Analysis

Problem statement
Exploratory analysis
Pre-processing
Modelling
Summary

Oleg Makarovskiy

Problem Statement

The ski resort has to boost its revenue to remain profitable this year despite increased operating expenses caused by installation of an additional chair lift. Analyze a potential change that will either cut costs or will support a higher ticket price. Increasing ticket price, increasing number of visitors by either attracting more people or staying open higher number of days. Additionally, we may consider increasing price for various amenities, including food and drinks.

Context

Big Mountain Resort, a ski resort located in Montana. Access to 105 trails. Avg visitors per year - 350,000. Lifting equipment: 11 lifts, 2 T-bars, and 1 magic carpet for novice skiers. The longest run - Hellfire - 3.3 miles in length. The base elevation is 4,464 ft, and the summit is 6,817 ft with a vertical drop of 2,353 ft. Additional chair lift has been recently installed. This additional chair increases their operating costs by $1,540,000 this season. Pricing strategy has been to charge a premium above the average
price of resorts in its market segment. Cut costs without undermining the ticket price or even propose higher ticket price.

Criteria for success

The ski resort has to boost its revenue to remain profitable this year

Scope of solution space

Number of changes that will either cut costs without undermining the ticket price or will support an even higher ticket price. Increasing ticket price, increasing number of visitors by either attracting more people or staying open higher number of days

Constraints within solution space

Increased ticket price can negatively stand out from the market average. Adding lifting equipment will increase maintenance and operational costs.

Stakeholders to provide key insights

Jimmy Blackburn, Director of Operations;
Alesha Eisen, the Database Manager

Key data source

CSV file provided by Alesha Eisen. File contains current ticket price, number of days open, and other technical details. Perhaps, specific user level access should be granted to an SQL database or an S3 bucket.

1. Data Wrangling

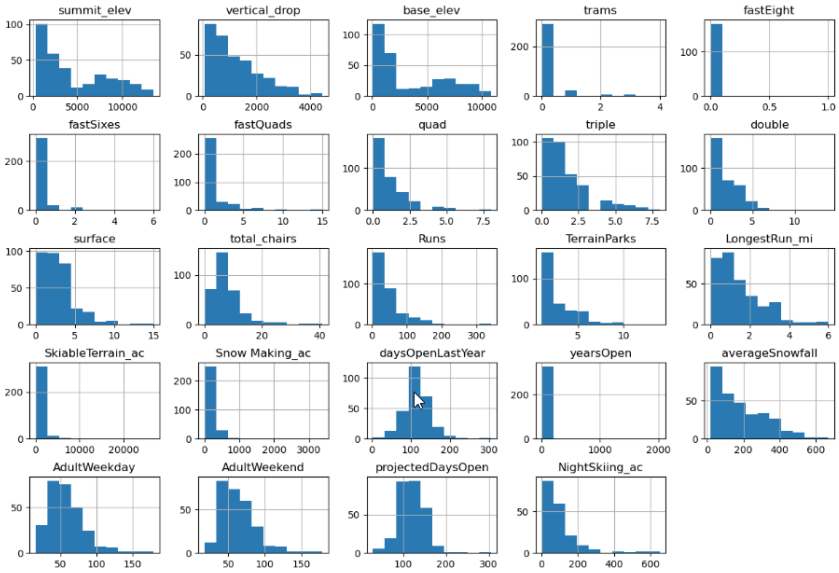
Original dataset included 330 rows and 27 columns. The resort of interest “Big Mountain Resort” is present in the dataset with the list of features to explore. Total of 330 ski resorts with total of 27 features described in columns. Columns with the following features have missing data:

	count	%
fastEight	166	50.303030
NightSkiing_ac	143	43.333333
AdultWeekday	54	16.363636
AdultWeekend	51	15.454545

Number of missing value can be higher as isnull() method used in analysis would miss numerically encoded ‘missingness’ such as ‘-1’, or ‘999’. More thorough analysis needed to increase accuracy. No duplicates found among resorts.

