Paper SAS4506-2020

Human Bias in Machine Learning: How Well Do You Really Know Your Model?

Jim Box, Elena Snavely, and Hiwot Tesfaye, SAS Institute

ABSTRACT

Artificial intelligence (AI) and machine learning are going to solve the world's problems. Many people believe that by letting an "objective algorithm" make decisions, bias in the results have been eliminated. Instead of ushering in a utopian era of fair decisions, AI, and machine learning have the potential to exacerbate the impact of human biases. As innovations help with everything from the identification of tumors in lungs to predicting who to hire, it is important to understand whether some groups are being left behind. This talk explores the history of bias in models, discusses how to use SAS® Viya® to check for bias, and examines different ways of eliminating bias from our models. Furthermore, we look at how advanced model interpretability available in SAS Viya can help end users to better understand model output. Ultimately, the promise of AI and machine learning is still a reality, but human oversight of the modeling process is vital.

INTRODUCTION

Machine learning is essentially the practice of creating algorithms that ingest data to detect patterns to predict likely outcomes, identify patterns in the data, categorize like groups in the data or detect unexpected behaviors in the data. Models using these algorithms have been in existence for decades, but as computing power has grown along with the volume, velocity, and variety of data available, machine learning is becoming more and more popular. More importantly, algorithms are now being used to automate the decision-making processes that influence our access to information, jobs, loans, and much more. As the execution of this type of modeling gets easier with new software applications and methods, it's important to take a step back and think about the process of using data to make decisions and how human biases can be amplified by applying machine learning.

Defining bias in machine learning is a tall order. In this paper, we illustrate several examples of machine learning bias, most of which are examples of the unintended consequences and discriminatory predictions of machine learning models.

BIASED DATA GIVES BIASED RESULTS

The most important part of the machine learning process is not the software, or the algorithm used, but the data source. All machine learning models are trained on existing data, and the machine can only learn from experiences that the data provides. If the data itself has existing biases, those biases will be amplified by the use of an algorithm. It's vital to understand the biases in the data and to carefully account for it.

In 2014, online retailer Amazon was developing machine learning algorithms to help sort through the hundreds of job applicants they receive for every position (Dastin, 2018). The idea was to use the database of resumes and CVs received over 10 years of hiring to rate every new applicant on a one-to-five star scale, then to focus hiring efforts on the five-star applicants. The algorithm would read through the years' worth of applications that were scored as hired or not hired. This is just the sort of problem for which machine learning is made, and the algorithm was able to find applicants that matched the types that had been

hired. The problem was that Amazon, like many other tech companies at the time, had been mostly hiring men. By training itself on this sort of data, the algorithm taught itself that men were preferred candidates, and downgraded applications that were easily identifiable as women (for example, "member of the women's chess club"). It even went as far as to exclude applicants from two all-women universities. None of this was done explicitly. The model was not designed to favor men. The problem is that the people making decisions in the past had excluded women, and the algorithm picked it up in the training data.

Sometimes, the bias is more subtle. As part of this hiring effort, the Amazon team built models to evaluate posted resumes and profiles on the web to try to identify potential recruitment targets. The algorithm wound up favoring people who used more active and aggressive verbs like "executed," which show up more frequently on profiles about men. Eventually, Amazon abandoned the effort and shut down the project.

The algorithms did their job – they identified candidates that matched the historical data about who the company preferred. The problem was in the training data. Without thoughtful consideration of the data and the goals of the project, it's easy to see how this human bias can impact the results.

The bias in the training data can sometimes be less obvious and can be due simply to the sampling methods. An article in the Guardian (Devlin, 2018) points out how this can be a significant issue when it comes to developing genetic tests. The problem lies with concern that the bank of genetic material in the UK biobank is too ethnically homogenous, because the material is collected mostly from white Europeans. Genetic tests developed on this population may not be reliable when generalized to other populations. For example, the article cites a study where "a commonly used genetic test to predict schizophrenia risk gives scores that are 10 times higher in people with African ancestry than those with European ancestry. This is not because people with African ancestry actually have a higher risk of schizophrenia, but because the genetic markers used were derived almost entirely from studies of individuals of European ancestry."

Applying machine learning algorithms on data that is significantly different from the training data is a sure way to create biased or unreliable results.

MODELS CAN FIND UNEXPECTED CORRELATIONS

As more machine learning methods get easier to use, it is important to try to understand how they are making predictions. With methods like logistic regressions and decision trees, it's fairly straightforward to see how the model was developed and what types of variables it takes into account. With more complicated models and deep learning methodologies, this becomes a much bigger challenge. Models can produce results based on confounding variables that can create algorithms that explain the training data very well but that are not practically useful.

A good example of this concern comes from the field of radiology (Zech, 2018). The article looks at using neural networks to analyze x-rays to look for cardiomegaly (enlarged heart). Images of patients with and without cardiomegaly were run through a neural net to develop scoring code, and the resulting model did a very good job of identifying patients with the condition. Researchers looked at the details of the process to see which parts of the image were most highly correlated to a prediction of a positive result. The expectation was that the areas around the heart in the x-ray would be most predictive, and they were. Surprisingly, there were positive contributions on the edge of the image, far away from the heart. it turned out that there were markings on the edge of the image that indicated which machine was used for the x-ray. The algorithm figured out that a marking that indicated the x-ray machine was portable was highly correlated to the indication of cardiomegaly (see Figure 1). Because portable x-rays were used on patients who were so sick that they could

not travel to the radiology lab, the model essentially figured out that sick patients were more likely to be sick. This conclusion was factually correct, but it wasn't terribly useful in applying the model generally. Other possible confounding variables in this case could be patients from a particular specialist who saw only the harder cases, or clinics that saw only patients who did not have adequate primary care.

