Toward Human-Centered Machine Learning

©Patrick Hall*

H_2O.ai

June 5, 2019

Contents

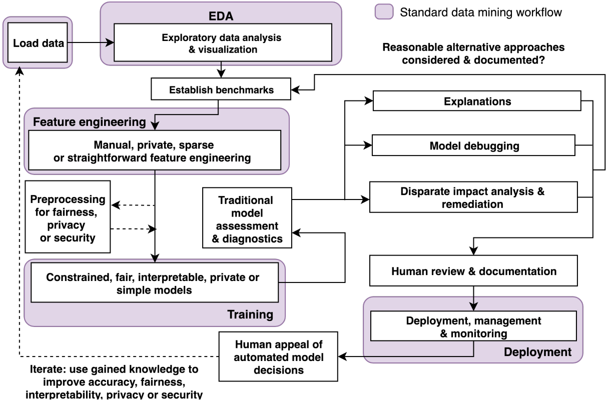
Blueprint

This mid-level technical document provides a basic blueprint for combining the best of AutoML, regulation-compliant predictive modeling, and machine learning research in the sub-disciplines of fairness, interpretable models, post-hoc explanations, privacy and security to create a low-risk, human-centered machine learning framework.

Based on guidance from leading researchers and practitioners.

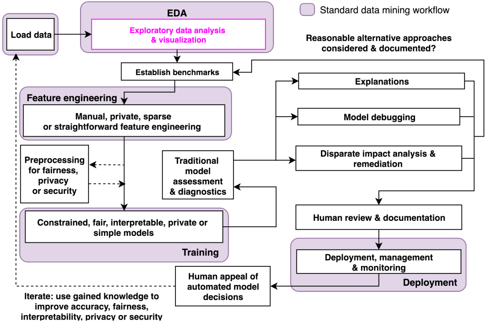
This document does not seek to redefine *human-centered machine learning* which has a notable presence in the human computer interaction (HCI) literature and is often associated with interactive software for enabling greater human control and understanding of ML processes (Gillies et al., 2016). This document simply seeks to draw attention to mostly newer, but certainly some pre-existing or established, ML techniques that may further the goals of human-centered ML.

Blueprint†



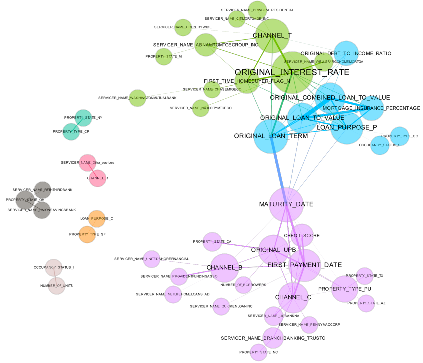
† This blueprint does not address ETL workflows.

EDA and Data Visualization

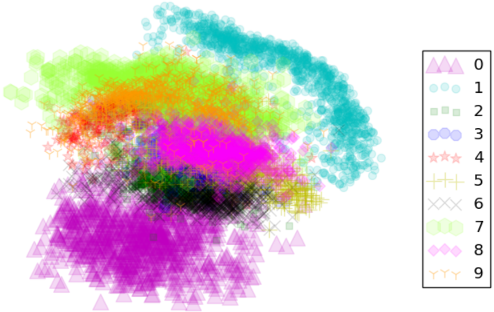


- Know thy data.
- OSS: H2O-3 Aggregator
- References: Visualizing Big Data Outliers through Distributed Aggregation; The Grammar of Graphics

Interlude: My Favorite Visualizations



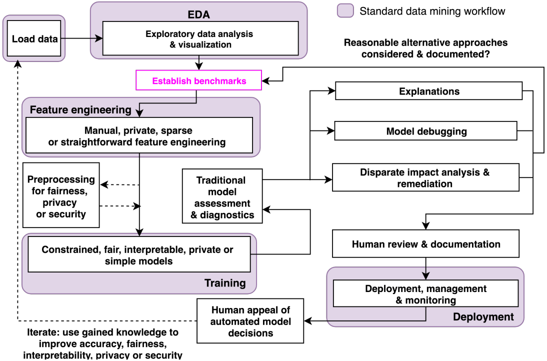
A network graph capturing the Pearson correlation relationships between many *columns* in a lending dataset.



An autoencoder projection of the MNIST data. Projections capture sparsity, clusters, hierarchy, and outliers in *rows* of a dataset.

