Guided Capstone Project: Predicting Big Mountain Resort Ticket Price.

1. Introduction

Big Mountain recently installed an additional chair lift which increases their operating costs by $1,540,000 this season. The question now is: what are the changes the company need to implement to optimize their ticket price and increase their capitalization by next season ?
The company is looking for guidance to assess how important some facilities are compared to others, to select a better value for their ticket price. The Big Mountain resort suspects it may not be maximizing its returns, relative to its position in the market. It also does not have a strong sense of what facilities matter most to visitors, particularly which ones they're most likely to pay more for. The purpose of this project is to build a predictive model for ticket prices based on a few facilities, or properties, boasted by resorts (at the resorts). This model will be used to provide guidance for Big Mountain's pricing and future facility investment plans.

2. Data Wrangling

We have been provided with a ski data DataFrame having: 330 rows and 27 columns and the Big Mountain was present in the data. No columns have been removed at this point, but the focus has been placed on those columns that are relevant to resort ticket price: TerrainPar, SkiableTerrain, daysOpenLast and NightSkiing_ac.

After looking at the distributions of features value, some features commanded a closer look for various reasons. We have dropped 14.33% of rows with no Price Data which is 49 rows, so we are left with 281 rows. Weekend prices have the least missing values of the two possible target features, so we dropped the weekday prices and then kept just the rows that have weekend prices. According to the data available, the Weekend price (**Adult Weekend** column in our data) is the most suitable feature to predict ticket price.

3. Exploratory Data Analysis (EDA)

After performing dimension reduction, we saw the same distribution of states as before, but with additional information about the average price. We couldn’t find an obvious pattern in our representation of the ski summaries for each state, which accounts for some 77% of the variance of ticket price.

We can offer some justification for treating all states equally, and work towards building a pricing model that considers all states together, without treating anyone particularly specially. We haven't seen any clear grouping at this point.

Turning our attention to our target feature, **Adult Weekend** ticket price, we saw quite a few reasonable correlations. Finally, some further features may be useful in that they relate to how easily a resort can transport people around. These are the numbers of various chairs, and the number of runs, but we don't have the ratio of chairs to runs. It seems logical that this ratio would inform us how easily, and so quickly, people could get to their next ski slope!

We have seen an exclusive versus mass market resort effect; if a resort doesn’t have so many chairs, it can charge more for its tickets, although with fewer chairs the resort will inevitably be able to serve fewer visitors. The price per visitor is high but the number of visitors may be low. Something very useful that's missing from the data is the number of visitors per year. It also appears that having no fast quads may limit the ticket price, but if your resort covers a wide area then getting a small number of fast quads may be beneficial to ticket price.

4. Preprocessing and training

The EDA gave us a clean data set and we started to build a machine learning model to predict the adult weekend ticket price of our Big Mountain resort.
We started by considering how useful the mean value is as a predictor. This first model using the mean as a predictor is a baseline performance comparator for any subsequent model. We calculated the mean (average) of the dependent variable (**Adult Weekend**) and obtained a value of 63.81.

The Mean Absolute Error (MAE) of this Dummy Regressor was 19.14 suggesting that on average, we might expect to be off by around $19 if we guessed ticket price based on an average of known values.
After imputed missing values and scaling the data we trained a Linear Regression (LR) model, then we tested a Random Forest model (RF) using the same exact steps as those used for the LR model. The MAE of the LR model was 11.79 while the MAE of the RF model was 9.53.

We then conclude that the Random Forest (RF) model has a lower cross-validation Mean Absolute Error (MAE) by almost $2. It also exhibits less variability. Verifying performance on the test set produces performance consistent with the cross-validation results.

Therefore, the RF model is the one we have decided to use going forwards because between two predictive models the best one is often the one producing a lower error.

5. Modeling

In this use case we already have plenty of data. We're often led to believe more data is always good, but gathering data invariably has a cost associated with it. We Assessed this trade off by seeing how performance varies with differing data set sizes. 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.

Besides the additional operating cost of the new chair lift, operating cost on vertical drop, snow making area, fast quads, runs, trams and skiable terrain area may also be useful. The data is missing information about visitor numbers.

The reason why the modeled price for Big Mountain is so much higher than its current price is probably due to the validity of our model which lies in the assumption that other resorts accurately set their prices according to what the market (the ticket-buying public) supports. It's reasonable to expect that some resorts will be "overpriced" and some "underpriced." or if resorts are pretty good at pricing strategies, it could be that our model is simply lacking some key data. Certainly, knowing more about operating costs, of key important features (fast quads, runs, snow making area and vertical drop) in the model would help.

This model could be made available for business analysts to use and explore via a Web Application. The goal of the application will be to execute the selected model in different scenarios or to test a new combination of parameters in a specific scenario and facilitate the decision process.