Evaluating Explainable Artificial Intelligence: Algorithmic Explanations for Transparency and Trustworthiness

Utsab Khakurel and Danda B. Rawat

Department of Electrical Engineering and Computer Science
Howard University, Washington DC 20059, USA

ABSTRACT

Explainable AI (XAI) is the capability of explaining the reasoning behind the choices made by the machine learning algorithm. It is the domain that maintains the transparency of the decision-making capability of the machine learning algorithm. Comparing the decision-making process of the machine with humans can further elaborate. Humans make thousands of decisions every day in their lives. Every decision an individual makes, they can explain the reasons behind why they made the choice. Nonetheless, it is not the same in the case of machine learning. Furthermore, Explainable AI was non-existent until suddenly the topic was brought forward and has been one of the most relevant topics in AI as of now. Explainable AI tries to provide maximum transparency to a machine learning algorithm by answering questions about how models effectively come up with the output. Machine learning models with Explainable AI will have the ability to explain the rationale behind the result, understand the weaknesses and strengths in the model, and be able to see how the model will behave in the future. In this paper, we will investigate Explainable AI for algorithmic trustworthiness and transparency. We will evaluate Explainable AI using some example use cases. The paper will also demonstrate the use of the SHAP library and visualize the effect of features individually and cumulatively in the prediction process.

Keywords: Artificial Intelligence, Explainable AI, Interpretability, Transparency, Trust, Ethics

1. INTRODUCTION

In modern times, machine learning has found it's new success and reached greater heights. With the availability of faster processing units, large chunks of data, and cheap data storage, machine learning has been able to achieve a lot in the fields of medicine, education, autonomous transportation, criminal justice, financial risk analysis, and many more. However, the more advanced machine learning has become, the harder it has been for humans to better understand, explain and control them. The effectiveness of these systems is limited by machines inability to explain their decisions to humans. We mostly judge the predictions and results of the model but never consider how those results were obtained and on what basis. We do not know what features contributed to the prediction or which single factor made a big impact on the decision. The lack of explainability and transparency in Machine Learning algorithms supports the rhetoric among the crowd who have doubts about the technology. This ultimately leads to users not completely trusting the AI systems.[1,2]

Explainable AI is an emerging field in the machine learning realm.[3,4] The aim of Explainable AI is to explain the decisions made by machine learning models. Most of the complex prediction models that are in use in the field are Black Box models giving users no idea of what the algorithm is actually doing.[2,5–7] In recent times, there have been cases of bias present in the machine learning algorithm. One of the most controversial cases is related to law and justice. The algorithm, called COMPAS, helps decide whether the defendant is too dangerous to be released on bail or not. The algorithm was found to be biased against black defendants according to the research done by ProPublica.[8] It has also been proved that machine learning algorithms can be easily fooled into misclassifying images with minor perturbations in the pixel through one-pixel and multi-pixel attacks.[9] Due to the inability to see through the decision-making process of the algorithm, these issues have been emerging with unexpected impacts in the real world.

Utsab Khakurel: E-mail: utsab.khakurel@bison.howard.edu, Telephone: 1 202 660 3139
Danda B. Rawat: E-mail: danda.rawat@howard.edu, Telephone: 1 202 806 2209

Machine learning algorithms have come a long way from Regression models, Decision Trees to Support Vector machine, Ensemble, and Neural Network. Although the accuracy of the prediction algorithms has been progressive, the explainability of the algorithms has been reduced hugely, causing a larger gap between users and the use case of an algorithm. Transparency is very vital in the context of machine learning because it can provide insight and a good understanding of the model and it's decision-making. Instead of completely agreeing with the model for its decision without a question, the machine's reasoning should be questioned so that we can simplify and better understand complex algorithms.

Explainable AI can be achieved by understanding simpler machine learning models like linear models or decision trees first. In addition, complex models can be further investigated using a form of layer to interpret the decision and features better. Decision tree models are very easy to understand as they are basically splitting the data recursively into smaller parts using features to distinguish them. The results produced can have an explanation. The ensemble of decision trees creating a Random Forest Classifier will produce a better result using uncorrelated decision trees. However, it is harder to explain the logic behind the prediction resulting from Random Forest.

Humans are worried about AI systems taking over the workforce and the entire human race. These horror stories are mentioned and portrayed as fictional works too. As we better understand algorithms, we will be better at creating algorithms, which can only make good decisions, ultimately avoiding the ever dreadful AI apocalypse.

The advantages of Explainable AI can be numerous. As we better understand machine learning algorithms, it is easier for us to accept the system and fully trust them. The increase in trust in AI systems will result in faster widespread adoption of the technology. Furthermore, in some critical areas like medicine, it is absolutely necessary for the system to be fully trusted. Enhanced understanding can produce better AI models which can ultimately help save more lives. This can also boost impact in the business or research field, helping us make better decisions.

