Mphasis
The Next Applied

Responsible AI in Large Scale Machine Learning Systems

Whitepaper by Pathik Chamaria,
Senior Associate Business Analyst – Mphasis NEXT Labs

Contents

Not long ago, credit card applications were processed by operators based upon their experience and judgment. In order to standardize and automate the process, rule-based systems were created. It was easy to understand how they worked and hence, predicting how they behaved with an application was possible. The problem was manual creation and maintenance of these rules, which became very tedious with time. Machine Learning (ML) came to the rescue as it was able to learn the rules from data, but at the cost of understandability of how they work. With time, deep learning based black box model took over as they provided better performance, but led to a complete loss of understandability of how the system works.

Today, ML has entered the business mainstream. It is helping businesses automate decisions in fully autonomous systems, leading to a boost in productivity and innovation, touching and shaping our everyday lives. Although, deep learning based techniques perform better, the output generated is black box models, which makes their functioning extremely hard or impossible to understand.

The scale and complexity of ML systems being developed have increased over time, ranging from demand predictions to fraud detection. The applications have become autonomous and are being used by a large number of users. Since there is no human intervention involved in processing the results further, any irregularity in the result of the ML model directly affects the customer. Such a system in a movie recommender scenario may be acceptable but not in the case of a credit score prediction. Questions regarding their trustworthiness, biases and veracity cannot be answered without understanding how the model works. In the absence of such understanding, it is hard to evaluate and diagnose a model properly and hence, it is possible that the model might not work as intended when deployed. Below are some examples where black box models were used in large scale systems and did not perform as intended:

- Apple launched a credit card in August 2019, and within days it was reported that even with similar applications, women were getting less credit rating. In some cases, the difference was as big as ten times. Soon after, it was on the radar of regulators for breach of financial rules[1].
- K1, an AI-based system was managing funds for a Hong Kong real estate tycoon rather than boosting the funds; it lost $20 million daily. The tycoon has now sued the K1 makers [2].
- IBM Watson was used to predict the best remedy for cancer patients. Medical specialists found out on several instances that it touted an unsafe or incorrect treatment. IBM lost a total of $62 million[3].
- Amazon’s hiring tool was identified to be downplaying female applicants and as a result, worthy resumes were not recommended for interviews[4].
- Several facial recognition offerings have been found to be unable to recognize the difference between humans of color and chimpanzees, or failed to identify the difference in the open or closed eye in Asians[5].

These problems happened because a black box AI technique was utilized just by checking its accuracy on the test data, but not how and why it works. This is a deviation from the past when the models were validated against how humans work. The due diligence along with a human layer between the customer and the ML model, could have avoided the above problems. Not just the black box nature of the model is a concern, these techniques are hungry for data, and in general, more the data, the better it is. Also, with everything going digital, there is no dearth of data. The historical data may contain biases towards a particular group, race or gender, since it represents the decisions humans have taken in each specific scenario in the past.
This increases the concerns and questions being raised for the use of black box models, calling for an urgent need to overhaul how ML solutions are developed. Bringing explainability and trustworthiness to the center stage in solutions being developed, is therefore the need of the hour.

At Mphasis, we have built the Responsible AI framework with a vision to create ML solutions which are high performing, privacy preserving, interpretable, transparent, explainable, auditable, bias-free and fair. In this paper, we will go through what is Responsible AI, how it can be implemented at scale, and what we are doing to make our solutions responsible.

1.
Responsible AI

The concept of Explainability of AI models predates that of Responsible AI. With more and more complex and powerful algorithms being developed, model debugging became hard. Why a model works in some cases and not in other cases in real life environment, was becoming a challenge to understand and explain. There were different methods and techniques available to explain the working of a model and in addition, explainability. The guidelines behind Responsible AI established that fairness, accountability and privacy should also be considered when implementing AI models in real environments[6]. Essentially, a solution conforms to Responsible AI framework if, while satisfying the business needs, it is also able to -

- Explain how and why the model works in simple terms
- Associate every prediction with a particular outcome for that input
- Account for the presence of fairness/bias towards a particular gender, class or section along with the privacy of all stakeholders
- Document every step - in development, evaluation and prediction

The above actions, at the broad level, can be divided into Model Explanation and Data Conditioning (in terms of bias and/or privacy) with audibility/documentation being a common theme.

