The ML Data Prep Zoo: Towards Semi-Automatic Data Preparation for ML

Vraj Shah Arun Kumar
University of California, San Diego
{vps002,arunkk}@eng.ucsd.edu

ABSTRACT

Data preparation (prep) time is a major bottleneck for many ML applications. It is often painful grunt work that is handled manually by data scientists, reducing their productivity and raising costs. It is also a roadblock for emerging AutoML platforms. We envision a new line of community-driven research to tackle this bottleneck based on a simple philosophy: *use ML to semi-automate data prep for ML*. For impactful research on this problem, we believe the major impediment is not new algorithms or theory but rather common task definitions and benchmark labeled datasets. To this end, we formalize a few major data prep tasks for ML over structured data as applied ML tasks. We discuss research challenges in scaling up data labeling, defining accuracy metrics, and creating practical tool support. We present a case study of our progress on a key data prep task: ML schema inference. Finally, we propose a public "zoo" of labeled datasets and pre-trained ML models for data prep tasks to act as a community-led repository for further research on this problem.

1 INTRODUCTION

Surveys of data scientists show that ML data prep often dominates their time and effort, even up to 80% [11]. It is tedious grunt work involving tasks such as identifying feature types and extracting feature values. Today, it is performed mostly manually in tools like Python and R, reducing data scientists' productivity and raising costs. Modern datasets also often have 1000s of columns, worseing this issue. Furthermore, SalesForce, Google, and other cloud vendors are starting to offer end-to-end AutoML platforms for enterprises; manual data prep at this scale of millions of datasets is untenable [12].

Challenge: Semantic Gap. While the DB community has long studied data cleaning/prep for SQL analytics, little work has studied the peculiarities of ML data prep. The *semantic gap* between what an *attribute* is in a DB/data file and what a *feature* is for ML means many tasks have fallen through the cracks. Thus, a pressing grand challenge for the DEEM community is to construct a shared understanding/terminology of such tasks, understand why they are hard to automate, and standardize evaluation of (semi-)automated tools.

Our Vision. To meet the above challenge, we envision a community-driven effort for semi-automating ML data prep. Our philosophy is to *abstract specific ML data prep tasks and cast them as applied ML tasks*. This raises 3 questions. What are the tasks and what is their role? How to cast them as applied ML tasks? How to create benchmark datasets for comparing tools? In particular, we believe the critical limiting factor for impactful and replicable research in this space is not fancier algorithms or theory but the *availability of large high-quality labeled datasets* for ML data prep tasks. As an analogy, the formalization of the ImageNet task and dataset spurred major recent advances in ML-based vision.

This Paper. We present our vision of the *ML Data Prep Zoo*, a repository of common ML data prep task definitions, benchmark labeled datasets, and pre-trained ML models. Figure 2 illustrates 6 tasks we have defined so far based on our conversations with data scientists. In Section 2, we explain these tasks and how to cast them as applied ML tasks. In Section 3, we discuss key research questions in realizing this vision and explain our plans. In Section 4, we present a case study of our progress on the first task: *ML schema inference*. For instance, our labeled data-based applied ML approach yielded a whopping 30% *lift* for identifying numeric features among attributes compared to existing rule-based approaches in Python Pandas and TensorFlow DataValidation [2]. In Section 5, we describe the ML Data Prep Zoo repository for our datasets and models and announce competitions for community contributions. Section 6 discusses related work.

2 DATA PREP TASKS FOR ML

2.1 Current Scope

Our current focus is on relational/tabular data, the most commonly analyzed form of data in practice [11]. Such datasets are typically stored with DB schemas in RDBMSs or data warehouses or as "schema-light" files (CSV, JSON, etc.) on data lakes and filesystems. Either way, we assume the dataset is a single table with all column names available. To build ML models on such data, the first thing most data scientists do is to load it into a Python or R "dataframe." This is when the laborious process of preparing this dataframe for an ML training library (e.g., Scikit-learn) begins. This stage is the focus of our work. We leave other scenarios of procuring

