Towards Transparency in Black-Box Models: Investigating Methods in Explainability for AI Systems

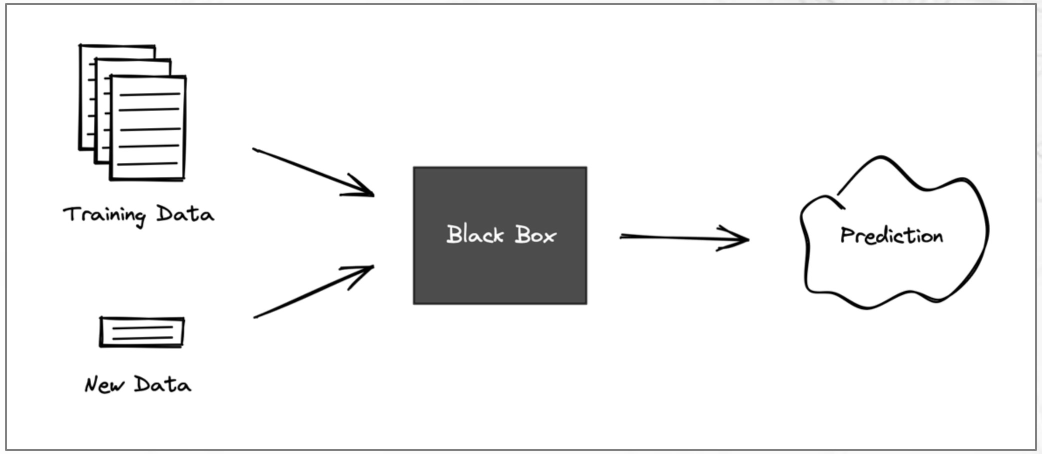
Presenter: Saleh Alkhalifa, Senior Manager of Data Science

Introduction

• Artificial Intelligence and Machine Learning models are increasingly deployed across **critical domains**, including pharmaceutical, healthcare, manufacturing, and finance, to automate complex **decision-making** processes.

• Despite their often high accuracy, many AI/ML models operate as "**black boxes**," where the reasoning behind their predictions remains unclear, creating challenges in trust, reliability, and accountability with regulators

• **Explainability** is crucial to bridge this gap by providing insights into model behavior, enabling developers and stakeholders to validate decisions, and ultimately ensure compliance with ethical and regulatory standards, and build trust in AI-driven systems