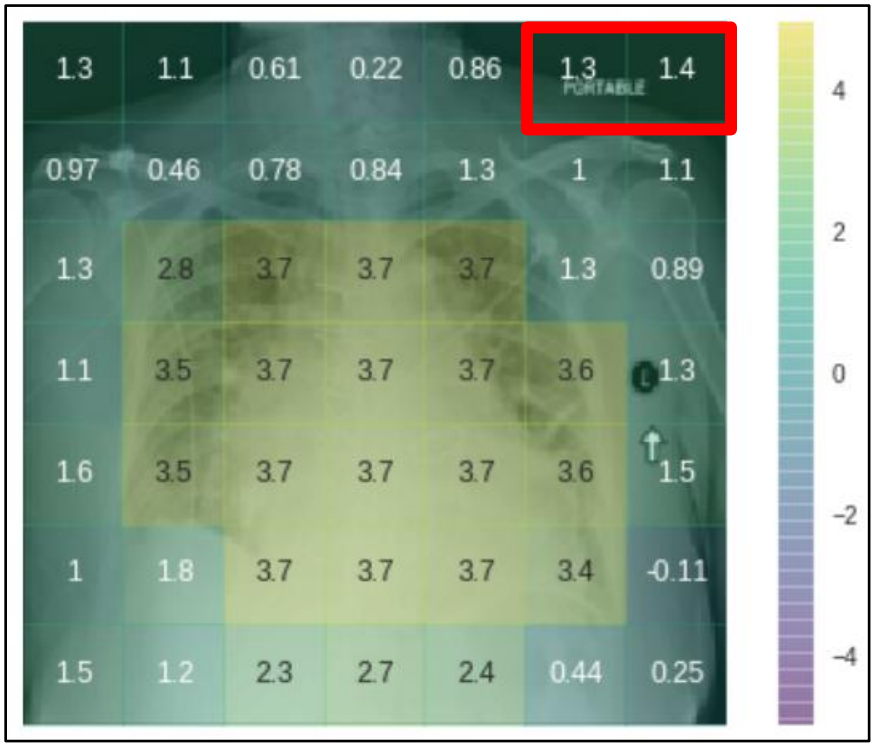

Figure 1. Chest X-ray Showing Areas of the Image Associated with a Higher Prediction for Cardiomegaly

It is crucial to understand how a model is using the training data, especially when using an algorithm that is computationally complicated. As more stand-alone "black box" software methods are available that claim to run hundreds of models to find the best solution, vigilance on the data used is essential to minimize bias. This is especially important in health care applications. There are also practical implications to developing complicated models. Netflix (Johnson, 2012) spent $1 million on an algorithm that was so complicated that they could not successfully implement it.

WHAT TO DO TO MINIMIZE BIAS

As trite as it might sound, awareness of the impact of human bias on machine learning is a huge first step in mitigating its impact. Bias exists in most data collected by humans and can sometimes be hard to detect. Some steps to consider:

FRAME THE PROBLEM

Understand the technical question being asked. Make sure that the training data is representative of the population to which you are going to apply the model. Understand the goals – are you trying to find the best possible job candidates, or are you trying to find people like you already have?

DEFINE FAIRNESS

Think about how you are using the data. Are you using data that won't be applicable to some of your population? Are there regulatory or ethical requirements about the types of inputs you could use? Codifying these decisions can be uncomfortable; be sure people understand why you are doing this.

CHECK FOR BIAS IN THE TRAINING DATA

A great first step to mitigating bias in your machine learning models is to thoroughly interrogate your training data set. Ensuring that the training data is a representative sample of the population is a good place to start. For instance, a predictive model aimed at identifying patients among the entire patient population who are likely to get readmitted into the hospital should not be trained on only female patients. The goal of this exercise is to acknowledge and act on the understanding that your training data is a snapshot in history that may have biases you would not want to carry forward into the future.

In SAS® Visual Data Mining and Machine Learning 8.5, you have access to the Data Exploration Node which handles some of the common and often time-consuming effort of data exploration that is required before feeding features into a model tournament. (see Figure 2)

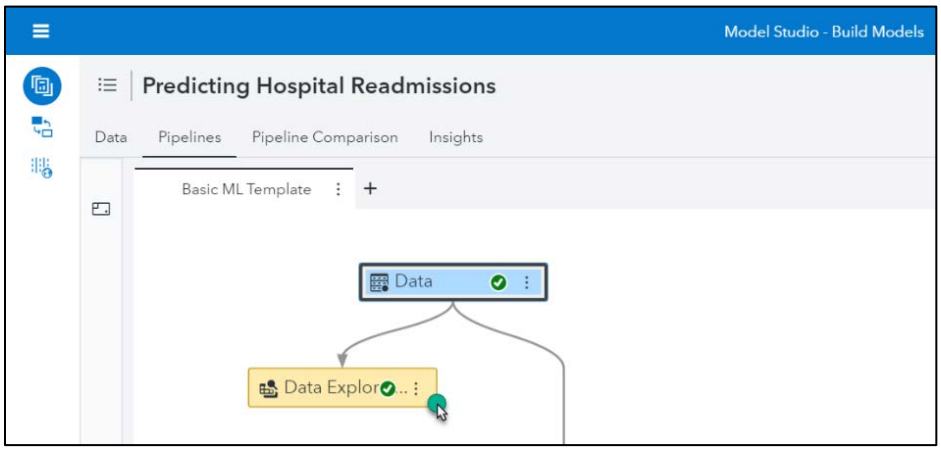

Figure 2. Model Studio Project Showing the Data Exploration Node

Some of the output from the Data Exploration Node can be seen in Figures 3 and 4. In Figure 3, you can assess the distribution of features in the training data. In this example, the training data has a relatively higher representation of male patients, which might not be representative of the patient population. Figure 4 shows the distribution of readmissions (the target variable) across gender, which allows the data scientist to discern if there is a potential association between gender and the target variable before performing any statistical test of association. "Understanding and interpreting your data set" by Ilknur Kaynar Kabul provides more in-depth mathematical assessment of your training data set, that includes dimension reduction and unsupervised machine learning techniques (Kabul, 2018).

