Governance and Regulations Implications on Machine Learning

Sima Nadler, Orna Raz, and Marcel Zalmanovici

IBM Research, Haifa
sima@il.ibm.com, ornar@il.ibm.com, marcel@il.ibm.com

Abstract. Machine learning systems' efficacy are highly dependent on the data on which they are trained and the data they receive during production. However, current data governance policies and privacy laws dictate when and how personal and other sensitive data may be used. This affects the amount and quality of personal data included for training, potentially introducing bias and other inaccuracies into the model. Today's mechanisms do not provide a way (a) for the model developer to know about this nor, (b) to alleviate the bias. In this paper we will show how we address both of these challenges.

Keywords: Data governance · implications · privacy · Machine learning

1 Introduction

More and more of today's computer systems include some kind of machine learning (ML), making them highly dependent on quality training data in order to train and test the model. The ML results are only as good as the data on which they were trained and the data they receive during production. On the other hand, data governance laws and policies dictate when and how personal and other sensitive data may be used. For some purposes it may not be used at all, for others consent is required, and in others it may be used based on contract or legitimate business. In the cases where personal data or other sensitive information cannot be used, or requires consent, the data sets used to train ML models will by definition be a subset of the data. If the data set doesn't include, for example, age, race, or gender, then there is no way to know that the data is not representative of the real target population. This has the potential to introduce bias into the model as well as other inaccuracies — without the solution creator having any idea of the potential problem. Data sets are sometimes augmented with meta data describing what is included in the data set, but currently that meta data has nothing about what has been excluded and why. There are no current methods for alleviating governance induced bias in ML models. The main contributions of this paper are:

1. Defining the potential implications of governance laws and policies as they pertain to machine learning based on governed data.
2. Demonstrating the encoding of governance implications via a governance enforcement engine as meta data that can be utilized for further implication analysis.
3. A set of techniques for governance implications impact analysis and the experimental feasibility demonstration of the first of these techniques, using US government Census data.

This line of research raises many challenges. We show a solution for one of these challenges related to identifying potential bias in an ML model. Other challenges will be addressed in future work.

We provide a short background on governance laws and policies and their potential impact on machine learning. Countries and industries have differing laws regulating how data, especially personal and other types of sensitive data, may be used. Europe's new General Data Protection Regulation (GDPR) went into effect in May 2018. It aims to strengthen data subject privacy protection, unify the data regulations across the member states, and broaden the territorial scope of the companies that fall under its jurisdiction to address non-EU companies who provide services to EU residents. In the United States privacy laws have typically been industry based with, for example, HIPAA governing the health care industry as it relates to digital health data. However, in June 2018 the state of California passed a new privacy law that will go into effect in 2020. Since each state could theoretically do the same, the US could find itself with 50 different privacy laws. As a result there are now discussions about creating a federal privacy law in the United States. Similar trends can be seen in other

countries around the world. While the laws and standards differ, they tend to be similar in their goals of (1) ensuring transparency about what personal data is collected and/or processed and for what purpose, (2) providing more control to data subjects about the purposes for which their personal data may be used, (3) enabling the data subject to receive, repair, request the deletion of personal data in some situations, and (4) data minimization. There are of course situations where such control is not in the control of the data subject, such as when the data must be retained for contractual or legal purposes, for public benefit, etc.

When creating a machine learning system, the goal which it aims to achieve is in essence the purpose. If personal, or other regulated, data is needed to train and/or use the ML system, either all or a subset of the original data will be made available based on the purpose. Likely this will vary from country to country and/or state to state, based on local regulations and company regulations. For example, if one is creating a model to predict the need for public transportation in a given neighborhood one could use information from current public transportation use, population size, and other relevant data. However, there may be laws restricting the use, for example, of location data and transportation use of minors. Thus, the training set for the ML-based solution for transportation planning would not include data about people under the age of 18. This introduces bias into the model, since the transportation patterns of children would not be included. There are many other well known fields where such bias is introduced. When the bias introduction is known, it can be accounted for and corrected. A well known example is the pharmaceutical industry, where pregnant women and children are rarely included in clinical trials for drug development. Another example, in which bias was not known in advance, is automated resume review systems (e.g., [7]), where the populations currently employed are the ones for which the machine learning system naturally is biased.