This paper discusses the Explainable AI systems and what tools can be used to explain the rationale behind the predictions made by the predictive model. The remainder of the paper is organized as follows. Section II presents the overview of XAI and the relevant work on the topic. Section III discusses the data set, tools, the classifier used, and the approach taken to explain the model. Section IV focuses on the experiments done and presents the findings. Finally, conclusions are presented in Section VI.

2. OVERVIEW AND PROBLEM STATEMENT

As the interest in Artificial Intelligence has grown over the years due to the availability of abundance of data and processing capability, the performance of the machine learning systems has also significantly improved. This advancement was achieved through complex deep learning models designed to mimic the human brain. However, as the model's predictive power has grown, the questions regarding the trust, transparency, and ethics of these models have been raised. There have been several cases of machine learning system being biased and prejudiced based on race, gender and other sensitive human features. Explainable AI addresses these concerns by generating human understandable explanations improving trust, transparency, and identifying reasons behind the biased predictions of the model.[10]

Explainable AI is largely defined by how well the model is able to explain the reasoning behind the results it predicts. Explainable AI answers 'wh' questions such as 'why', 'where', 'when', etc.[3] Explainable AI model extracts information or metadata from the data set to answer questions like "Which feature had the most impact on the prediction?", "Why model predicted this and why not else?", "What are the correlation between this and this feature?" and more. This ability of XAI system is recognized as Interpretability.

Interpretability in general is the degree to which humans can understand the reasoning behind the decision. One of the popular definition of interpretability defines it as "the ability to explain or to present in understandable terms to human.".[11][12] defines an ideal XAI as both complete and interpretable. It defines the goal of interpretability to describe the internals of the system and the goal of completeness to accurately describe the operation of the system.

However, the paper,[13] linardatos et al. mentions that there is a clear trade-off between the performance of the machine learning model and it's ability to produce explainable and interpretable predictions. XAI systems can produce different results on the same dataset compared to the traditional AI systems. The performance and prediction of the model can vary. Therefore, XAI systems are useful if it preserves the performance of the system and can produce reasonable arguments to back those predictions. Moreover, the system should have the ability to not just present the cause but also be able to defend the rationals presented to achieve robust XAI.

[14] helps us clearly define the life cycle of Interpretability in XAI. These three steps help scientists and researchers simplify the problem and make analysis to make the machine learning system interpretable. The first step is defined as defining the problem domain and collecting data to study following machine learning practice. Based on the problem domain and the data collected, a predictive model is then identified to work with the data. Finally, post hoc analysis is done to analyze the model using plots and diagrams and answer the questions regarding the decisions the model has made.

Machine learning and AI systems have become an integral part of a human life. It tends to dramatically improve efficiency at workplace and replace human labor effectively. Although, the rise of AI has assisted us in an unimaginable ways, the concerns regarding the ethics of these systems are not to be ignored. There is a dire need of an AI system that can make accurate and interpretable predictions. XAI is the key to building transparent systems that people can comprehend and trust.

3. RESEARCH APPROACH

3.1 Datasets

This paper uses two datasets for the experiment. We chose these datasets based on the simplicity of features and the context it provides to the readers. Both the dataset helps reader understand how the features are contributing individually and cumulatively toward the target variable.

The dataset used for the first experiment is the Titanic dataset provided by the Vanderbilt Department of Biostatistics. The dataset gives information on passengers and whether they survived the crash. The dataset has 891 instances and 12 attributes in it. The numbers of feature are important as more the features better we can analyze the effect of these features on the model prediction. The dataset has features providing passenger's name, sex, age, ticket number, fare, cabin number, Passenger Id, Pclass, SibSp, Embarked, and Parch. Pclass refers to passenger class and is a proxy for socio-economic class. SibSp defines the number of siblings or spouse present on the ship. Parch is the number of parents or children present on the ship. Embarked represents the port of entry among three embarkation points. The target variable of the dataset is Survived which tells whether or not the passenger survived the catastrophe. name, sex, ticket, cabin, and embarked are the features that fall into a categorical variable, while all the rest of the attributes are numeric.

The dataset used for the second experiment is the Bike Sharing dataset is provided by Capital Bikeshare, Washington D.C, USA.[15] The dataset provides a daily count of rental bikes between the years 2011 and 2012 with the corresponding weather and seasonal information. The dataset has 731 instances and 14 attributes. The dataset has features such as dteday, season, yr, mnth, holiday, weekday, workingday, weathersit, temp, atemp, hum, windspeed, casual, and registered. dteday is the date. season signifies one of the four seasons. yr and month represent year and month in that order. holiday tells us whether or not it was a holiday. They numbered weekday 1 - 7 notifying which day of the week. workingday is defined by a day that is neither weekend nor a holiday. weathersit is a complex attribute identifying the weather that day, i.e. clear, few clouds, light snow, heavy rain, Mist, Mist + cloudy, and such. temp and atemp are the normalized temperature, and the normalized temperature felt, respectively. hum and windspeed tell us the normalized humidity and wind speed. Finally, casual represents the number of casual riders, and registered represents the number of registered riders. The target variable for this experiment is cnt which is the count of total bike rentals including both casual and registered. All the attributes in the dataset are in numeric value except for the dteday.