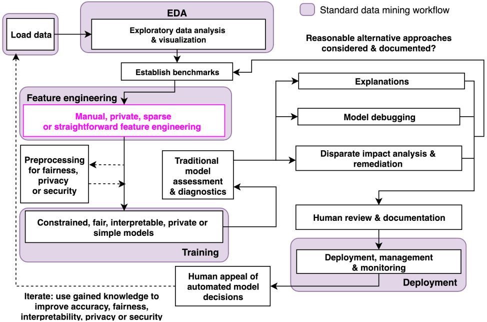
Both of these images capture high-dimensional datasets in just two dimensions.

Establish Benchmarks



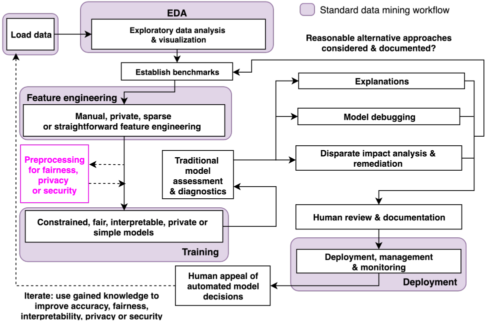
Establishing a benchmark from which to gauge improvements in accuracy, fairness, interpretability or privacy is crucial for good ("data") science and for compliance.

Manual, Private, Sparse or Straightforward Feature Engineering



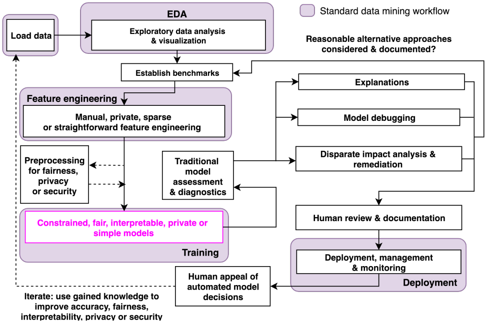
- OSS: Pandas Profiler, Feature Tools
- References: Deep Feature Synthesis: Towards Automating Data Science Endeavors; Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering; *t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-diversity

Preprocessing for Fairness, Privacy or Security



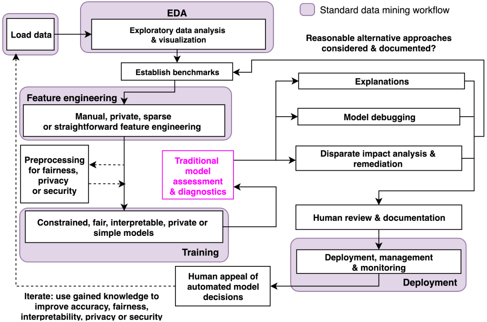
- OSS: IBM AIF360
- References: Data Preprocessing Techniques for Classification Without Discrimination; Certifying and Removing Disparate Impact; Optimized Pre-processing for Discrimination Prevention; Privacy-Preserving Data Mining; Differential Privacy and Machine Learning: A Survey and Review

Constrained, Fair, Interpretable, Private or Simple Models



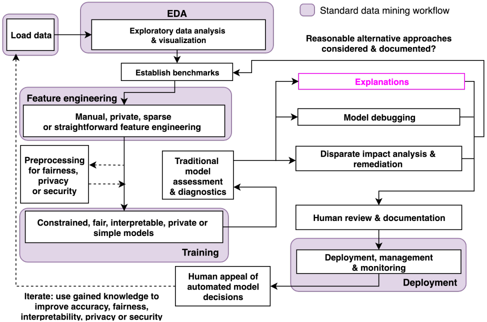
- OSS: Monotonic gradient boosting machines in H2O-3 or XGBoost
- References: Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP); Explainable Neural Networks Based on Additive Index Models (XNN); Scalable Private Learning with PATE; Scalable Bayesian Rule Lists (SBRL); Learning Fair Representations (LFR)

Traditional Model Assessment and Diagnostics



Residual analysis, Q-Q plots, AUC and lift curves etc. confirm model is accurate and meets assumption criteria.

Post-hoc Explanations

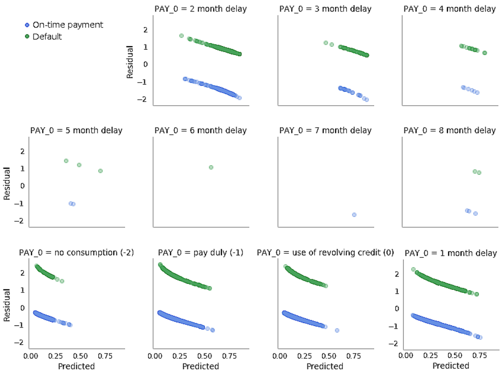


- Explanations enable *understanding* and *appeal* … *not trust*.
- OSS: lime, shap
- References: Why Should I Trust You?: Explaining the Predictions of Any Classifier; A Unified Approach to Interpreting Model Predictions; Please Stop Explaining Black Box Models for High Stakes Decisions (criticism)