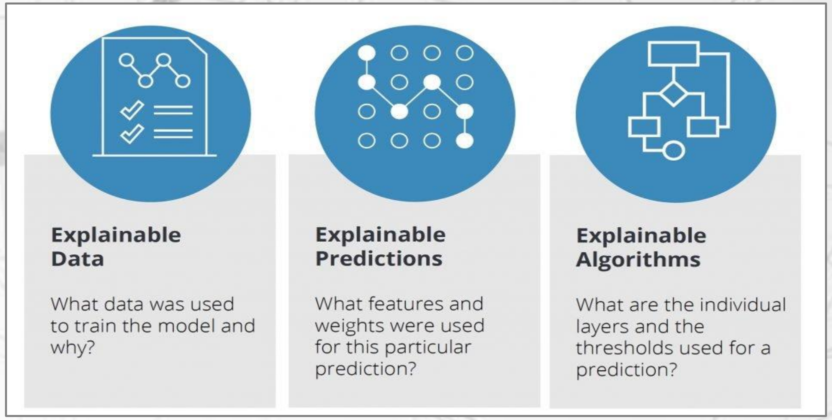


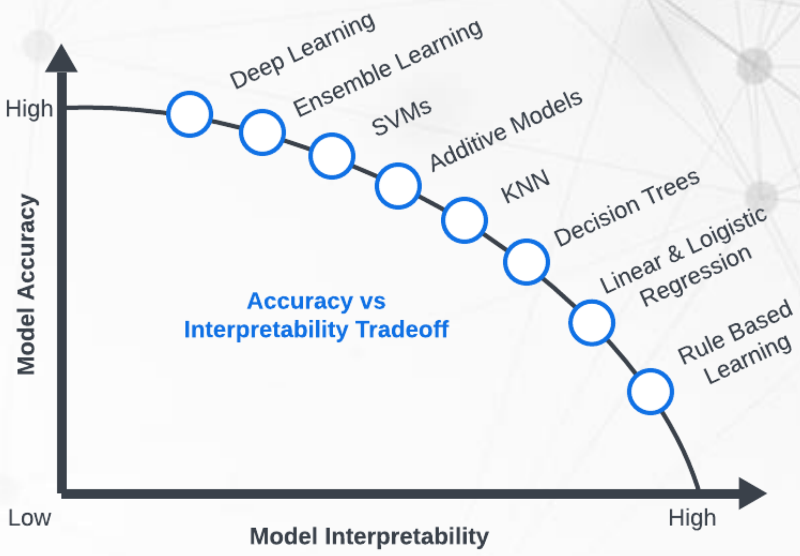
Approach & Methods

- Explainability can be the focus on a few different areas of a given experiment, such as the input data, the weights and features, or even the algorithms

- A diverse set of **datasets** (tabular, text, image), **models** (discriminative, generative, open-source, closed-source), and **interpretability methods** were selected.

- Aim to highlight the balance between model **accuracy and interpretability**, tailoring model and method choices to specific tasks for optimal transparency and performance of the model

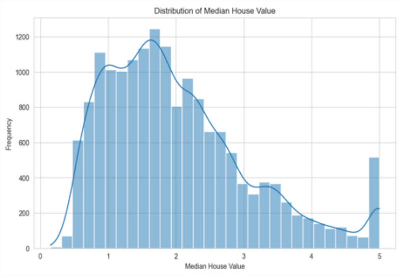




Data

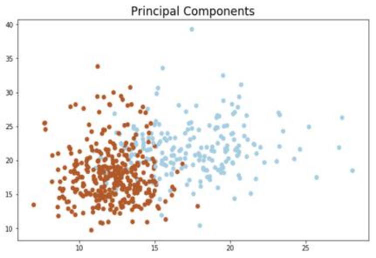
California Housing

- First published in 1990
- Regression Task
- Tabular Data
- 16813 observations
- 9 features in total
- Predict median value



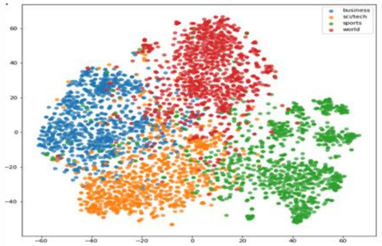
Wisconsin Breast Cancer

- First published in 1992
- Classification Task
- Tabular Data
- 569 observations
- 29 features in total
- Predict diagnosis



AG News

- First published in 2004
- Classification Task
- Textual Data
- 127000 observations
- Max_Len is 64
- Predict Category

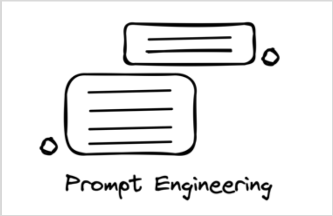


MNIST

- First published in 1994
- Classification Task
- Image Data
- 60,000 observations
- 64 x 64 pixels
- Predict Number

Custom Prompts

- Custom Made
- Classification Task
- Text Data
- 10 observations
- 10-15 words
- Prompt Engineering

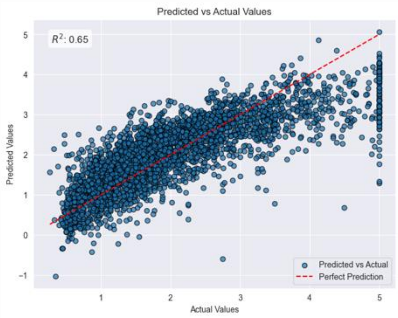


Models

California Housing

Ridge Regression

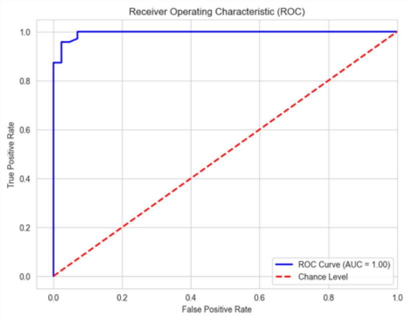
- Linear Regression Model
- L2 Regularization to reduce overfitting
- Predict median house price
- Feature weights available