In this paper we propose to alleviate governance induced bias in ML models by first capturing, and providing as meta data by a governance enforcement engine, information about what has been excluded and why. Then, such information can be used for identifying and alleviating governance implications on ML models. Section 2 provides a summary of related work. Section 3 details the meta data we propose, as well the impact analyses that may utilize it. Section 4 provides examples of extracting meta data and of utilizing it to identify governance implications on a ML model trained on US Census data. Section 5 summarizes our findings and discusses future work.

2 Related work

We are unaware of any work that addresses the issues of capturing information about data excluded due to governance regulations and policies, work that utilizes such information to characterize their impact on machine learning, nor work that suggests to utilize the impact analysis to improve the machine learning models.

There is a lot of work about capturing data governance and privacy policies, proper management of consent, and identification and handling of bias in ML. There are also various data governance tools available. To provide background and context to our work, we summarize some of the relevant papers and tools here. As our work captures governance implications as meta data, we also include work that addresses providing meta data for data sets and utilizing it for learning.

Ethics and Data Science As concerns have arisen regarding the ethics of the use of ML and other data science techniques, books such as [13] provide ethical guidelines for the development of ML based systems. They emphasize "The Five Cs": consent, clarity, consistency and trust, control and transparency, and consequences. A lot of emphasis is put on obtaining truly informed consent for use of personal data.

Princeton's Dialog on AI and Ethics [10] presents five case studies from different fields such as healthcare, education, law enforcement and hiring practices in which issues such as transparency, consent, data quality and how it introduces bias, and paternalism are discussed.

Ethics to Correct Machine Learning Bias There is also work about using ethics to correct ML bias. For example [17]. There are three primary ways that ethics can be used to mitigate negative unfairness in algorithmic programming: technical, political, and social. One interesting approach to prevent bias is Counterfactual Fairness [12].

The problem of exclusion of sensitive data under GDPR is introduced in [18], but no solution is discussed.

Privacy's Influence on Decision Making There is work assessing the value of privacy and how it affects people's decisions [8, 3]. Other work concentrates on social media data and how it may be utilized, mainly to improve health and well being. In [6] they highlight the cultural and gender differences with regard to willingness to disclose very sensitive information about their mental health as they vent and/or look for support via the social network.

Metadata Providing metadata about data sets used for training machine learning models is a well known practice. Information is often provided about the data set size, the type and format of the data, its source, and the organization providing the data. Such metadata is sometime a simple text [14], but can also be in JSON or XML format making it more machine readable [4].

The authors in [11] discuss the importance of metadata and its incorporation into the building of machine learning models. They conclude by highlighting the importance of using features from both the metadata and the data in building the machine learning models to increase the efficacy of the models.

Recognizing the challenges associated with sharing sensitive data, the Egeria open source metadata project [5] tackles the problem of how to discover appropriate data sets, and how to share information about data sets to make them easier to discover. Its focus on metadata is very relevant, but no mention is made about including information about data excluded from a data set, only information describing what is in it, its structure, source and how to sync the metadata with its main source.

Data governance tools The handling of personal and other sensitive data is a complex task. First, such data needs to be identified and cataloged by its nature, level of sensitivity, and where it is stored and processed. Tools such as IBM Guardium, IBM StoredIQ, BigID Discovery tools do that. Catalogs such as Infosphere Governance Catalog can help manage the location and metadata associated with the data stored by the enterprise, and even policies associated with it. Then the enterprise must also identify for what purposes the data is used and the legal basis for using it. Tealium and TrustArc are examples of tools for managing consent for digital marketing.

The hard part comes when sensitive and personal data is accessed and the policies and consent associated with it must be enforced. Apache Ranger and Atlas are examples of open source projects that address data governance, but lack the ability to do data subject level enforcement. IBM's Data Policy and Consent Management (DPCM) tool supports the modeling of purposes and policies, the collection and management of data subject consent, and the enforcement of them all when data is accessed. It also logs all governance decisions. This is the tool that we use in our experiments, as described in Section 4.

3 Method

We define governance implications and suggest to implement them as meta data to be added to the output of governance enforcement tools. Section 3.1 details the governance implications data and how it can be extracted. There are two major types of excluded data: excluded records and excluded features. These types differ in terms of how they can be identified and alleviated. Section 3.2 discusses approaches that we suggest for handling excluded records. Section 3.3 discusses approaches that we suggest for handling missing features. Section 4 presents experimental results of following the methods we describe here.