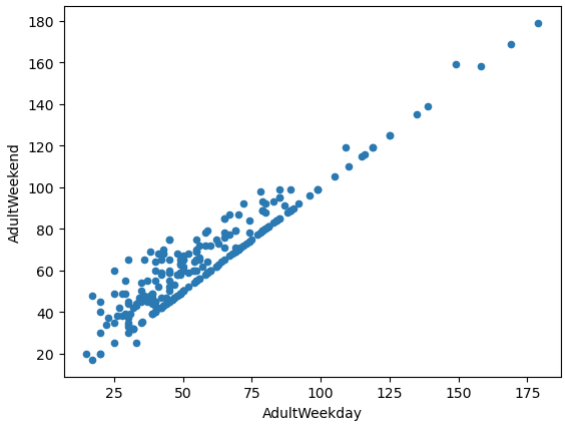
There are two resorts with the same name ‘Crystal Mountain’, but they are located in different states.

Distribution of Feature Values after preliminary data cleaning.



The following features have possible cause for concern about:

- *SkiableTerrain_ac* because values are clustered down the low end,
- *Snow Making_ac* for the same reason,
- *fastEight* because all but one value is 0 so it has very little variance, and half the values are missing,
- *fastSixes* raises an amber flag; it has more variability, but still mostly 0,
- *trams* also may get an amber flag for the same reason,
- *yearsOpen* because most values are low but it has a maximum of 2019, which strongly suggests someone recorded calendar year rather than number of years. Skiable terrain for Silverton Mountain has been corrected from 26819 to actual 1819.



There were 14% of resorts missing ticket price info in both *AdultWeekend* and *AdultWeekday*. All of them were used to review distribution of other features and then dropped. Graph shows that Weekend and Weekday price are mostly equal. Since Weekend price has fewer missing data, it should be selected as a target feature.

Column ‘AdultWeekday’ and ‘fastEight’ have been dropped. Resulting shape of the dataset after the data wrangling is (277, 25)

