DNSC 6279
Spring 2019
Assignment 3

In this assignment we will use the data created in assignment 2 to build gradient boosting and multilayer perceptron classifiers to predict whether a shopper will become a repeat shopper. Before beginning the steps below, check the key from assignment 2 to ensure your datasets are correct. You should begin from the end of your assignment 2 code.

Target encoding:

Use target encoding in h2o to conveniently perform a leave-one-out target encoding on the high-cardinality chain and market columns. Fit the target encoding on the training data, then transform the train, validation, and test data. My Python code to train and fit the target encoder on the training data looks like this:

```
# train target encoder
e_columns = ['market', 'chain']
te_ = TargetEncoder(x=e_columns, y=y)
train[y] = train[y].asfactor()
_ = te_.fit(train)
e_train = te_.transform(frame=train, holdout_type='loo', seed=12345)
```

Be sure to use the same seed as me for your three target encoding transformations.

GBM:

Use an h2o grid search to train a GBM. You should use the input variables from assignment 2 and your two new target encoded features as model inputs. Your grid search should contain at least 50 different models and should search over at least the parameters: ntrees, max_depth, sample_rate, and col_sample_rate. You may search over more models and more parameters. For more information about grid search for GBM with h2o, please see: https://www.h2o.ai/blog/h2o-gbm-tuning-tutorial-for-r/. Be sure to set the same seed as above for the grid search and for model training.

MLP:

Use an h2o grid search to train an MLP. Your grid search should contain at least 50 different models and should search over at least the parameters: hidden, l1, l2, input_dropout_ratio. You may search over more models and more parameters. Because neural networks do not, in general, have variable selection mechanisms, train your MLP grid search on the 10 most important input variables from the GBM. For more information about grid search for MLP with h2o, please see: https://www.h2o.ai/blog/deep-learning-performance/. Be sure to set the same seed as above for the grid search and for model training.

DNSC 6279
Spring 2019
Assignment 3

Answer the following questions:

1. (1 pt.) What is the value for chain_te for row id 258692579?

   0.124
   In range (0.149, 0.099) - 1 pt.
   In range (0.165, 0.083) - ½ pt.

2. (1 pt.) What is the value of market_te for the same row?

   0.121
   In range (0.146, 0.097): +1 pt.
   In range (0.162, 0.081): +½ pt.

3. (1 pt.) What is the test AUC of your best GBM model?

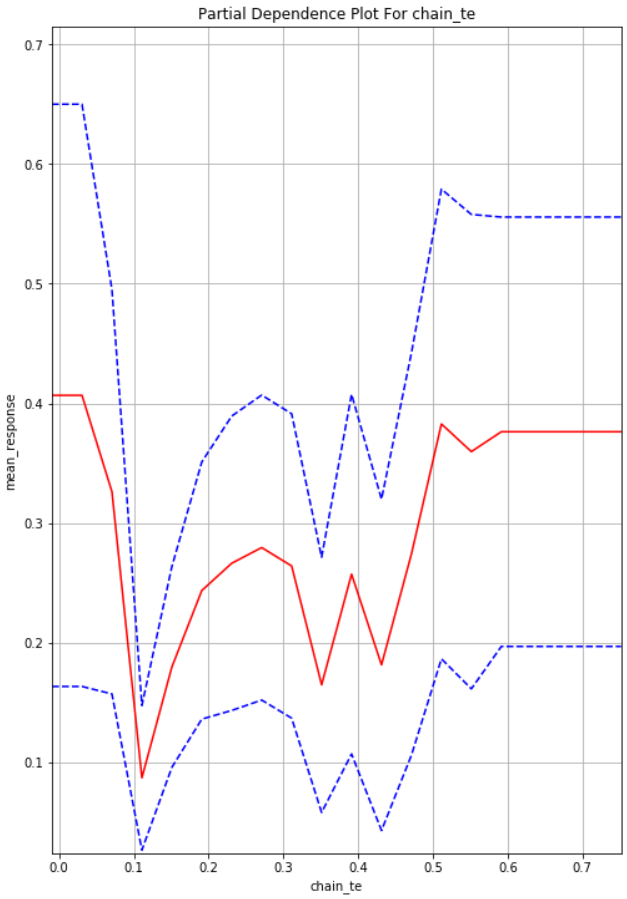
   0.703
   In range (0.774, 0.633): +1 pt.
   In range (0.809, 0.598): +½ pt

4. (3 pts.) What are the three most important variables for your best GBM? Paste a screenshot of your output here.