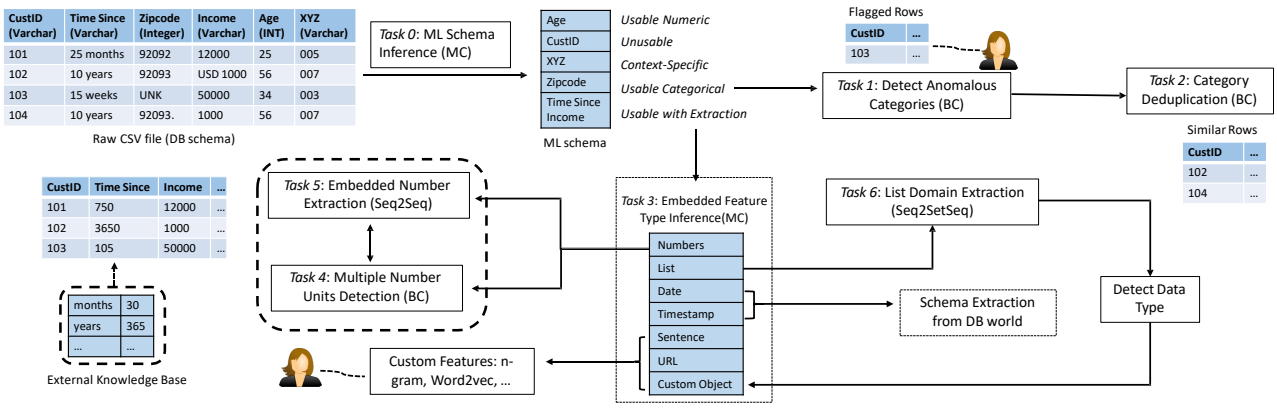

Figure 1: Illustrating major data prep tasks. The user loads a customers table to train, say, a churn predictor. BC stands for binary classification. MC stands for multi-class classification. Seq2Seq stands for sequence-to-sequence learning. Seq2SetSeq stands for sequence-to-set-of-sequence learning.

or transforming data for ML to future work. Note that our focus is *not* on feature engineering over prepared data (e.g., binning, outlier detection, one-hot encoding, etc.). Also, to avoid ambiguity, we call the ML model(s) to be trained on the prepared data the "target model(s)." For example, one might load a customer table to train a target model for predicting customer churn.

Figure 2 shows a typical data prep workflow on a dataframe before feature engineering or target model training begins. We dive into a few major steps in this laborious workflow to abstract its logic, explain what the human intuition's offers, and discuss how we could cast it as an applied ML task.

2.2 Task 0: ML Schema Inference

Description and Example. The very first step is to infer the "ML schema" from the DB schema. This task today requires careful human attention to every single column. It is hard to automate because the DB schema is *syntactic*–it tells us the datatype of a column, e.g., integer or string (say, VARCHAR), while the ML schema is *semantic*–it tells us what type of a feature a column is. Almost all ML models recognize only two types of features: *categorical* (a discrete set) and *numeric* (a continuous set) [7]. Alas, the semantic gap between DB and ML schemas means reading syntax as semantics often leads to nonsensical results. For example, consider the `Zipcode` column in Figure 2. It is usually stored as integers, e.g., `92093`. A tool like Python Pandas will thus treat it as a numeric feature. Thus, the human has to manually convert it to a categorical feature for the target model. This issue is ubiquitous in real-world dataset, since many applications encode categories as integers (e.g., disease code, product type, etc.).

Casting as an ML Task. Real-world datasets often have hundreds to thousands of columns. Asking a data scientist to spend even 1min to infer the feature type of a column could easily lead to hours, if not days, of pure grunt work! Thus, we cast this task as an ML classification task to leverage the ability of ML models to bridge the semantic gap. The input to such a classifier is the entire column, including its name, e.g., the whole `Zipcode` column in our example. But as we find, there may not be enough information in just the data file to predict the class correctly. Thus, we need more classes. We have made substantial progress on this front and achieved much higher prediction accuracy for ML schema inference than prior rule-based and syntax-based approaches in Pandas and TensorFlow's Data Validation tool [2]. We describe the prediction vocabulary of our ML task. We defer other details of how we tackled this problem to our case study in Section4.