Wisconsin Breast Cancer

Random Forest Classification

- Ensemble-Style Model
- Combines multiple trees for robustness
- Classify tumors as malignant of Benign
- Feature importance metrics available



AG News

DistilBERT Language Model

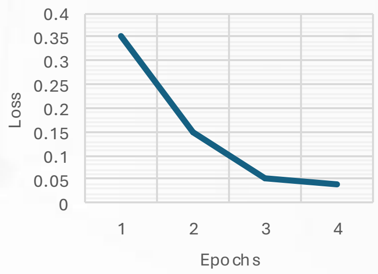
- Transformer-based Model
- Small-Medium language model
- Classifying news articles
- Contextual embeddings available

Feature	Value
Training Loss	0.44
Validation Loss	0.38
Accuracy	0.88

MNIST

Deep Learning

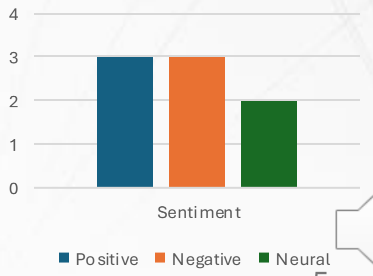
- Feed-Forward Neural Network
- Uses fully connected layers and ReLU
- Classifies hand-written digits
- Pixel-level contributions available



Custom Prompts

GPT-4o LLM

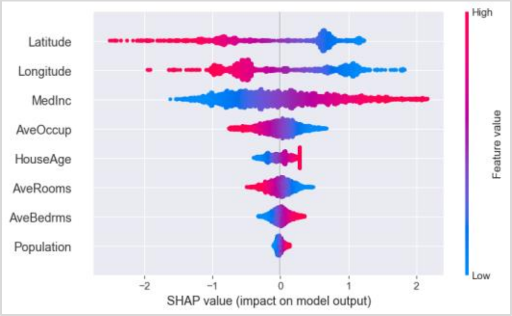
- Closed-source Large Language model
- Multi-modal language model
- Sentiment Analysis for text
- No access to weights, only input/output



Explainability

California Housing	Wisconsin Breast Cancer	AG News	MNIST	Custom Prompts
Ridge Regression	Random Forest Classification	DistilBERT Language Model	Deep Learning	GPT-4o LLM
SHAP/LIME		Token-Level Attribution using Integrated Gradients		Token-Level Importance

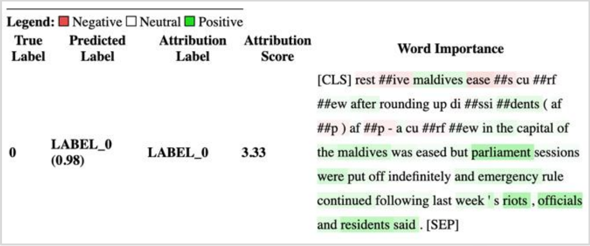
California (SHAP)



Cancer (LIME)

Feature	Value
Worst Area	-0.35
Worst Perimeter	-0.33
Worst Radius	-0.26
Mean Radius	-0.47
Worst Concavity	-0.04

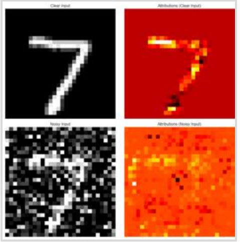
AG News (TLA)