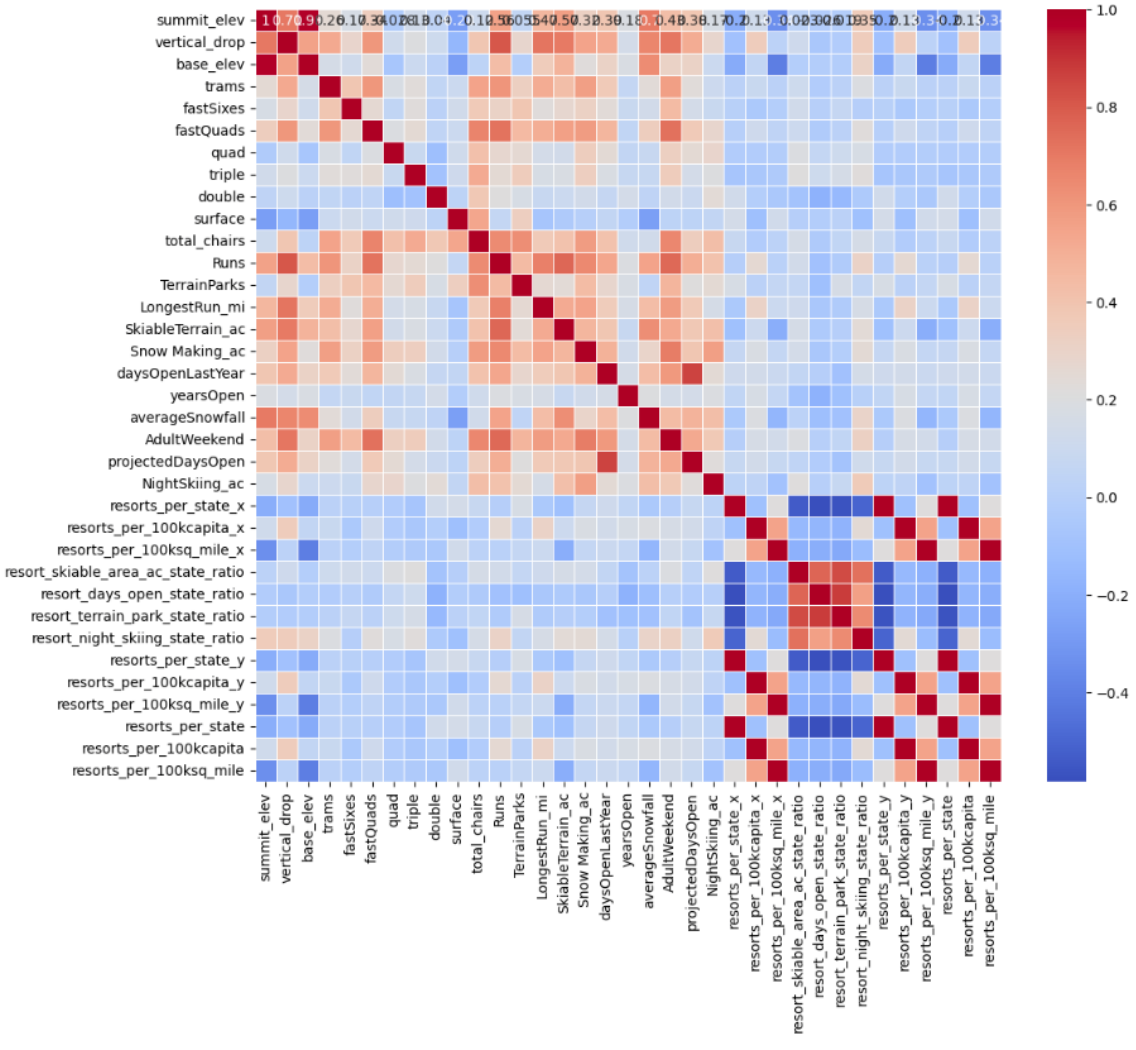
2. Exploratory Data Analysis

There are 3 categorical features such as Name (resort name), Region and State. 22 numerical data, including number of days open, skiable area, weekend ticket, runs, etc

No obvious pattern was found in a relationship between state and ticket price.

Having merged state summary features into the ski resort data, and adding "state resort competition" features, lead us to decide regarding features to use in subsequent modeling:

- ratio of resort skiable area to total state skiable area
- ratio of resort days open to total state days open
- ratio of resort terrain park count to total state terrain park count
- ratio of resort night skiing area to total state night skiing area



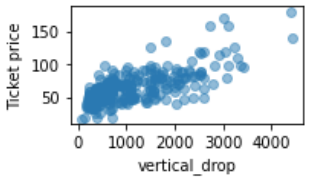
There is a lot to take away from this. First, summit and base elevation are quite highly correlated. This isn't a surprise. You can also see that you've introduced a lot of multicollinearity with your new ratio features; they are negatively correlated with the number of resorts in each state. This latter observation makes sense! If you increase the number of resorts in a state, the share of all the other state features will drop for each. An interesting observation in this region of the heatmap is that there is some positive correlation between the ratio of night skiing area with the number of resorts per capita. In other words, it seems that when resorts are more densely located with population, more night skiing is provided.

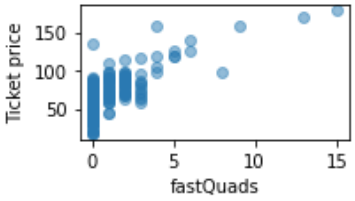
Turning our attention to our target feature, AdultWeekend ticket price, we see quite a few reasonable correlations. fastQuads stands out, along with Runs and Snow Making_ac. The last one is interesting. Visitors would seem to value more guaranteed snow, which would cost in terms of snow making equipment, which would drive prices and costs up. Of the new features, resort_night_skiing_state_ratio seems the most correlated with ticket price. If this is true, then perhaps seizing a greater share of night skiing capacity is positive for the price a resort can charge.

As well as Runs, total_chairs is quite well correlated with ticket price. This is plausible; the more runs we have, the more chairs we'd need to ferry people to them! Interestingly, they may count for more than the total skiable terrain area. For sure, the total skiable terrain area is not as useful as the area with snow making. People seem to put more value in guaranteed snow cover rather than more variable terrain area.

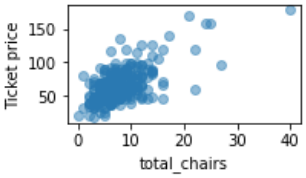
The vertical drop seems to be a selling point that raises ticket prices as well.

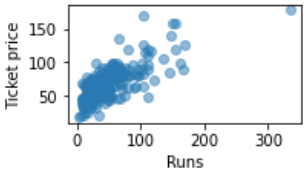
In the scatterplots we see what some of the high correlations were clearly picking up on. There's a strong positive correlation with vertical_drop.



