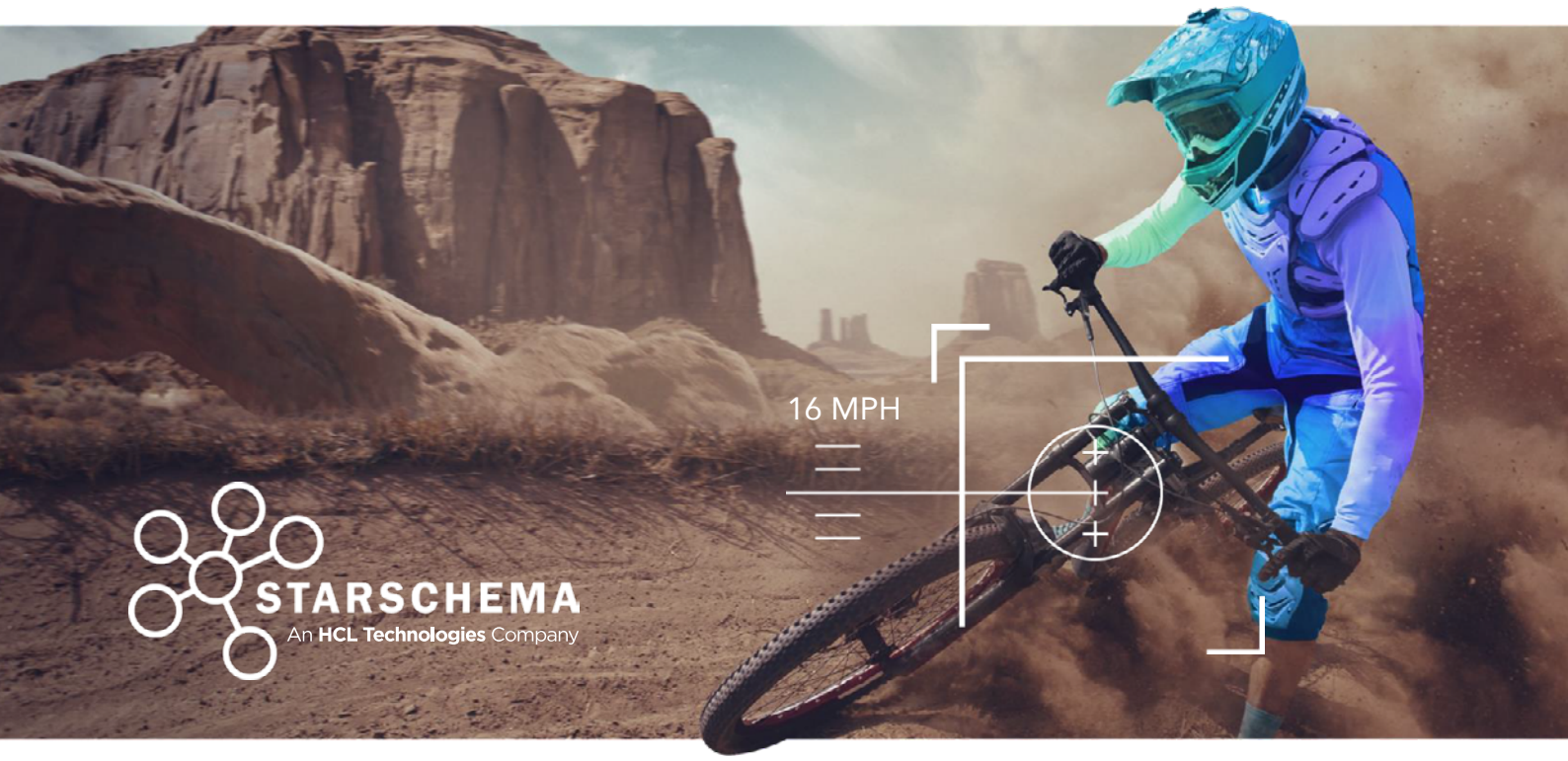


Opening the Black Box

Explaining complex machine learning models

Eszter Windhager-Pokol

Head of Data Science

Contents

Starschema, Inc. 2221 South Clark St., Arlington, VA 22202 +1 415 944-5455

Introduction

For people without a data science background, Artificial Intelligence (AI), and Machine Learning (ML) in particular, solve seemingly insoluble problems. Whether it's predicting with high probability when a machine will break down or determining with almost disconcerting accuracy what type of music or films you'll enjoy, algorithms can sift through mountains of data to find startling insights. In some cases, ML models can be so complex that they seem like a magical black box with inputs and outputs, but little understanding of how the outputs are derived.

