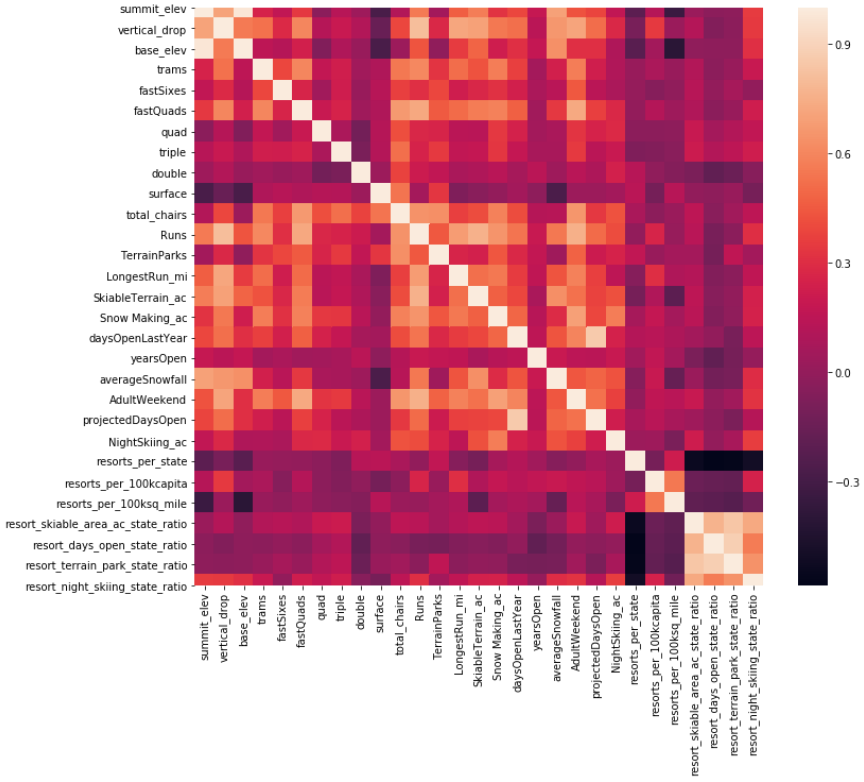
Introduction

Big Mountain Ski Resort has recently installed new chair lifts that increased operating costs by $1.54 million. In order to offset this cost, a team has been assigned to comb through ski resort comparison data to find a better ticket pricing model or reduce operating costs in other areas.

Data Wrangling

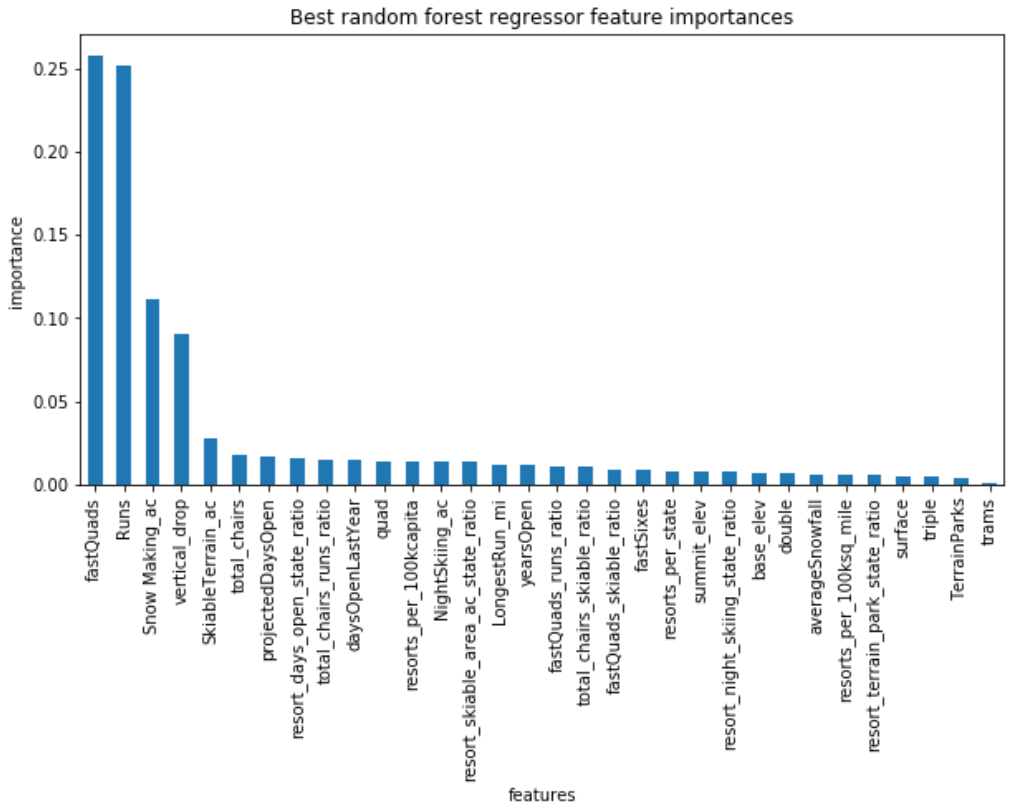
The original data contained 330 rows and 27 columns. Missing values were dropped, accounting for about 14% of those rows. Other data that was dropped included 'yearsOpen' for erroneous data, and the year '2019' for being indeterminate data. Another issue column was the 'fastEight' one, which contained many missing or zero values. The 'AdultWeekday' plot was too similar to 'AdultWeekend' one and was deleted. The 'SkiableTerrain_ac' data was changed for accuracy. After these edits, only 277 of the original 330 rows remain.

