

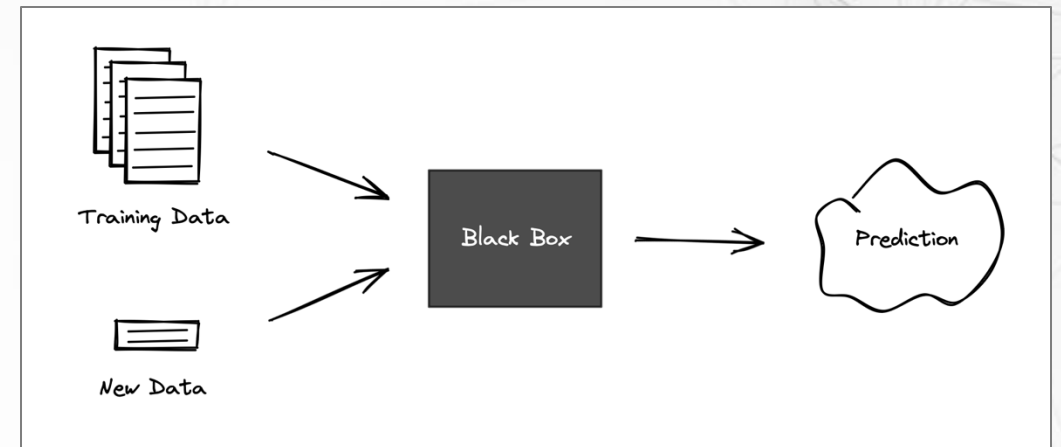
Towards Transparency in Black-Box Models: Investigating Methods in Explainability for AI Systems

Presenter: Saleh Alkhalifa, Senior Manager of Data Science



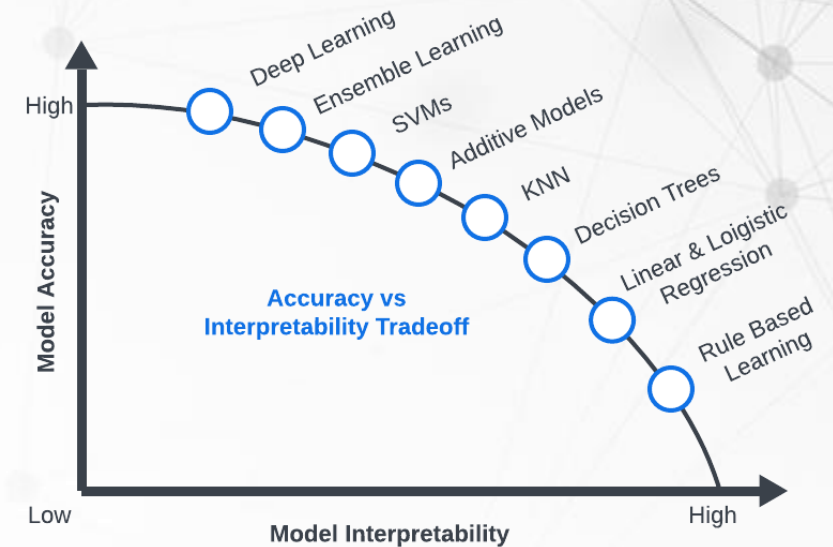
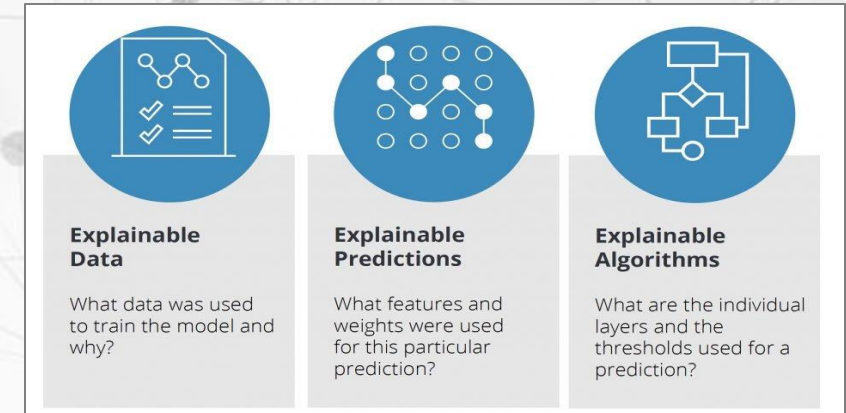
Introduction

