The Growing Dependence of Explainable AI: Techniques, Applications, and Model Interpretations with SHAP

Prepared for

Dr. John Anderson

Comp 3190: Introduction to Artificial Intelligence

By

Ian Maxwell

Student, [Student ID removed]

Computer Engineering, University of Manitoba

Final Project

April 4th, 2025

0. Abstract

As artificial intelligence (AI) systems continue to grow, influencing decision-making across various industries, the demand for transparency and interpretability in machine learning models has never been more critical. This project explores the significance of Explainable AI (XAI) in addressing the growing need for model interpretability, focusing specifically on SHAP (SHapley Additive exPlanations) values for interpreting complex black-box ML models. We first explore explainable AI, demonstrating its importance, application and benefits. We highlight the different XAI techniques and then zone on a mathematically supported fair technique called SHAP. After exploring the mathematical background of SHAP we demonstrate its implementation. We provide local and global insights into model predictions, offering clear visualizations such as waterfall and beeswarm plots to enhance understanding. Through this implementation involving a regression model predicting vehicle miles per gallon (MPG), we show how XAI techniques can uncover the relationships between features and outcomes, ensuring better trust, accountability, and fairness in AI-driven decisions. The findings underscore the importance of explainability in fostering user confidence and guiding further advancements in AI deployment, highlighting XAI's growing role in industries ranging from healthcare, cybersecurity and autonomous vehicles.

1. Introduction - What is explainable AI (XAI)

Explainable Artificial Intelligence (XAI) is an increasingly vital subfield of AI dedicated to making the decision-making processes of machine learning models transparent and comprehensible to humans. XAI aims to uncover how these AI models work and why

they produce the outcomes they do. In essence, XAI comes down to offering humans interpretability, functionality, and understanding of artificial intelligence systems. As AI systems scale their usage of complex "black box" models such as deep neural networks or ensemble methods, their internal workings become opaque, obscuring how inputs produce their respective outputs. These black box models often outperform white box models in terms of prediction accuracy, especially on complex tasks like image recognition and natural language processing (NLP), however, they lack transparency and interpretability [Arrieta et al., 2020]. Thus, “the key limitation of today’s intelligent systems is their inability to explain their decisions and actions to human users. A lack of explainability hampers our capacity to fully trust AI systems” [Mehta et al., 2023]. Fortunately, these black-box model challenges have nourished an environment which both benefits and necessitates the development of XAI. Our ability to trust AI stems directly from the depth of transparency, interpretability and explainability to which XAI can provide.

2. Applications & Importance of XAI

The importance of XAI application stems from both practical and ethical imperatives. Although any black-box system can benefit from XAI integration, the significance varies throughout different fields and applications. In particular, “XAI is especially important for risk-sensitive applications, such as security, clinical decision support, or autonomous driving” [Mehta et al., 2023].

XAI's potential in healthcare cannot be understated. "It goes without saying that AI models' recommendations to help clinicians categorize crucial diseases using structured parameters or unstructured data such as medical imaging have far-reaching implications. If an AI system predicts and also explains why it came to that conclusion, it will be far more beneficial than if it predicts and then allows clinicians to spend an equal amount of time (with or without AI judgments) determining whether the AI system's decision is accurate and trustworthy. Because lives are on the line in healthcare, XAI is critical" [Aluvalu et al., 2024].

Furthermore, XAI can also be used to fight the battle against malware-infected systems. "A recent study reported that over 33,412,568 unique malicious files were detected in 2020 with an average revenue loss of $2.6 million per organization [Li et al,. 2021]. Clearly, there is an urgent need to develop efficient malware detection techniques to enable trustworthy computing" [Pan and Mishra, 2023]. Machine Learning (ML) algorithms are widely used in cybersecurity vulnerability detection due to its outstanding performance in both supervised and unsupervised scenarios. The flexibility of ML models also enables their different variations to be successfully employed in diverse applications. Since most ML algorithms are implemented in the software layer, they are also vulnerable toward malicious attacks [Pan and Mishra, 2023].

The third crucial domain of XAI application is autonomous driving. Machine learning vehicle positioning problems in particular have key ethical implications necessitating trust and comprehension of artificial intelligence. Despite the remarkable performance of

machine learning on the vehicle positioning problem, there is the requirement of transparency and a higher level of accountability from the machine learning-based system design. Explanations for machine learning model's decisions and estimations are thus needed to justify their reliability. This requires greater interpretability, often requiring an understanding of the mechanism underlying the operation of the algorithms. Unfortunately, the black box nature of neural networks is still unresolved, and many estimates are still poorly understood. "Commonly, the eXplainable Artificial Intelligence (XAI) procedures consist of ensemble runs, random sampling and Monte-Carlo simulations, which are quite common methods in engineering. XAI may comprise of a systematic perturbation of some components of the model, which enables it to observe how it affects the model's estimates, mostly using sensitivity analysis. Due to the safety critical nature of the autonomous vehicle navigation systems, interpretability of the vehicular navigation machine learning-based models is necessary and provides sufficient argument for the suitability of a model for use on the road as well as sufficient argument when communicating anomaly behaviours to insurance stakeholders, Original Equipment Manufacturers (OEM) and other relevant stakeholders" [Onyekpe et al., 2023]. Onyekpe outlines that XAI's thorough communication to all those involved in autonomous driving processes is paramount. Through healthcare, security and autonomous driving, XAI presents a theme of extreme importance in domains where it is directly responsible for lives at stake. As technology advances XAI must be utilized as a tool of transparent, efficient, clear and thorough communication allowing for responsible employment of artificial intelligence in these pivotal applications.

