1. CLEANML DATASETS

Citation: This dataset [1] consists of titles of 5,005 publications from Google Scholar and DBLP. Given a publication title, the classification task is to determine whether the paper is related to Computer Science or not. This dataset contains duplicates.

EEG: This is a dataset[4] of 14,980 EEG recordings with 14 EEG attributes. The classification task is to predict whether the eye-state is closed or open. This dataset contains numerical outliers.

Marketing: This dataset [6] consists of 8,993 records about household income from a survey. Each record has 14 demographic attributes including sex, education, etc. The classification task is to predict if the annual household income is less than $25,000. This dataset contains missing values.

Movie: This dataset [10, 5] consists of 9,329 movie reviews, which we obtained by merging data from IMDB and TMDB datasets. Each record has seven attributes including title, language, score, etc. The classification task is to predict the genre of the movie (romance or comedy). It contains duplicates and inconsistent representations of languages.

Company: The original dataset [2] contains over 2.5 million records. We randomly sampled 5% records (128,889 records) from the original dataset. Each record has seven attributes including company name, country, city, etc. The classification task is to predict whether the public sentiment about a company is negative or not. This dataset contains inconsistent company names.

Restaurant: This dataset [7] contains 12,007 records about restaurants, which we obtained by merging data from the Yelp and Yellowpages datasets. Each record has 10 attributes including city, category, rating, etc. The classification task is to predict whether the price range of a restaurant is "$" or not. This dataset contains duplicates and inconsistent restaurant names and categories.

Titanic: This dataset [9] contains 891 records and 11 attributes from the Titanic including name, sex, etc. The task is to determine whether the passenger survived or not. This dataset has a significant number of missing values.

Credit: This dataset [3] consists of 150,000 credit records with 10 attributes including monthly income, age, then number of dependents, etc. The classification task is to predict whether a client will experience financial distress in the next two years. This dataset has a class imbalance problem with only 6.7% records in the minority class. We follow standard procedure to over-sample the minority and down-sample the majority class before training, and we use F1 score as the performance metric for evaluation. This dataset contains missing values and numerical outliers.

Sensor: The original [8] sensor dataset contains 928,991 sensor recordings with eight attributes including temperature, humidity, light, etc. We only used recordings from sensor 1 and sensor 2 and sampled the dataset to include 1 observation per hour for each sensor. The sampled dataset contains 62,076 records. The classification task is to predict whether the readings came from a particular sensor (sensor 1 or sensor 2). This dataset contains outliers.

University: This dataset [11] contains 286 records about universities. Each record has 17 attributes including state, university name, SAT scores, etc. The classification task is to predict whether the expenses are greater than $7,000$ for each university. This dataset contains inconsistent representations for states and locations.

USCensus: This dataset [12] contains 32,561 US Census records for adults. Each record has 14 attributes including age, education, sex, etc. The classification goal is to predict whether the adult earns more than $50,000. This dataset contains missing values.

Airbnb: This is our own dataset with 42,492 records on hotels in the top 10 tourist destinations and major US metropolitan areas, scraped from Airbnb.com. Each record has 40 attributes, including the number of bedrooms, price, location, etc. Demographic and economic attributes were scraped from city-data.com. The classification task is to determine whether the rating of each hotel is 5 or not. This dataset contains missing values, numerical outliers, and duplicates. We will release this dataset in the code repository.

BabyProduct: The dataset [13] contains 10,718 records on baby products of different categories (e.g., bedding, strollers). There are many missing values on the *brand* attribute. We randomly select a subset of categories with 4,019 records and we designed a classification task to predict whether a given baby product has a high price or low price based on other attributes (e.g. weight, brand, dimension, etc). For records with missing brand attribute, we also perform a Google search using the product title to obtain the product brand.

Clothing: This dataset contains 5,000 records which is down-sampled from the original dataset [14]. The original dataset is an image dataset about clothing with 14 different classes. There are many real mislabel errors in the dataset and the ground truth label are provided by human cleaning. To train our models on the image data, we use a pretrained ResNet-18 model as a feature extractor and extract 512 features for each image as the training data.

