Eric Chen

Part 1

1.) In this question we are going to design a machine learning problem for credit card fraud detection. You have to predict, based on certain attributes, whether a transaction is fraudulent or legitimate. Answer the following questions.

• 1 a. Is this a classification problem or a regression problem? (2 pts)

Classification Problem because the problem is indicating whether a transaction is fraudulent or legitimate which is the two possible outcomes. It’s classifying whatever the transaction is legitimate or not.

• 1 b. Describe at least 4 discrete features of each transaction that is going to help in our

problem. (8 pts)

Credit Card Number, Card Verification Value, Person’s Name, and Expiration Date

• 1 c. Describe at least 3 continuous features of each transaction that is going to help in our problem. (6 pts)

Balance in credit card, amount of money spent on the transaction, and transaction fees.

• 1 d. If we were to use an ML algorithm that only takes discrete features as input, what can we do to the continuous features you described in 1 (c) so that they can be used by the ML algorithm? (4 pts)

We can create intervals with ranges of possible values like [0-25], [25-50], [50-75], etc, and putting the values from the continuous features into the corresponding intervals like putting 18 into interval [0-25].

• 1 e. How would you know if your model suffers from overfitting? If you are using decision tree in your problem, how would you deal with overfitting? (4 pts)

I would know my model is suffering from overfitting if my decision tree is getting too big and/or working with too much data or attributes. Another way is if the ML model takes extra minors features that may not be relevant. I would deal with this by removing those features that are irrelevant and applying pruning to lower nodes in the decision tree, that is, if the gain of the best attribute at a node is below a threshold, stop and make this node a leaf rather than generating its children nodes.

• 1 f. In Table 1, observe the data and use your own intuition to draw a decision tree that can classify the data into Class 1 and Class 2. (6 pts)