Exploratory Data Analysis



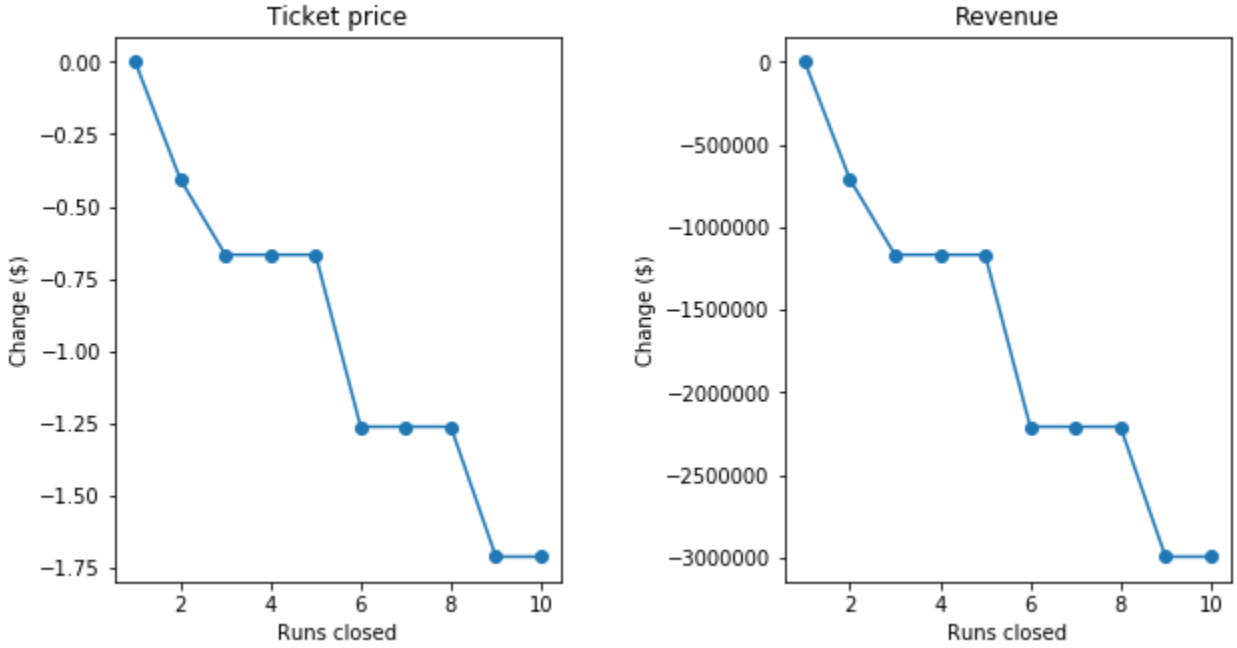
Numerical features included things relating to the measurements of the slopes and included things such as summit and base elevation, vertical drop, types of trails, and other information such as average snowfall. Categorical features included specified things such as the name, region, and state for the resorts. Some correlations include ticket prices with features such as quads, runs, chairs, and total drops. Other features to remain wary of are chairs to runs and quads to resort area. The states label will be included as a generic overarching feature, as there isn't any standout from initial data analysis.

Preprocessing and Training



The top four categories were fastQuads, runs, snow making_ac, and vertical_drop. After loading and filtering the useable data, I calculated the mean to see if it was a good predictor. The test model R2 was -0.003, comparing with the R2 of 0 in the base model, which means that the mean was not good for the data set. To create a better model, missing data values were imputed with median and mean. Data was scaled with StandardScaler. The results between the the median and mean model were not different. SelectKBest and f_regression was used to refine the model. The results showed that Skiable Terrain had the least impact on ticket prices, and vertical drop had the most impact on prices. The more efficient model used was Random Forest. This resulted in finding vertical drop to also be important, but the other three categories as well. Finally, the learning curve highlighted that the amount of data we had was sufficient.

Modeling



One area that could explain the model price difference would be the reported ticket price. These numbers could potentially not be accurate, given how the other resorts reported their prices. Also, the cost of gear and inflation may change over the seasons. This in conjunction with cost of maintenance of this gear could explain the difference. This data could be made useful by picking one of the strategies detailed above. Since the models are picked from a combination of the best variables and features, it should be enough for now without exploring other combinations of parameters.