3.2 Tools and Approach

Model explainability is a necessity for the modern machine learning pipeline. The public will trust the transparent system which has the ability to explain the predictions it makes. Adding an explainability layer to a complex machine learning model can help us better visualize what aspects are playing a better part in the role of prediction.

SHAP[16] is a game theoretic approach to explain the output of any machine learning model. It has optimized functions to interpret tree-based models and black-box models for which the predictions are unknown. It helps us create the SHAP values which are used to create a force plot that presents which features in the domain are having a larger impact on the label of the machine learning model and which features have a lesser impact on the result.

Game theory is the method of modeling strategic interaction between two or more players bound by set rules and outcomes. In game theory, a coalition of players cooperate and obtain a certain overall gain out of the coalition. Shapely values give a numeric representation of the contribution every individual player has towards the reward. The Shapely value is the average marginal contribution of a feature from the dataset.

In our context, features are the players and the outcome is the target variable in each of our datasets. The higher the value of SHAP, the greater the feature is contributing toward the target variable. The pink color represents positive SHAP values, while the blue color represents negative SHAP values. We can further dig deep and discover a few vital features among the many trivial ones. A better understanding of data and its role in the prediction process certainly explains a lot about the reasoning the machine learning algorithm is using.

There are three major advantages to using SHAP. It provides the global interpretability of the model. It shows the overall impact of each feature on the label over the dataset. The summary plot and the dependence plot are some examples of the global interpretability function provided by SHAP. It helps us visualize the importance each feature has on the target.

SHAP provides local interpretability. It can visually provide the effect of each of the features on the target label for an individual instance using a force plot. Finally, SHAP can generate SHAP values for any tree-based model. For these reasons, this paper uses SHAP as the tool to interpret the decisions made by the predictive model in the experiments.

Catboost library[17] is an open-source library developed by Yandex. It is one of the state-of-the-art, high accuracy algorithms while providing an inbuilt function to handle categorical features in the dataset. It can handle multiple categories of data, such as audio, text, and images. The performance of the Catboost library is competitive and provides the best-in-class accuracy. The need for extensive hyper-parameter tuning is not necessary while using Catboost, although it can be done.

One of the primary reasons to choose Catboost as the classifier for our experiment is because it provides state-of-the-art results without extensive data training. The datasets in our experiments are not very large, hence using Catboost seemed like a good idea. It also eliminated the process of converting and scaling categorical variables during the feature engineering phase. In the experiments, the Catboost library is used to handle categorical features and convert the dataset into SHAP values needed for visualization purposes.

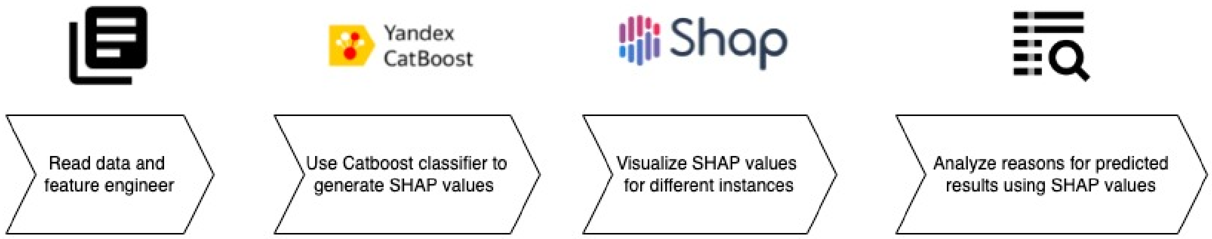

Yandex
CatBoost
Shap
Read data and feature engineer
Use Catboost classifier to generate SHAP values
Visualize SHAP values for different instances
Analyze reasons for predicted results using SHAP values

Figure 1. Flow diagram for Explainable AI experiments

Figure 1 shows the workflow model of the experiments done in the paper. For each experiment, the dataset is first observed and the features significant enough to make an impact on the target variable are selected. The features are then feature engineered and fed to the Catboost classifier. Catboost classifier generates the SHAP values. We can visualize the SHAP values using tools provided by the SHAP library. Finally, the results are analyzed to logically identify the reasons for the prediction made by the model.