- 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

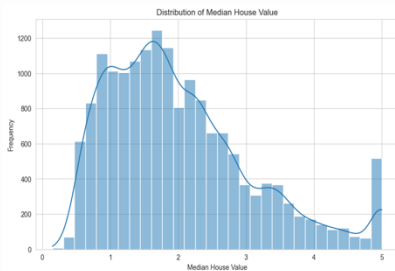
- Explainability can be the focus on a few different areas of a given experiment, such as the input data, the weights and features, or even the algorithms
- A diverse set of **datasets** (tabular, text, image), **models** (discriminative, generative, open-source, closed-source), and **interpretability 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

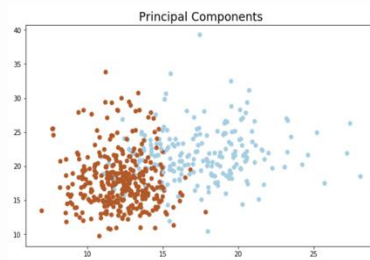
California Housing

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



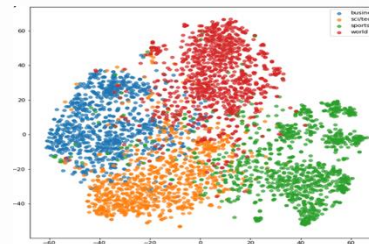
Wisconsin Breast Cancer

- 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



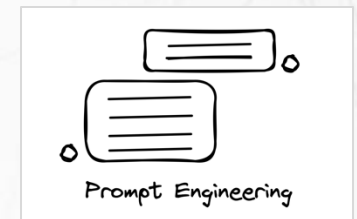
MNIST

- 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

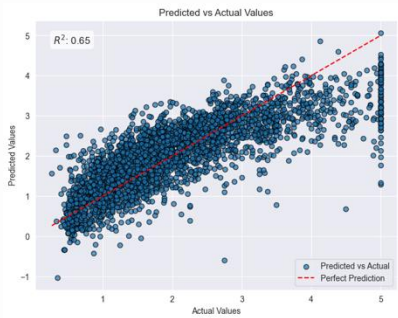


Models

California
Housing

Ridge
Regression

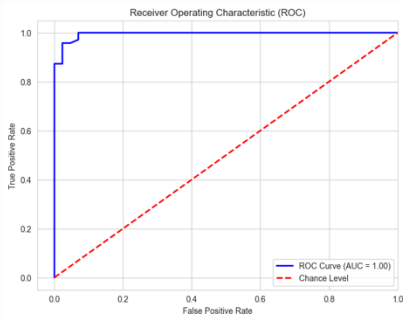
- Linear Regression Model
- L2 Regularization to reduce overfitting
- Predict median house price
- Feature weights available



Wisconsin
Breast Cancer

Random Forest
Classification

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



AG News

DistilBERT
Language Model

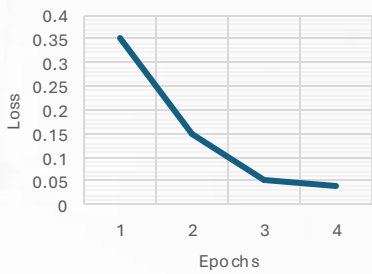
- Transformer-based Model
- Small-Medium language model
- Classifying news articles
- Contextual embeddings available

Feature	Value
Training Loss	0.44
Validation Loss	0.38
Accuracy	0.88

MNIST

Deep Learning

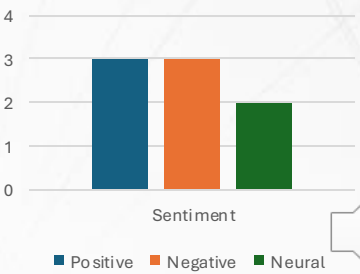
- Feed-Forward Neural Network
- Uses fully connected layers and ReLU
- Classifies hand-written digits
- Pixel-level contributions available