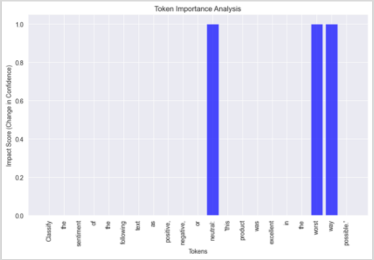
Legend: Negative Neutral Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
0	LABEL_0 (0.98)	LABEL_0	3.33	[CLS] rest ##ive maldives ease ##s cu ##rf ##ew after rounding up di ##ssi ##dents (af ##p) af ##p - a cu ##rf ##ew in the capital of the maldives was eased but parliament sessions were put off indefinitely and emergency rule continued following last week ' s riots , officials and residents said . [SEP]

MNIST (TLA)



Sentiment (TLI)



Results

California Housing	Wisconsin Breast Cancer	AG News	MNIST	Custom Prompts
Ridge Regression	Random Forest Classification	DistilBERT Language Model	Deep Learning	GPT-4o LLM
SHAP/LIME		Token-Level Attribution using Integrated Gradients		Token-Level Importance

- **SHAP** highlighted the most impactful features (number of rooms, etc...) on housing price predictions, confirming global linear trends in the data.
- **LIME** provided local explanations for specific predictions, enabling a clear view of how individual data points influenced the model's output (as we see previously)
- **SHAP** offered a global understanding of feature importance, while **LIME** allowed for instance-specific analysis, helping identify potential outliers (as seen before)

- **Integrated Gradients** revealed which tokens in AG News contributed most to classification decisions
- **TLA** Identified tokens that overly influenced predictions helped pinpoint potential biases in the dataset or model
- **Integrated Gradients** visualized the specific pixels in MNIST critical to recognizing digits, such as the horizontal and vertical strokes in "7" or "4" digits

- **Sensitivity Analysis** was used by removing specific tokens from prompts and observing output changes quantified the importance of individual words or phrases in sentiment classification.
- **Prompt Optimization** highlighted which parts of the input had the most significant impact

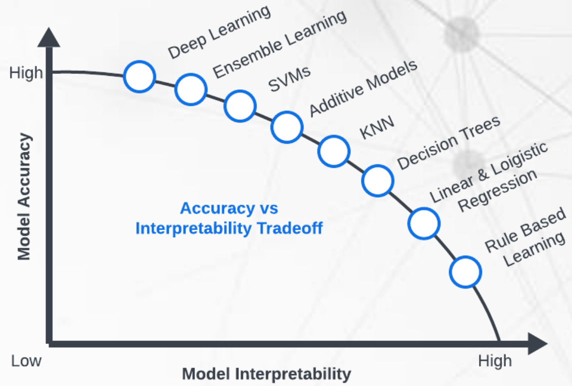
Conclusion

- **Explainability** has become crucial as AI and ML models become increasingly integrated into critical **decision-making processes**, particularly in sensitive domains like pharmaceutical, healthcare, finance, and regulatory environments.

- Different **models** and **data types** require tailored explainability approaches. For example, **SHAP** and **LIME** excel in structured datasets by providing both global and local insights, while **Integrated Gradients** shines in high-dimensional unstructured data like images and text.

- Our study demonstrated that even **highly complex** or **proprietary models** like GPT-4o can benefit from targeted explainability techniques.

- A key takeaway from our experiments is the **tradeoff between model complexity and interpretability**. While deep learning models like neural networks offer robust accuracy for tasks like digit classification, simpler models like Ridge Regression provide more transparent decision-making processes.

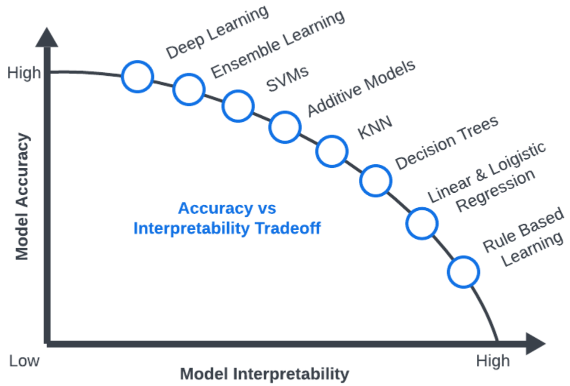


Future Work

- **Expand** the number of datasets, models, and explainability methods and in different combinations so that we can better understand the entire landscape.