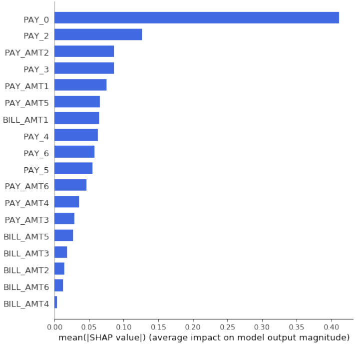
Interlude: The Time–Tested Shapley Value

1. **In the beginning**: A Value for N-Person Games, 1953
2. **Nobel-worthy contributions**: The Shapley Value: Essays in Honor of Lloyd S. Shapley, 1988
3. **Shapley regression**: Analysis of Regression in Game Theory Approach, 2001
4. **First reference in ML?** Fair Attribution of Functional Contribution in Artificial and Biological Networks, 2004
5. **Into the ML research mainstream, i.e. JMLR**: An Efficient Explanation of Individual Classifications Using Game Theory, 2010
6. **Into the real-world data mining workflow ... *finally***: Consistent Individualized Feature Attribution for Tree Ensembles, 2017
7. **Unification**: A Unified Approach to Interpreting Model Predictions, 2017

Interlude: Explaining Why Not to Trust



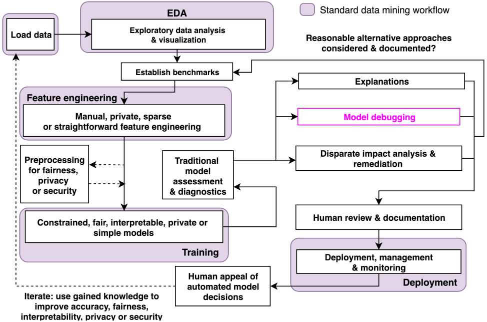
These residuals show a problematic pattern in predictions related to the most important feature, `PAY_0`.



This model over-emphasizes the most important feature, `PAY_0`.

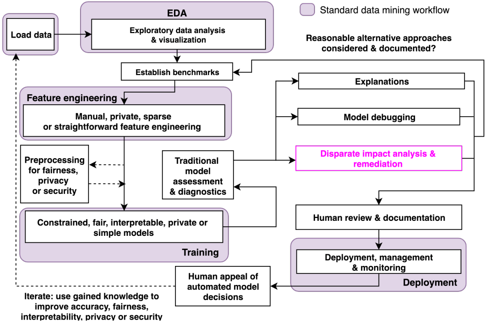
While this model is *explainable*, it's probably not *trustworthy*.

Model Debugging for Accuracy, Privacy or Security



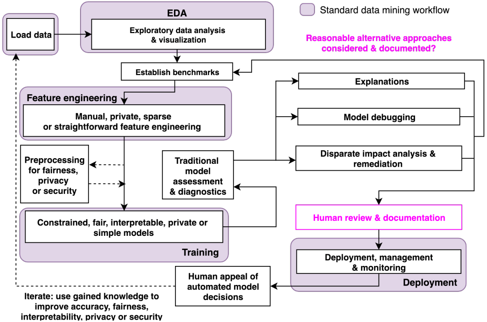
- Eliminating errors in model predictions by testing: adversarial examples, explanation of residuals, random attacks and "what-if" analysis.
- OSS: cleverhans, pdpbox, what-if tool
- References: Modeltracker: Redesigning Performance Analysis Tools for Machine Learning; A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private; The Security of Machine Learning

Post-hoc Disparate Impact Assessment and Remediation



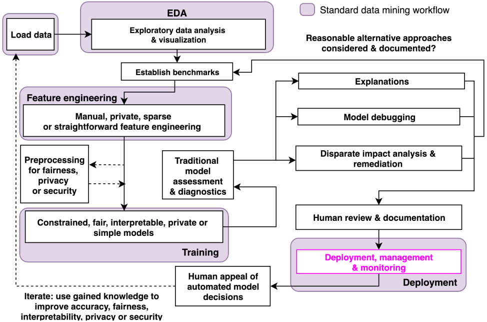
- Disparate impact analysis can be performed manually using nearly any model or library.
- OSS: aequitas, IBM AIF360, themis
- References: Equality of Opportunity in Supervised Learning; Certifying and Removing Disparate Impact

