Guided Capstone Project Report

Step1: Problem Statement Worksheet

Project Context: Big Mountain Resort is a famous ski resort in northwestern Montana. Every year it has about 350,000 people. It recently has installed an additional chair lift to increase the efficiency to distribute visitors. The additional cost with the chair is $1,540,000. We need to determine if the extra efficiency the additional chair brings can make the business still profitable.

Solution Scope: there can be a lot of solutions including increase the lift ticket price, increase open days in a year and change the resort condition. But in this exercise, we only focused on Adult weekend price to see if there is room to increase it without investing extra money on resort condition improvement.

Data Source: we get data exported from SQL database containing 330 rows with 27 columns (variables) (Table 1)

Step 2: Data Cleaning

2. Handle the missing and NA values.

2.1 ‘AdultWeekday’ and ‘AdultWeekend’ price is replaced with mean value.

2.2 Other NA values in columns 'daysOpenLastYear', 'TerrainParks', 'projectedDaysOpen', 'Snow Making\_ac', 'averageSnowfall', 'LongestRun\_mi', 'Runs','SkiableTerrain\_ac' ,'yearsOpen' are filled with zero.

Step 3: Exploratory Data Analysis

3.1 Delete duplicated column. Delete column ‘Region’ as it is the same as column ‘state’

3.2 Analyse anomalies and outliers. Use 1.5 IQR as a criterion to check outliers, and find 173 instances, more than half of resorts are gone, including the Big Mountain resort. In order to have enough data points for modelling, I decided to keep all the instances without filtering the outliers.

Step 4: Pre-Processing

4.1 Try to find the relationship between feature and variables. Set ‘AdultWeekend’ as a variable of response.

4.2 Create dummy features for categorical variable ‘state’. This impact the model results a lot.

4.3 Standardize the magnitude of numeric features.

4.2 Split 330 data instances into training (75%) and testing (25%) set, training is 75%data

Step 5: Modeling

5.1 Run 3 different linear regression models with selected features. Use explained variance and mean absolute error as parameters to quality control model performance (Table 2).

5.2 Pick model 3 to be the final model because it has it has smallest variance and absolute mean errors.

Table 1 Column name and description

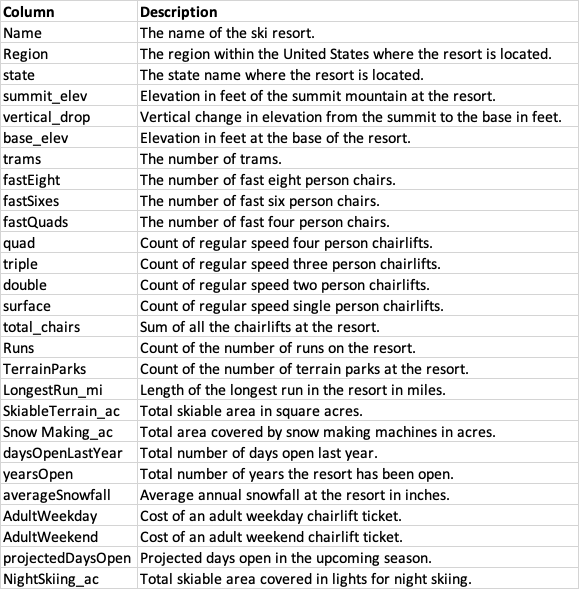


Table 2 Model Performances Comparison

| **Model** | **Explained Variance** | **Mean Absolute Error** | **Features Dropped** |
| --- | --- | --- | --- |
| Model 1. | 0.9383886568462728 | 5.024931713620744 |  |
| Model 2. | 0.9319995669341605 | 5.152252616611725 | 'state' |
| Model 3. | 0.9309648542109354 | 5.191281316415176 | 'state','summit\_elev' |

Step 6: Documentation

6.1 Fit model with training data set

6.2 Review results. Model predicted adult weekend price is about $89. The current price is $81.

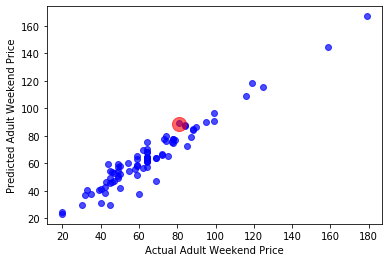


Fig.1 Comparison between predicted adult weekend price versus actual price. The red is Big Mountain data.

Summary

I recommend to increase the adult weekend price from $81 to 89$.