# Homework: Worksheet for Chapters 1-3:2

Student Name:

### Reading Assignment:

Skim Chapter 1:1 through Chapter1:3 (15 mins). Book: Reproducible Machine Learning for Credit Card Fraud Detection – Practical Handbook by Borgne et al. available free through CC BY-SA 4.0 license at [https://fraud-detection-handbook.github.io/fraud-detection-handbook/Foreword.html#](https://fraud-detection-handbook.github.io/fraud-detection-handbook/Foreword.html)

Read all of Chapter 2 (30 mins).

Read part of Chapter 3, specifically Chapter 3:1 ‘Introduction’ and Chapter 3:2 ‘Transaction data simulator’ (60 mins)

Answer the various questions below in a sentence or two. You may summarize in your own words or copy and paste from the book. These questions are to help draw your attention to items Dr. Humpherys considers important. The purpose is not to force you to memorize anything or have you write long-winded essays. This worksheet is to help you learn the book’s content. Do the various learning activities listed below, e.g., run code as instructed. Explore and experiment. The more you explore from the book and experiment with code, the more you will learn!

### Deliverable:

Submit this worksheet with your embedded answers and submit one pickle file you generated to WTclass\cidm6356\lessons\Week 12\Homework: Worksheet from Chapters 1-3\

## Chapter 1:3

Do this learning activity to learn how to run the code found in this book. Nothing to turn in yet.

1. Navigate to Chapter 1.3 ‘How to use this Book’.
2. Click the rocket icon (top right, See Figure 1).

Figure 1. The Rocket icon allows you to run the book’s code in Google Colab.

A screen shot of a computer

Description automatically generated

1. Choose ‘Collab’. The book chapter will be opened inside of Google Colab.
2. In Colab, scroll down to section ‘3.3.4 Try It’.
3. Click the run button next to the code cell print(‘Hello World’). See Figure 2.

Figure 2. Run code inside the book.

A screenshot of a computer

Description automatically generated

Did you successfully run code from the book in Google Colab’s environment? Yes or No

1. You can run code, add new code, experiment with code, but you cannot save anything to the original book. If you want to save your experiments, click File, Save a Copy in [Google] Drive.
2. Be sure to terminate your Google Colab session when done to not waste resources.

## Chapter 2:2

1. What do these terms mean CP and CNP? Write your answer here. You may even copy and paste from the book.
   1. Card-present
      1. Card is physically used
   2. Card-not-present
      1. Card is not physically used
2. Which accounts for more fraud CP or CNP?
   1. CNP

## Chapter 2:3

1. Look at Figure 1 in Chapter 2:3. Observe how automated systems are combined with human investigators. Just observe. Nothing to turn in. Figure 3 below show examples of different types of credit card terminals that the book talks about.

Figure 3. Examples of different types of credit card terminals.

A close-up of a device

Description automatically generated

1. What are transaction-blocking rules? Write your answer here. You may even copy and paste from the book.
   1. Transaction-blocking rules are if-then (-else) statements meant to block transaction requests that are perceived as frauds
   2. These rules use the information available when the payment is requested, without analyzing historical records or cardholder profiles.
2. How are scoring rules different from transaction-blocking rules?
   1. Scoring rules assess risk by assigning a score to each transaction based on the transactions factors. Transaction-blocking rules automatically block and decline a transactions that meet predefined criteria.
3. What is a data-driven model (DDM)?
   1. This layer is purely data-driven and adopts a classifier or another statistical model to estimate the probability for each feature vector to be a fraud.
   2. This probability is used as the fraud score associated with the authorized transactions.

## Chapter 2:4

1. Define the three groups of transaction data:
   1. account-related features
      1. They include for example the account number, the date of the account opening, the card limit, the card expiry date, etc
   2. transaction-related features
      1. They include for example the transaction reference number, the account number, the transaction amount, the terminal (i.e., POS) number, the transaction time, etc. From the terminal, one can also obtain an additional category of information: merchant-related features such as its category code (restaurant, supermarket, …) or its location.
   3. customer-related features
      1. They include for example the customer number, the type of customer (low profile, high profile, …), etc.
2. What is feature engineering (i.e., feature transformation, etc.?
   1. also known as feature transformation, feature extraction, or data preprocessing
   2. Feature Engineering transforms raw data into a format for machine learning algorithms by selecting, modifying, and/or creating new features.
3. What is a loss function in machine learning?
   1. A way to assess the performance of a prediction model
4. Why is zero/one loss function not a good measure for credit card fraud?
   1. It is for binary classification problems
   2. due to the high-class imbalance (much more genuine than fraudulent transactions)
5. What are the challenges of using machine learning for credit card fraud detection? Write a one sentence description for each.
   1. Class imbalance
      1. Transaction data contain much more legitimate than fraudulent transactions well under 1%
   2. Concept drift
      1. Transaction and fraud patterns change over time
   3. Near real-time requirements
      1. Fraud detection systems must be able to quickly detect fraudulent transactions.
   4. Categorical features
      1. Transactional data typically contain numerous categorical features
   5. Sequential modeling
      1. Each terminal and/or customer generates a stream of sequential data with unique characteristics.
   6. Class overlap
      1. The last two challenges can be associated with the more general challenge of overlapping between the two classes.
   7. Performance measures
      1. Standard measures for classification systems are not well suited for detection problems due to the class imbalance issue, and the complex cost structure of fraud detection.
   8. Lack of public datasets
      1. For obvious confidentiality reasons, real-world credit card transactions cannot be publicly shared.