Human Review and Documentation



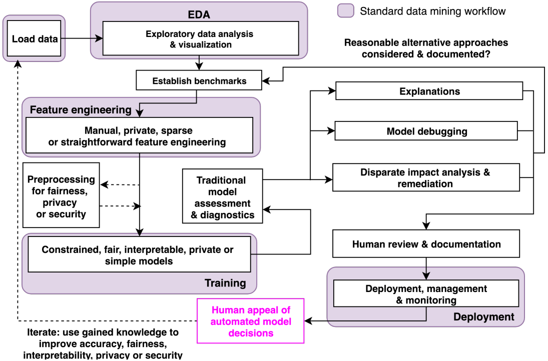
- Reference: Model Cards for Model Reporting
- Documentation of considered alternative approaches typically necessary for compliance.

Deployment, Management and Monitoring



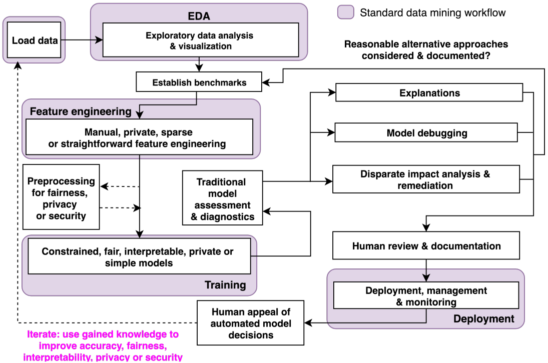
- Monitor models for accuracy, disparate impact, privacy violations or security vulnerabilities in real-time; track model and data lineage.
- OSS: mlflow, modeldb, awesome-machine-learning-ops metalist
- Reference: Model DB: A System for Machine Learning Model Management

Human Appeal



Very important, may require custom implementation for each deployment environment? Related problems exist *today*.

Iterate: Use Gained Knowledge to Improve Accuracy, Fairness, Interpretability, Privacy or Security



Improvements, KPIs should not be restricted to accuracy alone.

Open Conceptual Questions

- How much automation is appropriate, 100%?
- How to automate learning by iteration, reinforcement learning?
- How to implement human appeals, is it productizable?

References

In-Depth Open Source Interpretability Technique Examples:
https://github.com/jphall663/interpretable_machine_learning_with_python

"Awesome" Machine Learning Interpretability Resource List:
https://github.com/jphall663/awesome-machine-learning-interpretability

References

Agrawal, Rakesh and Ramakrishnan Srikant (2000). "Privacy-Preserving Data Mining." In: *ACM Sigmod Record*. Vol. 29. 2. URL: http://alme1.almaden.ibm.com/cs/projects/iis/hdb/Publications/papers/sigmod00_privacy.pdf. ACM, pp. 439–450.

Amershi, Saleema et al. (2015). "Modeltracker: Redesigning Performance Analysis Tools for Machine Learning." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. URL: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/amershi.CHI2015.ModelTracker.pdf. ACM, pp. 337–346.

Barreno, Marco et al. (2010). "The Security of Machine Learning." In: *Machine Learning* 81.2. URL: https://people.eecs.berkeley.edu/~adj/publications/paper-files/SecML-MLJ2010.pdf, pp. 121–148.

Calmon, Flavio et al. (2017). "Optimized Pre-processing for Discrimination Prevention." In: *Advances in Neural Information Processing Systems*. URL: http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf, pp. 3992–4001.

Feldman, Michael et al. (2015). "Certifying and Removing Disparate Impact." In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. URL: https://arxiv.org/pdf/1412.3756.pdf. ACM, pp. 259–268.

References

Gillies, Marco et al. (2016). "Human-Centred Machine Learning." In: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. URL: http://research.gold.ac.uk/16112/1/HCML2016.pdf. ACM, pp. 3558–3565.

Hardt, Moritz, Eric Price, Nati Srebro, et al. (2016). "Equality of Opportunity in Supervised Learning." In: *Advances in neural information processing systems*. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf, pp. 3315–3323.

Hu, Linwei et al. (2018). "Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP)." In: *arXiv preprint arXiv:1806.00663*. URL: https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf.

Ji, Zhanglong, Zachary C. Lipton, and Charles Elkan (2014). "Differential Privacy and Machine Learning: A Survey and Review." In: *arXiv preprint arXiv:1412.7584*. URL: https://arxiv.org/pdf/1412.7584.pdf.

Kamiran, Faisal and Toon Calders (2012). "Data Preprocessing Techniques for Classification Without Discrimination." In: *Knowledge and Information Systems* 33.1. URL: https://link.springer.com/content/pdf/10.1007/s10115-011-0463-8.pdf, pp. 1–33.

