DATA-448
Exam I

Name: _____

Exam Instructions

1. **Show all your work** and write complete and coherent answers.
2. Show all of the steps that you used to get your final answer. If you do not show your work I can not give partial credit in the case of incorrect answers.
3. **Please strive for clarity and organization.**
4. **You are not allowed to discuss any of the exercises in exam 1 with others.**
5. **Late submission will not be accepted, regardless of the circumstances.**

Take the time to carefully read all the questions on the exam. GOOD LUCK!

1. (20 points) **True or False**.

   (a) In class imbalanced classification tasks, the goal is to find the model that has the highest accuracy. (**True** **False**)

   (b) Over-fitting is the biggest issue with over-sampling. (**True** **False**)

   (c) Under-fitting is the biggest issue with over-sampling. (**True** **False**)

   (d) A model with a large number of input variables always have a good performance. (**True** **False**)

   (e) If we run the SMOTE sampling technique multiple times on an imbalanced dataset, we will always obtain the same synthetic dataset. (**True** **False**)

   (f) Having access to good data is better than having a good model. (**True** **False**)

   (g) RFE works with any support vector machine model. (**True** **False**)

   (h) RFE doesn't work with any support vector machine model. (**True** **False**)

   (i) A typical approach to engineer features is to consider interactions because most interactions help models to generate good predictions. (**True** **False**)

   (j) The support vectors in support vector machine models can be used to identify important variables/featues. (**True** **False**)

2. (5 points) How does the Recursive Feature Elimination (RFE) algorithm work? Be specific.

3. (5 points) Explain one-hot encoding. Be specific.

4. (5 points) If you have a `date` column in your data-frame, then how will you perform feature engineering on the `date` column? List at least three features that you will engineer from `date`. Be specific.

5. (5 points) In what scenarios, would you prefer to use the precision-recall curve instead of ROC curve to measure the performance of a classifier? Be specific.

6. (7 points) Explain the biggest drawback of one-vs-one multi-class classification when compared to one-vs-all multi-class classification. Be specific.

7. (4 points) A data scientist is building a linear regression model. One of the input variables is a categorical variable with three labels. So, he decided to use the one-hot encoding approach to transform the categorical variable into dummy variables. How many dummy variables does he need to include in the linear model?

   (a) 0

   (b) 1

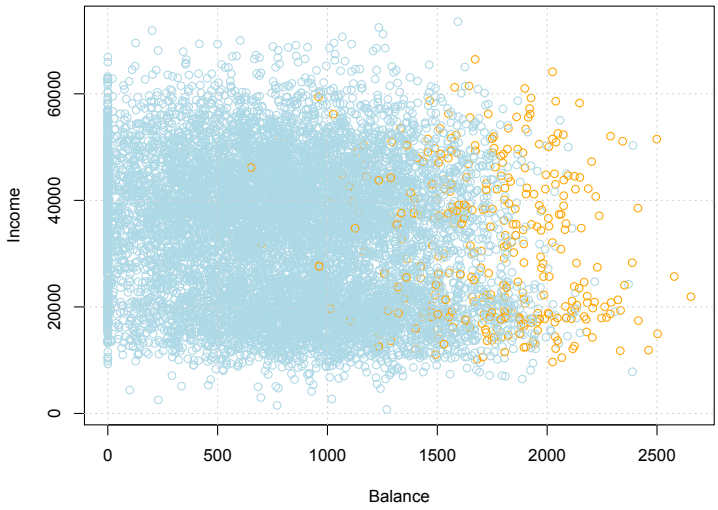
   (c) 2

   (d) 3

   (e) None of the above

8. (7 points) Suppose you are building a fraud detection system for major US bank. You have access to all the transaction data for the past week for users (date, location, and amount). What kind of new features can you engineer? Be creative and list at least three features that you would engineer for the fraud detection system.

9. The chart below was constructed using data related to credit card payments of college students. On the x-axis, we have the credit card balance, and on the y-axis we have the college students' incomes. The light-blue circles represent payments that didn't default; on the other hand, the orange circles represent credit card payments that default.



Let's assume that you are the data scientist in charge of this project. The goal is to build a classification system that can flag future credit card payments that are likely to default. Answer the following:

   (a) (3 points) Is the dataset imbalanced? Be specific.

   (b) (5 points) Using the above chart, engineer one feature for your classification model. Be specific.

10. (4 points) Which of the following is/are **TRUE** feature subset selection?

   (a) Subset selection can substantially decrease the bias of support vector machine models.

   (b) Ridge regression frequently eliminates some of the features.

   (c) Subset selection can reduce over-fitting.

   (d) Finding the true best subset takes exponential time.

11. (4 points) A data scientist is building a regression model. A few of the input variables are categorical, and he has not looked at the distribution of the categorical data in the test data. The data scientist wants to apply the one-hot encoding on the categorical features. What challenges he may face if he have applied one-hot encoding on a categorical variable of the train dataset?

(a) All categories of categorical variable are not present in the test dataset.

(b) Frequency distribution of categories is different in train as compared to the test dataset.

(c) Train and test always have same distribution.

(d) (a) and (b)

(e) (a) and (c)

(f) (b) and (c)

(g) None of the above.

