1. Introduction

One of the most important aspects when developing **data mining** solutions is the **quality** of the data that will be mined. This is particularly true when developing systems to **detect** and **prevent fraud**. When an **incorrect metric** is included in the training set, it could lead to **false positives** or even a **missed fraudulent transaction**. The **Credit Card Fraud dataset** was created by the **ULB Machine Learning Group.** It consisted of **249,100 transactions** during **2 days for** European cardholders. There were only **455 transactions** that were **identified as fraudulent**. As studied in week 11 of our class, **preparing the data** is the **most critical** component of **data mining**. Data quality matters because machine learning models rely heavily on **accurate, representative, and unbiased data**. In this project, I examined three data quality variables that are particularly relevant to data mining models:

- **Sampling Bias & Representativeness**
- **Extreme Class Imbalance**
- **Labelling Validity & Fairness Limitations.**

2. Data Quality Analysis

Sampling Bias & Representativeness

The **limited time span** of data and the **lack of representation of other regions** make it difficult to fully understand the **fraud patterns** that may occur outside this small sample size. The **fraud patterns** for a given merchant or region may be **vastly different** than those in this sample. Fraud patterns and methods also **differ between seasons** and **by merchant type**, and therefore, a model based solely **on two days'** worth of transactions **would not generalise** to **all transactions globally for an extended period**. In addition, models must consider the longitudinal and evolving nature of fraud (**time-variant, in warehousing terms**), and therefore, a two-day snapshot may become **outdated quickly** and would **violate the timeliness dimension** of data quality. While the **lack of missing values** in the dataset shows **strong technical completeness**, the lack of **comprehensive geographic** and **seasonal representation reduces the dataset's usefulness for Data mining**.

Extreme Class Imbalance

There is a significant imbalance in the dataset between the two classes, **as legitimate transactions make up 99.82%** of the data while the **fraudulent class occupies only around 0.18%**, causing issues in both the **accuracy and validity** of the data quality; this low probability of predicting a fraudulent transaction makes **accuracy misleading**. For example, a naive classifier that classified every transaction as non-fraudulent would report an overall accuracy **of approximately 99.8%**, which would **incorrectly lead** us to assume that this classifier performed well but would also **miss all fraudulent transactions**. Additionally, the class imbalance of the dataset in the context of a **CRISP-DM** model makes it difficult for a model to build classifiers and evaluate them; for a given classifier, in general, it is possible to overfit to the **majority class (Legitimate).** More specifically, when trying to learn to predict, the model will tend to be **biased toward the legitimate class** and thus produce a **high number of false negatives**, which are the worst error when trying to detect fraudulent transactions. This is closely related to the topic we discussed in Week 11 regarding how **class imbalance will often lead to a biased performance measure**.

Labelling Validity and Fairness Limitations

As explained in my presentation, since the dataset is labelled as **fraud only** when fraud can be confirmed, **undetected/unreported fraud may be classified as legitimate**. This is an example of a **false negative**, which creates a **labelling bias and affects the data validity** of the model's learning. This creates an opportunity for **inaccuracies** and **miscalibrations** to result from this **false negative identification**.

Additionally, without accessing **demographic information** (e.g., age, income, geographic location), it is **impossible to evaluate** how fairly each demographic group is being represented as it relates to **fraud detection**. This is an ongoing topic of discussion amongst many ethical AI systems, which heavily rely upon protecting the individual's right to privacy whilst creating bias audits to gauge fairness.

---

3. Recommendations

1. Address Class Imbalance (Critical)

Use resampling techniques like **SMOTE**, **under-sampling**, or **adjusting algorithmic class weights** to help increase minority detection. Also, use **evaluation metrics** for fraud detection, such as **precision**, **recall, F1 score, and AUC-PR,** instead of accuracy. This recommendation was directly related to addressing an imbalance identified.

2. Broaden Data Scope to Reduce Sampling Bias (Critical)

Increase transaction records collected from many **different regions** and issuers over **many months or years** to improve **representativeness** and **reduce sampling bias**, **making the model more robust**. This relates directly to the **timeliness** and **completeness dimensions**.

3. Refine Labelling Processes

Incorporate human-in-the-loop review, **Merchant-reported anomalies**, and **Unsupervised Detection** of Anomalies to **identify potential fraud cases** that were **not originally identified**. In addition to being better labelled, this also improves the **overall validity** and **accuracy** of the reporting. This aligns with our Week 11 discussion regarding the **influence of label noise** on the quality of reporting.

4. Ensure Fairness Evaluation Once More Attributes Are Available

If future Datasets include **demographic- or region-based attributes**, it is important to implement **fairness evaluation tools** that will detect **disparate impact** across groups. This will help ensure that **artificial intelligence** is developed in accordance with **ethical principles**. In addition, **regular audits** will allow **ongoing monitoring** for potentially **unintended biases**.

Conclusion

This project identifies serious **data quality issues** with established **benchmark datasets** that negatively impact data mining results. Problems may include **sampling bias, class imbalance, and labelling constraints**. The project emphasizes that the measure of data quality is based on **representativeness, fairness**, and **model validity**. The project conveys that **high-quality data**, as well as **advanced algorithms**, operate together to determine the **effectiveness** of **detecting fraud**. Improving the quality of data will result in **increased accuracy, ethics, and reliability of the fraud detection models** being used by organizations through the **proper addressing of identified issues** and the implementation of specified recommendations.