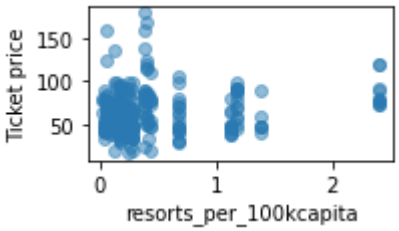


fastQuads seems very useful. Runs and total_chairs appear quite similar and also useful.





resorts_per_100kcapita shows something interesting that we don't see from just a headline correlation figure. When the value is low, there is quite a variability in ticket price, although it's capable of going quite high. Ticket price may drop a little before then climbing upwards as the number of resorts per capita increases. Ticket price could climb with the number of resorts serving a population because it indicates a popular area for skiing with plenty of demand. The lower ticket price when fewer resorts serve a population may similarly be because it's a less popular state for skiing. The high price for some resorts when resorts are rare (relative to the population size) may indicate areas where a small number of resorts can benefit from a monopoly effect. It's not a clear picture, although we have some interesting signs.



We should remain wary of the following features when we come to perform feature selection for modeling:

- total_chairs_runs_ratio
- total_chairs_skiable_ratio
- fastQuads_runs_ratio
- fastQuads_skiable_ratio

Two key points that must be addressed are the choice of target feature for our modelling and how, if at all, we're going to handle the states labels in the data. Number of fast quads may limit the ticket price. If we don't have so many chairs, we can charge more for our tickets, although with fewer chairs we're inevitably going to be able to serve fewer visitors. Our price per visitor is high but our number of visitors may be low. Something very useful that's missing from the data is the number of visitors per year**.**

3. Pre-Processing and Training Data

After verifying data types (only numeric) and splitting the data into train and test (70/30 correspondingly), the first method we tried to define a predictor was using the mean, assuming the best guess is the average price.

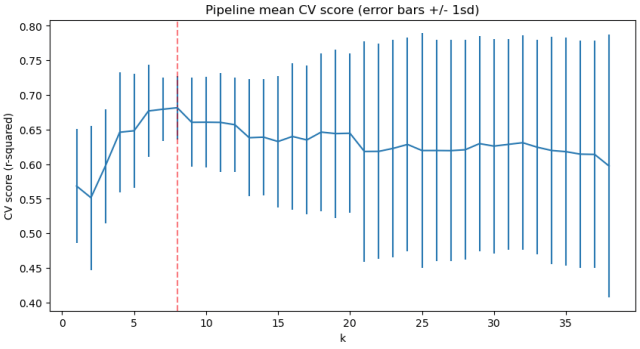
Mean absolute error (MAE) would be around $19 if we guessed ticket price based on an average of known values.

Median MAE with Linear Regression(LR) would be around $9, and MAE for LR was ~$11, which is much better than $19 when we were using the simple average as a predictor.

MAE when using Random Forest model was around $9, which is more accurate than LR, but insignificantly.

Refining the liner model did not give any significant improvements.

Best k for cross-validation was defined as 8.



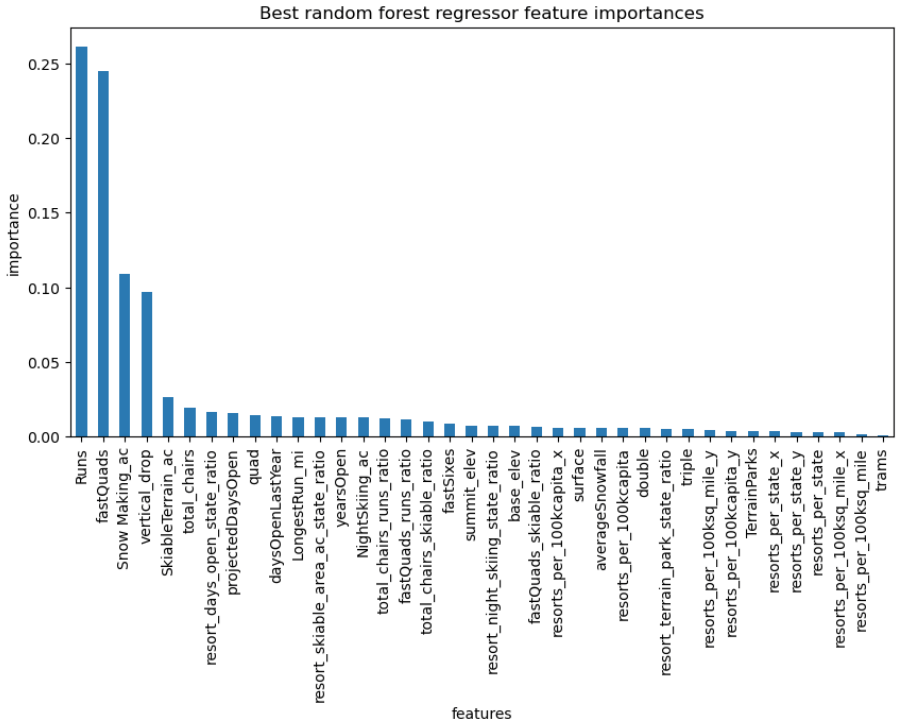
Based on the matching feature names from the column names of the dataframe, we got the list of linear model coefficients sorted on their values:

- vertical_drop 10.767857
- Snow Making_ac 6.290074
- total_chairs 5.794156
- fastQuads 5.745626
- Runs 5.370555
- LongestRun_mi 0.181814
- trams -4.142024
- SkiableTerrain_ac -5.249780

These results suggest that vertical drop is the biggest positive feature. This makes intuitive sense and is consistent with what we saw during the EDA work. Also, we see the area covered by snow making equipment is a strong positive as well. People like guaranteed skiing. The skiable terrain area is negatively associated with ticket price! This seems odd. People will pay less for larger resorts? There could be all manner of reasons for this. It could be an effect whereby larger resorts can host more visitors at any one time and so can charge less per ticket. The data are missing information about visitor numbers. We need to keep in mind that the coefficient for skiable terrain is negative for this model. For example, if we kept the total number of chairs and fastQuads constant, but increased the skiable terrain extent, we might imagine the resort is worse off because the chairlift capacity is stretched thinner.

Performance improvements caused by cross-validation are marginal.

Modeling using Random forest regressor showed the following feature importance that is common with our linear model.



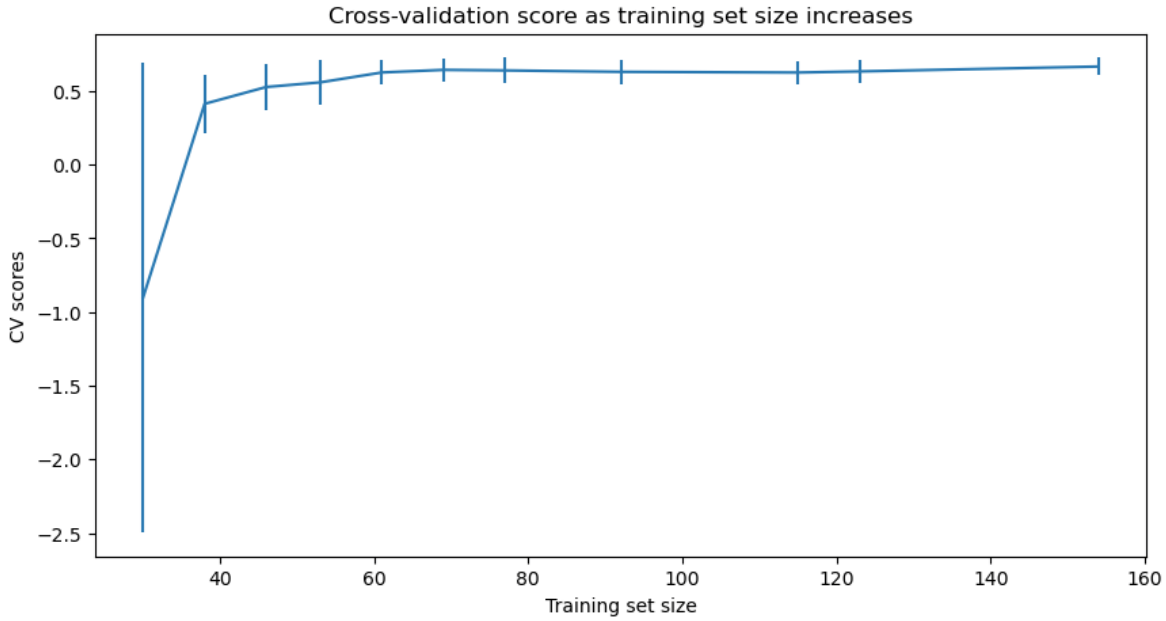
Decision-Making and model usage guidelines