Although healthcare, security, and autonomous driving are prosperous domains for the implementation of XAI, its potential reaches into all corners of AI. This includes but is not limited to running in people's smartphones to do various tasks, in banks to manage investment and loan decisions in law enforcement to help officials recover evidence and make law enforcement easier, in the military of many countries, in insurance organizations to determine risk. Moreover, many organizations are actively trying to integrate AI and XAI into their workflows due to its remarkable performance, which competes with human performance in a wide variety of tasks [Ali et al., 2023]. Ironically, XAI can also be used to improve artificial intelligence. XAI aids data scientists in debugging models by revealing unexpected feature influences or biases. The importance of XAI application stems from both practical and ethical imperatives. When the stakes are high, trust in AI predictions is paramount.

3. Benefits of XAI

The benefits of XAI can be broken down into three main categories: trust, model improvement and data exploration [Sullyds, 2023]. The artificial intelligence model's explanations and justifications for its conclusion(s) allow us to distinguish whether the basis for the judgement is rational and reasonable. Verification of a model's rationale garners trust in the AI model's capabilities. This trust is preeminent for safety, reassurance and reliability upon implementation. In high-stakes domains mistakes can be costly thus, the transparency and trust derived from XAI is priceless. Ethically, this transparency ensures accountability, especially as regulations like the European Union's GDPR enshrine a "right to explanation" for automated decisions [Casey et al., 2019].

Secondly, the explanations allow for better model improvement and debugging. If the black-box model produces ambiguous outputs, explainable AI can highlight the exact issue that is responsible. By outlining the root issue, XAI reduces the debugging time spent training the model. For example, in image recognition or vision systems, should the model become dependent on certain undesired variables such as the background colour, XAI can use pixel importance heat mapping to highlight these background input pixels as very significant [Aluvalu et al., 2024]. This visual explanation technique makes the model's issue evident and consequently, can be efficiently fixed.

Lastly, XAI can be a tool for data exploration. XAI fosters user adoption by demystifying AI, making it a cornerstone of responsible deployment in real-world systems. Oftentimes, the explanations provided by XAI will discover unforeseen patterns used by the model in the training datasets or unforeseen connections between seemingly independent variables [Sullyds, 2023]. We can learn from these discoveries and explore their patterns and connections further. Instead of machines learning from humans, humans can learn new methods from artificial intelligence machines!

4. Categories, Techniques and Methodologies of XAI

There are many techniques by which XAI can attempt to interpret an AI model. How a model can be interpreted differs based on the model's complexity and scope. Therefore, approaches within model interpretability differentiate between local and global decisions and model-specific or model-agnostic decisions. Local interpretability explains a single step or decision a model takes in its overall decision-making process, whereas global

interpretability explains all steps. Model-specific interpretability provides explanations for a specific type of machine learning model, whereas model-agnostic interpretability provides an explanation for any type of machine learning model. One can think of model-agnostic interpretability as a generalized approach to explaining a model [Pan and Mishra, 2023].

Explainable AI methods can be divided into six broad categories: model interpretability, knowledge extraction, saliency maps, integrated gradients, Shapley value analysis, and layer-wise relevance propagation [Pan and Mishra, 2023]. Although this encompasses most of XAI, a few items are missing from this list so we will generalize these into the six XAI categories that follow:

1. Knowledge Extraction: Rule-based or simplified model distillation.
2. Counterfactual Explanations: Input perturbation for "what-if" insights.
3. Intrinsic Interpretability: Simple, transparent models (e.g., linear regression, decision trees).
4. Feature Attribution: SHAP, PFI, and other importance-ranking methods.
5. Local Surrogate Models: LIME and similar approximations.
6. Gradient-Based Attribution: Integrated gradients, saliency maps, LRP (grouping neural network methods).

Knowledge Extraction: Rule-based or simplified model distillation

Knowledge extraction is about pulling out understandable insights from complex systems. “When first extracted, this knowledge will be in a form that machines can

understand. Therefore, it is necessary to convert knowledge into a form that humans are able to comprehend. There are two sub-approaches within knowledge extraction: rule extraction and model distillation” [Pan and Mishra, 2023]. They involve creating simple logical rules (like "if this, then that") or distilling a complicated model into something easier to grasp. For example, taking a complex AI model and summarizing its behavior into a few key patterns or decisions that humans can follow. .