3.1 Generating a Data Governance Impact Summary

The role of data governance is to enforce proper usage of personal and/or sensitive data as defined by policies, and data subject preferences. As data is accessed, stored, or transferred the governance module is responsible for invoking the governance policies on the data. Such function might be to filter out certain data, obfuscate the data, or to allow the data to be used as is. While doing this the governance module logs what it has done and on what the decision was based.

Apache Ranger [2] is an open source example of such a governance module. It logs its governance decisions to what it calls an access log, containing the information that Figure 1 depicts.

Search Criteria	Description
Access Enforcer	Access enforcer indicates who made the decision to allow or deny. In case of HDFS, the enforcer would XA (Ranger) or Hadoop.
Access Type	Type of access user has for e.g read,write
Start date,End date	Time and date is stored for each access. A date range is used to filter the results for that particular date range.
Service Name	The name of the service which the user tries to access
Service Type	The type of the service which the user tries to access
Result	This shows whether the operation was successfull or not
User	Name of the user which tried to access the resource
Client ip	Ip address of the user system which tried to access the resource

Fig. 1: Apache Ranger governance log data.

Similarly IBM Guardium [9] provides governance over many different types of data stores, both databases and files. As it enforces the policies, it logs the access decisions to a single, secure centralized audit repository from which compliance reports may be created and distributed. This may be used as the starting point for the creation of the data governance impact summary.

However, few if any commercial data governance solutions include enforcement of both policies and data subject level preferences. IBM Research's Data Policy and Consent Management (DPCM) does provide this capability. It stores the policies and laws as well as the preferences of the data subjects indicating

their willingness or lack thereof for their personal data to be used for different purposes. In our experiment (Section 4.2) we generate the data governance impact summary using DPCM.

To create the governance implication summary we parse the governance decisions log, and generate a summary containing: Original vs derived data set size, list of features removed from the data set and the removal reasons (policies), percentage of data subjects included in the derived data set, and affect on features included in the derived data set — ex: x% of people over age 60, y% of people from California.

The governance impact summary then provides important additional information which is taken into account when building and running machine learning models, as described in the following sections.

3.2 Governance Impact on Features Included in Derived Data

It is important to understand whether the excluded data records introduce bias into the ML model. Figure 2 depicts such a situation. The governance implications summary that our technique creates includes 'affect on features included in the derived data set'. We call the features in this part of the summary 'Affected-features'. Affected-features are a potential cause for bias. Our technique flags an affected-feature as suspected of bias if it is also a high-importance feature in the model trained using the given data. A simple algorithm identifies the bias-suspected features as those that belong to the intersection of the affected-features group and the important-features group. The latter can be computed once a model has been trained, based on the model itself, or via black box techniques such as LIME [16]. Section 4.3 provides an example resulting from our experiments.

Once bias-suspected features are identified, they can be sent to an entity that holds the entire data set, such as the data owner, for bias detection. Notice that identifying the suspected features is an important input for directing the bias detection techniques. Our technique provides this information so that any bias detection technique can utilize it.

3.3 Governance Impact on Features Not Included in Derived Data

Learning can only utilize the data that exists. When features are excluded from the data, they are, by definition, excluded from any model that uses that data for training. Figure 3 depicts such a situation. This might have dire implications if the excluded features are relevant to the learning goal. It is highly challenging to check such implications as the ML model naturally does not contain these features. However, we suggest to utilize the fact that the entire data set is available to its owner in order to ask the data owner or any other entity that has access to the full data set to run risk assessment code on the full data set, based on the ML model learned from the partial data set. This may be done, for example, as follows: the governance implications technique creates a feature relevance function and sends it to the owner of the original data set for execution. The owner

runs the function on the full data set, and returns a score for each feature that was removed from the derived data set. A high score indicates that the removed feature affects the model.

There are many possible implementations for a feature relevance function. One implementation is to investigate the distribution of values of the excluded features per the model classification or prediction categories and raise the implication score if it is significantly different per category. Another possible implementation involves retaining the model. This may not always be possible. If it is, then the function would require the owner to re-train a model while adding a single excluded feature every time and/or all excluded features. The implication score is then raised if the results are significantly different compared to those of the model trained on the data without the excluded features. We plan to explore these various approaches, especially in order to better understand their trade-offs. For example, we expect the second approach to be more informative compared to the first one, yet require more resources both in terms of compute resources and human attention resources.