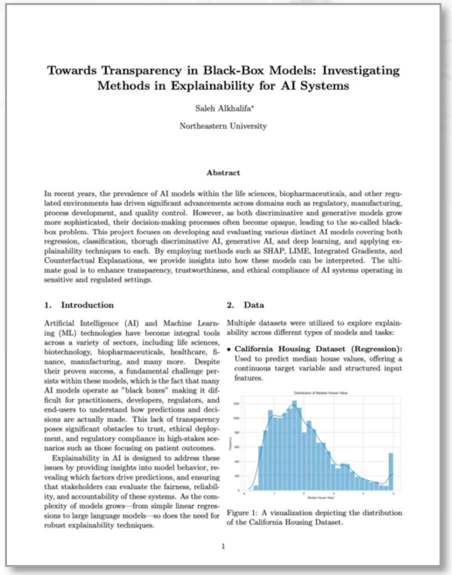
- Implement models at **different accuracy levels** (low, medium, high) and investigate the impact this would have on explainability.

- Investigate other areas such as video and audio to expand on the work done in this investigation.

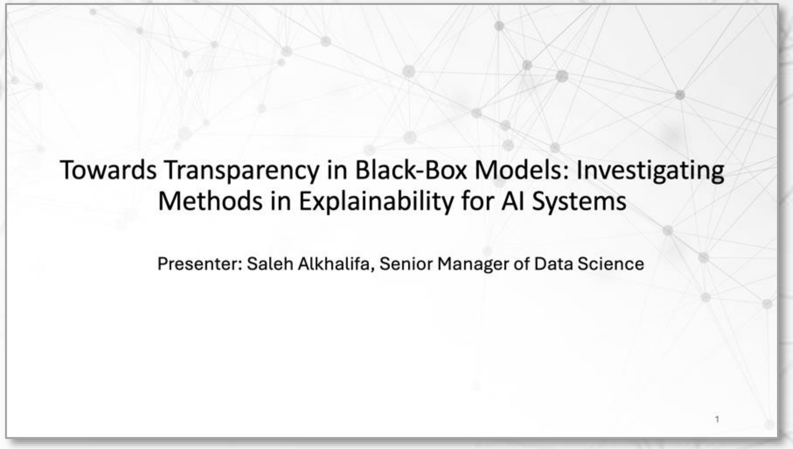


Deliverables

Full Codebase



Detailed Report



Summary Presentation

References

[1] Lundberg, S. M., & Lee, S.-I. (2017). A Uni- fied Approach to Interpreting Model Predic- tions. Advances in Neural Information Process- ing Systems.
[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD.
[3] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In- ternational Conference on Machine Learning (ICML).
[4] Kokhlikyan, N., et al. (2020). Captum: A Model Interpretability Library for PyTorch. arXiv preprint arXiv:2009.07896.
[5] Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. arXiv preprint arXiv:1906.05714.
[6] Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining Machine Learning Classifiers,Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*20).
[7] Geada, R., et al. (2021). TrustyAI Explainabil- ity Toolkit. arXiv preprint arXiv:2104.12717.
[8] Lin, Z. Q., et al. (2019). Do Explanations Re- flect Decisions? A Machine-centric Strategy to Quantify the Performance of Explainability Al- gorithms. arXiv preprint arXiv:1910.07387.
[9] Salih, A. M., et al. (2023). A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. Artificial Intelligence and Systems.
[10] Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting Artificial Intelligence Models: A Systematic Review on the Application of LIME and SHAP in Alzheimer's Disease Detection. Brain Informatics.
[11] Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localiza- tion. Proceedings of the IEEE International Conference on Computer Vision (ICCV).
[12] Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.
[13] Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards Robust Interpretability with Self- Explaining Neural Networks. Advances in Neu- ral Information Processing Systems (NeurIPS).