- If Interpretability is Crucial: Prefer Linear Regression.
- If the Data is Non-linear: Prefer Random Forest.
- If Computational Efficiency is Needed: Prefer Linear Regression.
- If Handling Many Features or Missing Values: Prefer Random Forest.
- If the Goal is Quick Prototyping and Simplicity: Prefer Linear Regression.

When the performance difference between Random Forest and Linear Regression is marginal, the choice depends on the specific needs and constraints of your project. Weighing the factors above will help you make an informed decision.

Since in this particular problem Interpretability, quick prototyping and simplicity are of an essence, Linear Regression should be a preferred model to use.

CV score vs Training set size shows we seem to have plenty of data. There's an initial rapid improvement in model scores as one would expect, but it's essentially levelled off by around a sample size of 40-50.



4. Modeling

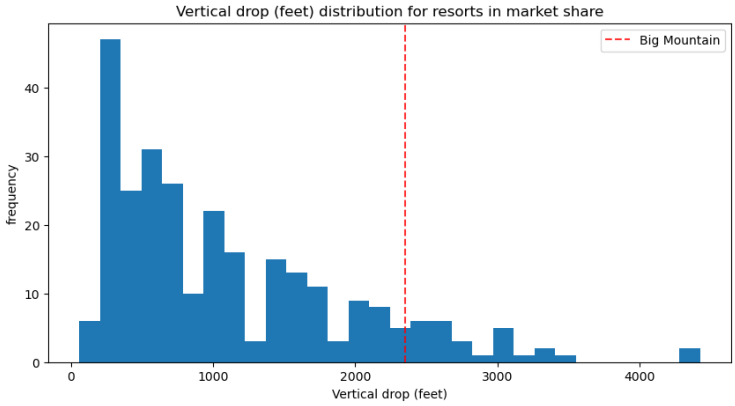
This summary should provide a quick overview for someone wanting to know quickly why the given model was chosen for the next part of the business problem to help guide important business decisions.

Currently Big Mountain charges $81 per ticket.

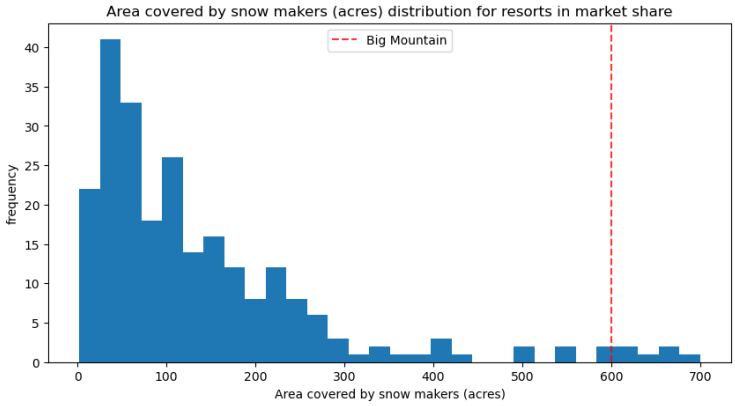
Based on important features analysis, modeling supports price increase of $1.20 per ticket. Assuming skiers on average spend 5 nights at the resort, proposed increase would lead to ~$2,1mln of additional revenue per season.