2. FEATURE ERROR RATIOS

We report the error ratios in the corresponding attributes for each error type as follows in Tables 1, 2, 3 and 4.

For inconsistencies, outliers and duplicates, we do not have ground truth. Those numbers are the results from various detection algorithms. We also calculate the average error prevalence for each dataset. It is obvious that the error prevalence varies largely by detection methods.

Table 1: Percentage of Mislabels in Datasets

dataset	Flipping Labels
Clothing	36.00%
Credit_major	4.65%
Credit_minor	0.35%
Credit_uniform	16.55%
EEG_major	2.76%
EEG_minor	2.24%
EEG_uniform	5.00%
Marketing_major	2.88%
Marketing_minor	2.11%
Marketing_uniform	4.99%
Titanic_major	3.28%
Titanic_minor	1.64%
Titanic_uniform	4.92%
USCensus_major	3.76%
USCensus_minor	1.24%
USCensus_uniform	5.00%

Table 2: Percentage of Missing Values in Datasets

dataset	Empty Entries
Airbnb	20.69%
BabyProduct	83.55%
Credit	19.82%
Marketing	23.54%
Titanic	79.46%
USCensus	7.37%

Table 3: Percentage of Inconsistencies in Datasets

dataset	OpenRefine
Company	50.64%
Movie	51.15%
Restaurant	5.21%
University	16.67%

Table 4: Percentage of Outliers in Datasets

dataset	SD	IQR	IF	avg
Airbnb	32.58%	66.48%	15.99%	38.35%
Credit	10.46%	33.26%	7.23%	16.98%
EEG	0.59%	20.87%	4.57%	8.68 %
Sensor	0.05%	19.92%	4.27%	8.08 %

Table 5: Percentage of Duplicates in Datasets

dataset	Key Collision	AutoER	avg
Airbnb	13.22%	2.59%	7.91%
Citation	18.91%	5.21%	12.06 %
Movie	45.51%	16.42%	30.97%
Restaurant	11.32%	6.94%	9.13 %

3. REFERENCES

[1] Citation dataset. `https://sites.google.com/site/anhaidgroup/useful-stuff/data`. Accessed: September 11, 2020.

[2] Company dataset. `https://www.kaggle.com/jacksapper/company-sentiment-by-location`. Accessed: September 11, 2020.

[3] Credit dataset. `https://www.kaggle.com/c/GiveMeSomeCredit/data`. Accessed: September 11, 2020.

[4] EEG dataset. `https://archive.ics.uci.edu/ml/datasets/EEG+Eye+State`. Accessed: September 11, 2020.

[5] IMDB movie dataset. `https://data.world/popculture/imdb-5000-movie-dataset`. Accessed: September 11, 2020.

[6] Marketing dataset. `https://sites.google.com/site/anhaidgroup/useful-stuff/data`. Accessed: September 11, 2020.

[7] Restaurant dataset. `https://sites.google.com/site/anhaidgroup/useful-stuff/data`. Accessed: September 11, 2020.

[8] Sensor dataset. `http://db.csail.mit.edu/labdata/labdata.html`. Accessed: September 11, 2020.

[9] Titanic dataset. `https://www.kaggle.com/upendr/titanic-machine-learning-from-disaster/data`. Accessed: September 11, 2020.

[10] TMDB movie dataset. `https://www.kaggle.com/tmdb/tmdb-movie-metadata`. Accessed: September 11, 2020.

[11] University dataset. `https://archive.ics.uci.edu/ml/datasets/University`. Accessed: September 11, 2020.

[12] USCensus dataset. `https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29`. Accessed: September 11, 2020.

[13] S. Das, A. Doan, P. S. G. C., C. Gokhale, P. Konda, Y. Govind, and D. Paulsen. The magellan data repository. `https://sites.google.com/site/anhaidgroup/projects/data`.

[14] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.