4. EXPERIMENTS AND FINDINGS

4.1 Titanic XAI

The machine learning problem for the Titanic dataset is a binary classification problem. First, the dataset is aggregated to exclude the attributes insignificant to prediction. Ticket, Cabin, Name, and PassengerId are the features that do not have a direct effect on passenger survival. Hence, these features are dropped for the experiment. We then categorize the features into categorical and numerical variables. In case there are any missing values among any numeric variable, the median value is calculated for each feature and substituted where required. The post-processed dataset is split into features and labels. We use these features with the Catboost classifier to generate the dataFrame containing SHAP values.

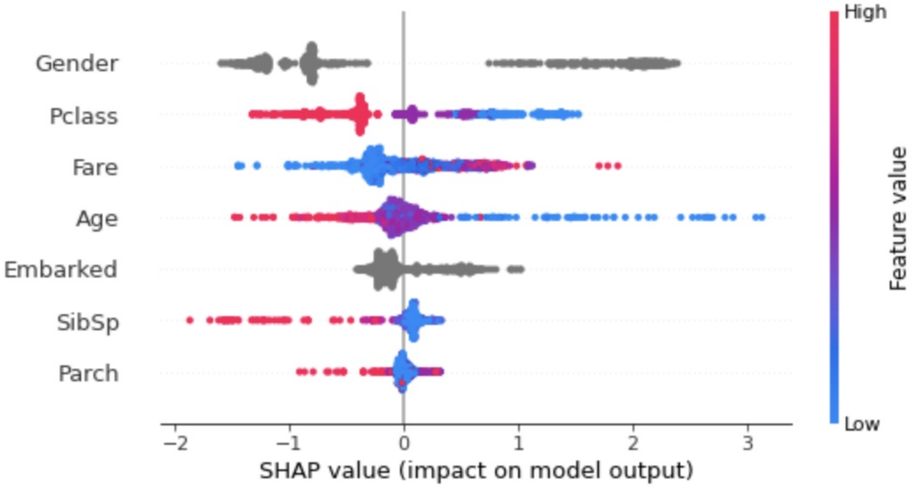


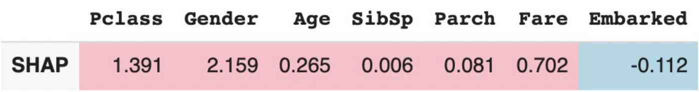
Figure 2. Summary plot for Titanic dataset

Each dot in the summary plot represents one prediction. The summary plot has each feature ordered in descending order based on the impact each has on the prediction. It represents the feature importance. The x-axis represents the SHAP value. A higher SHAP value has a positive effect on the prediction, and a lower SHAP value has a negative effect on the prediction. The color represents the real data value for the instance. The higher value is represented by pink and the lower value is represented by blue.

Figure 2 shows the summary plot of the Titanic dataset. Gender feature seems to have the highest impact on a passenger surviving or not surviving. There is a clear distinction in SHAP values for each gender, where one has a positive impact on survival and the other has a positive impact on death. Unfortunately, Catboost hasn't yet implemented visual aid to colorize categorical variables. Hence, Gender is represented with the neutral color grey.

The higher value of Pclass seems to have a negative effect on the chance of survival. Higher Pclass signifies lower socio-economic status, which reduces the chance of survival. Similarly, we can visualize the effect each feature has on the prediction through the summary plot. The summary plot accomplishes global interpretability by helping interpret the dataset as a whole.

To show local interpretability, we chose two different instances from the dataset to visualize the SHAP value and analyze the result.



	Pclass	Gender	Age	SibSp	Parch	Fare	Embarked
SHAP	1.391	2.159	0.265	0.006	0.081	0.702	-0.112

Figure 3. SHAP values representing each feature of the instance 3

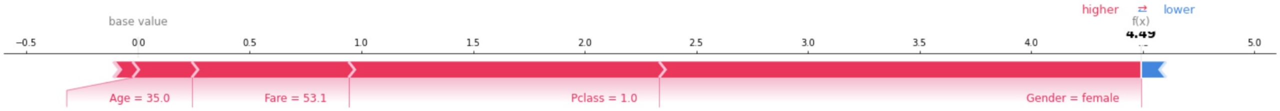
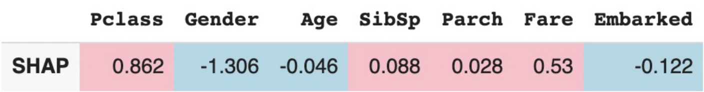


Figure 4. Force plot for instance 3 of Titanic dataset

Figure 3 and Figure 4 show the impact each feature has on prediction for the instance 3 from the dataset. It shows Gender has the most positive impact on the chance of survival. Pclass, Fare, and Age follow Gender in that order. The passenger, in instance 3, is a female passenger who also has a high socio-economic status. As per,[18] 75% total women survived the catastrophe compared to 19% male survival rate.[18] mentions upper-class passengers have a 62% survival rate compared to lower-class passengers with a survival rate of 25%. Hence, it makes sense for Gender and Pclass to be the two high-impact features for passenger survival. The fare was also $53.1 which is above average for the trip.