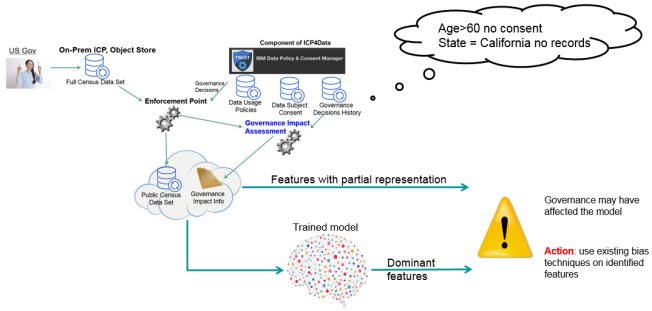

Fig. 2: Utilizing data governance impact summary when data records are excluded.

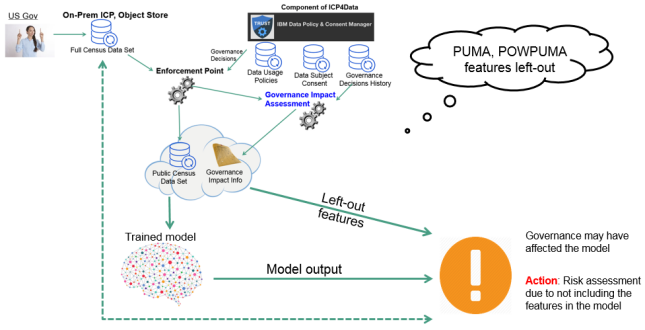

Fig. 3: Utilizing data governance impact summary when data features are excluded.

4 Experiments

We demonstrate the feasibility of our techniques on a US government Census data set. Our experiments show that:

1. It is possible to extract governance implications from a governance enforcement engine and encode them as meta data.
2. These governance implications can be effectively utilized to alert on data issues that negatively affect the ML model trained on the governed data subset. We demonstrate this for excluded data records, simulating no-consent situations.

Section 4.1 provides details about the data and governance policies that we experiment with. Section 4.2 and Section 4.3 provides experimental results that support the above claims, respectively.

4.1 Data and Governance policies

We ran experiments on the US American Community Survey from the US Census Bureau [1]. In this survey, approximately 3.5 million households per year are asked detailed questions about who they are and how they live. Many topics are covered, including education, work, and transportation. The result is over 600MB data set with over 280 features. We concentrated on transportation and followed a common usage of the data in which a classifier is trained to predict the transportation means that a person uses in order to get to work. This is a multi-label classification task and includes labels such as car, train, ferry, bus, taxi, and motorcycle.

We defined example governance policies for our experiments:

1. California residents information was excluded due to new strict privacy law. This resulted in 18233 records being excluded from the governed data subset.
2. People over 60 nationwide tended not to provide consent for their information to be included in the public data. This resulted in 14554 records being excluded from the governed data subset.
3. PUMA, POWPUMA codes (granular location information) were excluded entirely to prevent easy cross reference. These specific features of all records were excluded from the governed data subset.

4.2 Creating Governance impact summary

We show that it is possible to extract governance implications and encode them, using IBM's DPCM governance engine. For our experiment about the use of US census data for public transportation planning, we created a general purpose called "Commercial Use of Census Data" shown in Figure 4.

For this experiment we defined that due to strict privacy regulations in California, no personal data of US citizens living in California is allowed to be

Purpose Commercial Use of Census Data (ACTIVE) , Version 1 , State: ACTIVE - Latest Version

ID	3
Version	1
Name	Commercial Use of Census Data
State	ACTIVE
Access types	Process
Type	
Geography	

Description

Purpose Data

Data	Access Types	Retention	Mandatory
Age	Process		Yes
City	Process		Yes
Gender	Process		Yes
Occupation	Process		Yes
PUMA	Process		Yes
Salary	Process		Yes
State	Process		Yes

Fig. 4: Purpose indicating commercial use of census data.

included in the data subset that will be used for by our ML model. Figure 5 details this policy.

In our experiment we also note that people over the age of 60 years do not tend to provide consent for the use of their personal data. We capture this in DPCM users' consent settings, some examples of which are shown in Figure 6.