Counterfactual Explanations: Input perturbation for "what-if" insights

Counterfactual Explanations are about answering "what if" questions by tweaking inputs to see how outcomes change. Imagine changing one thing, like a word in a sentence, and seeing how it flips the result. It helps explain a model’s decision by showing what would need to happen for a different outcome, making the reasoning clearer [Aluvalu et al., 2024].

Intrinsic Interpretability: Simple, transparent models

Intrinsic interpretability focuses on using models that are naturally easy to understand from the start, like linear regression (where each factor has a clear weight) or decision trees (where choices follow a simple flowchart). The idea is that the model itself is transparent, so you don’t need extra tools to figure out what it’s doing. Like the name suggests, intrinsic interpretability can be fairly simply deciphered into a whitebox model and thus XAI implementation is straightforward.

Feature Attribution: SHAP, PFI, and other importance-ranking methods

This category ranks which parts of the input or "features" matter most to a model's decision. It includes methods like SHAP (SHapley Additive exPlanations) or PFI (Permutation Feature Importance) to calculate how much each input, like a word, number, or pixel contributes to the output. It allocates each input a responsibility to the output's result [Aluvalu et al., 2024. Later in this project we will explore an implementation of SHAP.

Local Surrogate Models: LIME and similar approximations

Local surrogate models such as LIME (Local Interpretable Model-agnostic Explanations) involve taking a complex model and creating a simpler "stand-in" model to explain a single prediction. These XAI methods zoom in on one example such as why an image was labeled "cat", and build an easily understandable approximation for that specific case. These smaller simplified models shed light on the bigger model's logic [Aluvalu et al., 2024].

Gradient-Based Attribution: Integrated gradients, saliency maps, LRP

This group of XAI is mostly for neural networks and uses math (like gradients) to track how inputs affect outputs. "Saliency maps are often used in the computer vision domain for images, where the gradients for the loss function are used to determine the important pixels in the image to the model's overall decision" [Pan and Mishra, 2023]. The integrated gradients or saliency maps methods essentially highlight which parts of the input "light up" the model's decision . Layer-wise Relevance Propagation (LRP)

does something similar by working backward through the network. It is similar to a heat map of importance [Pan and Mishra, 2023].

Method Comparison

Clearly, there are a wide range of XAI techniques. This begs the question of which methodology is most appropriate for implementation. Common forms of XAI include both intrinsic and post-hoc approaches. Intrinsic methods, such as decision trees, are naturally interpretable but limited in scope. Post-hoc techniques, like SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Permutation Feature Importance (PFI), apply to pre-trained models, offering flexibility across model types [Aluvalu et al., 2024]. SHAP, rooted in game theory, provides consistent feature attributions globally and locally, while LIME approximates local behavior with simpler surrogate models, whereas PFI measures feature impact via perturbation. These methods vary in scope, global explanations reveal overall model logic, while local ones clarify individual predictions.

This project's implementation centers on SHAP values due to the theoretical elegance and practical utility. We will later show how SHAP has mathematical theory backing it's utility as a tool to distribute fairly. Comparing SHAP implementation aims to evaluate their interpretability, efficiency, and robustness, contributing to a deeper understanding of XAI's potential. In doing so, it addresses the pressing need to refine tools that make AI not just powerful, but comprehensible. Before we delve into SHAP implementation it

is essential to first understand explainable AI's (XAI) history leading up to its development and SHAP's theoretical background.

5. SHAP & Shapley background

SHAP, short for SHapley Additive exPlanations, is a method created by Lundberg and Lee in 2017 to explain individual predictions. SHAP is based on the game-theory optimal Shapley values. However, To understand why SHAP is a thing and not just an extension of the Shapley values we must explore a bit of history. In 1953, Lloyd Shapley introduced the concept of Shapley values for game theory. In the context of explaining machine learning, Shapley values were suggested for the first time by Štrumbelj and Kononenko in 2011, however, they didn't become so popular. It wasn't until a few years later, in 2017, that Lundberg and Lee proposed SHAP, which was basically a new way to estimate Shapley values for interpreting machine learning predictions, along with a theory connecting Shapley values with LIME and other post-hoc attribution methods, and a bit of additional theory on Shapley values [Molnar, 2025].

Although it appears that SHAP is just a rebranding of Shapley values (which is true), it is important not miss the fact that SHAP also marks a change in popularity and usage of Shapley values, and introduced new ways of estimating Shapley values and aggregating them. Interestingly, SHAP also brought Shapley values to all machine learning blackboxes, including text and image models [Molnar, 2025].

6. Explaining SHAP

Before understanding SHAP we will first understand the Shapley value method first introduced by Lloyd Shapley in 1953. Shapley values are a method from coalitional game theory to fairly distribute the total gains or losses among a group of players who have collaborated. In essence, Shapley values tell us how to fairly distribute the "payout" among the features. In game terms, the features could represent players and in machine learning the features could represent inputs to the model.

The Shapley values formula is below in Figure 1.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

Figure 1. Shapley Value Calculation [Lundberg and Lee, 2017].