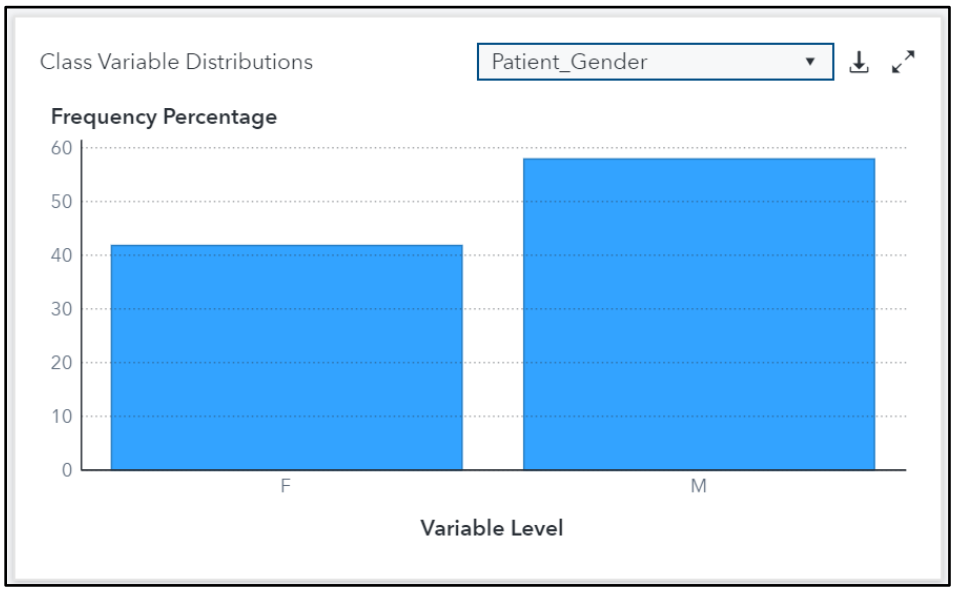

Figure 3. Assess the Distribution of Sensitive Features in the Training Data Set

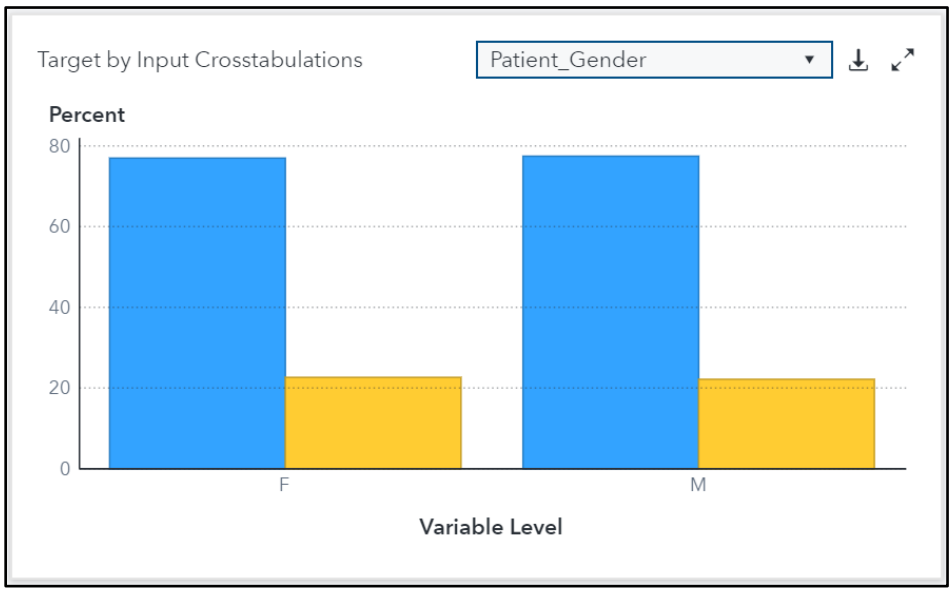

Figure 4. Assess Crosstabulations Between Sensitive Features and the Target Variable

However, there will be instances where different groups of people need to be treated differently by the model. To continue with the example of the hospital readmission model, older patients or patients with a specific diagnosis may have an entirely different health profile and treatment protocol than other groups of patients. For cases like this, leaving a sensitive feature in the model makes sense. You might even consider building separate models for different groups of patients if their health profiles and treatment protocols are expected to be significantly different.

Although removing features that may cause unintended discrimination is a best practice, bias shows up in unexpected ways. An example of this was seen in the model Optum built to prioritize patients into care management programs (Morse, 2019). The model was built on data that contained patients' medical history and how much money was spent treating

them. By including historical data for the amount that was spent on patients' health care, the data scientists who built the model might have assumed that more money is typically spent on sicker patients across the entire population. They later discovered that in practice, significantly less money was being spent on black patients with the same level of need and illness as their white counterparts. The model that was trained on this data assigned a lower risk score (and a lower chance to qualify into a care management program) to black patients with the same level of illness as white patients due to historically lower spending level for black patients. In this example, historical health care spending was not a good measure of severity of illness. Instead, it introduced bias to the model in the form of racial bias. The key is taking into consideration the association that may exist between seemingly unbiased features such as health care spending and features such as race and gender. Before removing features that might cause unintended discrimination, you might want to test whether any other feature in your training data can predict these features.

Figure 5 shows how you can use SAS Visual Data Mining and Machine Learning 8.5 to create a project where the target variable is a feature such as gender. From here, you can use the Data Exploration Node to assess the most important features in the data that can accurately predict gender. In this case, it appears that the feature "Marital_Status" is a good predictor of gender, so it should probably be rejected in the model building processes.