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

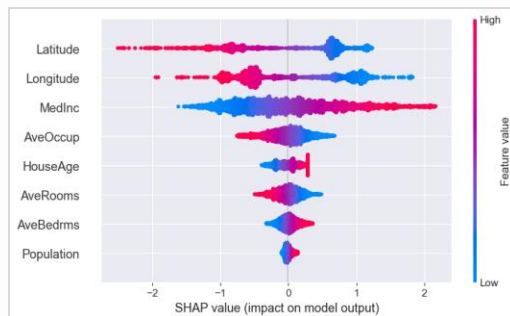
GPT-4o LLM

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

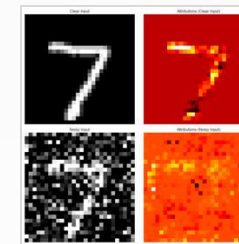
AG News (TLA)

Legend: ■ Negative □ Neutral ■ Positive			
True Label	Predicted Label	Attribution Label	Attribution Score
0	LABEL_0 (0.98)	LABEL_0	3.33

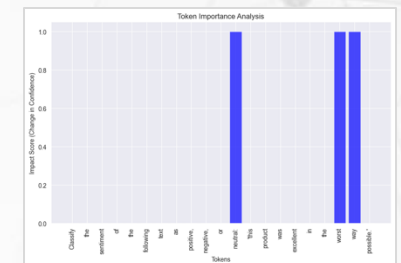
Word Importance

[CLS] rest ##five 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]

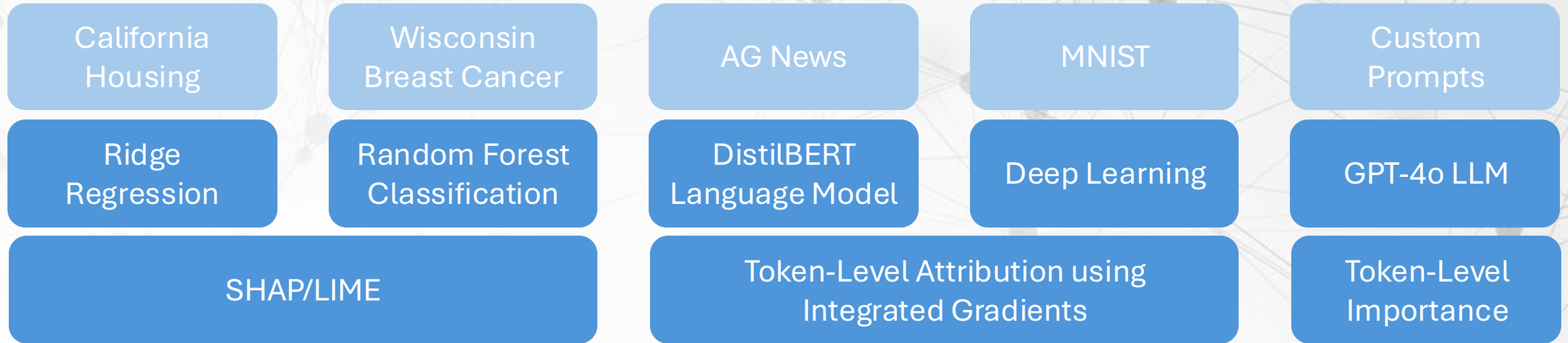
MNIST (TLA)



Sentiment (TLI)



Results



- **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)

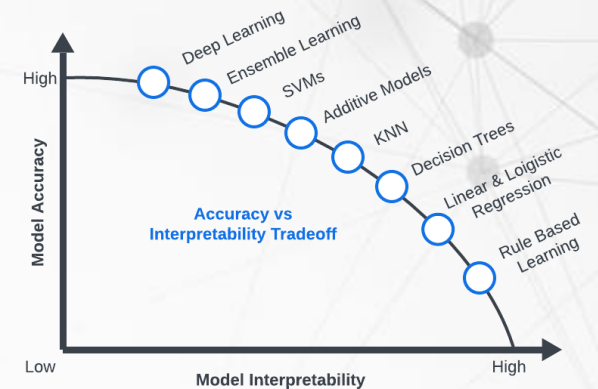
- **Integrated Gradients** revealed which tokens in AG News contributed most to classification decisions
- **TIA** Identified tokens that overly influenced predictions helped pinpoint potential biases in the dataset or model
- **Integrated Gradients** visualized the specific pixels in MNIST critical 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