References

Kanter, James Max, Owen Gillespie, and Kalyan Veeramachaneni (2016). "Label, Segment, Featurize: A Cross Domain Framework for Prediction Engineering." In: *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*. URL: http://www.jmaxkanter.com/static/papers/DSAA_LSF_2016.pdf. IEEE, pp. 430–439.

Kanter, James Max and Kalyan Veeramachaneni (2015). "Deep Feature Synthesis: Towards Automating Data Science Endeavors." In: *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*. URL: https://groups.csail.mit.edu/EVO-DesignOpt/groupWebSite/uploads/Site/DSAA_DSM_2015.pdf. IEEE, pp. 1–10.

Keinan, Alon et al. (2004). "Fair Attribution of Functional Contribution in Artificial and Biological Networks." In: *Neural Computation* 16.9. URL: https://www.researchgate.net/profile/Isaac_Meilijson/publication/2474580_Fair_Attribution_of_Functional_Contribution_in_Artificial_and_Biological_Networks/links/09e415146df8289373000000/Fair-Attribution-of-Functional-Contribution-in-Artificial-and-Biological-Networks.pdf, pp. 1887–1915.

Kononenko, Igor et al. (2010). "An Efficient Explanation of Individual Classifications Using Game Theory." In: *Journal of Machine Learning Research* 11.Jan. URL: http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf, pp. 1–18.

References

Lipovetsky, Stan and Michael Conklin (2001). "Analysis of Regression in Game Theory Approach." In: *Applied Stochastic Models in Business and Industry* 17.4, pp. 319–330.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee (2017). "Consistent Individualized Feature Attribution for Tree Ensembles." In: *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*. Ed. by Been Kim et al. URL: https://openreview.net/pdf?id=ByTKSo-m-. ICML WHI 2017, pp. 15–21.

Lundberg, Scott M and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf. Curran Associates, Inc., pp. 4765–4774.

Mitchell, Margaret et al. (2019). "Model Cards for Model Reporting." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. URL: https://arxiv.org/pdf/1810.03993.pdf. ACM, pp. 220–229.

Papernot, Nicolas (2018). "A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private." In: *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*. URL: https://arxiv.org/pdf/1811.01134.pdf. ACM.

References

Papernot, Nicolas et al. (2018). "Scalable Private Learning with PATE." In: *arXiv preprint arXiv:1802.08908*. URL: https://arxiv.org/pdf/1802.08908.pdf.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "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*. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf. ACM, pp. 1135–1144.

Rudin, Cynthia (2018). "Please Stop Explaining Black Box Models for High Stakes Decisions." In: *arXiv preprint arXiv:1811.10154*. URL: https://arxiv.org/pdf/1811.10154.pdf.

Shapley, Lloyd S (1953). "A Value for N-Person Games." In: *Contributions to the Theory of Games* 2.28. URL: http://www.library.fa.ru/files/Roth2.pdf#page=39, pp. 307–317.

Shapley, Lloyd S, Alvin E Roth, et al. (1988). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. URL: http://www.library.fa.ru/files/Roth2.pdf. Cambridge University Press.

"*t*-Closeness: Privacy Beyond *k*-Anonymity and *l*-diversity" (2007). In: *2007 IEEE 23rd International Conference on Data Engineering*. URL:http://www.utdallas.edu/~mxk055100/courses/privacy08f_files/tcloseness.pdf. IEEE, pp. 106–115.

References

Vartak, Manasi et al. (2016). "Model DB: A System for Machine Learning Model Management." In: *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. URL: https://www-cs.stanford.edu/~matei/papers/2016/hilda_modeldb.pdf. ACM, p. 14.

Vaughan, Joel et al. (2018). "Explainable Neural Networks Based on Additive Index Models." In: *arXiv preprint arXiv:1806.01933*. URL: https://arxiv.org/pdf/1806.01933.pdf.

Wilkinson, Leland (2006). *The Grammar of Graphics*.

— (2018). "Visualizing Big Data Outliers through Distributed Aggregation." In: *IEEE Transactions on Visualization & Computer Graphics*. URL: https://www.cs.uic.edu/~wilkinson/Publications/outliers.pdf.

Yang, Hongyu, Cynthia Rudin, and Margo Seltzer (2017). "Scalable Bayesian Rule Lists." In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. URL: https://arxiv.org/pdf/1602.08610.pdf.

Zemel, Rich et al. (2013). "Learning Fair Representations." In: *International Conference on Machine Learning*. URL: http://proceedings.mlr.press/v28/zemel13.pdf, pp. 325–333.