DNSC 6279
Spring 2019
Assignment 2

In this assignment we will use the data created in assignment 1 to build a logistic regression classifier to predict whether a shopper will become a repeat shopper.

Before beginning the steps below, check the key from assignment 1 to ensure your datasets are correct. Load the labeled training data and the unlabeled test data into a Python or R session. Both datasets must include the three new columns created in assignment 1. The unlabeled test data will now be referred to as the “score” dataset.

To begin the labeled training data should have 160057 rows and 14 columns. The score data should have 151484 rows and 13 columns.

The following steps can be completed using any packages you choose.

- Recode the target variable, repeater, such that t=1 and f=0.
- Set any level of the variable category that is in the labeled train set but not in the score set to ‘unknown’ in both sets. Set any level of the variable category that is in the score set but not in the labeled train set to ‘unknown’ in both sets.
- Set any level of the variable brand that is in the labeled train set but not in the score set to ‘unknown’ in both sets. Set any level of the variable brand that is in the score set but not in the labeled train set to ‘unknown’ in both sets.
- Set any level of the variable company that is in the labeled train set but not in the score set to ‘unknown’ in both sets. Set any level of the variable company that is in the score set but not in the labeled train set to ‘unknown’ in both sets.
- Set any level of the variable offer that is in the labeled train set but not in the score set to ‘unknown’ in both sets. Set any level of the variable offer that is in the score set but not in the labeled train set to ‘unknown’ in both sets.

  (We ignore the variables category and market for now as their cardinality is too high to be used in an interpretable manner in a logistic regression model without further data preparation steps, e.g. binning based on domain knowledge.)

Now load the two datasets into h2o.

- Use the h2o month() function to create a new categorical variable from the offerdate variable. Create this input variable in the labeled train and score sets.
- Use the h2o week() function to create a new categorical variable from the offerdate variable. Create this input variable in the labeled train and score sets.
- Use the h2o dayOfWeek() function to create a new categorical variable from the offerdate variable. Create this input variable in the labeled train and score sets.
- Use the h2o day() function to create a new numeric variable from the offerdate variable. Create this input variable in the labeled train and score sets.
- Split the labeled training data into labeled training, validation, and test sets using the h2o split_frame() function. Create a 40%-30%-30% split using the seed 12345.

DNSC 6279
Spring 2019
Assignment 2

We will now train a penalized logistic regression model using several categorical and numeric predictors.

The categorical inputs should be: offer, category, company, brand, whether the exact item was bought in the past, month, week, and day of week.

The numeric inputs should be: offervalue, the average dollar amount purchased in-category, the average quantity purchased in-category, and day of month.

The target is repeater.

Use iteratively reweighted least squares as your solver, use 3 CV folds in your training partition, standardize your inputs, conduct a search for the best lambda parameter, and use a seed of 12345.

Download the POJO for the logistic regression classifier after training.

Answer the following questions:

1. (2 pts.) By default h2o finds the cutoff that maximizes the F1 statistic in the validation data. What is the cutoff that maximizes the F1 statistic in the validation data?
   0.21 - 0.26 -- 1 pt.
   0.19 - 0.28 -- ½ pt.
   What is the validation data AUC?
   0.63 - 0.69 -- 1 pt.
   0.59 - 0.73 -- ½ pt.
2. (1 pt.) What is the accuracy in the validation data at this cutoff?
   0.53 - 0.59 -- 1 pt.
   0.50 - 0.62 -- ½ pt.
3. (1 pt.) What is the cumulative lift for the model at the 10th decile in the validation data?
   1.73 - 1.92 -- 1 pt.
   1.64 - 2.01 -- ½ pt.
4. (1 pt.) What offer day of the week is most positively associated with becoming a repeat shopper?
   Saturday
5. (2 pts.) What is the interpretation of the coefficient associated with a customer having bought the exact item in the past?
   The customer buying the exact item before is associated with a X.XX factor change in the odds of becoming a repeat shopper holding all else constant.

   “Holding all else constant” - 1 pt.
   1.12 < X.XX < 1.15 -- 1 pt.
   1.10 < X.XX < 1.17 -- ½ pt.

“On average” - -½ pt. (penalized regression does not model the mean of the conditional distribution of y|X, penalized regression is biased, but usually more accurate)

“Against the reference level” - -½ pt. (penalized regression does not require reference levels)

6. (1 pt.) What is the probability for ID 13584134 in the score set to become a repeat shopper according to your model.
   0.085 - 0.115 -- 1 pt.
   0.07 - 0.13 -- ½ pt.

   Bonus: What potential modeling problem did our data preprocessing in assignment 1 cause?
   Target leakage/leakage - 1 pt. (Because we treated the entire labeled training set as one data set. To be more careful, we should have split the labeled data into three partitions in assignment 1.)

To receive full credit for this assignment, turn in your group’s answers to these questions, your R or Python script/notebook (1 pt.), and your POJO (1 pt.) in one zipped file to Blackboard. Answers will be allowed to fall into ranges of values.