DSBA 6156
Spring 2023
Midterm
02/22/2023
Time Limit: 165 Minutes

Name (Print): _____

This exam contains 11 pages (including this cover page) and 16 problems. Check to see if any pages are missing. Enter all requested information on the top of this page, and put your initials on the top of every page, in case the pages become separated.

- You may *not* use your books, notes, or any calculator on this exam.
- You may not consult with fellow classmates during the exam.
- You may not use your phone, laptop, or other electronic device during the exam.
- If you need more space, use the back of the pages; clearly indicate when you have done this.

Do not write in the table to the right.

Problem	Points	Score
1	4	
2	4	
3	4	
4	2	
5	2	
6	2	
7	2	
8	2	
9	12	
10	12	
11	8	
12	4	
13	10	
14	8	
15	14	
16	10	
Total:	100	

1. (4 points) A business stakeholder gives you a file of customer data and asks you to find patterns in the data that they can turn into audiences for marketing purposes. This is an example of what?

    A. Supervised Learning

    B. Unsupervised Learning

    C. Reinforcement Learning

2. (4 points) A business stakeholder gives you a file of labelled customer data and asks you to learn a model that can predict which of tomorrow's customers will cancel their orders. This is an example of what?

    A. Supervised Learning

    B. Unsupervised Learning

    C. Reinforcement Learning

3. (4 points) A business stakeholder asks you to build a recommender that will learn through experience which advertisement to show to each user. This is an example of what?

    A. Supervised Learning

    B. Unsupervised Learning

    C. Reinforcement Learning

4. (2 points) What types are the objects defined below:

```
{'a': 1, 'b': 2, 'c': 3}
['a', 'b', 'c']
('a', 'b', 'c')
{'a', 'b', 'c'}
```

    A. set, tuple, list, dictionary

    B. dictionary, tuple, list, set

    C. dictionary, list, tuple, set

    D. list, tuple, set, dictionary

5. (2 points) If run, what would be printed below?

```
def foo(x):
    y = (x+2)/3
print(foo(4))
```

   A. 2
   B. None
   C. True
   D. 3

6. (2 points) What is the purpose of cross validation?
   A. To automate the task of feature engineering.
   B. To reduce the bias of a learner.
   C. To estimate model quality.
   D. To identify correlated features.

7. (2 points) What is the purpose of a ColumnTransformer?
   A. To apply pipelines selectively to certain columns
   B. To identify which columns add value to your model
   C. To estimate model quality.
   D. To identify correlated features.

8. (2 points) What is the purpose of a train/test split?
   A. To reduce the bias of a learner.
   B. To apply pipelines selectively to certain columns
   C. To identify which columns add value to your model
   D. To help understand generalization error and whether your model is overfit.

9. A fellow data scientist at a bank has created a classification model to identify fraudulent credit card transactions. The business will use the model to flag payments and reject ones the model deems fraudulent, where a 1 indicates fraud and 0 indicates a normal transaction.

   The model is a random forest classifier and predictions were generated by using the `predict_proba` method and then applying a prediction threshold of 0.2. Below is the confusion matrix of the model scored on a held out test set:

	actual 0	actual 1
predicted 0	73	4
predicted 1	1024	598

   (a) (3 points) What is the main problem with the model predictions?

   (b) (3 points) What will be the effect on the bank process if this model is deployed?

   (c) (3 points) How should the prediction threshold be changed to make the problem better?

   (d) (3 points) How will your change to the decision threshold most likely affect the quantities in the confusion matrix? That is, will you have more or fewer true negatives, true positives, false negatives, and false positives?

10. A data scientist is using a single deep decision tree classifier to model a binary classification task. The model is excellent on the training set, but has very low quality on the test set.

    (a) (3 points) Using the terms bias and variance, describe what is wrong with the model.

    (b) (3 points) How can he adjust the depth of the tree to most likely improve the problem?

    (c) (3 points) If he does not want to adjust tree depth, how can he use multiple such trees to solve the problem?

    (d) (3 points) Do you think a logistic regression classifier would likely experience the same issue on this data? Why or why not?

