OpenL2D Fraud Detection Dataset

Anonymous Author(s)
Affiliation
Address
email

Motivation

Q1: **For what purpose was the dataset created?**

A1: The target of this dataset, comprised by a *learning to defer* (L2D) training scenario with limited expert predictions and a set of expert predictions, is to contribute to the development and evaluation of L2D algorithms. We focus particularly in testing fairness and performance in dynamic conditions, in order to *stress-test* L2D methods.

Q2: **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

A2: Currently anonymous for double-blind review process.

Q3: **Who funded the creation of the dataset?**

A3: The dataset is synthetically generated using the OpenL2D framework. There was no specific funding for the creation of this dataset.

Composition

Q4: **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**

A4: The OpenL2D Fraud Detection Dataset is comprised of:

- Input dataset - we utilize the bank-account-fraud (BAF) dataset's base variant. Each instance in the bank-account-fraud dataset represents a synthetic, feature-engineered bank account opening application in tabular format. For more information on this dataset, please consult `https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022?datasetId=2673949`.
- Limited expert decision dataset - Subset of the input dataset, where each instance is associated with a decision from either the model or an expert. It was used to develop our L2D methods with limited human decision data.
- Expert decision table - This table contains every expert's decision for every instance in the BAF dataset.
- Capacity constraint tables - For any given subset of the data (e.g. test split), each capacity constraint is defined by a pair of tables:
  - Batch tables - define which batch every instance of the input dataset belongs to.
  - Capacity tables - define the maximum number of cases that can be deferred to each expert for a given batch.

Q5: **How many instances are there in total (of each type, if appropriate)?**

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

A5:
- The input dataset has 1M instances
- The limited expert decision dataset has 506118 instances
- The expert decision table contains 1M instances
- The batch tables contain the same number of instances as the data for which they are generated (e.g batch table generated for the limited expert decision dataset contains 506118 instances). The capacity tables contain one instance for each batch.

Q6: **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**

A6:
- The input dataset was used in its totality
- The limited expert decision dataset includes instances from a subset of the BAF dataset, corresponding to instances from months 4 to 7.
- The expert prediction table contains all instances

Q7: **What data does each instance consist of?**

A7: In the limited expert decision dataset, the generated fields are:

- **model_score** (numeric) : Model score obtained with our ML Model.
- **batch** (categorical): Batch to which the instance belongs, defined by capacity constraints.
- **assignment** (categorical): Entity (expert or model) to which instance's decision was deferred.
- **decision** (numeric): Decision on said instance. In case of deferral to an expert, it is either 0 (accept), or 1 (reject). In case of deferral to the ML model, it is equal to that model's score for said instance.

In our expert decision table, every instance is matched to each entity's decision. Each column corresponds to one of the 51 decision making entities: (50 experts) and one ML Model. Experts are identified by their group and a numeric identifier, in the format "{group_name}#{id}". The ML model is identified as "model#0". Expert decisions are binary, while the model's decision is comprised of the model score for each instance, allowing for posterior thresholding.

In the batch tables the columns are:

- **case_id** - identifier for each instance in the input dataset
- **batch_id** - identifier of the batch that the instance belongs to

In the capacity tables the columns are the same as the expert decision table, and each instance is identified by the "batch_id".

The fields of the input dataset, related to information on the bank account application, are the same as the base BAF variant. These are also present in the limited expert decision dataset:

- **income** (numeric): Annual income of the applicant (in decile form). Ranges between $[0.1, 0.9]$.
- **name_email_similarity** (numeric): Metric of similarity between email and applicant's name. Higher values represent higher similarity. Ranges between $[0, 1]$.
- **prev_address_months_count** (numeric): Number of months in previous registered address of the applicant, *i.e.* the applicant's previous residence, if applicable. Ranges between $[-1, 380]$ months (-1 is a missing value).
- **current_address_months_count** (numeric): Months in currently registered address of the applicant. Ranges between $[-1, 429]$ months (-1 is a missing value).
- **customer_age** (numeric): Applicant's age in years, rounded to the decade. Ranges between $[10, 90]$ years.