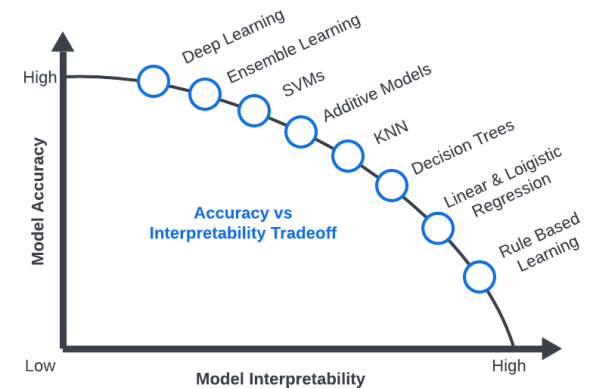
Conclusion

- **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

CS 6140 Machine Learning Project:

```
In 1: # I needed to disable col to download the dataset
2: import os
3: os.environ["XDG_CACHE_HOME"] = os.path.expanduser('~/.cache')
```

Table of Contents:

1. Base Model of Implementation
2. Ridge Regression using California Housing Data (LIME, SHAP)
3. Random Forest Classification using Breast Cancer Data (LIME, SHAP)
4. Open-Source GenAI Classification using AG News (Attributions)
5. Closed-Source GenAI Classification using Prompting (Masking)
6. Random Forest Classification (Counterfactuals)
7. Deep Learning Classification using MNIST

1. Base Model

```
In 10: import os
11: from sklearn.metrics import mean_squared_error, r2_score
12: from sklearn.linear_model import LinearRegression
13: from sklearn.preprocessing import StandardScaler
14: from sklearn.model_selection import train_test_split
15: from sklearn.datasets import load_breast_cancer
16: from sklearn.metrics import r2_score
17: from sklearn.metrics import mean_squared_error
18: from sklearn.metrics import r2_score
19: from sklearn.metrics import mean_squared_error
20: from sklearn.metrics import r2_score
```

Full Codebase

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

Saleh Alkhalifa
Northeastern University

Abstract

In recent years, the prevalence of AI models within the life sciences, biopharmaceuticals, and other regulated environments has driven significant advancements across domains such as regulatory, manufacturing, process development, and quality control. However, as both discriminative and generative models grow more sophisticated, their decision-making processes often become opaque, leading to the so-called black-box problem. This project focuses on developing and evaluating various distinct AI models covering both regression, classification, through discriminative AI, generative AI, and deep learning, and applying explainability techniques to each. By employing methods such as SHAP, LIME, Integrated Gradients, and Counterfactual Explanations, we provide insights into how these models can be interpreted. The ultimate goal is to enhance transparency, trustworthiness, and ethical compliance of AI systems operating in sensitive and regulated settings.

1. Introduction

Artificial Intelligence (AI) and Machine Learning (ML) technologies have become integral tools across a variety of sectors, including life sciences, biotechnology, biopharmaceuticals, healthcare, finance, manufacturing, and many more. Despite their proven success, a fundamental challenge persists within these models, which is the fact that many AI models operate as "black boxes" making it difficult for practitioners, developers, regulators, and end-users to understand how predictions and decisions are actually made. This lack of transparency poses significant obstacles to trust, ethical deployment, and regulatory compliance in high-stakes scenarios such as those focusing on patient outcomes. Explainability in AI is designed to address these issues by providing insights into model behavior, revealing which factors drive predictions, and ensuring that stakeholders can evaluate the fairness, reliability, and accountability of these systems. As the complexity of models grows—from simple linear regression to large language models—so does the need for robust explainability techniques.

1

2. Data

Multiple datasets were utilized to explore explainability across different types of models and tasks. • **California Housing Dataset (Regression):** Used to predict median house values, offering a continuous target variable and structured input features.

Figure 1: A visualization depicting the distribution of the California Housing Dataset.

Detailed Report

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

Presenter: Saleh Alkhalifa, Senior Manager of Data Science

Summary Presentation



References

- [1] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing 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. *International 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 Explainability Toolkit. *arXiv preprint arXiv:2104.12717*.
- [8] Lin, Z. Q., et al. (2019). Do Explanations Reflect Decisions? A Machine-centric Strategy to Quantify the Performance of Explainability Algorithms. *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 Localization. *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 Neural Information Processing Systems (NeurIPS)*.