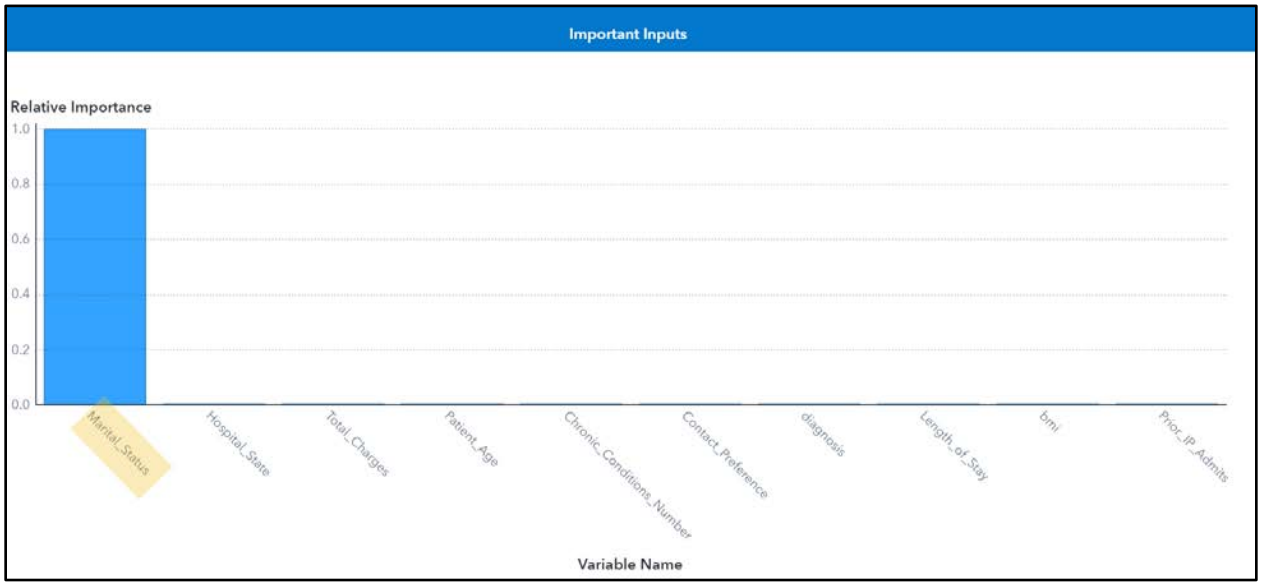

Figure 5. Variable Importance List in Predicting a Sensitive Feature Such as Gender

An alternative approach is to ensure that the model results do not change unexpectedly due to factors such as age, race, or gender. You can plot the distribution of your model results (predictions) segmented by sensitive features and check to see if the differences in the distributions can be explained.

MODEL INTERPRETABILITY

Traditional statistical models have a reputation of being easily interpretable. For example, the coefficients of a linear regression tell you the magnitude and direction (positive or negative) of the correlation between the target variable and the input variable. Now, with the advent of more advanced modeling techniques (such as neural networks, random

forests, gradient boosting, and so on), there is not a similarly straight-forward interpretation available for the relationship between the input variables and the target. Variable importance plots can tell us which inputs were the highest drivers of a prediction, but not how those variables actually affected the prediction. For example, if age is the highest driver for a readmission model, that doesn't necessarily mean as age goes up, so does the likelihood of readmission.

SAS provides several ways to interpret "black-box" machine learning models – either at the global or local level. Global explanations provide insight into a model's behavior across the entire data set (variable importance & PD plots), whereas local explanations provide insight at an individual observation level (ICE plots, LIME plots & Shapley values).

Partial Dependence and Individual Conditional Expectation Plots

Partial dependence (PD) and individual conditional expectation (ICE) plots are both model-agnostic techniques that help to interpret the relationship between input variables and the target predictions. They are model-agnostic because they are run after the model has been trained and they can be used on any type of model. In contrast to variable importance, these plots help to explain how predictions vary based on input values.

Ray Wright (2018) does an excellent job of introducing both techniques. The major difference between partial dependence and ICE plots is the level of detail they drill into. While partial dependence plots show the overall relationship (linear, step, and so on) of inputs to a prediction, ICE plots can drill to individual observations, possibly revealing subgroup variations and interactions between inputs.

For example, a PD plot may show that there is no overall relationship between X1 and a model's predictions:

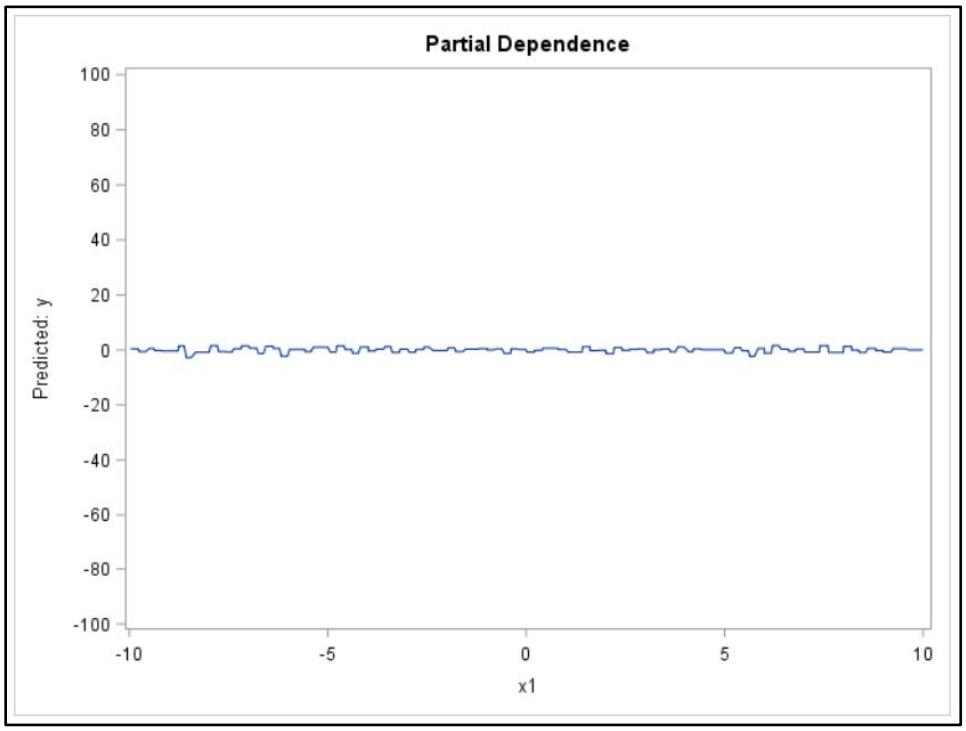

Figure 6. Partial Dependence Plot Generated Through SAS Code

The ICE plot might reveal that in fact, X1 is related to the predicted value, but that the relationship varies between subgroups. In this example, in one subgroup X1 has a strong positive relationship with the predicted values, while in the other subgroup, X1 has a strong negative relationship with the predicted values. In aggregate, the overall relationship is as

seen above in the Partial Dependence Plot:

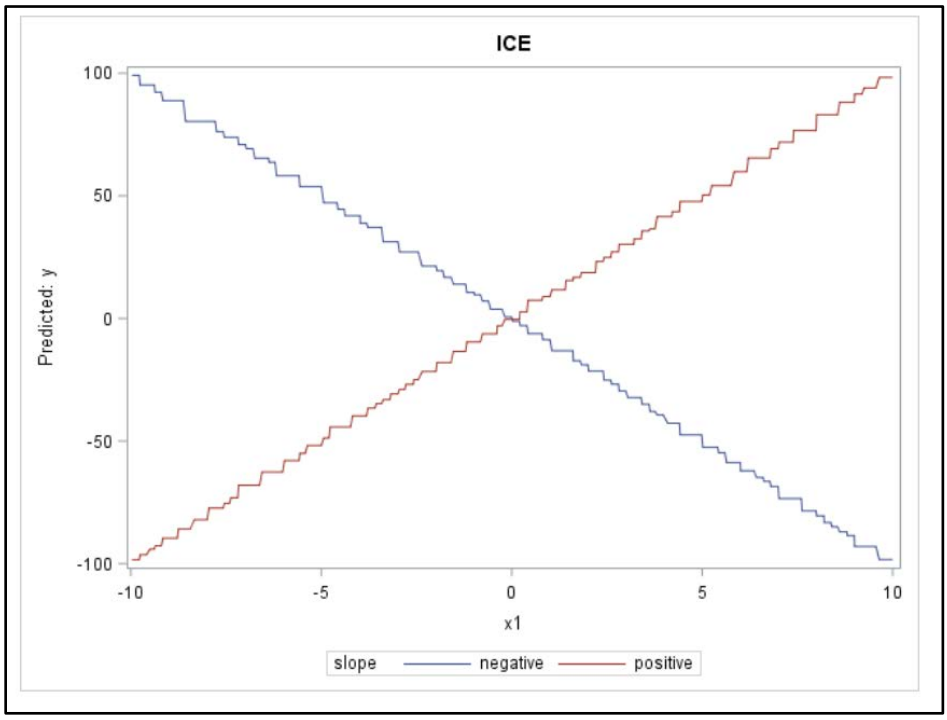

Figure 7. ICE Plot Generated Through SAS Code

While Wright's paper discusses in detail how to create these plots using coding methods, Model Studio now includes options to automatically compute these plots. Use these steps to compute these plots:

1. Open the Properties page for each node.
2. In the Post-training Properties area, open the Model Interpretability entry.
3. Select PD plots under the Global Interpretability entry and ICE plots under the Local Interpretability entry.

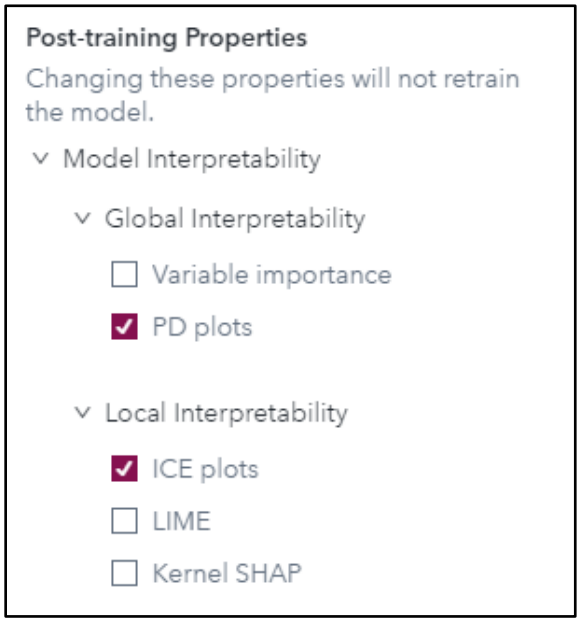

Figure 8. Post Training Properties Options for Supervised Learning Models in Model Studio

After you make these selections, you now have a Model Interpretability tab available when you look at a node's results.

The PD plot from Model Studio shows the relationship between selected input (in this case, Patient Age) and the predicted value of the target, averaging out the effects of the other inputs. It displays values of the input variable on the X axis and the corresponding average prediction for the target variable on the Y axis. Clicking on the information icon (the i in a circle) provides an explanation of what the plot is showing and how best to interpret it.