Based on comparative analysis, important features of Big Mountain resort are strongly positioned in the upper range of distributions

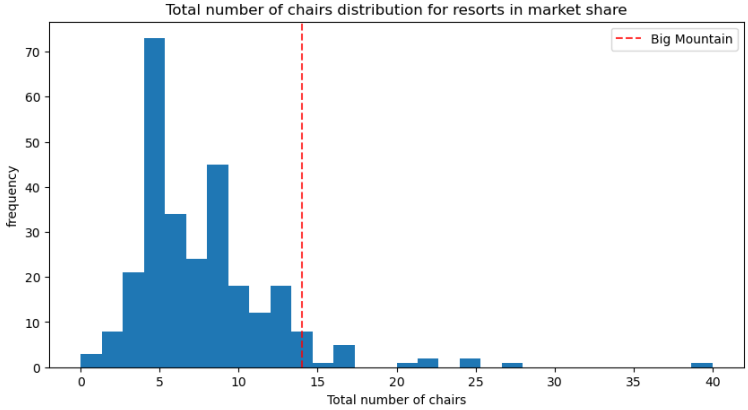
- Vertical drop



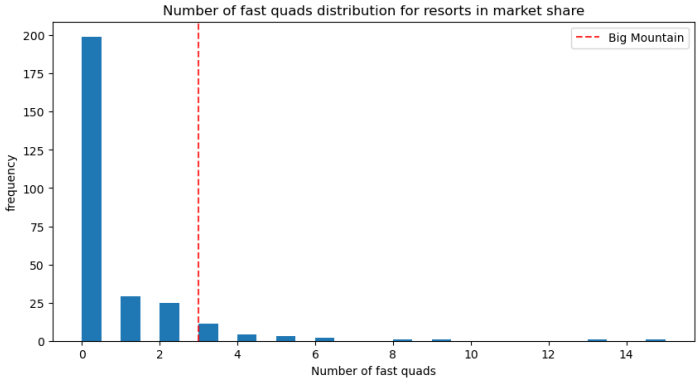
- Area covered by snow makers



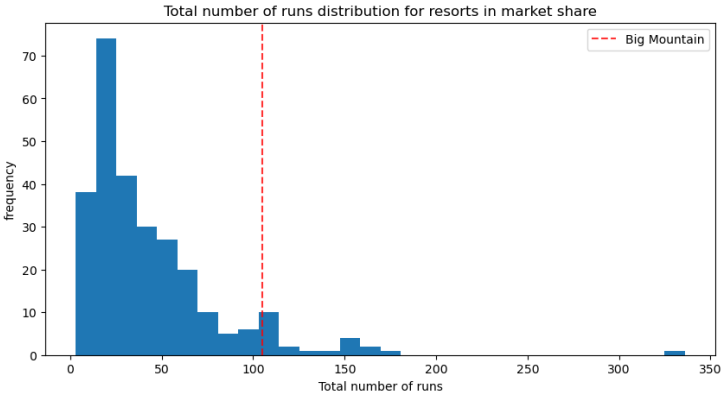
- Total number of chairs



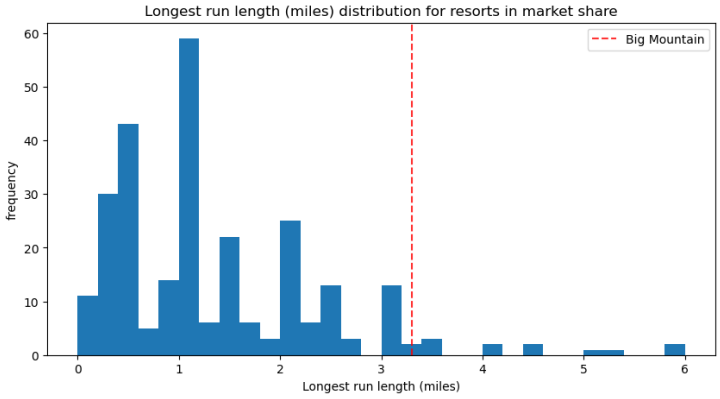
- Number of fast quads



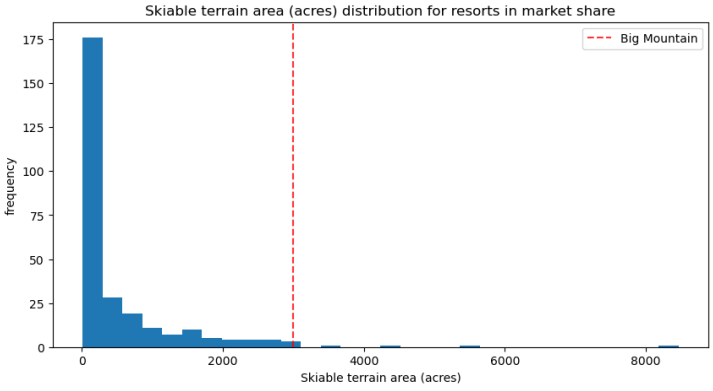
- Number of Runs



- Longest run length



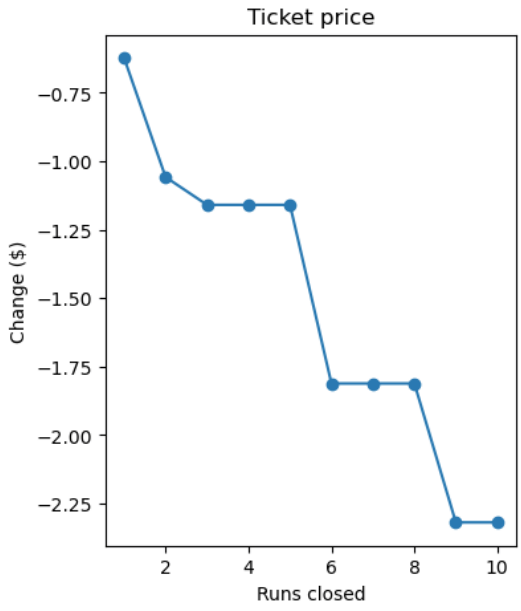
- Skiable terrain

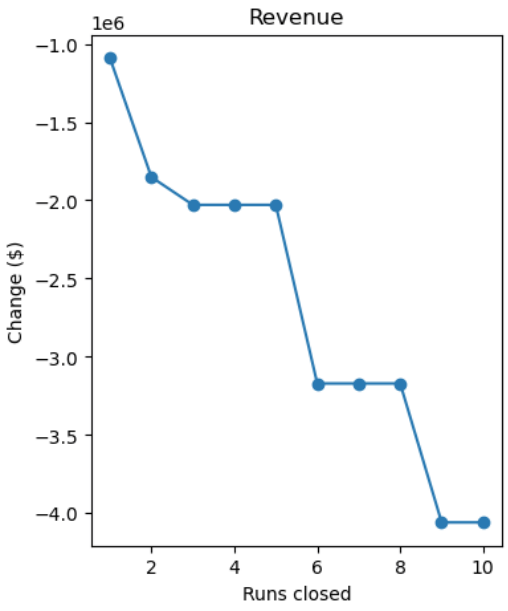


Even though Big Mountain is already the most expensive resort in Montana, its important features are at the same level or higher of resorts with significantly higher prices from other States.

Installation of an additional chair lift increases operating costs by $1.5mln this season, but at the same time adds value and support to the ticket price. Additional chair lift even more strengthens Big Mountain position among top tier resorts across the country. Key point is, the proposed $1.2 price increase will generate $2.1mln of additional revenue per season that will be sufficient to cover increased expenses related to a chair lift installation and operation ($1.5mln).

Additionally, comparative analysis shows that up to 4-5 runs can be closed, which will reduce support of a ticket price but at the same time significantly decrease operating costs with minimum effect on revenue.





Other options to be considered in the future are:

- Increase the vertical drop by adding a run to a point 150 feet lower down but requiring the installation of an additional chair lift to bring skiers back up, without additional snow making coverage
- Same as previous, but adding 2 acres of snow making cover
- Increase the longest run by 0.2 mile to boast 3.5 miles length, requiring an additional snow making coverage of 4 acres

One of the biggest limitations is missing information about operating costs of ski resorts across the country. Plus we don't know if there are resorts that are 'overprices' or 'underpriced', that also affects our model.

Missing information on weekday prices can also significantly change the outcome of our analysis.

With all the features and potential for a price increase, Big Mountain is still the most expensive facility in Montana. That can be a determining factor for executives to be conservative on ticket price.

In order to find out if business executives are surprised with a mismatch shown by a model, start a presentation with an open question on what executives believe about current price and how it compares to the facilities across the country.

It is important to have well defined and maintained pipeline for a model is in place. That would allow business analysts to add any missing data down the road and implement further ticket price or operating cost related decisions in the future.

Detailed model description and instructions should be handed over to local business analysts. Training sessions to be conducted with local team to make sure they are comfortable with conducting additional analysis as soon as more data is gathered. Me, as a data scientist developing and implementing original model, can be available for consultations moving forward.