- **days_since_request** (numeric): Number of days passed since application was done. Ranges between $[0, 79]$ days.
- **intended_balcon_amount** (numeric): Initial transferred amount for application. Ranges between $[-16, 114]$ (negatives are missing values).
- **payment_type** (categorical): Credit payment plan type. 5 possible (annonymized) values.
- **zip_count_4w** (numeric): Number of applications within same zip code in last 4 weeks. Ranges between $[1, 6830]$.
- **velocity_6h** (numeric): Velocity of total applications made in last 6 hours *i.e.*, average number of applications per hour in the last 6 hours. Ranges between $[-175, 16818]$.
- **velocity_24h** (numeric): Velocity of total applications made in last 24 hours *i.e.*, average number of applications per hour in the last 24 hours. Ranges between $[1297, 9586]$
- **velocity_4w** (numeric): Velocity of total applications made in last 4 weeks, *i.e.*, average number of applications per hour in the last 4 weeks. Ranges between $[2825, 7020]$.
- **bank_branch_count_8w** (numeric): Number of total applications in the selected bank branch in last 8 weeks. Ranges between $[0, 2404]$.
- **date_of_birth_distinct_emails_4w** (numeric): Number of emails for applicants with same date of birth in last 4 weeks. Ranges between $[0, 39]$.
- **employment_status** (categorical): Employment status of the applicant. 7 possible (annonymized) values.
- **credit_risk_score** (numeric): Internal score of application risk. Ranges between $[-191, 389]$.
- **email_is_free** (binary): Domain of application email (either free or paid).
- **housing_status** (categorical): Current residential status for applicant. 7 possible (annonymized) values.
- **phone_home_valid** (binary): Validity of provided home phone.
- **phone_mobile_valid** (binary): Validity of provided mobile phone.
- **bank_months_count** (numeric): How old is previous account (if held) in months. Ranges between $[-1, 32]$ months (-1 is a missing value).
- **has_other_cards** (binary): If applicant has other cards from the same banking company.
- **proposed_credit_limit** (numeric): Applicant's proposed credit limit. Ranges between $[200, 2000]$.
- **foreign_request** (binary): If origin country of request is different from bank's country.
- **source** (categorical): Online source of application. Either browser (INTERNET) or app (TELEAPP).
- **session_length_in_minutes** (numeric): Length of user session in banking website in minutes. Ranges between $[-1, 107]$ minutes (-1 is a missing value).
- **device_os** (categorical): Operative system of device that made request. Possible values are: Windows, macOS, Linux, X11, or other.
- **keep_alive_session** (binary): User option on session logout.
- **device_distinct_emails** (numeric): Number of distinct emails in banking website from the used device in last 8 weeks. Ranges between $[-1, 2]$ emails (-1 is a missing value).
- **device_fraud_count** (numeric): Number of fraudulent applications with used device. Ranges between $[0, 1]$.
- **month** (numeric): Month where the application was made. Ranges between $[0, 7]$.
- **fraud_bool** (binary): If the application is fraudulent or not.

Q8: **Is there a label or target associated with each instance?**

A8: Yes, the label is contained in the **fraud_bool** field. A positive value (fraud_bool=1) represents a fraudulent bank account application. A negative value (fraud_bool=0) represents a legitimate bank account application.

When accepted, all accounts are opened with access to credit.

For additional information on how the labels were obtained, consult the BAF datasheet, available at

Q9: **Is any information missing from individual instances?**

A9: There is no missing information from individual instances.

Q10: **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**

A10: There are no relationships between individual instances.

Q11: **Are there recommended data splits (e.g., training, development/validation, testing)?**

A11: The performed data splits are based on the temporal information of the dataset. To this end, we use the column **month**. Practitioners can test different temporal L2D development strategies (e.g. Training an ML Model, training an assignment system, deploying the system).

Q12: **Are there any errors, sources of noise, or redundancies in the dataset?**

A12: Not applicable to the synthetically generated decisions. There may be sources of error associated with the BAF dataset used as input to generate the expert decisions. Please refer to BAF datasheet for more details `https://github.com/feedzai/bank-account-fraud/blob/main/documents/datasheet.pdf`.

Q13: **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?**

The generated expert predictions are self-contained. The input dataset is the publicly available BAF dataset `https://www.kaggle.com/datasets/sgpjesus/bank-account-fraud-dataset-neurips-2022`

Q14: **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?**

A14: There is no confidential data in this dataset.

Q15: **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?**

A15: No.

Q16: **Does the dataset relate to people?**

A16: This dataset relates to synthetically generated human experts, it does not relate to real people.

Q17: **Does the dataset identify any subpopulations (e.g., by age, gender)?**

A17: The synthetically generated dataset using the OpenL2D framework does not identify any subpopulations. The BAF dataset used as input dataset identifies age groups on the synthetically generated bank account applications. Please refer to BAF datasheet for more details `https://github.com/feedzai/bank-account-fraud/blob/main/documents/datasheet.pdf`.

Q18: **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**

A18: No, there is no information that allows the identification of individuals.

Q19: **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**

Our synthetically generated expert decisions do not relate to characteristics of specific individuals.