12. (5 points) What is the difference between feature engineering and feature selection? Be specific.

13. (5 points) Based on the discussions from Chapter 4, list at least two benefits of feature selection. Be specific.

14. (6 points) Consider the F1-scores of three different classifiers in a 5-folds cross-validation setting (using the same test dataset) is shown below.

fold	Model 1	Model 2	Model 3
1	0.76	0.81	0.75
2	0.77	0.78	0.73
3	0.78	0.78	0.76
4	0.79	0.80	0.75
5	0.76	0.79	0.73

Based on the F1-score, what model would use to make predictions? Be specific.

15. Consider the `train.csv` and `test.csv` data files. The `train.csv` data file contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. There are 25 variables:

- `LIMIT_BAL`: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- `SEX`: Gender (1=male, 2=female)
- `EDUCATION`: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- `MARRIAGE`: Marital status (1=married, 2=single, 3=others)
- `AGE`: Age in years
- `PAY_0`: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above)
- `PAY_2`: Repayment status in August, 2005 (scale same as above)

- `PAY_3`: Repayment status in July, 2005 (scale same as above)
- `PAY_4`: Repayment status in June, 2005 (scale same as above)
- `PAY_5`: Repayment status in May, 2005 (scale same as above)
- `PAY_6`: Repayment status in April, 2005 (scale same as above)
- `BILL_AMT1`: Amount of bill statement in September, 2005 (NT dollar)
- `BILL_AMT2`: Amount of bill statement in August, 2005 (NT dollar)
- `BILL_AMT3`: Amount of bill statement in July, 2005 (NT dollar)
- `BILL_AMT4`: Amount of bill statement in June, 2005 (NT dollar)
- `BILL_AMT5`: Amount of bill statement in May, 2005 (NT dollar)
- `BILL_AMT6`: Amount of bill statement in April, 2005 (NT dollar)
- `PAY_AMT1`: Amount of previous payment in September, 2005 (NT dollar)
- `PAY_AMT2`: Amount of previous payment in August, 2005 (NT dollar)
- `PAY_AMT3`: Amount of previous payment in July, 2005 (NT dollar)
- `PAY_AMT4`: Amount of previous payment in June, 2005 (NT dollar)
- `PAY_AMT5`: Amount of previous payment in May, 2005 (NT dollar)
- `PAY_AMT6`: Amount of previous payment in April, 2005 (NT dollar)
- `default payment next month`: Default payment (1=yes, 0=no)

Notice that the `test.csv` data file has the information as the `train.csv` data file but `default payment next month`. The goal is to predict `default payment next month` on the `test.csv` data file. In Python, answer the following:

(a) (4 points) Using the pandas library, read the `train.csv` and `test.csv` data files and create two data-frames, called them `train` and `test`.

(b) (3 points) Report the frequency table of `default payment next month` variable in the `train` data-frame.

(c) (10 points) Split the `train` data-frame intro `training` (80%) and `testing` (20%) (taking into account the proportions of 0s and 1s). Make sure the distribution of the categorical variables is the approximately the same in the `training` and `testing` data-frames. See the `this link` for reference.

(d) (25 points) Using the `training` data-frame, engineer at least five different features, that can help to predict `default payment next month`, using the given input variables. Engineer the same features, that you engineer on the `training` data-frame, on the `testing` and `test` data-frames. These are the rules to engineer the features:

- You can't use the Box-Cox transformation.
- You can't use neither 0-1 scaling nor z-score standardization.
- You can engineer at most two features using *strong heredity* principle. For reference, see homework assignment 5.

(e) (20 points) Using the `training` data-frame (including the engineered features from part (c)), run the RFE algorithm to identify important variables. Run the RFE (using the `RFECV` function) with a base estimator of your preference, `step = 1`, `min_features_to_select = 2`, and `cv = 3`. Run the RFE 100 times and extract the support of each of the features. Combine the results and rank the features. *Tip related to features to be included in the RFE algorithm*: if you engineered a feature based only on a single raw input variable, you should only include the engineered feature not both (engineered feature and raw feature).

(f) (10 points) Using the top 5 variables from the part (e) and the same base estimator from part (e), build a model on the `training` data-frame. Then, use this model to make predictions on the `testing` data-frame. Use the provided `precision_recall_cutoff.py` (posted under the Exam 1 link) file to estimate the optimal cutoff value. Compute the classification report of this model.

(g) (10 points) Using the top 6 variables from the part (e) and the same base estimator from part (e), build a model on the `training` data-frame. Then, use this model to make predictions on the `testing` data-frame. Use the provided `precision_recall_cutoff.py` (posted under the Exam 1 link) file to estimate the optimal cutoff value. Compute the classification report of this model.

(h) (3 points) Using the results from parts (f) and (g), what model would you use to predict `default payment next month`? Be specific.

(i) (10 points) Based on your answer from part (h), predict the likelihood of `default payment next month` on the `test`. Submit the likelihoods in a csv file. Also submit the associated cut-off value.

[1]

[1]The best answer for question 15(i) will receive 10 extra points on Exam 1. The second best answer for question 15(i) will receive 5 extra points on Exam 1. Answers will be evaluated using the F1-score.