Springboard: Data Science Career Track
Guided Capstone Project: Project Report
By Lucien Meteumba
June, 2020

PROBLEM IDENTIFICATION OVERVIEW

Big Mountain Resort is a company that was created in 1947, located in northwestern Montana. It offers spectacular views of Glacier National Park and Flathead national Forest. Despite the increase in operational costs by $1,540,000 due to an additional chair lift installed to help increase the distribution of visitors across the mountain, the investors want to keep the business profit margin at 9.2% given that every year about 350,000 people ski or snowboard at Big Mountain.

As a Data Scientist my contribution was to find the predicted cost of an adult weekend chairlift ticket based on the data given to us by the database manager. The data is the recorded information about other resorts in different states. We did not have any information regarding when the data was created or the data timeframe. Since the target variable was continuous we chose to use a Linear Regression Model for our prediction. We generate an Excel spreadsheet containing the coefficients of our model and the model performance metrics.

DATA PREPROCESSING STEPS OF NOTE

Our first big step was to handle the missing values. We identified the percentage of missing values for all the columns, then based on the nature of the column we replace the missing values by the mean value or zero.We replaced the missing value by the mean when we know that there is supposed to have a value for that resort in the column. For example, every resort should have a value for ‘LongestRun_mi’ feature (the Length of the longest run in the resort in miles). It will be therefore reasonable to replace the missing value with the mean, which is only one method among many. Replacing the missing value with zero comes therefore in place when we know that not all resorts must have a value for that feature. For example, 'NightSkiing_ac' represents the number of acres of night skiing available at this resort. Not all resorts have night skiing because it requires stadium lights and is not that popular with skiers and snowboarders. From there, we checked to see if there are duplicate rows to remove. In this case it will be necessary to remove duplicate values because each row represents a resort. We use the bar chart to show that the Region feature is nearly identical to the State feature. Since the two features were nearly identical, we decided to remove the Region feature. We used the correlation matrix displayed in the heatmap to select and remove collinear features. Setting the threshold for the correlation coefficient greater than 0.95, the collinear features we found was the summit_elev and the base_elev. We decided to keep summit_elev and drop base_elev. Based on the unsupervised

learning K-Means algorithm, we created another feature called cluster to find patterns in our data and to see the different groups of points in our data.

MODEL DESCRIPTION

The data used for modeling has 26 columns and 330 rows.Some of the features are AdultWeekday, averageSnowfall, total_chairs, summit_elev, vertical_drop and AdultWeekend. We used the StandardScaler()[1] method of the preprocessing package to scale the features of our dataframe except the name of the resort, the state, and the summit_elev feature, then call the fit()[2] method with parameter the input variable. Then we used the train_test_split() [3] method from the sklearn.model_selection train and test our data. Following that, we used the linear regression to fit our model.

MODEL PERFORMANCE

The model performance we used for this project is the explained variance score (R-Squared) and the mean absolute error (MAE). The R-Squared will help us measure the discrepancy between our model and the actual data and the MAE score will help us measure the average of the absolute values of the errors, measured as the difference between the actual values and the predicted values. We got 0.93653 for the explained variance score and 5.095 for the mean absolute error. The units of MAE are the same units of the target, in this case it is US Dollars.

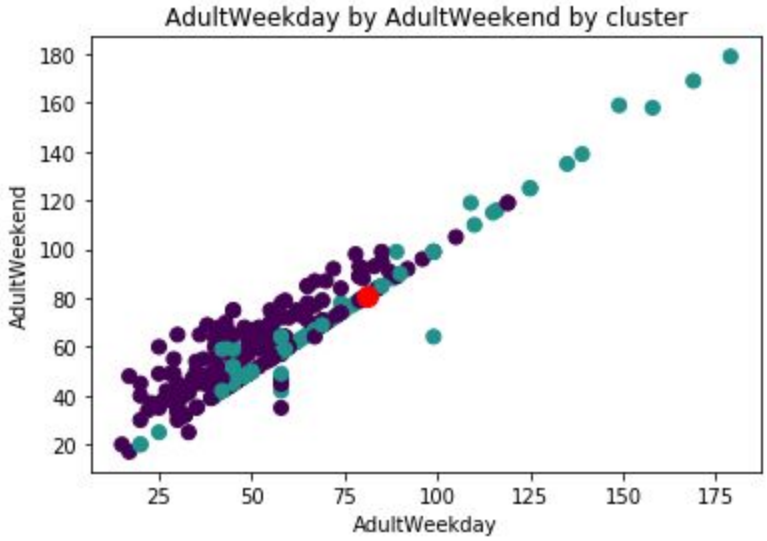
MODEL FINDINGS

Based on the model we built, the predicted cost of an adult weekend chairlift ticket will be closed to $88. This is $7 more than the actual cost.

Also based on the figure below we can say that there is almost an increasing linear relationship between AdultWeekday and AdultWeekend. So we can conjecture that with the same model that the expected cost of an adult weekday will be greater than the actual cost. We can also fit a regression model to find out the specific relationship between both ticket prices, so knowing the ticket price for one, one can compute the other.

---

[1] https://scikit-learn.org/stable/modules/preprocessing.html
[2] https://scikit-learn.org/stable/modules/preprocessing.html
[3] https://scikit-learn.org/stable/modules/preprocessing.html



NEXT STEPS

We saved the model as a pickle file ready to be used. Also we also saved the model performance metrics, the coefficients and the y_intercept in an excel spreadsheet in case we want to refer to them later.
We could use projectedDaysOpen as the response variable to predict the expected projected days open in the upcoming season.