This presents several issues for data scientists and decisions makers:

- A lack of trust from stakeholders that don't understand data science.
- Difficulty complying with government regulations, particularly in financial services, that require risk and trading models to be documented and explained.
- Increased risk if the models take a long time to be validated due to slow data velocity, like credit risk models, rather than fast data velocity, like webshops.
- Reluctance to implement major business process changes based on insights from models stakeholders don't understand.

In this white paper, we'll explain the concept of interpretability and several methods to address the problems posed by black box models.

What is model interpretability?

Interpretability is the extent to which humans can understand and explain how a model's outputs are derived from its inputs. Depending on who is consuming the outputs of the data, interpretability can be critical. Humans need to understand the logic of models and the results.

Models using linear regression or decision trees are logical and easy to understand. A model predicting housing prices based on several factors can logically show the relative effects of variable such as size and location on house prices. In very complex ML models, such as recommender systems used by entertainment streaming services, the relation between inputs and outputs cannot be easily explained, even by the model creators. While these ML models can be very accurate, can they be trusted?

Starschema, Inc. 2221 South Clark St., Arlington, VA 22202 +1 415 944-5455

Why is interpretability important?

When the ML models only tell us which movie we might like or which songs to add to our playlist, model creators focus on the accuracy of the model; explaining how the model determines its recommendations is of secondary importance. However, many applications of ML have far greater consequences than those of the entertainment industry. In the field of healthcare, ML is being used in radiology to automate image review. In cases like this, all the stakeholders — healthcare practitioners and patients alike — need to trust in the algorithms producing the results. Without being able to understand and explain how models produce outputs, stakeholders will find it difficult to trust the solution.

This isn't merely an academic discussion. Governments, aware of the pitfalls of artificial intelligence and big data are beginning to address interpretability. The European Union has taken the lead in this regard. The General Data Protection Regulation (GDPR), for example, explicitly addresses automated data processing stating, "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her."

Automated processing in this context includes ML. Credit scoring for loan applications provides a real-world example. Retail banks could use a complex, black box ML model to determine the credit worthiness of a mortgage applicant. The model could produce very accurate results but also offer no way to explain why a person was rejected for a loan. In this scenario, the bank would be in violation of this GDPR requirement.

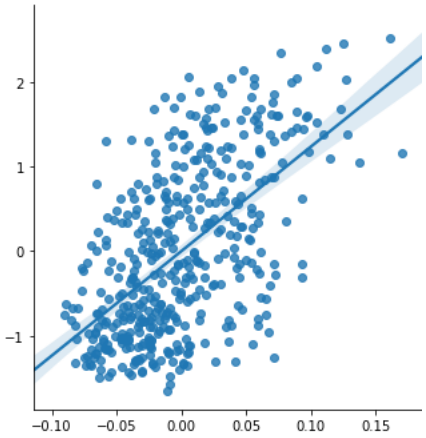
Types of interpretability

There are two types of methods used to understand models — global and local.

Global interpretability seeks to explain the relationship between the entirety of the inputs to a model. Local interpretability focuses on explaining the reasons of one particular prediction. If we take the example of a credit risk model, global interpretability tells us that a higher salary means a lower risk of default. Using the same example, local interpretability tells us why a specific loan application, one data point, was rejected. In this particular example, the age of the loan applicant, or something else like the amount of the loan, might have stronger impact on the rejection than the salary.

Global Methods

For simple ML models using linear or logistic regression algorithms, decision makers can see the relationship between input and outputs. These widely understood statistical techniques are easy to understand.



