DC-CHECK (WORKED EXAMPLE 1)

Project Description: A Dynamic Pipeline for Spatio-Temporal Fire Risk Prediction

Link: https://dl.acm.org/doi/abs/10.1145/3219819.3219913

Citation: Singh Walia, B., Hu, Q., Chen, J., Chen, F., Lee, J., Kuo, N., Narang, P., Batts, J., Arnold, G. and Madaio, M., 2018, July. A dynamic pipeline for spatio-temporal fire risk prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*

DATA

Q1: How did you select, collect or curate your dataset?

- *Have you conducted forensics on the dataset (i.e. provenance)?*
- *Did you assess the pertinence of the dataset for the task?*
- *Is your dataset curation a once-off?*

Data curation

We start by acquiring data sets from a variety of sources that contain data we hypothesize (based on prior work) to be relevant in predicting fire risk in non-residential properties.

1. ***Pittsburgh Bureau of Fire:*** Historical fire incident data from 2009-2017, updated on a weekly basis. All fire incident codes are included that have an associated address.
2. ***Allegheny County Office of Property Assessments (OPA):*** property assessment data, updated on a monthly basis, as well as a parcel dataset, which contains information about every parcel in the City of Pittsburgh.
3. ***Pittsburgh Department of Permits, Licenses, and Inspections (PLI):*** records of non-fire inspections and violations (e.g. noise or sanitation violations).
4. ***Residential risk model uses the following additional data:***

- Allegheny County Department of Court Records (DCR): tax lien data, updated monthly
- US Census Bureau: 2012-2016, 5-year estimate American Community Survey (ACS) data

Data selection

Datasets are chosen that are likely correlated with fires based on prior work, specifically data about income, occupancy, year built, and the year resident moved in.

Q2: What data cleaning and/or pre-processing, if any, has been performed?

- *Do you need to "clean" your data?*
- *Do you have any missing data?*

1. Commercial property risk model

Joining of data

Fire incidents were logged at the address level, and thus, to predict fire risk, all the other data first needed to be aggregated across the individual parcels in each address

Join the Allegheny County property assessment data set, with the PLI inspection violations dataset with the non-residential properties (by which we mean Commercial, Industrial, Governmental, or Utility), at the parcel level.

Data Cleaning:

- Stripping white spaces from text values
- Drop duplicate columns
- Drop rows with significant (85%) missing values for data
- Aggregate the parcel data at the address level by taking the mean of numeric features and using the most frequent category for categorical features.

2. **Residential property risk model**

Joining of data

Same procedure as above, except that we further aggregated the data to the census block group level by taking the sum for numeric features, which included tax lien, and PLI data, and the most common for categorical features, which included some property data.

Data Cleaning:

- Sum as opposed to the mean to capture the total number of violations and inspections in each block, as well as the total amount of unpaid taxes.
- Stripping white space
- Standardizing hyphens
- Standardizing street abbreviations

BOTH:

For both models, we then merged the resulting aggregated data frame with the fire incidents dataframe (for both the commercial and residential model).

Q3: Has data quality been assessed?

- Completeness, diversity, label quality
- Easy, Hard examples etc

Data quality has not been explicitly assessed, as the data used comes from high-quality governmental data sources with their own checks and balances.

Q4: Have you considered synthetic data?

Synthetic data has not been considered as:

- Data privacy is not an issue as it is an internal government deployed project.
- Data access is permitted. Hence, synthetic data is not required.
- Individual datasets are large >100k, hence no need to augment with synthetic data

TRAINING

Q5: Have you conducted a model architecture and hyperparameter search?

- *Have you specifically adapted the model for the task?*
- *Is there a potential to incorporate inductive biases?*

Model architecture:

Commercial risk model: Logistic Regression, Ada Boost, Random Forest, XG Boost compared with XG Boost chosen

Residential risk model: Ada Boost, Random Forest, XG Boost compared with Random Forest chosen

Hyperparameters:
Grid search is done for the following hyperparameters

- XGBoost commercial risk model, we searched max_depth, min_child_weight, subsample, colsample_by_tree and tuned the rest manually.