Prediction Vocabulary. We have 5 classes as shown below.

Usable-Numeric (resp. *Usable-Categorical*): This class is for columns that are directly usable as numeric (resp. categorical) features without (almost) no modifications. For example, `Age` in Figure 2 is *Usable-Numeric*, while *Zipcode* is clearly *Usable-Categorical*.

Usable-with-Extraction: This class is for columns whose syntax is "messy" and preclude direct use as categorical or numeric features. For example, `Income` and `TimeSince` in Figure 2 require some custom processing before they can be used as numeric features. Although such processing is hard to automate fully but later on, we explain a few recurring tasks with such columns that we can cast as ML tasks to reduce human grunt work.

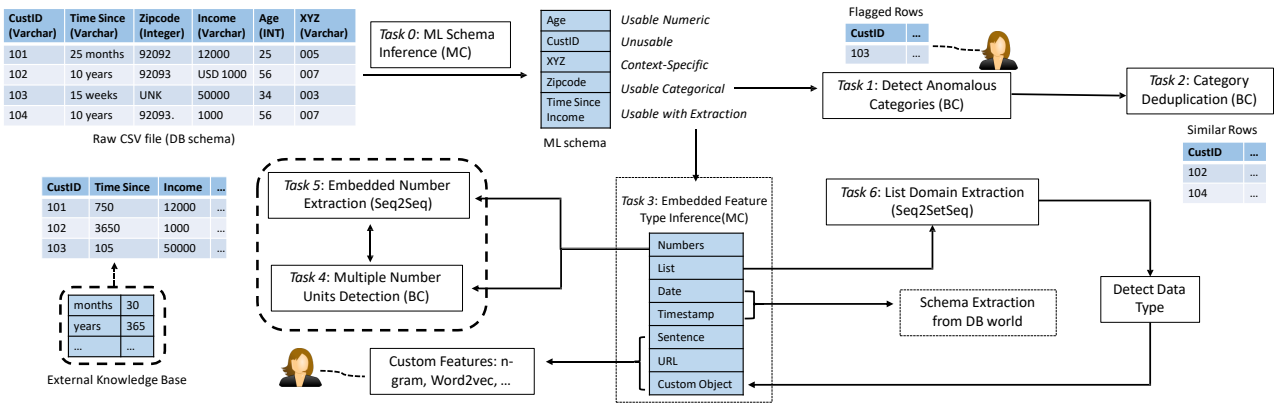

Figure 2: Data prep tasks. BC represents binary classification problem, MC represents multi-class classification problem, Seq2Seq represents sequence to sequence learning problem, and Seq2Set represents sequence to set of sequence learning problem.

Unusable: This class is for attributes that cannot be used as feature for the target model because they are not "generalizable." For example, `CustID` in Figure 2 belong to this class, since every future customer will have a new `CustID`.

Context-Specific: This class is for columns wherein the data file has not enough information even for a human to judge its feature type. For example, `XYZ` in Figure 2 has integer values but is it categorical (like `Zipcode`) or numeric (like `Age`)? Clearly, answering this question would require manually tracing down the provenance of how this column came to be using external "data dictionaries" maintained by the application or speaking to the data creator.

2.3 Tasks for Usable-Categorical

While a *Usable-Categorical* column can be used directly, data scientists often seek to resolve two issues with its domain to boost target model accuracy: *missing value categories* and *duplicate categories*. For instance, we saw both"-999" and "unknown" for missing values and both "CA" and "California" for California in real datasets. One may want to discard missing value categories and instead use statistical techniques for handling missing values. One may also want to deduplicate categories to reduce domain size, which helps in the bias-variance tradeoff. Thus, we formalize two new data prep tasks as binary classification: **Task 1: Detect Anomalous Categories** to flag missing value categories and **Task 2: Category Deduplication** to flag pairs of categories that are duplicates. The column name and its domain are the raw features for both tasks. Task 2 is an instance of the entity matching problem in the data cleaning literature but with much less metadata for devising similarity scores; one could consider Siamese neural networks for this task.