   Variables: [‘chain_te', 'week', 'avg_category_amount', 'avg_category_quantity']: + 1 pt. each
   Variables ['market_te', 'offer', 'offervalue', 'dayOfWeek']: + ½ pt. each

DNSC 6279
Spring 2019
Assignment 3

5. (1 pt.) Paste the partial dependence plot of the most important variable in the GBM model's test data here.



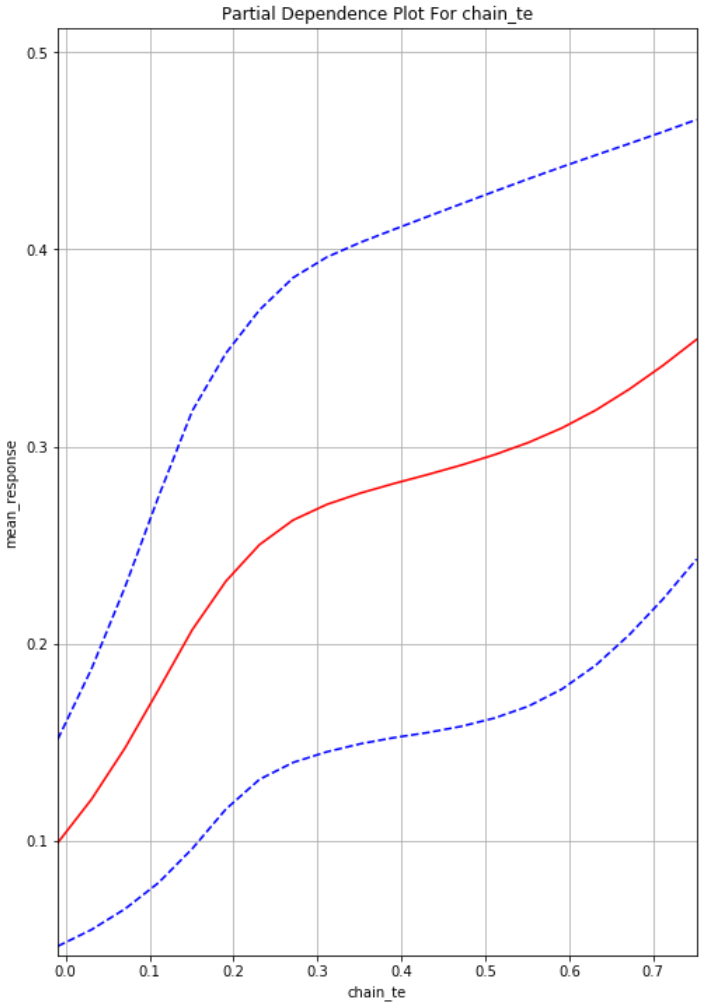
(Any reasonable plot accepted for + 1 pt.)

6. (1 pt.) What is the test AUC of your best MLP model? Paste a screenshot of your output here.

0.691
In range (0.760, 0.622): +1 pt.
In range (0.794, 0.587): +½ pt

DNSC 6279
Spring 2019
Assignment 3

7. (1 pt.) Paste the partial dependence plot of the same variable in #5, but for your best ANN model's test data, here.



(Any reasonable plot accepted for + 1 pt.)

Bonus: The best test AUC in the class for either model will receive 2 pts. extra credit. (Assuming you did the rest of the assignment correctly.)

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