When accessing or sharing data, a data governance enforcement point is used to generate a compliant subset of the original data based on (1) the purpose for which the data will be used, (2) the geography which determines the relevant policies and laws. The enforcement point filters and transforms the full data set based on the policies and data subject preferences. During this process, all decisions about which data is included and excluded is logged. Figure 7 shows examples of governance queries, as would be invoked by the enforcement point, and the governance decisions which are stored to the log. A log entry is made as a result of a request, for example, denying the use of John Doe's personal data because he resides in California. Sally Smith and Joe Harrison don't live in California, but they both are over 60 and both denied consent for the use of their personal data.

Each governance decision log entry contains the following information: Purpose, Data subject ID, Access Type (ex: process, share with 3rd party), Approved (yes, no), Data item (feature), Reason code for the approval or denial, Reason description, Policy on which the decision was based, and various other fields not relevant to this discussion.

Fig. 5: Policy indicating that personal data of Californians is now allowed.

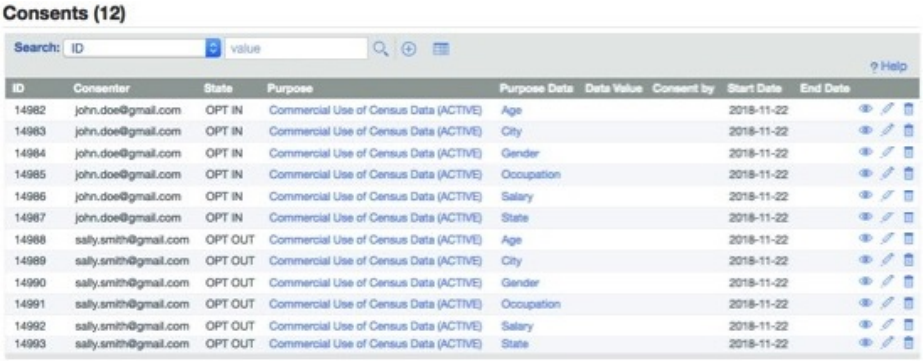

Fig. 6: Users' consent information.

For each feature (known as a data item in DPCM) there is an entry in the governance decision log. We parse the log, creating and interim data structure as defined in Figure 8

From this interim data we generate the a summary about what data was included and excluded, as Figure 9 shows.

However, the above may not include information about important features such as geography, age, gender, race and other such personal information which may not be part of the source data set, but could influence the ML model's results. If such information exists in a profile management system that can be cross referenced, then we can further generate this data by taking the interim summary and correlating it with information from the profile system, as shown in Figure 10.

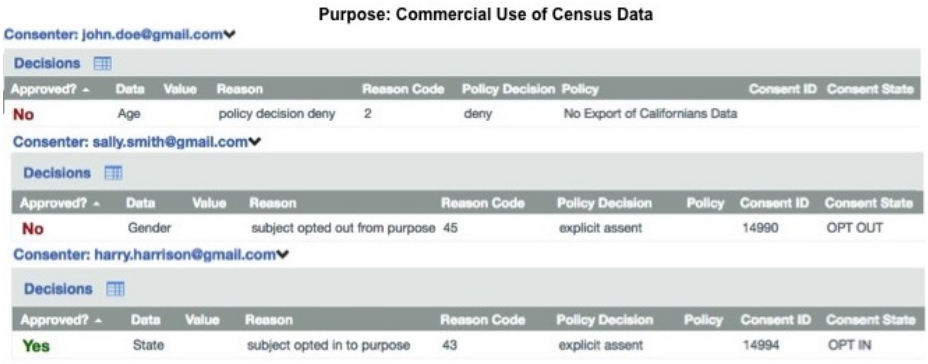

Fig. 7: Governance queries.

```
{
  "purpose_id": int,
  "percent_data_subjects_included": float,
  "data_items_completely_excluded": [ { "data_item_id": int } ],
  "data_items_included": [{
      "percent_included": float,
      "exclusions" : [{
          "num_data_subjects_excluded": int,
          "exclusion_reason_codes": [{
            "reason_code": int,
            "num_excluded_for_reason": int }],
        }],
      "inclusions" : [{
          "num_data_subjects_included": int,
          "inclusion_reason_codes": [{
            "reason_code": int,
            "num_included_for_reason": int }],
        }]
    }]
}
```

