Guided Capstone Project Report: Big Mountain Resort

Sonjoy Das, Ph.D.

May 22, 2021

Introduction: Big Mountain Resort, a ski resort located in Whitefish Montana, with spectacular natural scenery and access to 105 trails, serves about 350,000 people annually. The resort houses many exciting facilities for the skiers. It is the most expensive resort in the state of Montana. They currently charge $81.0 per adult. Eighty percent of the resorts in the USA charge less than this price. But, the resort management is suspicious that their tickets are underpriced because they are possibly not capitalizing well on all the exciting facilities that they offer to the visitors. The objective of this data science project is, therefore, to explore:

1) if the features and facilities, that control best the ticket price, can be identified through data analysis, and
2) how a business strategy can be rolled out, e.g., how an investment scheme can be planned to improve the identified facilities, to maximally increase the business profit by hiking the ticket price?

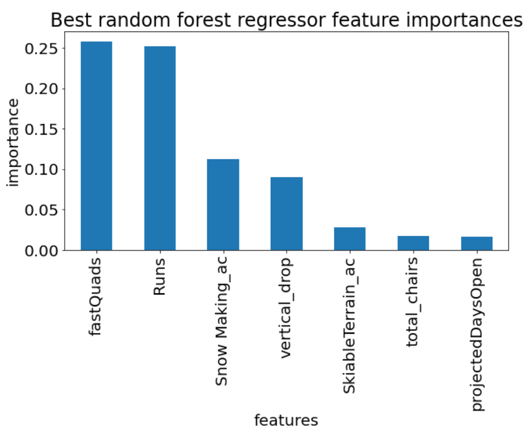
Data Analysis: The primary data source here is a single CSV file containing ticket prices and potential features of all resorts in the country. In "data wrangling and cleaning" phase, it is found that the ticket prices for weekdays and weekends are pretty much same. In particular, the resorts of Montana have exactly the same ticket prices for weekdays and weekends. It is also uncovered that the weekend prices have the least missing values of the two prices in the source CSV file. Therefore, the *weekend ticket price* is considered as the *target feature*. To discover any possible relationships among the state-wide market size, total population and total area (in square miles) of each state are extracted from Wikipedia and used as additional data. Both the datasets are cleaned and tidied.

In the "exploratory data analysis" (EDA), the Wikipedia dataset is used to derive some measures of resort density relative to the state population and size, which are more useful *features* than the state-specific data extracted from Wikipedia. All the original and derived features are finally merged together into one `DataFrame` for final analysis. One of the noteworthy conclusions from the EDA made through `Seaborn's` heatmap of *correlations plot* is that a few feature variables, in particular, `fastQuads`, `Runs`, `Snow Making_ac`, `total_chairs`, and `vertical_drop`, are very well correlated with *ticket price* which is the target feature.

The next step, "pre-processing and training data" phase, splits the combined dataset into train and test datasets (70/30 split). Two regression models, particularly based on `sklearn's` `LinearRegression` and `RandomForestRegressor` classes, are then considered to determine the best possible model. The optimum hyperparameters of each model class are computed only using the training data by applying a 5-fold cross-validated grid-search technique, `sklearn's` `GridSearchCV`. The top 4 important features (not necessarily in the same order) per both the optimum (best) linear regression model and random forest regression model are found to be `fastQuads`, `Runs`, `Snow Making_ac`, and `vertical_drop`. This is consistent with the result obtained in EDA.

	Best linear regression model	Best random forest regression model
R^2-error	0.5968	0.7515
MAE	$11.7935	$9.5377
RMSE	$15.2876	$12.0019

The performances of these two best models against the test split are then estimated in terms of three error-metrics, namely, R^2-error, mean absolute error (MAE), and root mean squared error (RMSE). The results are shown in the table above. The best random forest regression model wins against the best linear regression model. Therefore, the best random forest regression model is selected for rest of the analysis. A barplot of feature importances associated with this best model is shown to the right.



Conclusions and Recommendations: The finally selected random forest regression model predicts that Big Mountain's ticket price can be increased to $95.87 which will yield a *predicted revenue* of about $168 million. Based on the ticket price of $81, the current revenue is about $142 million. The revenue estimation here presumes that the expected number of visitors over the season is 350,000 and, on average, visitors ski for five days, i.e., each visitor buys five tickets.

Another information that was already available is that the Big Mountain recently installed an additional chair lift. This increases their operating costs by $1.54 million this season. To break even this operating cost, a small increase of $0.88 per ticket is necessary this season. However, implementing the new plan of increasing price straightway to about $95.87 without doing any in-depth customer research can backfire. No data is available to explore and support such customer research. Therefore, it may be a good idea to increase the price little conservatively, say, to about $88 this season to gauge skiers' initial reactions. If this pricing negatively affects the expected number of visitors, then the management needs to re-evaluate the pricing strategy with additional set of data that should possibly include new key features missing in the current data.

If, on the other hand, increase in revenue is reasonably made as anticipated, then it possibly implies that the random forest model can capture the key features that govern the current resort market. Accordingly, different modeling scenarios can be analyzed to determine an appropriate investment plan. Four different modeling scenarios are particularly considered. The one scenario, that stands out among these 4 possible modeling scenarios in the final “modelling” phase, suggests to 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. With this change, the predicted ticket price will jump to $104.5 increasing the *predicted revenue* to about $183 million.

Guided Capstone Project Report: Big Mountain Resort

Sonjoy Das, Ph.D.

May 22, 2021

Limitations of the Data Analysis: The predictive analysis above strongly relies on the assumption that the other resorts accurately set their prices according to the market. This may not be a reasonable assumption. There are many factors, that play a significant role in ticket pricing, are simply not available. For instance, we know nothing about the operating costs except the additional operating cost of a new chair lift. Are there any competitive resorts in proximity of the Big Mountain resort? What is economic condition of the people in the areas around the resorts? Are the visitors coming across from different states to the resorts? Is the state tax included in the ticket price for all the resorts? How can we incorporate the factor of spectacular natural scenery of the Big Mountain Resort in ticket pricing? There are many more questions! Nevertheless, the random forest model seems to provide some insight in dynamics of resort market. The management can use this model as their first cut model to explore further, which is likely to lead them to uncover more discrepancies and deficiencies of the existing dataset, and finally to an improved and next *avatar* of the current model.