# Homework: Worksheet for Chapters 1-3:2

Student Name: Ben Collier

### 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 written 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!

Watch a 5-minute video explaining the reading and coding assignments <https://www.screencast.com/t/2fWGthc5fV93>

### 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? Yes.

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. Card-present (CP) scenarios, refer to scenarios where a physical card is needed, such as transactions at a store (also referred to as a point-of-sale - POS) or transactions at a cashpoint (for instance at an automated teller machine - ATM). The second, called card-not-present (CNP) scenarios, refers to scenarios where a physical card does not need to be used, which encompasses payments performed on the Internet, by phone, or by mail.
2. Which accounts for more fraud CP or CNP? CNP accounts for far more fraud than CP. CNP fraud is more than half in most cases and up to 79% reported by the European Central Bank in 2020.

## 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 shows 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 is transaction-blocking rules? Write your answer here. You may even copy and paste from the book. Transaction-blocking rules are *if-then (-else)* statements meant to block transaction requests that are perceived as frauds. 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? Scoring rules operate on feature vectors and assign a score to each authorized transaction: the larger the score, the more likely the transaction is to be a fraud. Scoring rules are manually designed by investigators, which arbitrarily define their associated scores.
3. What is a data-driven model (DDM)? The data-driven model is trained from a set of labeled transactions and cannot be interpreted or manually modified by investigators. An effective data-driven model is expected to detect fraudulent patterns by simultaneously analyzing multiple components of the feature vector, possibly through nonlinear expressions.

## Chapter 2:4

1. Define the three groups of transaction data: account-related features, transaction-related features, customer-related features. Write your answer here. You may even copy and paste from the book.

* Account-related features: They include for example the account number, the date of the account opening, the card limit, the card expiry date, etc.
* Transaction-related features: 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.
* Customer-related features: They include for example the customer number, the type of customer (low profile, high profile, …), etc.

1. What is feature engineering (i.e., feature transformation, etc.? Feature engineering is the process of enriching data with additional variables to ideally improve the detection performance of a prediction model.
2. What is a loss function in machine learning? A loss function is how a prediction model’s performance is assessed. It compares the true label *y*  to the predicted label *y^*=ℎ(*x,*θ) for an input *x*.
3. Why is the zero/one loss function not a good measure for credit card fraud? Due to the high-class imbalance of the zero/one loss function, it is not suitable for credit card fraud detection.
4. What are the challenges of using machine learning for credit card fraud detection? Write a one sentence description for each.

* Class imbalance: Learning from imbalanced data is a difficult task since most learning algorithms do not handle well large differences between classes.
* Concept drift: Those who commit fraud are always looking for new ways to commit it, and the spending habits of credit card users change over time.
* Near real-time requirements: Due to the high volume of transaction data daily, classifying fraud may need to be complete within milliseconds of a transaction taking place.
* Categorical features: Machine learning algorithms are not suited to handle categorical features like customer ID and card type, they must be converted into numerical features.
* Sequential modeling: Modeling the stream of sequential data that comes with each terminal and/or customer is a challenge.
* Class overlap: Using only the raw information from a transaction, distinguishing between fraudulent and real transactions is almost impossible.
* Performance measures: Standard measures for classification problems are not suited for detection problems because of class imbalance as well as their complex cost structure.
* Lack of public datasets: The scarcity of public datasets involving fraud detection makes it difficult to reproduce research works and impossible to compare different techniques by independent researchers.

## Chapter 3:2

1. According to the book’s design choice, what are the values for legitimate transactions and value for fraudulent transactions? 0 for legitimate transactions, 1 for fraudulent transactions.
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\_profiles\_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 |
| 1 | 1 | 42.365480 | 64.589411 | 46.570785 | 23.285393 |
| 2 | 2 | 96.366276 | 38.344152 | 80.213879 | 40.106939 |
| 3 | 3 | 56.804456 | 92.559664 | 11.748426 | 5.874213 |
| 4 | 4 | 2.021840 | 83.261985 | 78.924891 | 39.462446 |
| 5 | 5 | 97.861834 | 79.915856 | 48.840539 | 24.420270 |
| 6 | 6 | 11.827443 | 63.992102 | 18.618562 | 9.309281 |
| 7 | 7 | 52.184832 | 41.466194 | 30.132783 | 15.066392 |
| 8 | 8 | 45.615033 | 56.843395 | 6.785031 | 3.392516 |
| 9 | 9 | 61.209572 | 61.693400 | 94.656067 | 47.328034 |

* 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 |
| 0 | 0 | 54.881350 |
| 1 | 1 | 60.276338 |
| 2 | 2 | 42.365480 |
| 3 | 3 | 43.758721 |
| 4 | 4 | 96.366276 |
| 5 | 5 | 79.172504 |
| 6 | 6 | 56.804456 |

* 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. The purpose of .apply() is to apply a function (custom or predefined) to each element, row, or column in a DF or Series.
2. List here a brief description of the three scenarios of fraud using in Chapter 3:2 ‘Fraud scenarios generation’.

* Scenario 1: Any transaction over the amount of 220 is considered fraudulent.
* Scenario 2: Two terminals are chosen at random, daily. These terminals will be marked as fraudulent for the next 28 days.
* Scenario 3: 3 customers are drawn at random, daily. In the next fourteen days, one third of their transactions have their amounts multiplied by 5 and marked fraudulent.

1. What does ‘class imbalance’ mean? Class imbalance is a situation in classification tasks where the distribution of classes of inequal. In the context of fraud detection, a class imbalance would be where genuine transactions (majority class) heavily outnumber the fraudulent transactions (minority class).
2. 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