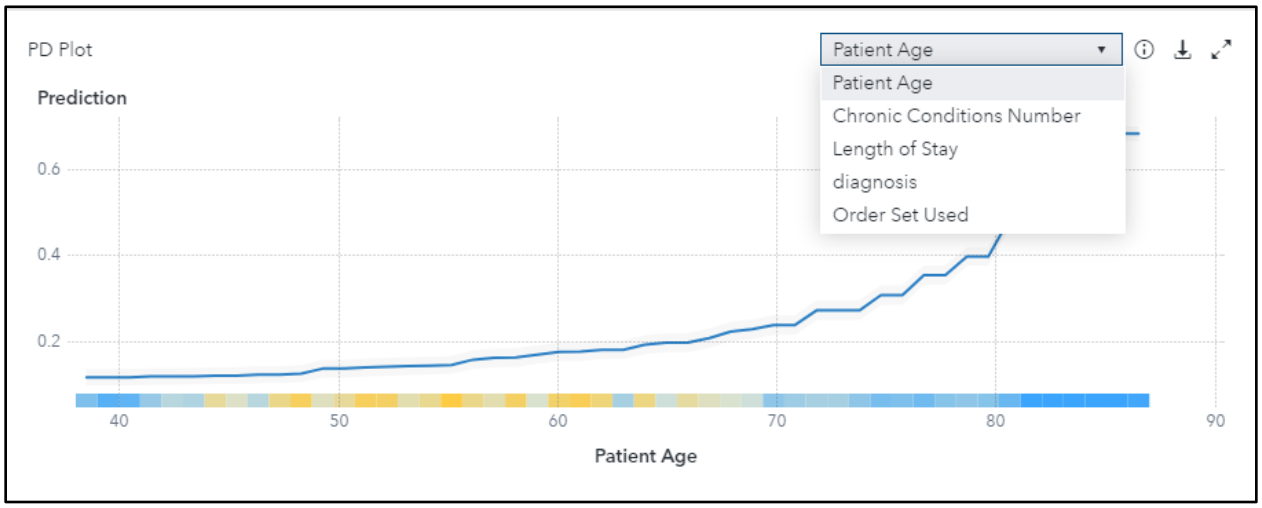

Figure 9. PD Plots Generated Automatically in Model Studio

In this instance, it appears that age has an almost exponential relationship to the predicted values. Below age 65 there is a small positive slope, but soon afterwards the effect appears to build.

From a bias and ethics point of view, having age as in input into a readmission model makes sense. However, in other instances age may be an inappropriate input (for example, for a predictive hiring algorithm).

For your ICE plot, you can either have randomly selected observations or specify up to five individual observations. The PD and ICE overlay plot in Figure 10 contains the same PD plot as in Figure 9 (colored blue in both instances) and compares it to up to 5 observations in the ICE plots. As in Figure 9, the specific input is selected from a drop-down list at the top of the graph. When the input variable is nominal, the graph is a bar chart, and when the input variable is interval, the graph is a line plot. Clicking on the information icon provides an explanation of what the plot is showing and how best to interpret it.

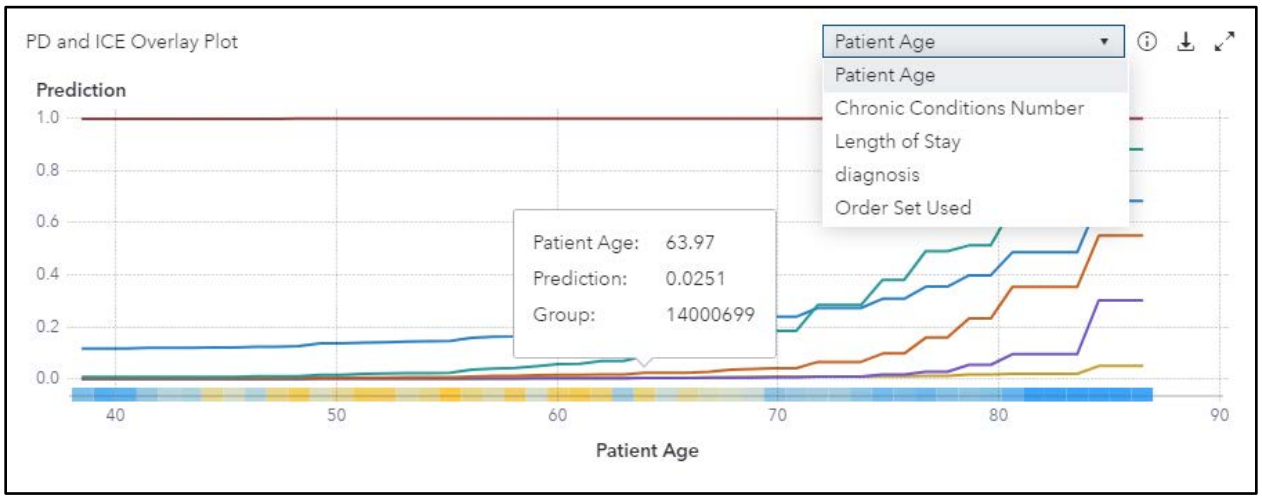

Figure 10. PD and ICE Overlay Plot Generated Automatically in Model Studio

LIME and Shapley Values

Local interpretable model-agnostic explanations (LIME) and kernel Shapley additive explanations (Kernel SHAP) both provide explanations for individual predictions. Ribeiro, et al. (2016) provide an in-depth explanation of LIME. Basically, LIME creates a linear model around a point to be interpreted, and it is the coefficients of this surrogate model that provide insight into what is probably driving the predictions of the "black box" model. Shapley values help you determine the relative importance of each variable to a given observation's prediction.

The input variables are ordered by significance such that the most important variable for the local regression model is at the bottom of the chart. A positive estimate indicates that the observed value of the input variable increases the predicted probability of the event.

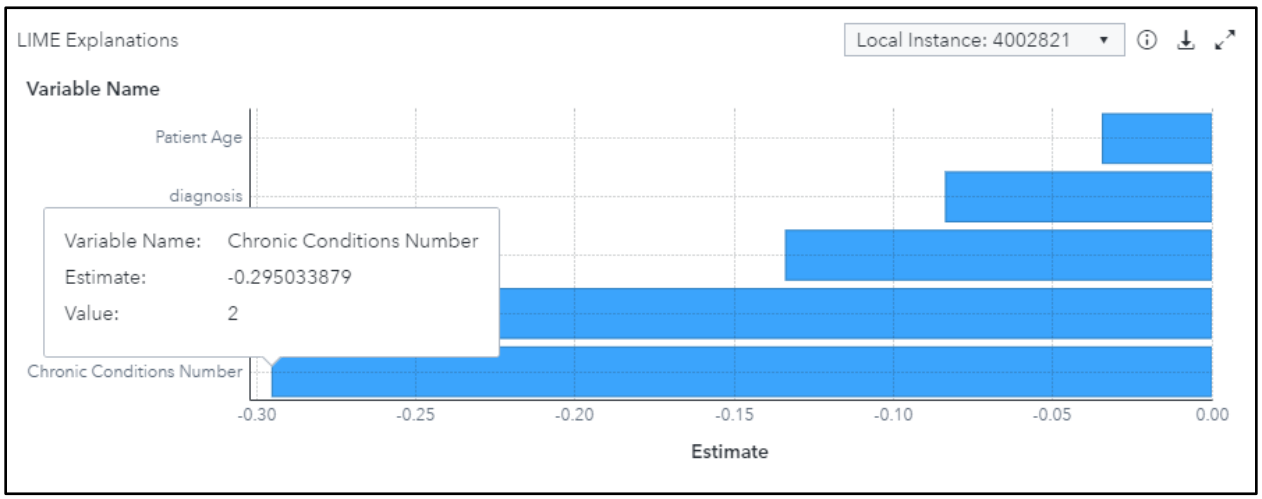

Figure 11. LIME Explanation Automatically Generated in Model Studio for Local Instance 4002821