Fig. 8: Interim governance data in JSON format.

```
{
  "purpose_id": int,
  "num_data_subjects_included": int,
  "num_data_subjects_excluded": int,
  "data_items": [{
      "data_subjects_included": [{
        "data_subject_id": int,
        "reason_code_for_inclusion": int,
        "policy_id": int
      }],
      "data_subjects_excluded": [
        "data_subject_id": int,
        "reason_code_for_inclusion": int,
        "policy_id": int ]
    }
  ]}
```

Fig. 9: Summary of included and excluded data in JSON format.

To the governance impact summary we then add the list of profile features, indicating for each the percentage excluded for each data item. The final summary is then shown in Figure 11.

4.3 Alerting on governance implications

We demonstrate that the governance implications summary can be effectively utilized to raise alerts regarding potential ML model under-performance. In our experiments we leverage governance implications detailed in Section 4.2. We trained a random forest classifier where the target was the transportation means feature and all other features were used to train the model.

Figure 12 shows the results of analyzing the model important features and intersecting the resulting group of features with the features that the governance implication method marked as affected. This method is relevant when the features exist in the data provided for training. The resulting features are indeed two of the features that were under-represented as a result of simulating

```
{
  "purpose_id": int,
  "num_data_subjects_included": int,
  "num_data_subjects_excluded": int,
  "data_items": [{
      "data_subjects_included": [{
          data_subject: {
            "data_subject_id": int, "street": string,
            "city": int,            "state": int,
            "country": int,         "age": int,
            "gender": int,          "race": int
          }
          "reason_code_for_inclusion": int,
          "policy_id": int
      }],
      "data subjects excluded": [
        # each would be assigned a profile_feature_id
        data_subject: {
          "data_subject_id": int,   "street": string,
          "city": int,              "state": int,
          "country": int,           "age": int,
          "gender": int,            "race": int
        }
        "reason_code_for_inclusion": int,
        "policy_id": int
    ]}
]}
```

Fig. 10: Summary with additional included and excluded data.

```
{
  "purpose_id": int,
  "percent_data_subjects_included": float,
  "data_items_completely_excluded": [ { "data_item_id": int } ],
  "data_items_included": [{
      "percent_included": float,
      "exclusions" : [
        {
          "num_data_subjects_excluded": int,
          "exclusion_reason_codes": [
            { "reason_code": int,
              "num_excluded_for_reason": int }
          ],
          "profile_feature_exclusions:": [
            feature: {
              profile_feature_id: int,
              percent_excluded: float
            }]
        }],
      "inclusions" : [
        {
          "num_data_subjects_included": int,
          "inclusion_reason_codes": [
            { "reason_code": int,
              "num_included_for_reason": int }
          ],
          "profile_feature_inclusions:": [
            feature: {
              profile_feature_id: int,
              percent_included: float
            }]
        }]
    }]
}
```

Fig. 11: Final governance impact summary.

governance policies on the full data. The meaning of model-important features is that any changes to the distribution of these features is likely to affect the model results. Because the model owners now know that there are important features that were under-represented, and they know which features they are, they actually know in advance that it is highly likely that the model is biased. In our example, people over 60 and people from California are under-represented and age and state are important model features. Assume that elderly people use public transportation more. The model is likely to miss that. Also, assume that people in California tend to bike and walk more. Again, the model is likely to miss that. However, because utilizing the governance impact analysis summary alerts on these governance implications, the model owners can run existing bias detection and alleviation techniques, as Figure 2 describes, to reduce the model under-performance for people over 60 and people from California.

```
Model features
Index(['SPORDER', 'ST', 'ADJINC', 'PWGTP', 'AGEP', 'CIT', 'COW', 'DDRS', 'DEAR', 'DEYE',
       ...
       'FSEMP', 'FSEXP', 'FSSIP', 'FSSP', 'FWAGP', 'FWKHP', 'FWKLP', 'FWKWP', 'FWRKP',
'FYOEP'], dtype='object', length=155)

Redacted features (Governance)
{'AGEP', 'ST', 'PUMA', 'POWPUMA'}

WARNING
Dominant model features are under-represented: {'AGEP', 'ST'}
WARNING
```

Fig. 12: Intersection of ML model important features and Governance implications features results in a warning about potential implications.