The variables in Figure 1 are defined as the following:

ϕ_i : The Shapley value ϕ of feature i.

i: an arbitrary feature

F: The set of all features

S: A subset of F of features except feature i

f(x): The model or game's function mapping features to outputs (called the value of coalition x)

$\sum_{S \subseteq F \setminus \{i\}}$:The summation over all subsets S

$\frac{|S|!(|F| - |S| - 1)!}{|F|!}$:Weighting based on the number of permutations of this feature coalition

$\left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$:The marginal contribution of the feature i. (Difference with feature i vs without feature i in the coalition)

It is worth noting that summation over |F|! in the divisor can be understood as a means of averaging the formula.

The Shapley value formula can be repeated for each feature i. This will result in a list of Shapley values for each input feature. The Shapley values map to Real numbers and can be positive, zero or negative. Shapley value explanations are not to be interpreted as local in the sense of gradients or neighborhood [Bilodeau et al., 2024]. For example, a positive Shapley value doesn't mean that increasing the feature would increase the prediction. Instead, the Shapley value has to be interpreted with respect to the reference dataset that was used for the estimation [Molnar, 2025].

To understand and contextualize the true representations that are Shapley values it is helpful to use an example. Suppose a group of up to four people were to build a startup company and then sell it for profit. The Shapley values are a method to distribute the profiting money. Those that were more responsible for the company's success would have higher Shapley values and thus receive more of the payout than those that contributed less. If someone were to have caused more harm than good then they would receive a negative Shapley value. Before utilizing Shapley values, we would train our model to predict the profits based on the combination of people who did and didn't join the team i.e. the inputs. We could then assign each individual a Shapley value based on the Shapley formula, utilizing different combinations of people as different coalitions for calculations, where each individual is a feature i in the Shapley formula.

Fortunately, in 2017 Lundberg and Lee developed SHAP (SHapley Additive exPlanations) to expand this coalition game theory mathematical equation into interpreting machine learning. SHAP, however, is uniquely focused as a method to explain individual predictions. In contrast to Shapley values, the goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. SHAP maintains its theoretical origin by computing Shapley values from coalitional game theory, however, the model is expanded upon to include feature values of a data instance as players in a coalition. Shapley values tell us how to fairly distribute the prediction's output or "payout" among the input features. A player can be an individual feature value, e.g., for tabular data or even a group of feature values. For example, to explain an image, pixels can be grouped into superpixels, and the prediction distributed among them. One innovation that SHAP brings to the table is that the Shapley value explanation is represented as an additive feature attribution method, a linear model. Interestingly, this view connects LIME and Shapley values [Molnar, 2025].

8. SHAP Strengths

The Shapley value is the only attribution method that satisfies the properties Efficiency, Symmetry, Dummy, and Additivity, which together can be considered a definition of a fair payout [Molnar, 2025].

The efficiency principle dictates that the sum of all players' Shapley values equals the full coalition value. In the context of SHAP implementation, The efficiency principle dictates that the sum of all players' Shapley values equals the difference between the prediction for x and the average/expected prediction. The symmetry principle dictates that if two players have the same contribution values for all coalitions S not containing the players i and j, their Shapley values are also equal. The Dummy principle states a feature j that does not change the predicted value regardless of which coalition of feature values it is added to should have a Shapley value of 0. Lastly, the Additivity principle states for a game with combined payouts v+w, the Shapley value of player j is the sum of their Shapley values in the individual games v and w [Molnar, 2025].

In SHAP, the difference between the prediction and the average prediction is fairly distributed among the feature values of the instance the Efficiency property of Shapley values. This efficiency property distinguishes the SHAP method from other methods such as LIME. LIME does not guarantee that the prediction is fairly distributed among the features. The Shapley values deliver a full explanation of this distribution. The Shapley value allows contrastive explanations. Instead of comparing a prediction to the average prediction of the entire dataset, you could compare it to a subset or even to a single data point. This contrastiveness is also something that local models like LIME do not have. The Shapley value is the only explanation method with a solid theory. The axioms, efficiency, symmetry, dummy, additivity, give the explanation a reasonable foundation. Comparatively, methods like LIME assume linear behavior of the machine learning model locally, but there is no theory as to why this should work [Molnar, 2025].

9. XAI and SHAP Weaknesses

Despite its promise, SHAP and XAI face significant challenges. A key tension exists between accuracy and interpretability: high-performing models often sacrifice clarity for complexity, while interpretable models may underperform [Wojciech et al., 2019]. Computational demands also pose hurdles; advanced methods like SHAP, though rigorous, can be resource-intensive, limiting scalability for large datasets or real-time applications [Aluvalu et al., 2024]. These challenges highlight the need to assess XAI techniques critically, balancing explanatory power with practical feasibility, a goal this study pursues through implementation and analysis.

In contrast, simpler models like linear regression offer intrinsic interpretability but often lack the predictive power needed for modern applications. XAI bridges this gap by developing techniques to elucidate model behavior, either through inherently interpretable designs or post-hoc explanations applied after training [Wojciech et al., 2019]. This project explores XAI through the lens of SHAP values, a prominent method, while evaluating its role alongside other interpretability techniques.

