Answers to Assignment 1 Part 2

Student ID: s3716113

Student Name: LiangyuNie

I certify that this is all my own original work. If I took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in my submission. I will show I agree to this honor code by typing "Yes": *Yes*.

Feature Engineering:

For the feature engineering we can convert the whole dataset to numeric values.

We convert the values of each column to numeric data separately, as follows:

"Seen a Star Wars film"
"No" = 0, "Yes" = 1

"Fan of Star Wars"
"No" = -1, "unknown" = 0, "Yes" = 1

Seen (Movie)
"No" = 0, "Yes" = 1

Rank for (Movie)
1-6, or 0 if missing value

View of (Character)
"Very unfavorably" = -2, "Somewhat unfavorably" = -1, "Neither favorably nor unfavorably (neutral)" = 0, "Somewhat favorably" = 1, "Very favorably" = 2
Note that we will later exclude the "Unfamiliar (N/A)" answer. From the start we set it -100.

"Which character shot first?"
"Han" = -1, "I don't understand this question" = 0, "Greedo" = 1

(Familiar with) the Expanded Universe?
"No" = -1, "?" = 0, "Yes" = 1

Gender
Male = 0, Female = 1

Age "18-29" = 1, "30-44" = 2, "45-60" = 3, ">60" = 4, "Missing Value" = 0

Household Income "$0 - $24,999" = 1, "$25,000 - $49,999" = 2, "$50,000 - $99,999" = 3, "$100,000 - $149,999" = 4, "$150,000+" = 5, "Missing value" = 0

Education "Less than high school degree" = 1, "High school degree" = 2, "Some college or Associate degree" = 3, "Bachelor degree" = 4, "Graduate degree" = 5, "Missing Value" = 0

Location (Census Region)
"South Atlantic" = 1, "West South Central" = 2, "West North Central" = 3, "Middle Atlantic" = 4, "East North Central" = 5, "Pacific" = 6, "Mountain" = 7, "New England" = 8, "East South Central" = 9, "Missing Value" = 0

Model Selection

There are several model we can choose “Decision Tree” “KNN” , however I choose the decision tree because decision tree is easy to use on categorical variable and our data is numeric and categorical.

Training the Model

First of all we can create an array to store each of the labels we attempt to classify, and another to store the features and we use in our classification.

The Decide Tree function should be:

1. As we can see in the training data, which have some feature variables and classification or regression output.

2. Then we determine the “best feature” in the data to split the data on;

3. As the third step, we should split the data into subsets that contain the possible values for this best feature. Because in this splitting basically defines a node on the tree i.e each node is a splitting point based on a certain feature from our data.

4. As the end we need to recursively generate new tree nodes by using the subset of data. We keep splitting until we reach a point where we have optimized, by some measure, maximum accuracy while minimizing the number of splits / nodes.

Model Validation

We can use Cross-validation. We approach modelling as follows:

1. We need to choose the appropriate model

2. As the second, we need to split the training set into a smaller training set and validation set.

3. We need to tune the parameters of each model by cross-validation on the smaller training set.

4. We need to choose the best parameter set for each model.We will test each model separately on the validation set.

5. At the end we will choose the best model according to your metric.

Applying the trained model to unseen future data

At the end, we need to predict the new observation called model scoring.

We will prepare a data set which is Gender, Age, Household Income, Education, Location (Census Region) and we need to apply the model on these five column, and take a look of this results in a prediction.