DataCamp-DataScienceAssociate

L Burleigh

2023-06-26

Task 1

The dataset contains 1500 rows and 8 columns [with missing values] before cleaning. I have validated all the columns against the criteria in the dataset table:

booking_id: same as description. No missing values. 1500 unique identifiers. integer. months_as_member: same as description. No missing values. 72 unique whole integers. converted to numeric for ease of use. weight: 20 missing values - replaced with overall average of column. lowest value is appropriately 55.41 and cells use 2 decimal places. numeric. days_before: 25 rows include 'days' after value, which I removed so column includes only number values. No missing values. Minimum value of 1. character converted to numeric for ease of use. days_of_week: No missing values. Inconsistent labeling of days [i.e., both "Mon" and "Monday" in column] so made all rows consistent with 3 letter day as listed in description, adjusting 71 incorrect rows. character. time: same as description. Every row includes either 'AM' or 'PM'. No missing values. character. category: Missing relevant data in 13 cells contain '-' so replaced these with 'unknown' as specified in description. No NAs. character. attended: Same as description. No missing values - each cell contains either 1 or 0. integer adjusted to factor for ease of use.

After the data validation, the dataset contains 1500 rows and 8 columns.

Task 2

From Graph 2 Classes Attended, the most attended class category was HIIT with 213, then followed by Cycling with 110. The attendance of classes is varied, however with the inclusion of Graph 2-2 Class Attendance, we can see that the number of people who attended and did not attend each class vary together as HIIT was also the highest not attended class, followed by Cycling, Strength, Yoga, Aqua, then unknown.

Task 3

The months_as_member variable is our target variable. Graph 3-1 Membership Months Distribution shows a positive skew distribution with the majority of members holding a membership for less than 50 months with only 4 over 100 months and an outlier at 148 months. With no negative or zero values, and a qqplot, Graph 3-2 QQplot Membership Length, further indicating a right skewed distribution, a logarithmic transformation was performed. Graph 3-3 Log Membership Months Distribution and Graph 3-4 QQplot Log Membership Length indicate the transformed distribution is much closer to a normal distribution.

Task 4

From Grahp 4-1 Membership and Attendance Relationship, the outlier identified in Graph 3-1 again provides a difficulty in interpreting the relationship properly. After removing the outlier of 148, Graph 4-2 shows that attended sign ups have a larger range of months as a member than sign ups that were not attended and a higher median of membership months with 20 months while the not attended sign ups have a halfway point of 10 months.

Task 5

Predicting whether members will attend a class is a classification problem in machine learning.

Task 6

Baseline Model - Logistic Regression Model

Task 7

Comparison Model - Random Forest Model

Task 8

I chose the Logistic Regression model as a baseline model because it is a popular, simple, and efficient model to predict a binary outcome. I chose the Random Forest model as a comparison model because it can robustly capture more complex interactions between variables, making predictions by combining multiple decision trees.

Task 9

I am choosing the evaluation metric(s) precision, recall, and F1 score due to the imbalance of the `attended` variable. Precision gives the proportion of true positives in all positive predictions, recall gives the proportion of true positives in all positive instances, and the F1 score combines precision and recall to give a balanced msaure of the model's performance.

Task 10

A larger precision, recall, and F1 score indicate a better performing model. While the Random Forest model has a higher recall, the Logistic Regression model has a higher precision and F1 score. The F1 score is a harmonic mean of the precision and recall, balancing between the two measures in which precision prioritizes cost of false positives and recall prioritizes cost of false negatives. Given the classification is regarding and imbalanced fitness class attendance and the consequences of false positives and false negatives for this variable are similar, I would prioritize the F1 score, indicating the Logistic Regression model performed better at predicting whether a member signed up will attend the class based on the variables collected in the data frame.