10. Methodology for SHAP implementation

To demonstrate the utility of SHAP for explainable AI it is best to see its implementation. We will utilize SHAP in python to attempt to explain a regression model which we will train on a csv file containing information regarding vehicles, and their features. In this implementation we utilize the Xgboost machine learning library and the SHAP python import and the [Lundberg and Lee, 2022]. We will also use common python libraries

such as pandas numpy matplotlib.pyplot and seaborn. The goal of this experiment is to apply SHAP (Shapley Additive Explanations) to a regression model predicting the miles per gallon (MPG) of vehicles. By using SHAP, we can gain insight into the individual contributions of each feature to the model's predictions, allowing us to explain the black-box nature of the trained XGBoost model. The CSV file was retrieved from a Waskom's dataset hosted on Github [Waskom, 2021]. The attributes/features of the csv file can be seen below as well as some sample data points.

	mpg	cylinders	displacement	horsepower	weight	acceleration	model_year	origin
0	18.0	8	307.0	130.0	3504	12.0	70	usa
1	15.0	8	350.0	165.0	3693	11.5	70	usa
2	18.0	8	318.0	150.0	3436	11.0	70	usa
3	16.0	8	304.0	150.0	3433	12.0	70	usa
4	17.0	8	302.0	140.0	3449	10.5	70	usa

Figure 2. CSV File outline

As seen in figure 2, the csv file contains vehicles with miles per gallon (mpg), engine displacement, horsepower, weight, acceleration, model year and country of origin as features. In total there are 398 vehicles with data in the dataset [Waskom, 2021]. We will train the model using this dataset and observe the relationships between the features and the target variable, MPG. Note for now that all features are continuous except the country of origin, we will end up adjusting for this categorical feature later.

To get an understanding of the dataset we see below a scatter plot of the mpg vs weight.

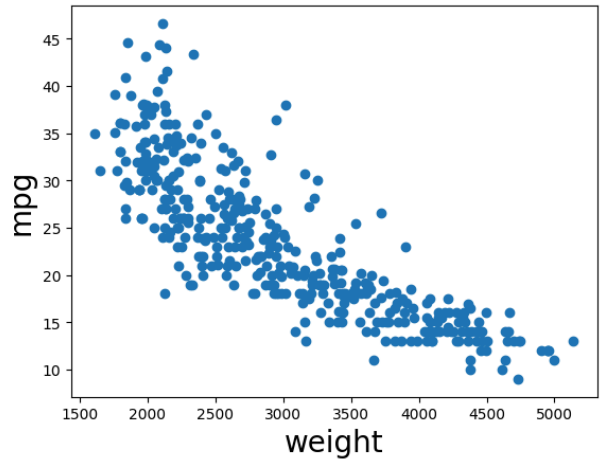

Figure 3. Miles per Gallon vs vehicle weight.

It can be clearly seen in figure 3 that the miles per gallon tend to decrease as the vehicle's weight increases. This intuitively makes sense.

```
data = pd.read_csv("mpg.csv", usecols=lambda column: column != "name")
```

```
y= data['mpg']
X = data[["cylinders", "displacement", "horsepower", "weight", "acceleration", "model_year", "origin"]]
X['origin.usa'] = [1 if s == 'usa' else 0 for s in X['origin']]
X['origin.japan'] = [1 if s == 'japan' else 0 for s in X['origin']]
X['origin.europe'] = [1 if s == 'europe' else 0 for s in X['origin']]
X=X.drop('origin', axis=1)
```

Figure 4. CSV file feature allocation to X and y for the model.

After reading our csv file, figure 4 shows, the features are separated into the inputs (X) and output (y), where y is the miles per gallon feature and input features are the rest. Since the country of origin feature is categorical with only 3 possibilities it is then split into three sub features noted as origin.usa, origin.japan and origin.europe and given binary values of 1 where it is the country of origin and 0 where not. Next, a XGB regression model that predicts country of origin is created and it is fit to the X and y, inputs and outputs.

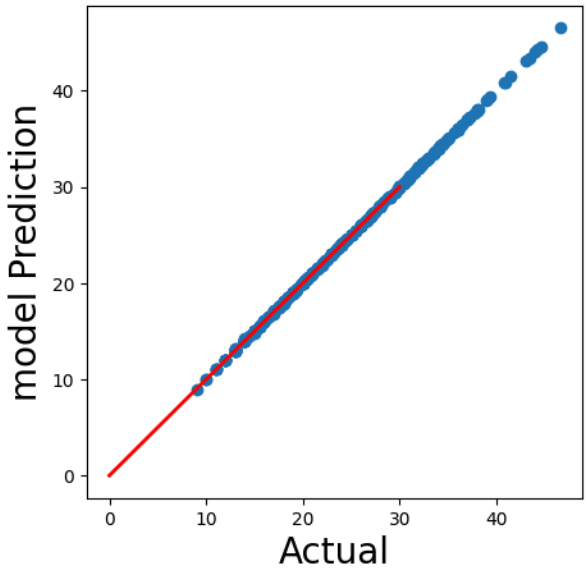

Figure 5. Regression model verification

We create and compare the prediction array with the actual data, plotting the results in Figure 5. We find that the results of the prediction model (red line) almost perfectly match the actual data points. The regression model thus appears to be properly trained.