The force plot shows almost all the features have a positive impact on prediction for the passenger in instance 3. The closer the feature is to the prediction value in the force plot, the more impact they have on prediction, whether positively or negatively.



	Pclass	Gender	Age	SibSp	Parch	Fare	Embarked
SHAP	0.862	-1.306	-0.046	0.088	0.028	0.53	-0.122

Figure 5. SHAP values representing each feature of the instance 55

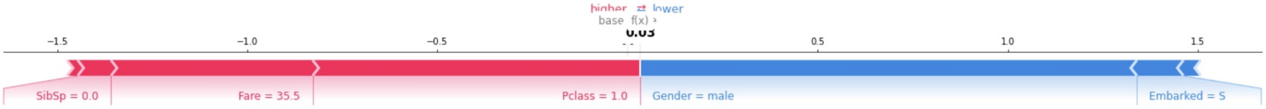


Figure 6. Force plot for instance 55 of Titanic dataset

Figure 5 and Figure 6 represent the impact each feature has on prediction for instance 55 from the dataset. The most contributing factor to the chance of survival in instance 55 is Pclass which is followed by Fare, SibSp, and Parch. The feature most negatively affecting the prediction is Gender in this case. The passenger, in instance 55, is a male. It makes sense for the SHAP value of Gender for this instance to be least, as the Gender factor is pushing prediction towards death. However, the male passenger for instance 55 survives. The reason behind the survival is that the passenger comes from an upper social class and has paid an above-average amount of fare of $35.5 for the trip.

The force plot shows Gender and Pclass to be the most contributing factor toward the decision the classifier makes. Fare also seems to have quite a significant impact on the prediction. Hence, a larger positive impact from the features led the passenger to survive the disaster.

The cumulative SHAP values can help us better select the few vital causes from the trivial many.

Figure 7 and Figure 8 represent graphs for the cumulative SHAP value sorted in ascending order, for instance 3 and instance 55 of the dataset. The passenger, for instance 3, was female and survived the disaster. The four major factors contributing to passenger 3's survival were Gender, Pclass, Fare, and Age in that order.