5 Conclusions and future work

Data governance is crucial for abiding by privacy laws, industry standards and enterprise policies. ML solutions, which are highly dependent on training and production data, potentially introduce bias and errors if the subset of data on which they train and run are not representative. By definition, this is the case when data governance filters or changes the data provided to the ML. Thus, it is crucial to capture information about data excluded due to the governance policies and data subject preferences to alleviate any resulting ML model undesired effects, such as bias.

In this paper we define what information to capture and how to compile the summary describing the excluded data, in order to understand governance impact. We further demonstrate the utilization of this governance impact summary to characterize governance impact on an ML model. We demonstrate the ability to raise alerts on potentially biased features, using US Census data for our use case.

This line of research raises many challenges. We show a solution for one of the challenges related to identifying potential bias in an ML model. In future work we will address other challenges. For example, capturing in the governance implication summary information about obfuscation methods used on the data. Another direction is to define actions based on the alerts that follow governance implications impact detection. Another challenge is to identify when data drift [15] is caused by governance implications.

References

1. Us census population data (2013), kaggle Data, www.kaggle.com/census/2013-american-community-survey
2. ApacheRanger: Apache ranger, accessed February 2018 ranger.apache.org/
3. Ayalon, O., Toch, E.: Not even past: Information aging and temporal privacy in online social networks. Human Computer Interaction **32**(2), 73–102 (2017)
4. Census, U.: Education finance data (1992), catalog.data.gov/harvest/object/890f9407-e5a9-4862-8e6d-c1e133aa0a64
5. Chessell, M., Grote, C.: Virtual data connector (vdc) demo (2018), egeria open source metadata project github.com/odpi/egeria/blob/master/open-metadata-resources/open-metadata-demos/virtual-data-connector/README.md
6. De Choudhury, M., Sharma, S.S., Logar, T., Eekhout, W., Nielsen, R.C.: Gender and cross-cultural differences in social media disclosures of mental illness. In: Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing. pp. 353–369. CSCW '17 (2017), doi.acm.org/10.1145/2998181.2998220
7. GIZMODO: Amazon's secret ai hiring tool reportedly 'penalized' resumes with the word 'women's' (2018), one of many reports on the topic, gizmodo.com/amazons-secret-ai-hiring-tool-reportedly-penalized-resu-1829649346
8. Hirschprung, R., Toch, E., Bolton, F., Maimon, O.: A methodology for estimating the value of privacy in information disclosure systems. Computers in Human Behavior **61**, 443 – 453 (2016)
9. IBM: Guardium, accessed February 2018 www.ibm.com/security/data-security/guardium
10. for Information Technology Policy, P.U.C., for Human values, U.C.: Princeton's dialog on ai and ethics (2018), aiethics.princeton.edu/
11. Jones, A., Bazrafshan, M., Delgado, F.P., Lihatsh, T., Schuyler, T., ajones: The role of metadata in machine learning for technology assisted review (2015)
12. Kusner, M.J., Loftus, J.R., Russell, C., Silva, R.: Counterfactual fairness. In: NIPS (2017)
13. Patil, D., Mason, H., Loukides, M.: Ethics and Data Science. O'Reilly Media, Inc. (2018)
14. Pham, S., Vergano, D.: Cocaine deaths are rising at an alarming rate, and its because of fentanyl (2018), buzzFeed News blog.bigml.com/2018/05/24/gdpr-compliance-and-its-impact-on-machine-learning-systems/ see also dataset with text metadata https://git.io/fh7e9
15. Raz, O., Zalmanovici, M., Zlotnick, A., Farchi, E.: Automatically detecting data drift in machine learning based classifiers. In: AAAI 2019 Workshop on Engineering Dependable and Secure Machine Learning Systems (EDSMLS19) (2019)
16. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. KDD '16 (2016)
17. Shadowen, N.: How to prevent bias in machine learning (2018), see Section 'Using ethics to solve machine bias' becominghuman.ai/how-to-prevent-bias-in-machine-learning-fbd9adf1198
18. Tabakovic, V.: Gdpr compliance and its impact on machine learning systems (2018), see Section 'Using ethics to solve machine bias' blog.bigml.com/2018/05/24/gdpr-compliance-and-its-impact-on-machine-learning-systems/