- Random Forest residential risk model, we searched n_estimators, max_depth, and max_features using a 1-year validation set.

Q6: Does the training data match the anticipated use?

- *Can we anticipate the data the model will be applied on?*
- *Could the model be exposed to data from different distributions/domains?*

Data matching anticipated use:
- The model will only be used to provide risk scores for properties in Pittsburgh.

Distribution/domain shift:
- There is no expected distribution or domain shift to account for - as we keep the location fixed and will always use the latest data.

Q7: Are there different data subsets or subgroups of interest?

- *Are the subgroups identifiable?*
- *Have you assessed fairness/bias introduced by training?*

Subgroups/subsets of data:
- There are different property types that can be considered as subgroups/subsets.

Training based on subgroup/subset:
- Currently, the models are optimized on average rather than per subgroup.

Q8: Is the data noisy, either in features or labels?

Presence of data noise:
- There is no specific data noise considered or assessed

Accounting for data noise:
- We do not account for data noise or potentially noisy labels, as the data does not come from crowd-sourcing, but rather from official sources.

TESTING

Q9: How has the dataset been split for model training and validation?

- *Are you using a benchmark dataset?*
- *Is the splitting random?*

Data split

- Used a walk forward time partition approach, to ensure past events are used to assess future events only
- Training set (6 years of data), validation set for feature selection (1 year of data), and test set (the final 1 year of data).

Q10: How has the model been evaluated (e.g. metrics & stress tests)?

- *What metrics are used to assess the model?*
- *Has the model been evaluated beyond average?*
- *Is there a potential to test specific aspects/sub-groups of the model (i.e. stress tests)?*

As the use case of fire prediction, we want to prioritize correctly classifying more of the positive class (i.e. fire) over minimizing false positives (which may result in more inspections, but would be less likely to lead to missed incidents).

Metrics:

- Main: kappa and recall
- Secondary: AUC and precision

Cross-validation was performed using the training set.

Assessment beyond average and/or subgroups:

- Not automated but “face validity” assessment conducted, where different property types model risk scores were assessed.
- It was found they agree with existing expert knowledge and Bureau of Fire risk reduction efforts

DEPLOYMENT

Q11: Are you monitoring your model?

- *Have you considered dimensionality in the monitoring?*
- *Is there lag in ground truth feedback?*

Ground truth availability/lag:

- Ground truth risk scores are not immediately available & would require domain expertise.

Monitoring:

- There is no explicit model monitoring, as the model is retrained weekly

(i) Metric being monitored: N/A

(ii) Monitoring method: N/A

Q13: Do you have mechanisms in place to address data shifts?

- *Have model updates been considered & when to retrain?*
- *Does your system provide actionable feedback upon failure?*
- *Are you able to characterize the type of shift?*
- *Do system failures inform dataset updates?*

Model updates

- Cron job runs every Saturday. Scrapes data sources for the latest dataset, retrains the model, and updates those risk scores on the map and dashboard.

Failure feedback:

- Our system does not provide feedback on failures

Characterizing the type of shift:

- We do not consider characterizing the type of shift

Failure informed dataset updates:

- The dataset is updated weekly, which while not failure informed would help to rectify issues, representing the latest data

Q14: Have you incorporated tools to engender model trust?

- *Do you require predictive uncertainty estimates?*
- *Does your ML system have explainability?*
- *Does your system account for OOD inputs?*
- *Do you assess issues of bias and fairness in deployment?*

Uncertainty estimates:

- Uncertainty estimates of the risk scores have not been included.

Explainability:

- Feature importance scores were computed for samples (Explainability)

OOD:

- OOD detection has not been included in the pipeline

Bias/Fairness:

- Assessment of model bias or fairness gaps (for example between property type subgroups) have not been considered