We now utilize the SHAP python packages SHAP explainer. We view the results for the first and hundredth vehicle instances in the figure 6 SHAP waterfall plots.

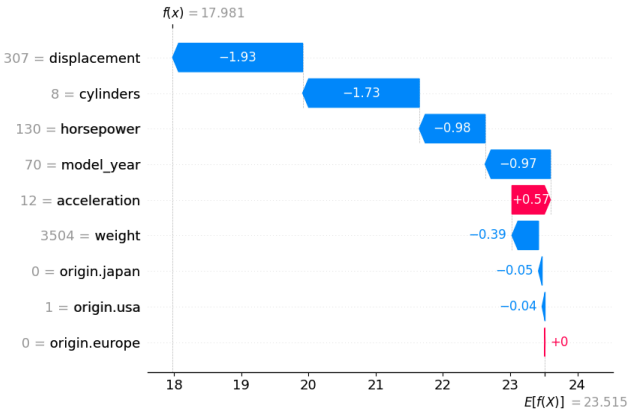

Figure 6. SHAP waterfall plots

In Figure 6, E[f(x)]=23.515 represents the average mpg prediction whereas f(x) represents the models predicted mpg. Each feature has a SHAP value that represents how the feature has contributed to the difference between f(x) and E[f(x)]. Note that the sum of figure 6's SHAP features sum to the difference between f(x) and E(F(X)]. This property is a key aspect to SHAP and remains true for any and all other vehicle instances.

These SHAP values are excellent explanations for local/individual instances, however, to get a more generalized interpretation we adjust our perspective to look over the dataset as a whole.

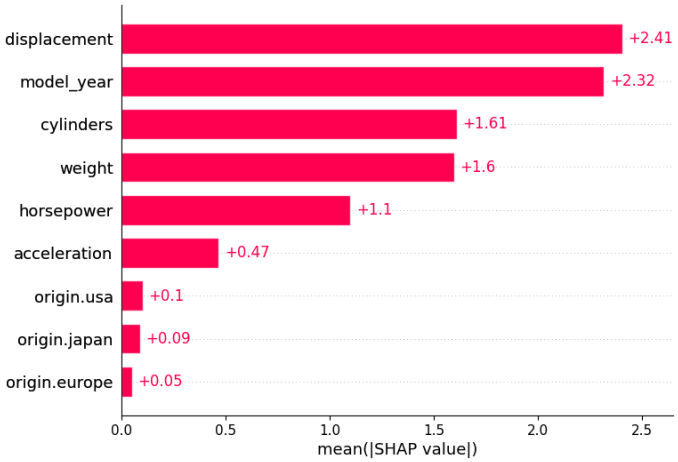

Figure 7. Mean SHAP plot.

Figure 7 demonstrates the mean of the absolute SHAP values for each vehicle feature across all testing instances. The interpretation from this graph can be a trend ranking the importance of features based on their influence on mpg. Here we find the engine displacement, then the model year, then the cylinders, then weight and so on represent the contribution significance of each feature. The engine displacement is the most

important feature to dictate the mpg prediction whereas the origin is the least significant feature. This XAI technique outlines the model's rationale which when verified logically makes sense as it is intuitive that larger engine displacements would most likely be associated with worse mpg as they tend to be used in larger vehicles which become less efficient. Thankfully, there are some more advanced SHAP plots that can help validate or reject these assumptions.