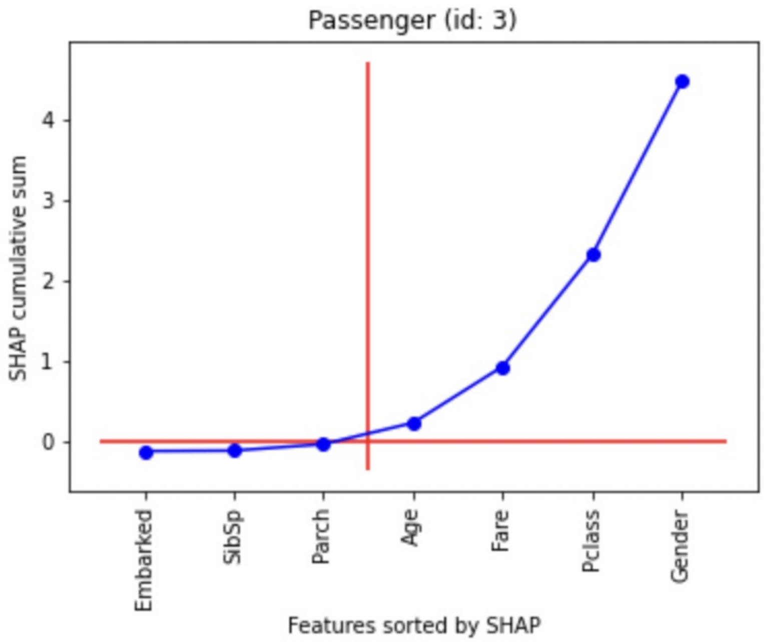


Figure 7. Cummulative SHAP values in ascending order of instance 3

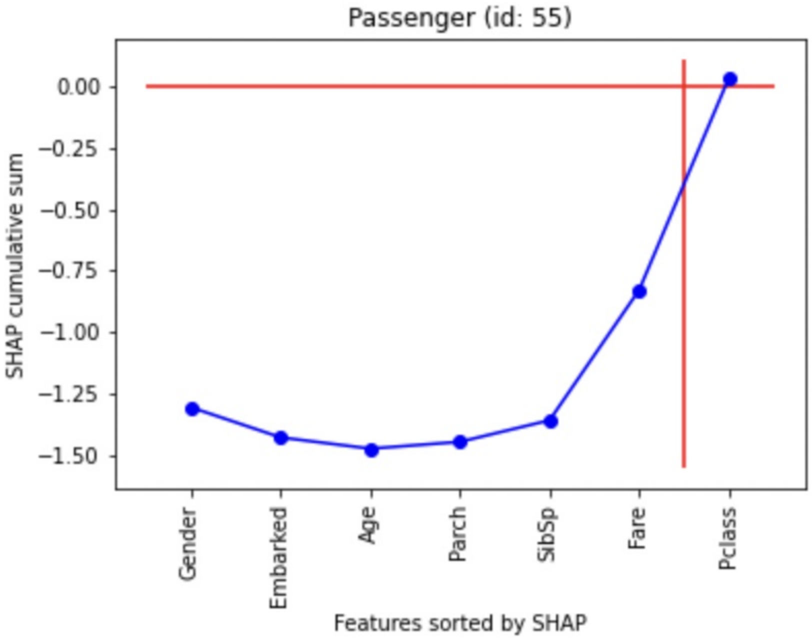


Figure 8. Cummulative SHAP values in ascending order of instance 55

The passenger, for instance 55, was male and survived the disaster. Pclass seemed to be the only factor that significantly helped him survive the crash after all.

4.2 Capital XAI

The Capital Bikeshare XAI problem is a regression problem that predicts the count of bike rentals for Capital Bikeshare for the entire years 2011 and 2012. Similar to feature engineering done on the Titanic dataset, features like registered, casual, and instant are removed for the experiment as it is not vital for prediction. Categorical and numerical variables are identified and substituting missing numeric variables with the median value is done. We then feed the processed data to the classifier to generate SHAP values.

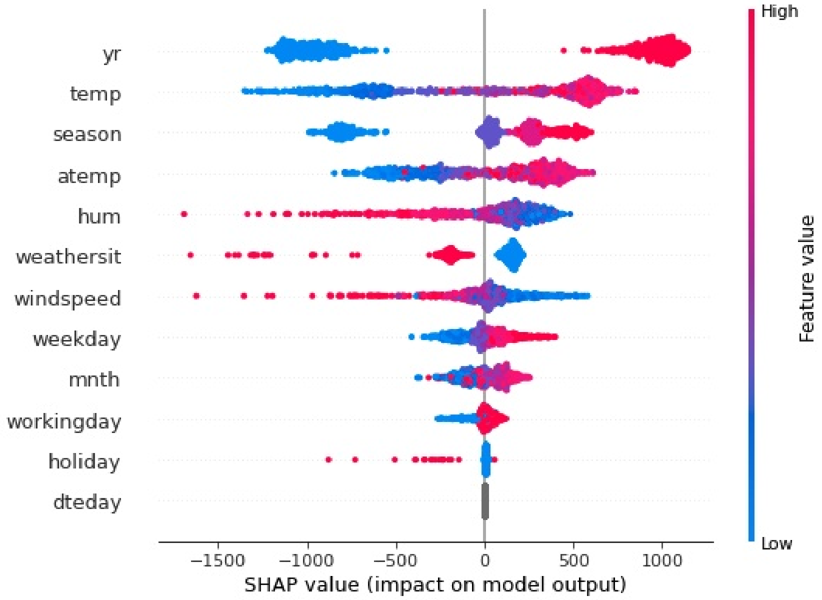
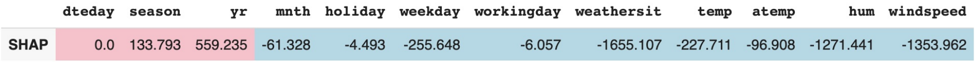


Figure 9. Summary plot for Capital Bikeshare dataset

Figure 9 shows the summary plot of the Capital Bikeshare dataset. The feature year seemed to matter the most towards the prediction of the count of bikes rented. 2012 has a huge positive impact on increased bike rentals, while 2011 has a negative impact on bike rental count. Subsequently, temperature, season, and average feeling temperature were highly significant features contributing to higher bike rental counts. Humidity certainly has a reverse relationship to bike rental counts.

To show local interpretability, we have chosen two instances over the years where one of the instances has the fewest bike rentals and the other has maximum bike rentals.



	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed
SHAP	0.0	133.793	559.235	-61.328	-4.493	-255.648	-6.057	-1655.107	-227.711	-96.908	-1271.441	-1353.962

Figure 10. SHAP values representing each feature of an instance with minimum bike rental count

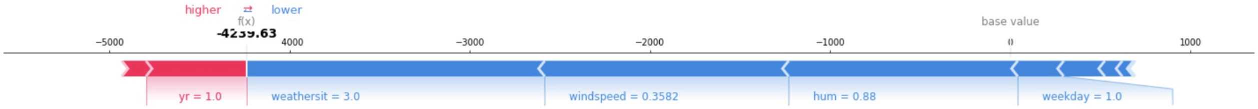
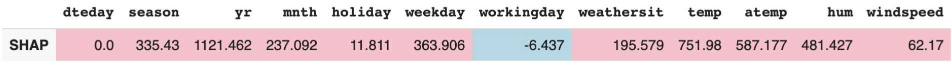