Collection Process

Q21: **How was the data associated with each instance acquired?**

A21: Our training dataset was obtained by associating an entity's decision with a given instance of the BAF dataset, respecting the generated work capacity constraints for each entity (See Section 3.3 of the paper). Our expert predictions were obtained by utilizing our synthetic expert decision generation method. The ML model's decision is represented by its score relating to the positive class (rejecting the application).

Q22: **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**

A22: The details of our synthetic expert generation method are available in Section 3.1 of the paper. These are generated by applying noise to the BAF dataset's label, with an instance-dependent noise approach.

The L2D training dataset was generated according to the previous answer.

Q23: **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

A23: The training dataset corresponds to instances from months 4 to 7 in the BAF dataset. This split is deterministic, and is done across months due to the temporal nature of the dataset.

Expert predictions are available for every instance in this dataset.

Q24: **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

A24: There was no data collection process. All the new data provided was synthetically generated.

Q25: **Over what timeframe was the data collected?**

A25: The generation of our synthetic expert decisions and generation of training scenarios is not time dependant.

Q26: **Were any ethical review processes conducted (e.g., by an institutional review board)?**

A26: No.

Q27: **Does the dataset relate to people?**

A27: This dataset relates to synthetic expert decisions and work capacity constraints.

Q28: **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

A28: Not applicable.

Q29: **Were the individuals in question notified about the data collection?**

A29: Not applicable..

Q30: **Did the individuals in question consent to the collection and use of their data?**

A30: Not applicable..

Q31: **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**

A31: Not applicable..

Q32: **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**

A32: No. The datasets are synthetic and should not be used to train fraud detection models or human-AI collaboration systems to be used in real-world fraud applications. The use of these datasets should be self-contained for L2D experimentation. Our synthetic expert predictions should not be a replacement for real human behaviour data.

Preprocessing/cleaning/labeling

Q33: **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**

A33: No.

Q34: **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?**

A34: Not applicable.

Q35: **Is the software used to preprocess/clean/label the instances available?**

A35: Not applicable.

Uses

Q36: **Has the dataset been used for any tasks already?**

A36: Our generated synthetic experts and training dataset have only been used for the experiments detailed in the paper.

Q37: **Is there a repository that links to any or all papers or systems that use the dataset?**

A37: There are still no applications of the presented datasets. We intend to keep track of its uses in the project GitHub repo [1].

Q38: **What (other) tasks could the dataset be used for?**

A38: These datasets should be used for the context of benchmarking L2D methods.

Q39: **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**

A39: Assuming the dataset is used exclusively for the evaluation or development of L2D algorithms, the composition of the dataset should not impact future uses.

Q40: **Are there tasks for which the dataset should not be used?**

A40: Using models trained in these datasets for real-world bank account opening fraud detection (or any other related application) directly should be avoided. The same applies to assignment systems trained in this dataset. The patterns and behaviours observed in these applications are highly dynamic and context-dependant, and using these models can result in unexpected low performances and biased decisions.

[1] `https://anonymous.4open.science/r/openl2d-7BD3`

Distribution

Q41: **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**

A41: Yes, the training dataset, expert prediction data, and the BAF base variant are all publicly accessible.

Q42: **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?**

A42: It will be distributed on GitHub.

Q43: **When will the dataset be distributed?**

A43: The suite is publicly available as of today on GitHub. It will be updated and code will be organized over the following days. There are no plans in removing the datasets from public usage.

Q44: **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?**

A44: The suite is licensed under the Creative Commons CC BY-NC-ND 4.0 license.

Q45: **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**

A45: No.

Q46: **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**

A46: No.

Maintenance

Q47: **Who is supporting/hosting/maintaining the dataset?**

A47: The dataset is supported and maintained by currently anonymous personnel.

Q48: **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

A48: The authors are currently anonymous.

Q49: **Is there an erratum?**

A49: No, there is no erratum as of yet. If necessary in the future, an erratum will be developed for the suite, as well as for this document.

Q50: **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**

A50: There are no current plans on updating the current version of the datasets. This can change in the future, to correct any undetected bug in the generated datasets.

Q51: **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**

A51: There are no applicable retention limits of the data.

Q52: **Will older versions of the dataset continue to be supported/hosted/maintained?**

A52: Currently, there is only the initial version. If any updates are published, previous versions will be available.

Q53: **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

A53: There are no current mechanisms to contribute to the suite of datasets. Novel ideas and variants of the dataset should be submitted via email to the authors or as an issue on GitHub.

Author Statement

The authors confirm the data in the BAF suite is under the Creative Commons CC BY-NC-ND 4.0 license. The authors bear responsibility in case of violation of copyrights.