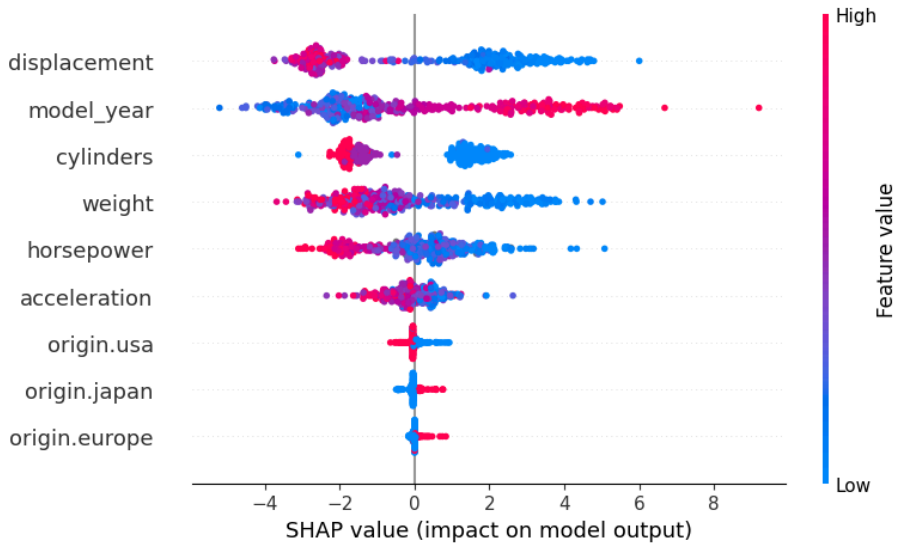

Figure 8. Beeswarm plot

In figure 8, we have three axes for each feature, the feature's name, the features SHAP value and the feature's value represented on a colour spectrum with red being relatively large values and blue as relatively small values. All vehicle testing instances for all features are plotted on this beeswarm plot. Essentially, this helps visualize all the testing data points. We can verify our previous assumption that large engine displacements decrease the mpg whereas low engine displacements increase mpg. The beeswarm plot confirms the validity of this assumption since the color distribution representing high vs low engine displacements has a clear separation between red (large engine

displacements) and blue (small engine displacements) across the SHAP value axis. Almost all of the red feature values have large negative SHAP values whereas almost all blue data points have large positive SHAP values. The clear separation between positive and negative SHAP values suggests that as the engine displacement decreases SHAP values increase and thus the mpg increases, validating our assumption. The centering of these SHAP values tend to cluster at decently elevated SHAP value magnitudes, highlighting the significance of the engine displacement on mpg. Interestingly, it is also worth noting other patterns such as that it appears that increasing the model year increases the SHAP value and thus the mpg. This also makes intuitive sense as we expect newer vehicles to have better mpg.

We can visualize the exact relationships we see in the beeswarm plot and check for linearity or exponentiality using

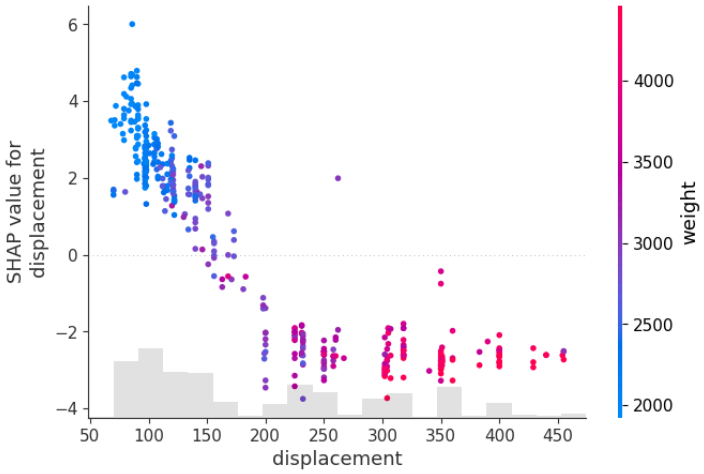

Figure 9. Dependence plot.

This scatter plot in Figure 9 outlines two relationships simultaneously. First, it shows that small engine displacements and small total vehicle weights are associated with large SHAP values. Secondly, we can see that the relationship between engine displacement

and SHAP values is not exactly linear as it has a curve similar to that of decreasing exponentials.

11. Results & Analysis

Through the use of SHAP values and multiple graph visualizations we have been able to interpret our model's function. The SHAP analysis reveals that engine displacement is the most influential feature in predicting MPG, with a clear negative relationship between engine size and fuel efficiency. As engine displacement increases, the predicted MPG decreases. Additionally, model year and vehicle weight also have notable influences on MPG, with newer vehicles and lighter cars being associated with higher fuel efficiency. One particular explanation we highlighted was the strong influence of the engine displacement having a decreasing exponential relationship with the miles per gallon. These XAI insights into the model's function are invaluable especially with more complex black-box models.

From the mean SHAP plot (Figure 7), we observe that engine displacement and model year are the most influential features, with larger engine displacements associated with lower MPG. This finding aligns with industry knowledge, as larger engines typically consume more fuel. Conversely, the country of origin, which was encoded as a categorical feature, had the least impact on the model's predictions, suggesting that geographical manufacturing differences have less of an effect on fuel efficiency.

The beeswarm plot (Figure 8) further supports our assumption that larger engine displacements result in lower MPG. The clear separation between red (large engine displacements) and blue (small engine displacements) indicates that higher engine displacement leads to lower SHAP values, thus predicting lower MPG. Additionally, the plot reveals an interesting trend where newer model years tend to correlate with higher SHAP values, indicating that more recent vehicles are likely to have higher MPG ratings."

While SHAP has provided valuable insights into the factors driving MPG predictions, the model could be further improved by incorporating additional features such as vehicle brand, type, or driving conditions. Furthermore, the impact of categorical features like 'country of origin' could be explored in greater detail, perhaps using alternative encoding techniques or domain-specific knowledge to refine the model. Additionally, exploring the model's performance in real-world scenarios could offer more practical insights into its applicability for real-time decision-making.

The SHAP analysis provides interpretable explanations for why certain vehicles are predicted to have better fuel efficiency than others. This can be useful for automotive engineers looking to design more efficient vehicles or for consumers seeking to choose a car with better MPG. Policymakers could also use this information to create regulations or incentives that encourage the development of more fuel-efficient cars.