## Chapter 3:2

1. According to the book’s design choice, what are the value for legitimate transactions and value for fraudulent transactions?
   1. A binary variable, with the value for a legitimate transaction, or the value for a fraudulent transaction.
2. Do the following learning activities:
   1. Run all the code cells in Chapter 3:2. Some code takes a minute or two to execute. Focus on understanding the general purpose of each function. You do not need to know what every line of code means.
   2. Change the code to generate ten customers and paste your new customer\_prfiles\_table here as evidence of success.

|  | **CUSTOMER\_ID** | **x\_customer\_id** | **y\_customer\_id** | **mean\_amount** | **std\_amount** | **mean\_nb\_tx\_per\_day** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 54.881350 | 71.518937 | 62.262521 | 31.131260 | 2.179533 |
| **1** | 1 | 42.365480 | 64.589411 | 46.570785 | 23.285393 | 3.567092 |
| **2** | 2 | 96.366276 | 38.344152 | 80.213879 | 40.106939 | 2.115580 |
| **3** | 3 | 56.804456 | 92.559664 | 11.748426 | 5.874213 | 0.348517 |
| **4** | 4 | 2.021840 | 83.261985 | 78.924891 | 39.462446 | 3.480049 |
| **5** | 5 | 97.861834 | 79.915856 | 48.840539 | 24.420270 | 3.122117 |
| **6** | 6 | 11.827443 | 63.992102 | 18.618562 | 9.309281 | 3.778676 |
| **7** | 7 | 52.184832 | 41.466194 | 30.132783 | 15.066392 | 3.096935 |
| **8** | 8 | 45.615033 | 56.843395 | 6.785031 | 3.392516 | 2.470542 |
| **9** | 9 | 61.209572 | 61.693400 | 94.656067 | 47.328034 | 2.727281 |

* 1. Change the code to generate 7 terminals and paste your new terminal\_profiles\_table here as evidence of success. Note that because we are changing the number of customers and numbers of terminals, the book’s prose may be different than our resulting data.

|  | **TERMINAL\_ID** | **x\_terminal\_id** | **y\_terminal\_id** |
| --- | --- | --- | --- |
| **0** | 0 | 54.881350 | 71.518937 |
| **1** | 1 | 60.276338 | 54.488318 |
| **2** | 2 | 42.365480 | 64.589411 |
| **3** | 3 | 43.758721 | 89.177300 |
| **4** | 4 | 96.366276 | 38.344152 |
| **5** | 5 | 79.172504 | 52.889492 |
| **6** | 6 | 56.804456 | 92.559664 |

* 1. Continue to run the rest of the code in Chapter 3:2.

1. In general, what is the purpose of Panda’s .apply() function (see Chapter 3:2.4 ‘Generation of Transactions’ for sample code)? You may search Google or get an explanation from Google Gemini.
   1. The Panda’s .apply() function allows you to apply a function along a column or row of the dataframe.
2. List here a brief description of the three scenarios of fraud using in Chapter 3:2 ‘Fraud scenarios generation’.
   1. Scenario 1
      1. Transactions that have an amount more than 220 is a fraud.
      2. This will provide a baseline for fraud.
   2. Scenario 2
      1. Every day, a list of two terminals is drawn at random. All transactions on these terminals in the next 28 days will be marked as fraudulent.
      2. This is used to simulate an active ciminal
   3. Scenario 3
      1. Every day, a list of 3 customers is drawn at random. In the next 14 days, 1/3 of their transactions have their amounts multiplied by 5 and marked as fraudulent.
      2. This is used to simulate card-not-present fraud where a customer’s information has been leaked.
3. What does ‘class imbalance’ mean?
   1. a mix of numerical and categorical features (with categorical features involving a very large number of values), non-trivial relationships between features, and time-dependent fraud scenarios.
4. After you run all the code, from Google Colab, download just one pickle file 2018-04-01.pkl. Upload 2018-04-01.pkl to WTclass as evidence of being able to generate credit card fraud data. See Figure 4.

Figure 4. How and where to download files from Google Colab

A screenshot of a computer

Description automatically generated

## Grading Rubric: 100 points.

1. Answered all the questions in this worksheet. 50 points
2. Submitted the appropriate pickle file. 50 points