11. You are given a data set $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$, where the x_i and y_i are real numbers (not vectors).

    (a) (2 points) Would you describe modeling y_i in terms of x_i as a regression or classification problem?

    (b) (3 points) What form would a linear regression model $f(x)$ take here? That is, $f(x) = \ldots$?

    (c) (3 points) If we wish to select parameters for our linear regression model by minimizing squared error, what function would we minimize? Write the function below.

12. Consider the code below

```
i = 1
j = 1
print(i)
print(j)
c =2
while c < 100:
    next_fib = i + j
    print(next_fib)
    i = j
    j = next_fib
```

The intention is to print the first 100 Fibonacci numbers. The first two numbers in the Fibonacci sequence are 1 and 1. The next number in the sequence is always the sum of the previous two. The code above is misbehaving.

(a) (2 points) What will the above code actually do?

(b) (2 points) How can we adjust the code to have it behave as intended?

13. Consider the code below

```
vec_1 = np.array([1, 2, 3, 4])
vec_2 = np.array([2, 4, 6, 12])

my_dictionary = {'integers': pd.Series(vec_1, index=['a', 'b', 'c', 'd']),
                 'even_integers': pd.Series(vec_2, index=['a', 'b', 'c', 'f'])}

my_df = pd.DataFrame(my_dictionary)

my_df['A'] = my_df['integers'] * my_df['even_integers']
my_df['B'] = my_df['A'] <= my_df['even_integers']
```

(a) (2 points) How many rows does `my_df` have?

(b) (2 points) How many columns does `my_df` have (not including the index)?

(c) (2 points) What type of object is `my_df` ?

(d) (2 points) What type of object is `my_df['A']` ?

(e) (2 points) What type of object is `my_df[['A']]` ?

14. We wish to create a logistic regression model on the input feature vector x and binary outcome feature y.

    (a) (4 points) What form does the model have? That is, the model f, takes the form $f(x) = \ldots$.

    (b) (4 points) The model parameters are learned by maximizing what function? That is, with data $\{(x_i, y_i)\}_{i=1}^N$ and model $f(x)$, what quantity does $f(x)$ maximize?

15. Consider the code below

```
numeric_tx = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
])
cat_tx = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])

preprocessor = ColumnTransformer(transformers=[
    ('numeric', numeric_tx, num_vars),
    ('categorical', cat_tx, cat_vars)
])

param_grid = {'learning_rate': [0.1, 0.01], 'n_estimators': [100, 1000]}

gbm = GradientBoostingClassifier()
clf = GridSearchCV(gbm, param_grid, scoring='roc_auc')
```

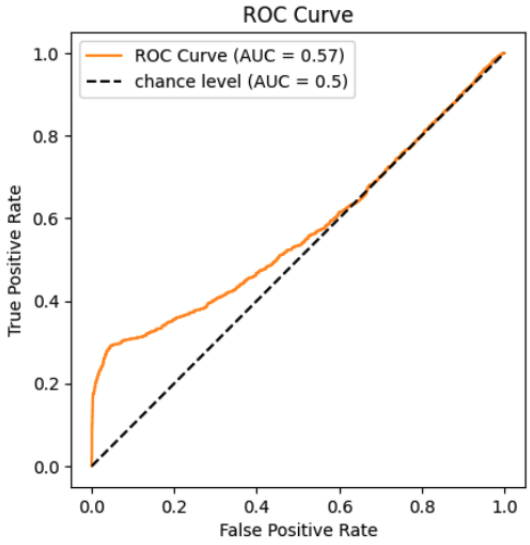
(a) (2 points) What are `learning_rate` and `n_estimators` examples of?

(b) (4 points) Describe how the appropriate `param_grid` values are chosen.

(c) (4 points) Why might the OneHotEncoder pipeline step be necessary?

(d) (4 points) What strategy is being employed by the classifier to improve high bias of the single shallow tree here?

16. Consider the roc curve below:



(a) (5 points) Would a large or small prediction threshold likely get the most value out of this classifier? Justify your answer.

(b) (5 points) Describe what this model could be potentially useful for? What would it not be useful for?