Overall, SHAP has proven to be an effective tool for explaining the decisions of our regression model, providing clear, interpretable explanations for individual predictions and overall feature importance. By utilizing these insights, we can better understand the model's behavior and make more informed decisions based on its predictions.

12. Conclusion:

This project demonstrated the effectiveness of SHAP (Shapley Additive Explanations) in making a regression model that predicts vehicle MPG more interpretable. By using SHAP, we gained insights into the importance of features like engine displacement, model year, and weight in the model's predictions. Engine displacement emerged as the most significant factor, with larger displacements generally leading to lower MPG.

Through SHAP's visualizations, we explained individual predictions and understood how each feature contributes to the model's overall behavior. This approach enhances transparency, making machine learning models more understandable and trustworthy, especially in complex tasks like predicting fuel efficiency.

The findings also highlight the value of explainable AI (XAI) in providing clear, actionable insights. While the model performed well, future work could involve adding more features or experimenting with different algorithms to improve accuracy and explainability.

Interestingly, the project tied to key concepts from this Comp 3190 course including problem-solving techniques and searching problem spaces, as we navigated through the dataset to find meaningful patterns. The concept of knowledge representation was

evident in how we structured the data and presented model outputs. Additionally, the use of SHAP links to both symbolic and subsymbolic AI. While the machine learning model itself is a subsymbolic system, SHAP's output provides a symbolic, human-readable explanation of the model's behavior.

This project not only highlighted the power of SHAP in explaining machine learning models but also demonstrated the importance of making complex, black-box models more transparent and accessible. It reinforced the value of explainable AI in ensuring that machine learning models are not just accurate but also understandable, which is crucial for trust and accountability in real-world applications.

13. Bibliography:

[Onyekpe et al., 2023] Onyekpe, U., Lu, Y., Apostolopoulou, E., Palade, V., Umo, E., and Kanarachos S., "Explainable Machine Learning for Autonomous Vehicle Positioning Using SHAP". *Explainable AI: Foundations, Methodologies and Applications,* 232:1:157-184, 2023.

[Ali et al., 2023] Ali, M., Javed, A. R., Khan, Z., Batool, A., Srivastava, G., and Gadekallu, T. R., "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence". *Future Generation Computer Systems*, 141:280–299, 2023.

Retzlaf et al., 2024] Retzlaf, R., Häußler, S., and Morik, K., "Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists". *Cognitive Systems Research*, 86:1-22, 2024.

[Ahmed et al., 2022] Ahmed, A., Kholidy, H.A., Raza, A., and Elhoseny, M., "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where". *IEEE Transactions on Industrial Informatics*, 18(5):3500–3516, 2022.

[Aluvalu et al., 2024] Aluvalu, R., Polinati, S., and Sri, S., *Explainable AI in Health Informatics*. Springer-Verlag, Heidelberg, 2024.

[Pan and Mishra, 2023] Pan, L., and Mishra, M., *Explainable AI for CyberSecurity*. Springer-Verlag, Heidelberg, 2023

[Wojciech et al., 2019] Wojciech, B., Samek, W., and Müller, K., *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer-Verlag, Heidelberg, 2019.

[Casey et al., 2019] Casey, B., Farhangi, A., and Gasser, U., "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise". *Berkeley Technology Law Journal*, 34(1):145–190, 2019.

[Arrieta et al., 2020] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI". *Information Fusion*, 58:82–115, 2020.

[Li et al., 2021] Li, Y., Huang, J., Wang, W., and Ren, J., "Arms Race in Adversarial Malware Detection: A Survey". *ACM Computing Surveys*, 54(7):1–36, 2021.

[Lundberg and Lee, 2017] Lundberg, S.M., and Lee, S., "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems (NeurIPS-17)*, 30:4765–4774, 2017.

[Lundberg and Lee, 2022] Lundberg, S.M., and Lee, S., "An Introduction to Explainable AI with Shapley Values". *SHAP Documentation*. Available at: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html, 2022.

[Molnar, 2025] Molnar, C., *Interpretable Machine Learning*. Self-published, 2025.

[Bilodeau et al., 2024] Bilodeau, B., Jaques, N., Koh, P.W., and Kim, B., "Impossibility Theorems for Feature Attribution". *Proceedings of the National Academy of Sciences*, 121(2):e2304406120, 2024.

[Mehta et al., 2023] Mehta, M., Palade V., Chatterjee I., *Explainable AI: Foundations, Methodologies and Applications,* 232:1:V (preface) -10, 2023.

[Sullyds, 2023] Sullyds, C. SHAP values for beginners | What they mean and their applications, *A Data Odyssey*. Available at: https://www.youtube.com/watch?v=MQ6fFDwjuco 2023.

[Waskom, 2021] Waskom, M. (2021) *Seaborn data repository: mpg dataset*. Available: https://raw.githubusercontent.com/mwaskom/seaborn-data/refs/heads/master/mpg.csv (Accessed March & April 2025).