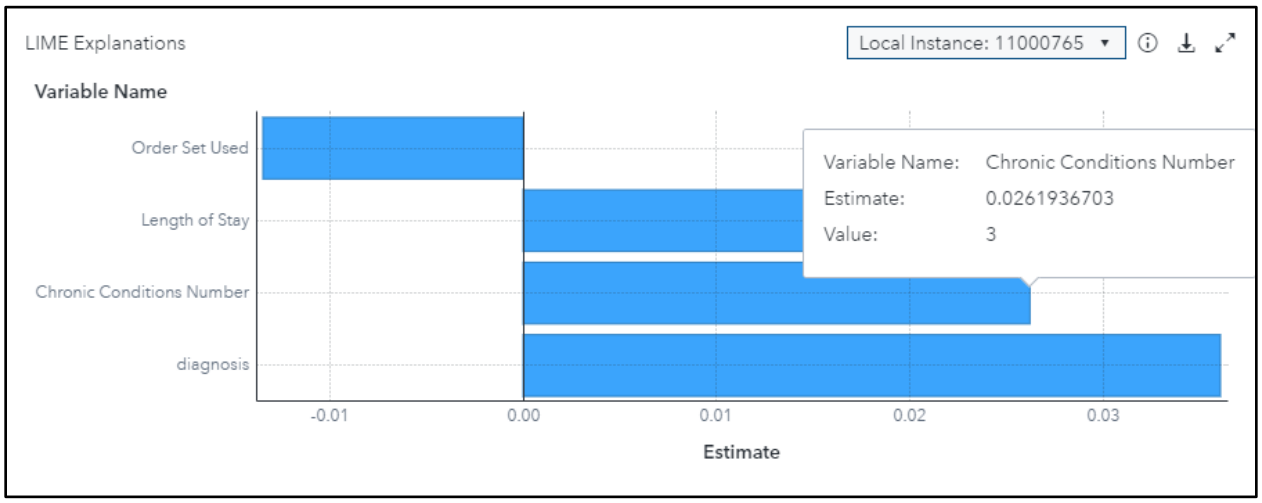

Figure 12. LIME Explanation Automatically Generated in Model Studio for Local Instance 11000765

Figure 11 (Local Instance 40002821) illustrates that all of the observed values for the variables listed (Patient Age, diagnosis, Order Set Used, Length of Stay, and Chronic Conditions Number) decrease the likelihood of readmission. For example, the value of 2 for the Chronic Conditions Number lowers the readmission prediction by 29%, while Figure 12 shows a value of 3 for Chronic Conditions Number increases the readmission risk by 2.6%.

While LIME allows you to understand how a change in a variables value impacts the model's prediction, Kernel SHAP values are used to explain the contribution of each variable to the prediction of a single observation. The sum of the Kernel SHAP values for all inputs is equal to the predicted value. The inputs are displayed in the chart ordered by importance according to the absolute Kernel SHAP values, as illustrated in Figure 13. SAS also provides the HyperSHAP method of generating shapley values that is accessible through SAS code (SAS Institute Inc., 2019).

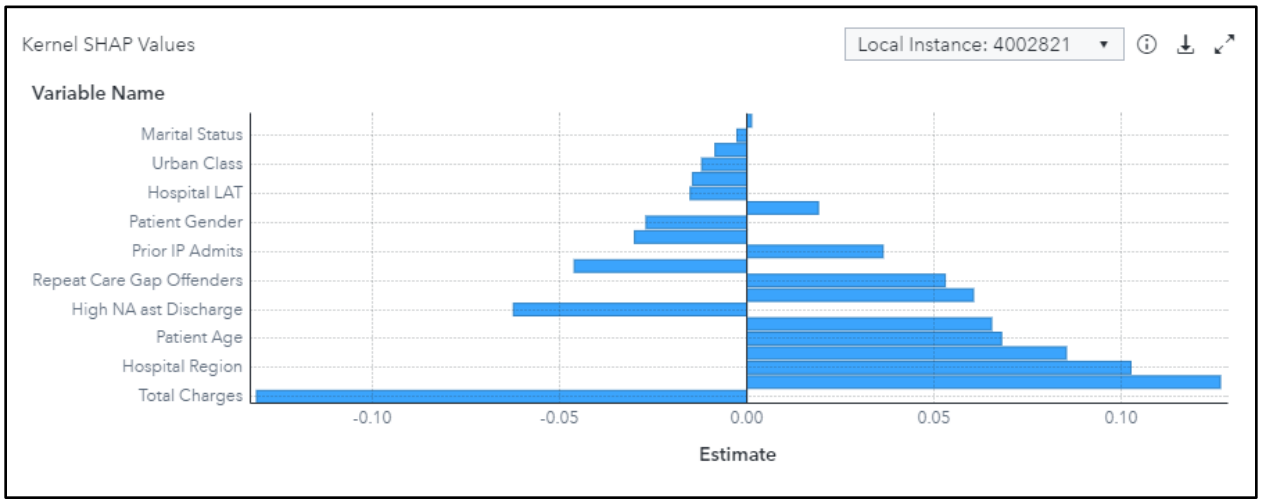

Figure 13. Kernel SHAP Values

These methods provide you the ability to understand not only which variables influence your predictions, but how they influence the predictions. Utilizing model interpretability tools and the insights they generate influence your model building process to become iterative, and you might need several models and cycles to feel comfortable with which inputs are driving

predictions.