Linear Regression Plot

For simple rule-based models with small decision trees, it's easy to visualize the models on a tree plot. When working with more complex models, incorporating larger decision trees, random forests, support vector machines and neural networks, we need tools to understand what is happening inside the model.

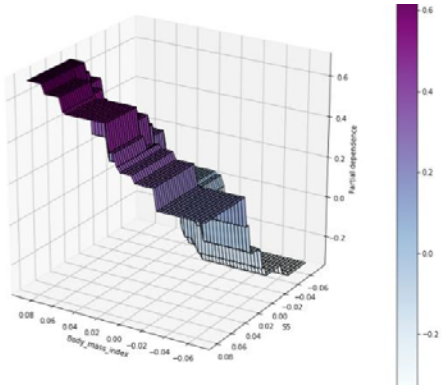
Feature Importance

In the context of machine learning, a feature is an individual measurable property or characteristic of what is being studied. In the credit risk example mentioned above, the features would be attributes such as payment history, salary, and age.

We can measure the importance of a feature by changing that data – replacing a column with different values – and seeing if this permutation affects the model's accuracy. If there is little change to the model's accuracy, then that particular feature is not deemed important. We can perform this operation on all of the models features, or columns in a data set, giving us a raking of the most important features in a model.

Partial Dependence Plots

Partial dependence plots show how one feature affects the prediction and the relationship between the target and the selected feature. It can be a useful addition to feature importance, showing not only the strength of the impact but also the direction. For example, salary and the amount of the loan are both important factors in a credit rick model. But higher salary means lower risk of default, but higher loan amounts means higher risk. Partial dependence plots show more details about the relationship between the feature and the target.



Partial Dependence Plot

Local Methods

Local interpretability methods can help explain how predictive models derive their outputs.

LIME (Local interpretable model-agnostic explanations)

The goal of this method is to understand why the ML model made a certain prediction on one data point. Similar to the feature importance method, but LIME only uses one specific case and measure the variables impact on that case only. With iteration, LIME can generate new data points by changing the values of the features of a given input instance and measure the effect on the prediction. This allows us to measure which element of the input data point has what kind of effect. LIME is model-agnostic, meaning that it can be applied to any machine learning model.

Conclusion

As Machine Learning models are increasingly deployed in a variety of applications and embeds itself into daily life, opening and explaining black box models will grow in importance. The methods and techniques mentioned above are integrated in data science tools to provide understanding of complex models. Being able to explain how complex models function has the following benefits:

- Build trust of stakeholders in complex models that have a significant impact on their lives.
- Easier compliance with government ever increasing regulations.
- Support from business leaders to make big, importance decisions based on ML models.
- More accurate ML models through greater quality control.

Business leaders can increase the value of their data science initiatives at their companies by taking the following steps:

- Fostering a culture of data literacy to help stakeholders understand the approaches taken by data scientists to help build trust.
- Recognizing the importance of interpretability when evaluating data science projects and their results.
- Selecting data science experts, both internal and external, that transparently communicate their approach and methods in a way that stakeholders understand.

About the Author

Eszter Windhager-Pokol, the Head of Data Science at Starschema, holds a MSc degree in Applied Mathematics from Eötvös Loránd University. She has eight years' experience in supporting data driven decision making as a consultant, four years research experience in the field of collaborative filtering and three years background in developing user behavior analytics product for IT security purpose.

Eszter holds data science trainings regularly for business users from the basic level to the advanced analytics courses and teaches Mastering the Process of Data Science at CEU as a visiting faculty instructor. She is an organizer of the R-Ladies Budapest meetup group and participates in program committees of several international data science conferences.

About Starschema

At Starschema we believe that data has the power to change the world and data-driven organizations are leading the way. We help organizations use data to make better business decisions, build smarter products, and deliver more value for their customers, employees and investors. We dig into our customers toughest business problems, design solutions and build the technology needed to compete and profit in a data-driven world.

Starschema, Inc. 2221 South Clark St., Arlington, VA 22202 +1 415 944-5455