Task 1: Detect Anomalous Categories The data is typically acquired from a combination of automated sources and manual data entry, thus, leading to inconsistency in the values. For example, in many real-world datasets, we observed values such as "-999" and "unknown" to represent missing data. Such values should not be interpreted semantically, but rather flagged so that a data scientist can decide upon how to handle missing values. We cast this problem as binary classification problem, where the learned model predicts if the categories is anomalous or not.

Task 2: Category Deduplication Data obtained from raw heterogeneous sources often contains duplicate categories referring to the same real world entity. For instance, Figure 2 shows the raw csv file containing duplicate categories in `ZipCode` column. Existing literature in entity deduplication relies on selecting a suitable similarity metric upon manual inspection of the underlying data. For example, similarity metric for names would be chosen differently than one for timestamp.

Instead of hand-tuning a similarity metric for every domain, we propose to derive a learned similarity metric adaptable to different domains. We cast category deduplication as ML binary classification problem, where the model learns a similarity metric to identify if a given category is duplicated or not. We plan to explore deep net architectures like Siamese Neural Network, that has recently been shown effective in identifying similarity between images [6] and text [8].

2.4 Tasks for Usable-with-Extraction

Usable-with-Extraction columns require more processing to extract numeric and/or categorical features, e.g., `Income`. Figure 2 has "USD" prefixing a number, while `TimeSince` has "months," "years," etc. suffixing numbers. Data scientists often write regular expressions or custom code to extract such values. While it is perhaps impossible to automate all such extractions, we identify three common tasks that can be cast as applied ML tasks.

Task 3: Embedded Feature Type Inference: What is the feature type embedded? Figure 2 shows our current taxonomy for embedded feature types. Dates and timestamps can be processed using standard DB techniques, while URLs and custom objects may require human intervention. One could consider character-level CNNs and RNNs for this task.

Task 4: Multiple Number Units Detection: Are the units of an embedded number the same? If not, we need to standardize the units, likely with human intervention and/or external knowledge bases about units. In Figure 2, `TimeSince` has multiple units. If yes, we get **Task 5: Embedded Number Extraction**: What is the embedded number? For instance, extract 1000 from "USD 1000." This can be seen as both a Seq2Seq task and a sequence-to-regression task. An encoder-decoder CNN/RNN may fit this task. One could also consider joint multi-task learning for Tasks 4 and 5.

Task 6: List Domain Extraction: Some columns have lists in a string separated by commas, space, semicolons, etc. Data scientists typically write custom code to *extract the domain* of the list values and use the domain to get new numeric/categorical features for the target model. This is a complex task that converts a sequence to a *set of sequences* representing domain entries. Once could consider more complex neural architectures for this task.

Other Featurization Routines. In our current scope, we leave other standard featurization routines for custom processing to the user. For instance, to process a full English sentence in a *Usable-with-Extraction* column, data scientists may want to use bag-of-words, n-grams, or embeddings like Word2Vec or Doc2Vec. Such feature engineering decisions are orthogonal to our focus and are often application-specific.

3 RESEARCH QUESTIONS AND OPTIONS

We now discuss a few major research questions in tackling the applied ML tasks we listed.

3.1 Metrics and Featurization

The accuracy metrics for Tasks 0 to 4 are standard, but for Tasks 5 and 6, we may need to define new metrics. For Task 5, edit distance and/or squared loss are candidates with differing results, e.g., "12" is closer to "$12.99" under the latter but not the former although edit distance helps sequence extractors. Task 6 has a complex structured prediction output, which may need a complex loss function (ideally, still differentiable) and multiple accuracy metrics. Even the featurization of the raw column is an open question, since the ML models for our tasks also need numeric, categorical, or string features. Several options exist: obtain n-grams or embeddings from column names and sample values, get summary statistics, and so on. Characterizing which of these features matter the most is also part of our research, since such featurization matters for both accuracy and inference latency at deployment time.