DIVERSITY IN MODEL BUILDING PROCESS

An alternative approach to mitigating unintended discriminatory model predictions is to include a diverse set of perspectives, ideas, and approaches in the model building process. Research shows that diversity in teams and in leadership roles leads to innovation and greater financial returns (Lorenzo and Reeves, 2018). Diversity is quickly being recognized as a business strategy. Researchers have also investigated the role of diversity across 2.5 million research papers and found that research teams with more diversity produced research papers that received more citations and publications in higher-impact journals (Freeman and Huang, 2014).

CONCLUSION

As machine learning models become more pervasive in our day to day lives and as the automation of the modeling process is more heavily practiced, we need to be increasingly vigilant of the of the role human biases. Education and determination are the keys to understanding and reducing that impact. It is a big problem, and one that can be easy to ignore and difficult to address.

Selecting the appropriate data for your problem is one of the most important aspects of applying machine learning. Algorithms created using data with hidden or overt biases produce models that are exceptionally adept at applying and, most importantly, amplifying these human biases It's vital that you be sure to think carefully about your data sources and how you are applying them to your scientific questions in order to ensure that you minimize the impact of human biases. Continuing the conversation with colleagues is an essential step in raising awareness of the potential pitfalls of utilizing machine learning without thoughtful oversight.

REFERENCES

Dastin, Jeffrey. "Amazon scraps secret AI recruiting tool that showed bias against women." Reuters, October 9, 2018. Available https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-iduskcn1mk08g

Devlin, Hannah. "Genetics research 'Biased towards studying white Europeans.'" The Guardian, October 8, 2018. Available https://www.theguardian.com/science/2018/oct/08/genetics-research-biased-towards-studying-white-europeans

Freeman, Richard. and Huang, Wei. "Collaboration: Strength in diversity." Nature, September 16, 2014. Available at https://www.nature.com/news/collaboration-strength-in-diversity-1.15912

Johnston, Casey. "Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs." Wired, April 16, 2012. Available https://www.wired.com/2012/04/netflix-prize-costs/

Kabul, Ilknur. "Understanding and interpreting your data set." Available https://blogs.sas.com/content/subconsciousmusings/2018/03/09/understanding-interpreting-data-set/. Last modified March 9, 2018. Accessed February 4, 2020.

Lorenzo, Rocio and Reeves, Martin. "How and Where Diversity Drives Financial Performance." Harvard Business Review, January 30, 2018. Available https://hbr.org/2018/01/how-and-where-diversity-drives-financial-performance

Hunt, Vivian, Layton, Dennis. and Prince, Sara. "Why diversity matters." McKinsey, January 2015. Available https://www.mckinsey.com/business-functions/organization/our-insights/why-diversity-matters

Morse, Susan. "Study finds racial bias in Optum algorithm." Healthcare Finance, October 25, 2019. Available https://www.healthcarefinancenews.com/news/study-finds-racial-bias-optum-algorithm

Piper, Kelsey. "Google cancels AI ethics board in response to outcry." Vox, April 4, 2019. Available https://www.vox.com/future-perfect/2019/4/4/18295933/google-cancels-ai-ethics-board/

Tan, Pei-Yi. "Improving model interpretability with LIME." Available https://blogs.sas.com/content/subconsciousmusings/2018/10/31/improving-model-interpretability-with-lime/. Last modified October 31, 2018. Accessed February 4, 2020.

Wright, Ray. 2018. "Interpreting Black-Box Machine Learning Models Using Partial Dependence and Individual Conditional Expectation Plots." *Proceedings of the SAS Global Forum 2018 Conference*. Cary, NC: SAS Institute Inc. Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1950-2018.pdf

Zech, John. "What are radiological deep models actually learning?" Medium, July 8, 2018. Available https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Ribeiro, Marco, et al. 2016 ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135-1144 doi: 10.1145/2939672.2939778

SAS Institute Inc. 2019. SAS® Visual Data Mining and Machine Learning 8.5: Programming Guide. Cary, NC: SAS Institute Inc.

RECOMMENDED READING

Nevala, Kimberly. 2017. SAS Institute white paper. "Machine Learning Primer." Available https://www.sas.com/en/whitepapers/machine-learning-primer-108796.html

Malliaraki, Eirini. "Toward ethical, transparent and fair AI/ML: a critical reading list." Medium, February 19, 2019. Available https://medium.com/@eirinimalliaraki/toward-ethical-transparent-and-fair-ai-ml-a-critical-reading-list-d950e70a70ea

Knight, Will. "Biased Algorithms Are Everywhere, and No One Seems to Care." MIT Technology Review, July 12, 2017. Available https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/

Torres, Orlando. "7 Short-Term AI ethics questions." Medium, April 4, 2018. Available https://towardsdatascience.com/7-short-term-ai-ethics-questions-32791956a6ad

Hudson, Laura. "Technology Is Biased Too. How Do We Fix It?" FiveThirtyEight, July 20, 2017. Available https://fivethirtyeight.com/features/technology-is-biased-too-how-do-we-fix-it/

Princeton University. "Pay no attention to that man behind the curtain." American Association for the Advancement of Science, January 18, 2017. Available https://www.eurekalert.org/pub_releases/2017-01/pu-na011317.php

Hao, Karen. "This is how AI bias really happens – and why it's so hard to fix." MIT Technology Review, February 4, 2019. Available

https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/

Lipton, Zachary. 2017. "The Mythos of Model Interoperability." University of Southern California, San Diego. Available https://arxiv.org/pdf/1606.03490.pdf

Pandey, Parul. "Is your Machine Learning Model Biased?" Medium, February 7, 2019. Available https://towardsdatascience.com/is-your-machine-learning-model-biased-94f9ee176b67

Baer, Tobias and Kamalnath, Vishnu. 2017. McKinsey and Company. "Controlling machine-learning algorithms and their biases." Available https://www.mckinsey.com/business-functions/risk/our-insights/controlling-machine-learning-algorithms-and-their-biases

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Jim Box
SAS Institute
Cary, NC, USA

Email:
jim.box@sas.com

Elena Snavely
SAS Institute
Cary, NC, USA

Email:
elena.snavely@sas.com

Hiwot Tesfaye
SAS Institute
Cary, NC, USA

Email:
hiwot.tesfaye@sas.com