Guided Capstone

Project Report

By Jeffrey P. Koskulics, Ph.D.

Introduction

The Guided Capstone project analyzes ski resorts in the United States. The client, Big Mountain Resort in Montana, has added a new ski lift that increases the resort's expenses. Big Mountain wishes to offset these expenses by considering changes to lift ticket pricing.

In this analysis, we create a ski lift ticket pricing model. This model is based on Big Mountain's peer resorts. The results show opportunities to increase the price for lift tickets to align more closely with Big Mountain's peers.

Data Wrangling

The peer resort data set is contained in a single comma separated value (CSV) file that lists 330 ski resorts. The data are loaded using Python/Pandas in a Jupyter notebook. A Github repository is established for online presentation of the notebooks. A file structure is established for storing notebooks, data files and figures.

Data are reviewed to determine the column names and data types. We examine each column to determine the number of unique data types, the range of data spanning numerical columns, and the unique codes contained in categorical columns.

The data are examined for null or missing values and duplicated rows. No duplicated rows are found, but there are several instances of null values. The null values are replaced by column means to enable analysis. The decision to support filing null values with the mean is consistent with a subsequent linear regression analysis using centered data using z-scores (these values will not contribute to changes in the output).

The data are saved in a new CSV file to avoid overwriting original data file. Subsequent analysis will be performed on the new CSV file.

Exploratory Data Analysis (EDA)

EDA begins with a tabular presentation of the columns using the Pandas function .describe( ). The resulting table provides the count, mean, standard deviation, minimum, maximum, and 25, 50 and 75th percentile values. Histogram plots for each numerical column are examined for outlying data.

Several outlying values are found. One resort has clearly mistaken values entered for the number of years open.

Bar charts plot column values for each state to look for statewide trends. Boxplots provide a graphical indication of outlying values.

Identification and creation of features

A Pearson correlation coefficient heat map indicates that there are several values which are highly correlated. This provides an indication that there are features that may be collinear. In particular, the summit elevation and base elevation are highly correlated with each other. The prices for weekdays and weekends are also highly correlated. These values stand out on the heat map as lightly colored points.

Next, we categorize the resorts using k-means, an unsupervised learning technique. In this step, we select an increasing number of categories to fit and examine an “elbow plot” that shows how the k-means “inertia” decreases with each added categories. This provides an indication of the diminishing returns of increased category number. The plot shows a sharp decrease in error from 1 to 3 categories with much smaller decreases thereafter. This indicates that we should consider either 2 or 3 categories.

For k-means, the guide indicates that the k-means categorization should take place using raw data, rather than data that has been scaled and centered. In doing so, variables with large absolute value tend to dominate the distance function. The results are unsurprising with k-means categories tracking closely with elevation (which is measured in feet and has values ranging into the 10,000s). When the analysis is repeated using scaled data, the k-means categories are spread more evenly across elevation.

Preprocessing

Categorical variables (states and k-means categories) are replaced with binary dummy variables. Greatly increases the number of variables, but enables their inclusion in numerical regression techniques. Feature data is centered and scaled to provide z-scores.

Features are selected for the model’s input and output. Data are split into training and test sets. Normally, splitting data is done to avoid “overfitting” where the model output tracks the input data exactly. Since we are using a linear regression, where the number of coefficients is smaller than the number of instances, this step seems questionable. Alternatively, we could train the model using all of the instances. This seems more fair in that the resulting model represents all peers.

Modelling

To model the ski lift ticket prices, we choose a simple linear regression. Linear regression creates a model where each feature is assigned a value (coefficient) that represents the feature’s contribution to the output. In this case, the values have a value in dollars.

We examine several models that include and exclude certain features. For example, we find that if the dummy variables for states are included as z-scores, the resulting coefficients weight them very heavily with values in the 10 trillion dollar range. Including the states as raw binary values brings them down to reasonable value, but we’re left with the task of explaining the value that a resort is located in a particular state. For most states, the value is just a few

dollars, but for the state of New Mexico we find that the value is a staggering $-43. This strongly suggests that we exclude the feature of state.

Conclusions

The final model is selected to exclude the state, and the base and summit elevations. The results indicate that the lift ticket price could be increased by a modest amount. We also report the value of certain features.

The findings indicate that night skiing acreage and total acreage have negative coefficients. This indicates that the market assigns little value to these features. It may be advantageous to consider reducing these features for both cost savings and a possible increase in value.