3.2 Creating Large Labeled Datasets

This is our central research challenge. To the best of our knowledge, there are no large benchmark labeled datasets for any of our 6 data prep tasks. So far, we have collected 360 CSV data files from Kaggle, UCI repository, etc., adding up to 9000 columns. Manual labeling for each task each could yield best accuracy but it is highly time-consuming and expensive. We plan to try 3 alternative approaches: crowdsourcing, active learning, and weak supervision.

Crowdsourcing labels is common in ML practice, but we face a major quality issue: most crowd workers are lay users, not data scientists who "get" data prep. In fact, our pilot run for crowdsourcing labels for Task 0 on the FigureEight platform resulted in too much noise even with 5 labels per example. Thus, how to structure crowd labeling questions better is an open research question. Active learning with a data scientist in the loop is another option we plan to explore. But a key disadvantage here is that we need to fix the task's ML model beforehand. Finally, weak supervision is a promising approach here, since it is often possible to write small labeling functions (LFs) to encode structural

heuristics and dictionary lookups for some tasks. A denoising framework like Snorkel [9] can potentially help boost accuracy over the LFs' outputs. We also plan to try Snuba-on-Snorkel [14] to automate the production of LFs for some classification tasks. But an open challenge is that Snorkel current does not support complex prediction outputs like in Tasks 5 and 6.

3.3 Creating Human-in-the-loop Tools

Our work cannot end at getting ML models for our tasks. To complete the loop,we need to integrate them for inference in popular data prep ecosystems. There are two main kinds of tools: programmatic (e.g., R, Pandas, and TFDV) and visual (e.g., Excel and Trifacta). Each presents its own set of interesting implementation challenges. For the former, we plan to introduce simple APIs to plug our trained ML models. For the latter, it is an open research question as to how to create appropriate interface mechanisms that can exploit both our ML models' predictions and human-in-the-loop correction capabilities. For instance, the user could "guide" an ensemble of ML models based on column semantics or specific column values they see. Looking even further out, we can integrate ML models with programming-by-example and program synthesis approaches, especially for Tasks 5 and 6 that require value extraction. This requires resolving ambiguity in the program search space and defining new ranking schemes aimed at reducing manual extraction effort.

4 CASE STUDY: ML SCHEMA INFERENCE

Data Labeling. We obtained over 360 real datasets as CSV files from several sources such as Kaggle, UCI ML repository and our prior work [10]. Each attribute (or column) of the csv file is just one example for our ML task. We manually labeled 9000 columns from the data files we collected into the one of five classes of Task 0. This process took about 10 man-weeks across 4 months.

Featurization. Based on the two raw features (column name and values), we extract several hand-crafted features to train classical ML models. Our feature set is diverse: n-grams from column name, summary statistics (mean, %ge NaNs, etc.), castability as number, length of a random sample value, etc.

Experimental Setup. We perform 5-fold nested CV with a random quarter of the train fold used for hyperparameter tuning. We compare logistic regression and RandomForest trained on our data against TF-DV and Pandas. TF-DV can infer only 2 types of features in our vocabulary: numeric or otherwise. Pandas can only infer syntactic types: int, float and object. Thus, we report the results on a binarization of our 5-class vocabulary: numeric (Num) and all non-numeric (N-Num). Figure 3 presents the test accuracy results.

	TF-DV		Pandas		LogReg		RandForest	
	Num	N-Num	Num	N-Num	Num	N-Num	Num	N-Num
Precision	0.5117	0.9876	0.5418	0.9382	0.9331	0.9450	0.9722	0.9360
Recall	0.9915	0.4166	0.9502	0.4849	0.9093	0.9598	0.8909	0.9843
Accuracy	0.6359		0.6667		0.9394		0.9508	

Figure 3: Test Accuracy Results.

Initial Results. We see a massive lift of 30% in accuracy for our approach against both TF-DV and Pandas. Interestingly, TF-DV and Pandas have high recall on numeric features but very low precision. This is because their rule-based heuristics are syntactic, leading them to wrongly classify many categorical features such as `ZipCode` as numeric. Our models have slightly lower recall on numeric features but much higher precision and overall accuracy.