Model Explanation

Be it the problem with Apple credit card, IBM Watson or K1, they all arise typically because the model is being evaluated on a specific metrics for accuracy, without investigating the reasons for the outcomes. If questions, similar to the ones mentioned below, were asked during the development or evaluation of the models, the failures could have been avoided.

- How does the model work?
- What are the key inputs?
- What inputs led to this output?
- What if a particular input is slightly tweaked, how would the model react?

Simple models like linear regression or decision tree are inherently explainable, but they are constrained by the scope of problems in which they can provide high accuracy. While a deep learning model can work in a large variety of problems, it lacks the explainability. To bridge the gap of explainability in black box models, several methods have been developed which allow asking questions for a specific outcome or on the overall model itself, such as SHAP, LIME.

These methods help to find out how much the different input variables contribute to a particular outcome, and how changing one of them affects the output. Moreover, they can help in quantifying the weightage the model is giving to a specific input, to be able to validate it with the subject matter expert. This helps in identifying features which the model should not use to comply with regulatory and societal norms. For instance, in the case of Apple credit card, if the allocation of credit limit for different test applications was questioned and then examined, the gender disparity would have been discovered. Similarly, in the case of IBM Watson, if the driving factors of the model were identified and cross-referenced with the experts earlier on, during model development or evaluation, the problems could have been avoided.

Data Conditioning

Issues with Amazon hiring tool or face recognition systems come inherently from the data, where past precedents or under-representation has affected the model. In the first case, the data to train the model was predominantly male-centric, and hence it favored this gender. In such a scenario, the minority class can be oversampled, or variables representing the class can be removed to create a dataset which is not biased towards a class or group. Similarly, the face recognition problem arises because the standard datasets available to train models for this task is predominantly of Caucasian origin, and hence all other races are under sampled. Balancing the dataset with respect to diversity should mitigate the problem.

The above examples demonstrate how the presence of bias in the solution creates a scenario, where ML models can cause unfair disadvantage to a particular section of the society.
With the large amount of data being used for the training of models, privacy of the data must be maintained, as it is not just ethical but legal requirement in several geographies. Different methods and techniques are used to mitigate bias and privacy concerns. This is in addition to sensitizing developers and decision makers.

Auditability

When a model is being developed, it goes through several iterations, not just in training but also in the data preparation. Each iteration provides different result, and possibly introduces a bug which is not detected in evaluations. In case all the iterations are not recorded in detail, it would be almost impossible to identify when the bug was introduced. Moreover, with time, the underlying pattern in the data changes and hence the model performance deteriorates. In order to maintain a consistent performance, it is imperative to retrain the model with new data. All of this is possible if over time the model's predictions are captured and audited. It is hence imperative to create an audit trail for all activities being performed for the development of AI solution. It is not a necessary requirement from legal perspective, but with several countries coming up with the national AI guidelines, it just a question of time when it would be.

2.
How to Achieve Responsible AI?

Implementation of Responsible AI is possible in both existing solutions and the ones being developed. It requires certain steps at each stage of the AI/ML lifecycle, based upon several parameters such as problem being solved, data collected, choice of models, and so on. There are ways to protect the privacy or remove biases in an already trained model without retraining, but they are limited by their capabilities. Some of the key steps to take, at different stages of AI/ML lifecycle are explained below:

- **Data Collection:** At this stage, the data being collected from different sources should be checked for Personally Identifiable Information (PII) data. If found, it should be either discarded, or anonymized, if the use of this data is imperative. For example, in case of the Apple credit card, employment, financial and other similar details are considered as PII as per several regulatory systems. However, since these are necessary for decision making, they should be anonymized in data used for training and testing purposes, to protect the identity of customers.
- **Data Preprocessing:** While processing the data for model training, it should be checked to identify if any bias towards a class or group is present. If found, appropriate techniques should be utilized to mitigate it. Also, any feature which should not be used as per regulatory standards, should be removed. Societal norms of where the model would be utilized should also be kept in mind. For example, if datasets available for training facial recognitions are looked from the perspective of diversity, than bias can be identified and removed by incorporating more samples from minority groups.
- **Model Training:** While selecting the model to be trained, the explainability needs should be kept in mind. Either a simple model should be chosen which is inherently explainable, or an appropriate framework like Google What If and SHAP should be utilized to make it explainable. It should also be checked if the model is ignoring any relevant features.