Figure 11. Force plot for instance with minimum bike rental count of Capital Bikeshare dataset

Figure 10 and Figure 11 show the impact each feature has on the day with the fewest bike rentals. The feature most contributing towards least bike rental is weathersit. The weathersit for the instance shows light rain and light snow for the day. The second most contributing factor to low bike rental is windspeed which is a normalized value of 0.36. The average normalized windspeed throughout two years time period is 0.12. The humidity was 0.88 which is way above the average of 0.63. It was Monday, and the normalized temperature was 0.44. Almost all the features determining the count of bike rentals were against the odd. It was a freezing winter Monday with high windspeed, high humidity, and light snow or rain. Hence, the bike rental count was very low.

The force plot shows all the features negatively impacting the number of bike rental predictions. The bike



	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed
SHAP	0.0	335.43	1121.462	237.092	11.811	363.906	-6.437	195.579	751.98	587.177	481.427	62.17

Figure 12. SHAP values representing each feature of an instance with maximum bike rental count

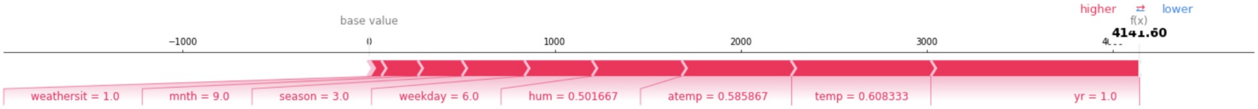


Figure 13. Force plot for instance with maximum bike rental count of Capital Bikeshare dataset

rental count for the day was 22. The model is pushing prediction towards low with all the features contributing to low bike rental count.

Figure 12 and Figure 13 show how features for the instance were aiding for high bike rental count for the day. The feature year was contributing the most for the bike rental count to be large. temperature was the second most contributing factor. The temperature for the day was 0.60 which is slightly above the average of 0.50. The feeling temperature 0.59 seems higher than the average feeling temperature of 0.47. The humidity is slightly lower at 0.50 compared to the average of 0.62. It was a Saturday of fall with clear weather, low humidity, and moderate feeling temperature. Hence, it was a perfect day to rent a bike and go around the city.

The force plot displays all the features positively impacting the prediction for bike rental counts except for the working day.

Figure 14 and Figure 15 represent graphs for cumulative SHAP values sorted in ascending order for the instance with minimum bike rental count and the instance for maximum bike rental count, respectively. No feature contributed positively towards a higher bike rental count, for the instance, with a minimum bike rental count. All the features have a vital contribution towards bike rental count being high for the instance with maximum bike rental count except for dteday and working day.

5. CONCLUSION

Explainable AI is a powerful aspect of machine learning as it answers the questions about AI systems and can answer concerns on ethics and trust. There have been several cases where the presence of human bias in the data and algorithm has been identified. With the increase in the availability of plenty of data, the applications of AI

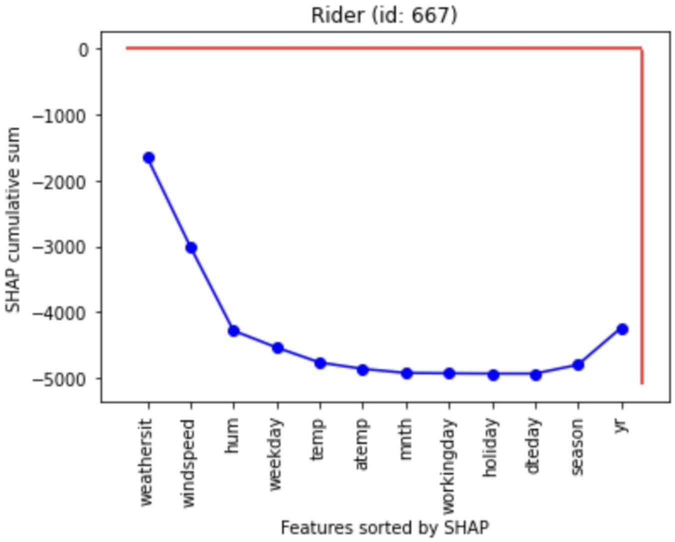


Figure 14. Cummulative SHAP values in ascending order of instance with minimum bike rental count

