Machine Learning Isn't Always a Black Box

There are a number of common myths about machine learning, which we touched on in a previous post. Now, we want to dive deeper into one of those myths: that machine learning is a black box.

While it’s true that many machine learning platform companies do not provide any insights about results or explain how their algorithms and models work, this does not have to be so. Some companies invest in making their machine learning predictions and results interpretable to users. Machine learning isn’t always a black box.

Interpretability is important

Companies that provide machine learning solutions should help users interpret the results. This instills a sense of trust for users, helping them feel that they’re not using an opaque system. For example, some fraud prevention platforms score transactions but don’t provide adequate explanations of the scores. Without these explanations, users won’t have confidence in the platform and may question its accuracy.

In a 2015 presentation, Sift Science CEO and Co-Founder Jason Tan explained that customers need to understand where scores come from, so they can trust the system enough to make automated decisions based on those scores.

“The nice property of an algorithm like Naïve Bayes or decision forest is that it’s very easy to visualize,” Tan said. “We’ve built out this whole console that our customers can log into to get insight into why someone was scored the way they were, and to get transparency.”

In addition to building trust between human and algorithms, another important aspect of interpretability is that it actually augments the human in decision making, helping focus human attention and intuition to a small set of important data, and thereby improving efficiency.

Interpretability is challenging

There are many different types of machine learning algorithms, models, and approaches – each with their strengths and weaknesses. One of the classic problems in the field of machine learning is telling a story about the results that come from those algorithms and models. Some models are extremely accurate but not very interpretable. Some models are highly interpretable, but not all that accurate. And AI researchers are starting to design models that are optimized for both – though skepticism remains about whether this is even possible.

Ultimately, regardless of what that algorithm or model surfaces, a challenge exists across all of them: how do you use human language to describe why the algorithm surfaced a particular example of fraud – for example, a particular user, order, or transaction – as being “bad.” It’s difficult to explain why an algorithm or model make a specific prediction. It’s challenging to describe or present the results in a way that humans can interpret and understand.

Explaining results with visualizations

Visualizations and dashboards are a great way to explain the results of a machine learning-based platform. Visualizations can take complex machine learning concepts and simplify them so that users can interpret and understand the results. For example, Sift Science provides dashboards where users can log in and find out why fraudulent transactions are scored the way they are.

Sift Science provides users with a numerical score that indicates how likely an order, user, or transaction is fraudulent. Users base their decisions on these scores, and it is important that they understand what those scores mean and whether or not a transaction should be allowed to proceed or should be reviewed to ensure it isn’t fraudulent.

Sift Science uses visualizations to tell the story of fraudulent transactions and suspicious signals. The visualizations simplify the results that come from the platform’s advanced machine learning algorithms and models. Users can also access the raw data and activity details behind the visualizations. Providing both visualizations and raw data helps users better understand different levels of fraudulent activity. Visualizations also help guide users so that they take the proper course of action.

Users must understand the results

Interpretability is key when it comes to machine learning. Users of fraud prevention and other machine learning platforms must understand the results to trust that the platform is doing its job. Plus, interpretability can help users make better, more efficient decisions. Visualizations are a great way to accomplish this.

Bottom line: machine learning platforms don’t have to be black boxes.

Tags: machine learning

Janet Wagner

Janet Wagner is a technical writer who specializes in creating well-researched, in-depth content about machine learning, deep learning, GIS/maps, analytics, APIs and other advanced technologies.