- **Model Evaluation:** The model should not be evaluated only on the metrics of accuracy, but also on whether the bias towards a class or group has been mitigated or the model has learned it somehow. For example, in the Amazon hiring tool, even without the gender input, bias was present in the model. It had learned it based on other inputs, such as - a woman's only college name was mentioned.
- **Feedback Loop:** Model Evaluation identifies the performance, based on which model retraining or data processing is done. In case bias is detected, mitigation remedies are performed. Hyperparameter tuning can also be done to improve the performance on any of the evaluation metrics. In some cases, the data may be processed to handle certain specific kind of error also.
- **Model Deployment:** The end user should always be made aware when they are interacting with an AI/ML interface, and wherever possible, explanations for why a decision was taken by the model, should also be provided. For example, frameworks like LIME, ELI5 should be used to provide explanations in case of complex models.
- **Drift Detection:** Over time, often the data distributions and patterns change, which is known as data drift. This causes old learnings getting applied to new data, and hence results in inappropriate outputs. It is therefore necessary to monitor for data drift, and retrain the model when it happens.
- **Retraining Loop:** Continuous monitoring of the model post deployment may lead to identification of certain cases that are not predicted properly. In such scenarios, model retraining allows for better performance. Also, more the data, the better it is for AI/ML algorithms.

While going through these stages, all actions should be documented and logged, starting from how and when the data is collected to every prediction made by the model. This allows for the auditability of the solution. It also helps in monitoring the model and data for performance and drift, so that appropriate actions are taken. Audit trails also become ready source of data for retraining loops as they contain the input data and associated predictions. Since all the data is in a single place, it reduces the time and effort for retraining the models.

Implementation of Responsible AI is not just using the right framework, appropriate soft skill development is needed from the developers to the executives. Skills related to understanding different biases and their mitigation is important. Societal norms vary from place to place, and hence proper cultural assimilation of the development team is necessary. Decision makers need to be sensitized about the importance of Responsible AI to allow for required resources for the extra effort needed. An organization with both the appropriate technical and soft skills in the domain of Responsible AI, is future proof.

3.

Mphasis Responsible AI Framework

Mphasis' proprietary solution, Responsible AI helps remove the black box nature of machine learning model predictions. The solution has been designed to address the previously discussed issues in black box models by utilizing targeted application of state-of-the-art algorithms. It helps in the complete lifecycle of model development, starting from identifying and removing data biases or privacy concerns, to tracking feature importance over time as well as drift detection algorithms. The framework extracts comparative feature importance, displayed as summary plots for both global and local explanations. Users also have the option of conducting 'what if' analysis by changing input values. The framework's counterfactual analysis feature helps in identifying representativeness and in-group biases, by highlighting how predictions alter with changes in single features like race, sex, etc.

Responsible AI can be utilized for incorporating components from the outset in the new AI initiatives, and added to the ones already developed. This helps in designing trustworthy, interpretable systems that can pass muster on openness and fairness concerns of regulatory bodies and civil society. The framework is modular, ensuring easy integration in any stage of the solution development, as per the need. It is generic enough to be applicable to all industries, while being adaptable to any kind of regulatory changes. The framework helps in developing models that perform better in terms of accuracy on the unseen data. Below are two use cases where Mphasis Responsible AI was used.

Use Case 1

A global provider of financial market data, as part of the due diligence process of an entity, looked for various news sources to identify any negative news. It is a highly, effort intensive task, and hence was automated. A news gathering and natural language processing-based classification system was created to collect and classify news in one of the 8 required categories, with over 85 percent accuracy. An explainability framework was used to highlight the keywords in the articles which were used by model for classification. It helped in further fine tuning of the model by eliminating certain words as part of pre-processing. Moreover, in articles where model had low confidence, the associates could quickly look through the result and correct the classification when needed.