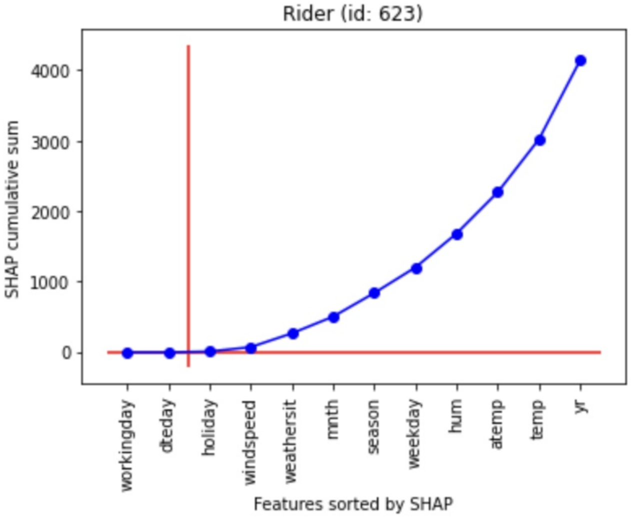


Figure 15. Cummulative SHAP values in ascending order of instance with maximum bike rental count

systems are expanding. However, the expansion is slowly uncovering the issues of ethics in AI systems as well. Designing AI systems without the ability to explain the reasoning behind the decision it makes is inadvisable at this point. XAI is of utmost importance as it will help humans gain trust in the system. As the public puts their trust in AI systems, it is inevitable for AI systems to be widely adopted.

Visualizing the features in data through visual tools and analyzing them can help us better understand the model. Post-hoc analysis methods are most common in XAI as it is easier to integrate and analyze. The process includes extracting information from the data and displaying them using plots and diagrams. The diagrams are then analyzed and the reasons for the predictions made by the classifier can be justified.

In this paper, we used the post-hoc analysis method to analyze the Titanic dataset and the Capital Bikeshare dataset. We used the SHAP library because of the diverse visual tools it provided for us to better analyze the dataset. The results were revealing and provided some crucial information on why the classifier made the predictions it did. The in-built plot tools SHAP provided were vital to successfully showing how an AI system can be made interpretable. Integrating such tools within the machine learning algorithm along with a well-designed user interface to make human-understandable explanations can help AI systems to be transparent, trustworthy, and accountable.

Acknowledgments

This research was funded by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory and in part by the US NSF grant CNS/SaTC 2039583. However, any opinion, findings, conclusions, or recommendations expressed in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

REFERENCES

[1] Wells, L. and Bednarz, T., “Explainable ai and reinforcement learning—a systematic review of current approaches and trends,” *Frontiers in Artificial Intelligence* **4** (2021).

[2] Rawat, D. B., “Secure and trustworthy machine learning/artificial intelligence for multi-domain operations,” in [*Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III*], **11746**, 1174609, International Society for Optics and Photonics (2021).

[3] Gohel, P., Singh, P., and Mohanty, M., “Explainable AI: current status and future directions,” *CoRR* **abs/2107.07045** (2021).

[4] Rawal, A., Mccoy, J., Rawat, D. B., Sadler, B., and Amant, R., "Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives," *IEEE Transactions on Artificial Intelligence* **1**(01), 1–1 (2021).

[5] Vilone, G. and Longo, L., "Explainable artificial intelligence: a systematic review," (2020).

[6] Edwards, D. and Rawat, D. B., "Study of adversarial machine learning with infrared examples for surveillance applications," *Electronics* **9**(8), 1284 (2020).

[7] Ghimire, B. and Rawat, D. B., "Recent advances on federated learning for cybersecurity and cybersecurity for federated learning for internet of things," *IEEE Internet of Things Journal* (2022). DOI: https://doi.org/10.1109/JIOT.2022.3150363.

[8] Larson, J., Mattu, S., Kirchner, L., and Angwin, J., "How we analyzed the compas recidivism algorithm," (2016).

[9] Su, J., Vargas, D. V., and Sakurai, K., "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation* **23**, 828–841 (Oct 2019).

[10] Das, A. and Rad, P., "Opportunities and challenges in explainable artificial intelligence (XAI): A survey," *CoRR* **abs/2006.11371** (2020).

[11] Doshi-Velez, F. and Kim, B., "Towards a rigorous science of interpretable machine learning," (2017).

[12] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. A., and Kagal, L., "Explaining explanations: An approach to evaluating interpretability of machine learning," *CoRR* **abs/1806.00069** (2018).

[13] Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S., "Explainable ai: A review of machine learning interpretability methods," *Entropy* **23**(1) (2021).

[14] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B., "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences* **116**, 22071–22080 (Oct 2019).

[15] Fanaee-T, H. and Gama, J., "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence* **2**, 113–127 (06 2014).

[16] Lundberg, S. M. and Lee, S., "A unified approach to interpreting model predictions," *CoRR* **abs/1705.07874** (2017).

[17] Dorogush, A. V., Ershov, V., and Gulin, A., "Catboost: gradient boosting with categorical features support," (2018).

[18] Henderson, J. R., "Demographics of the titanic passengers: Deaths, survivals, nationality, and lifeboar occupancy," (1998).