5 THE ML DATA PREP ZOO

We announce the ML Data Prep Zoo, a living public repository (on GitHub) of labeled data for ML data prep tasks [1]. We will release all datasets we create as CSV files. We will also release our trained ML models in Python for the defined tasks. Our first release will be for Task 0, with the base features being the column name, summary statistics, and 5 random sample values. Our trained models will include logistic regression, RandomForest, kernel SVM, and a character-level CNN. The Zoo will also tabulate the accuracy of the baselines and our models on each task. We invite contributions from the research community to augment these datasets, create new data for the other tasks, and/or define new tasks along with their own labeled data and models. We also plan to have leaderboards for public competitions on the hosted datasets with multiple accuracy and runtime metrics, inspired by the ImageNet competition. We invite researchers to use our data to create better featurization and models to semi-automate ML data prep tasks.

6 RELATED WORK

Data Prep and Cleaning. TFDV [2] is a tool for managing ML-related data in TensorFlow Extended. It uses

conservative rule-based heuristics to infer ML schema from column statistics. Our ML-based approach raises accuracy of ML schema inference substantially. DataLinter is a rule-based tool to inspect a data file and flag possible data quality issues to the user [3]. It still requires users to perform data transformations manually, which makes it orthogonal to our focus. There is growing work on reducing data cleaning effort using ML properties (e.g., [5]). Our work is part of this growing direction but our work specifically targets data prep tasks and casts them as applied ML tasks.

AutoML Platforms. Existing AutoML platforms such as Einstein [12] and AutoWeka [4] focus mainly on model selection, not ML-based ML data prep. Thus, the models we produce can enhance such platforms. OpenML [13] is an open-source platform for ML users to share and compare models, data, and analysis workflows. Our focus on *creating* high-quality labeled datasets for semi-automating ML data prep tasks is thus complementary. Our artifacts can be contributed to OpenML for spurring more research on end-to-end AutoML platforms. We could also get more analysis workflows from OpenML to enhance our work in the future.

REFERENCES

[1] Accessed March 15, 2018. The ML Data Prep Zoo Repository. https://github.com/pvn25/ML-Data-Prep-Zoo.

[2] Denis Baylor et al. 2017. Tfx: A tensorflow-based production-scale machine learning platform. In *SIGKDD*.

[3] Nick Hynes et al. 2017. The data linter: Lightweight, automated sanity checking for ml data sets. In *NIPS MLSys Workshop*.

[4] Lars Kotthoff et al. 2017. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *JMLR* (2017).

[5] Sanjay Krishnan et al. 2016. Activeclean: An interactive data cleaning framework for modern machine learning. In *SIGMOD*.

[6] Iaroslav Melekhov et al. 2016. Siamese network features for image matching. In *ICPR*.

[7] Tom Mitchell et al. 1990. Machine learning. *Annual review of computer science* (1990).

[8] Paul Neculoiu et al. 2016. Learning text similarity with siamese recurrent networks. In *RepL4NLP*.

[9] Alexander Ratner et al. 2017. Snorkel: Rapid training data creation with weak supervision. *PVLDB* (2017).

[10] Vraj Shah, Arun Kumar, and Xiaojin Zhu. 2017. Are key-foreign key joins safe to avoid when learning high-capacity classifiers? *Proceedings of the VLDB Endowment* 11, 3 (2017), 366–379.

[11] https://www.kaggle.com/surveys/2017. Accessed February 15, 2018. 2017 Kaggle survey on data science.

[12] https://www.salesforce.com/video/1776007. Accessed February 15, 2018. SalesForce Einstein AutoML.

[13] Joaquin Vanschoren et al. 2014. OpenML: networked science in machine learning. *ACM SIGKDD Explorations Newsletter* (2014).

[14] Paroma Varma et al. 2018. Snuba: automating weak supervision to label training data. *PVLDB* (2018).