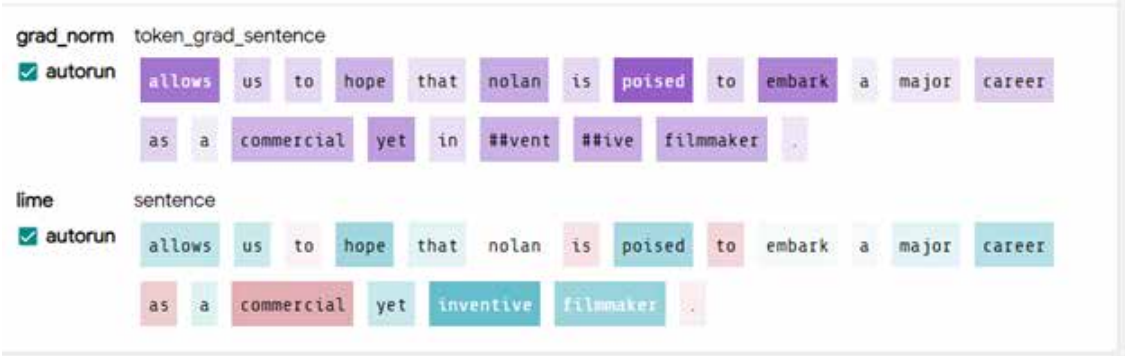

Fig.1: Explanations for a sample document

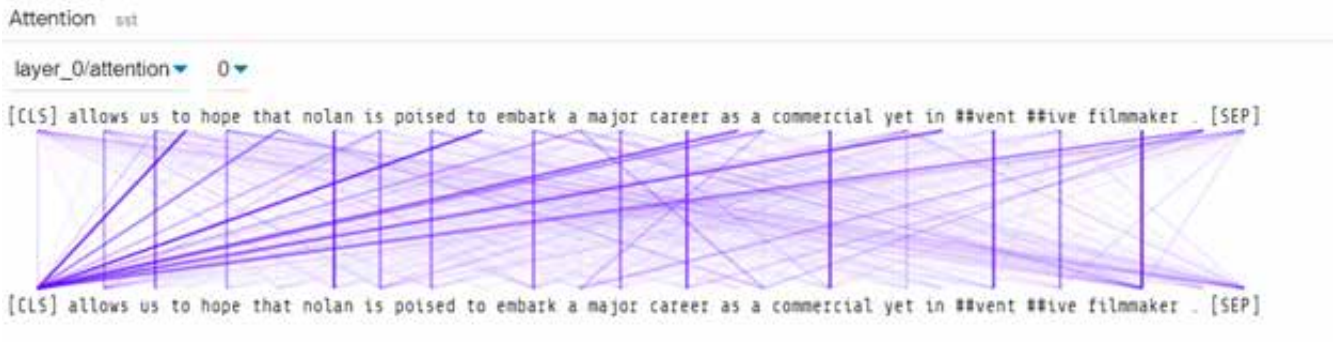

Fig.2: Attention map for a sample document

Use Case 2

A large logistics provider was trying to identify the damage shipments at each transit point in order to identify problem source, reduce damage shipment claim and increase customer satisfaction. A deep learning based image classification model was created to identify damaged products along with explainability module to highlight the areas being considered by the model to predict. The explanations were used to identify images which were predicted incorrectly, helping in debugging the model and achieve more than 90% accuracy overall, with all damaged shipments identified correctly.

