Guided Capstone Project Report_Meg Miller

Every year about 350,000 people ski or snowboard at Big Mountain Resort. The resort installed a chair lift to help distribute the visitors more evenly across the mountain. The new lift increased their operating costs by $1,540,000 for one season. To compensate for this price increase the resort has increased their costs above the market average. There is also a possibility for the resort to make more capital on their facilities as well. With solely the price increase the resort is losing out on opportunities in making wise investments in their facilities to increase capital elsewhere. The resort is needing to reduce operating costs by 30% within the next year.

Originally there were 330 rows in the data and Big Mountain Resort was present in the original data set. The fastEight column was dropped, half the values were missing and all but the others were value zero. A new column was added for price with a new categorical column that represents the ticket type. This made it easier to create boxplots of the ticket price distributions for each ticket type for each state and compare and see where Montana sat amongst others in the market. I am currently missing the ticket prices for some 16% of resorts. Over 82% of resorts have no missing ticket price, 3% are missing one value, and 14% are missing both, I dropped the 14% of rows with no ticket pricing information. For Silverton Mountain, I noticed that there was an error where the skiable terrain contained the value 26819, but the correct value needed to be 1819, therefore this change was made to Silverton Mountain.

I also created a new data frame by aggregating columns from the ski_data dataframe. This new dataframe called state_summary gives more information about each state. These columns include, resorts per state, total skiable area, total days open, total parks terrain, and total night skiing for each state. I also added more information to the state_summary data frame such as population and area information. I then separated and eliminated redundancies in the state data, then I was able to join the state data with the ski_data.

The categorical features in this data set are the State name, Resort name, and Region. There are several numerical features in the date set. Numerical features in the data include summit elevation, vertical drop, base elevation, trams', fast sixes, fast quads, quads, triples, doubles, surface, total chairs, runs, terrain parks, longest run, skiable terrain, snow making', days open last year, years open, average snowfall, days open', night skiing, resorts per state, resorts per 100,000 capita, resorts per 100,000 sq miles, skiable area to state ratio, resort days open to state ratio', resort terrain park to state ratio, and resort night skiing to state ratio. These features describe the various attributes of each resort and state. There was no clear pattern suggestive of a relationship between state and ticket price. Due to this, state labels were treated equally towards building a pricing model that considered all states together, without treating any one particularly special. The target feature is “AdultWeekend” for modeling of numerical data, these include Adult Weekend ticket prices, total chairs, runs, skiable terrain etc. I examined two states that have high average ticket prices (Vermont and New Hampshire) to see if they had any high numerical values in categories that other resorts did not. Both states have large numbers of resorts per 100,000 sq miles and Vermont also has a large number of resorts per 100k capita. Therefore I looked at all features compared to ticket prices across all states, as it appears that the state does not determine the ticket price, rather the other features of the resorts. Real insights were gained after calling a seaborn correlation

heatmap (Figure 1) on the original dataset that identified which features were more positively and negatively associated with ticket price. From here it was immediately clear that ticket price is heavily impacted by these primary features: FastQuads, Runs, Vertical Drop, Total Chairs; and then moderately impacted by the secondary features: Longest Run, Skiable Acres, Snowmaking Acres, and Night Skiing. It seems that the four primary features are in fact best for modeling ticket price.The features that remain of most interest for subsequent modeling are vertical drop, fast quads, runs, totals chairs.

For the preprocessing and model training, I obtained an idea of performance determining how well the average price did. I built a linear model to determine if I had enough data to move forward with the project or if more data was needed. The model determined that I have a sufficient amount of data. I used a random forest regressor. Some preprocessing steps that I found to be useful were calculating the R-squared, the mean absolute error, and the mean squared error. I also had to impute some missing values with the median. I decided to use the random forest model going forward. It had a lower cross-validation mean absolute error by nearly $1. It also exhibits less variability. Its performance on the test set is consistent with the cross-validation results.

Big Mountain Ski resort currently charges $81 for a weekend pass. The model suggests a ticket price of $95.87. I would approach suggesting this change in price by pointing out our strengths as a resort. Some features that correspond with other high ticket priced resorts are vertical drop, snow making, total chairs, fast quads, runs, longest runs, trams, and skiable terrain. Big Mountain has a high vertical drop, some of the highest snow making area, one of the highest numbers of total chairs, it has 3 fast quads (while majority of resorts have none), we have one of the highest number of runs, and we have one of the longest runs. Lastly, Big Mountain is amongst the resorts with the largest amount of skiable terrain. The only feature we do not have is trams. However the vast majority of resorts have no trams. Big Mountain's addition of the new chair lift and operating costs could be supported by increasing the ticket price by $11.16. Over the season this could be expected to amount to $19,529,239, since on average visitors typically ski for 5 days. This would cover the operating costs of the chair lift. The model showed that closing one run makes no difference. Closing 2 and 3 successively reduces support for ticket price. If Big Mountain closes down 3 runs, it seems they may as well close down 4 or 5 as there's no further loss in ticket price. Increasing the closures down to 6 or more leads to a large drop. The modeling scenario I would recommend is potentially closing down up to 4 or 5 of the least used runs. This doesn't impact any other resort statistics. If Big Mountain were to close 4 or 5 runs we would still be in the higher league for total number of runs when compared to other resorts.

In regards to the further work that can be done on this project, we could identify the historical information about revenue and facilities (particularly tracking facility changes) would be useful. The modeled data is based on pretty non-specific input data. There is no direct tie from operational changes to revenue. The modeled price being higher than the current price can come from a lot of sources. When was the last price increase? Did Big Mountain improve their facilities without bumping the price? Did Big Mountain not improve their facilities recently or run an analysis? Was there market pressure in the last couple of years to not bump the price? Interviewing the executives or one of the other stakeholders might yield some information around this. Our best model turns out to be a Random Forest Regressor and could easily be deployed as a web app on the intranet or a traditional application.

Figure 1. Heatmap of correlations