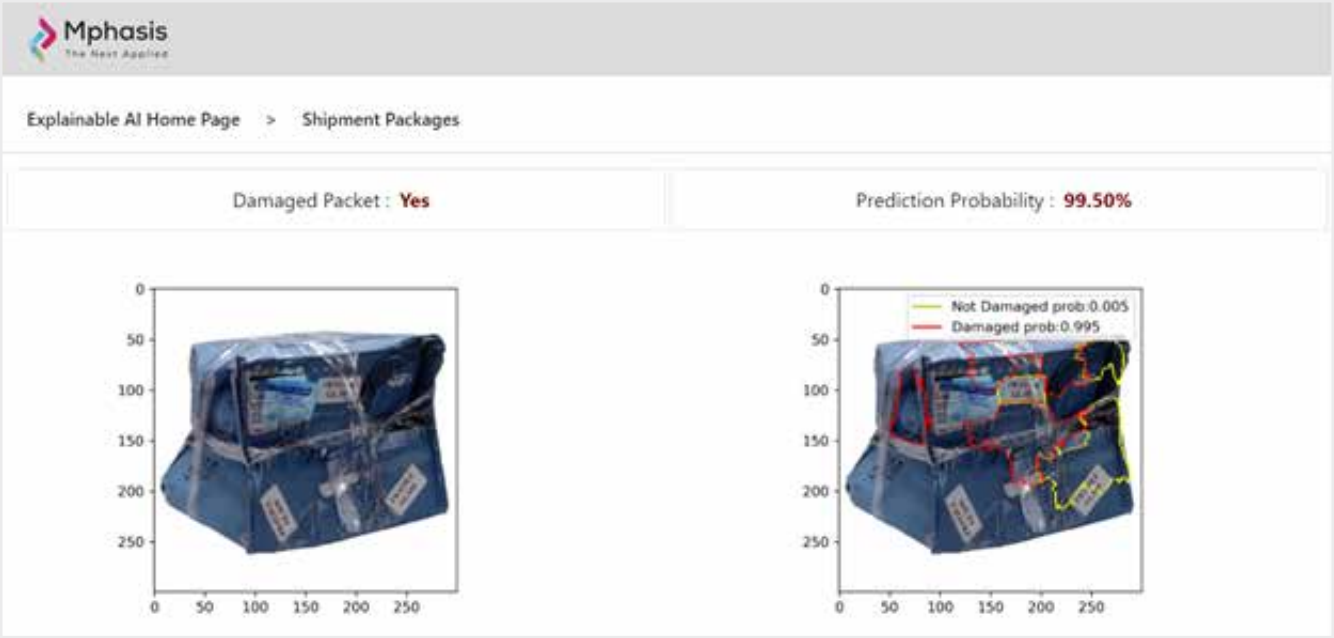

Fig.3: Sample output of the explainability module

4.
Conclusion

We, at Mphasis, believe that these aspects of Responsible AI should remain a necessary part of any AI solution being developed, becoming more refined with time. Responsible AI, today, may not be a legal or regulatory requirement, but it is not far off, given the way AI has started affecting our day-to-day life. New innovative solutions are getting developed with the help of powerful computing AI algorithms.

Incorporating Responsible AI in solutions helps the business in not just identifying the issues with models early on in the lifecycle, but also helps in adhering to regulatory requirements and avoiding issues with discrimination against gender, race or other groups. The explainability of the model helps in creating solutions that are understandable by users, and hence builds more confidence in them. All this results in a solution which is more trusted by users and is acceptable by society at large.

5.
References

[1] [Online]. Available: https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/. [Accessed 5 10 2020].

[2] [Online]. Available: https://www.bloomberg.com/news/articles/2019-05-06/who-to-sue-when-a-robot-loses-your-fortune. [Accessed 5 10 2020].

[3] [Online]. Available: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/. [Accessed 5 10 2020].

[4][Online]. Available: https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine. [Accessed 5 10 2020].

[5] [Online]. Available: https://www.bbc.com/news/technology-50865437. [Accessed 5 10 2020].

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina and R. Benjamins, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82-115, 2020.

Author

Pathik Chamaria

Senior Associate Business Analyst – Mphasis NEXT Labs

A data geek, Pathik is part of Mphasis' innovation group – NEXT Labs. He has been working in the corporate world for the last 3 years, helping businesses improve their bottom line by implementing machine learning and statistical algorithms. An MBA from IIT Kanpur has helped him to understand business intricacies and offer pragmatic solutions based on the data.

About Mphasis

Mphasis (BSE: 526299; NSE: MPHASIS) applies next-generation technology to help enterprises transform businesses globally. Customer centricity is foundational to Mphasis and is reflected in the Mphasis' Front2Back™ Transformation approach. Front2Back™ uses the exponential power of cloud and cognitive to provide hyper-personalized ($C = X2C^{2}_{TM} = 1$) digital experience to clients and their end customers. Mphasis' Service Transformation approach helps 'shrink the core' through the application of digital technologies across legacy environments within an enterprise, enabling businesses to stay ahead in a changing world. Mphasis' core reference architectures and tools, speed and innovation with domain expertise and specialization are key to building strong relationships with marquee clients. To know more, please visit www.mphasis.com

For more information, contact: marketinginfo.m@mphasis.com

USA
460 Park Avenue South
Suite #1101
New York, NY 10016, USA
Tel.: +1 212 686 6655

UK
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village
Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

www.mphasis.com

NR 24/11/20 US LETTER BAS1165566