Machine Learning Explainability: 1st lesson – Use Cases for Model Insights

***What types of insights are possible?***

Many people say machine learning models regarded as “black-boxes”, in the sense that they can make reasonable predictions but programmers can’t understand the logic behind those predictions. This statement is true in the sense that most data scientists don't know how to extract insights from models yet.

However, this micro-course will teach user’s techniques to extract the following insights from sophisticated machine learning models:

* What features in the data did the model think are most important?
* For any single prediction from a model, how did each feature data affect that particular prediction?
* How does each feature affect the model’s predictions in a big-picture sense, like what is its typical effect when considered over a large number of possible predictions?

***Why are these insights valuable?***

These insights have many uses, including:

* Debugging

The world has a lot of unreliable, disorganized and generally dirty data. Users add a potential source of errors as writing preprocessing code. Add in the potential for target leakage, and it is the norm rather than the exception to have errors at some point in a real data science project. Given the frequency and potentially disastrous consequences of bugs, debugging is one of the most valuable skills in data science. Understanding the patterns a model is finding will help users identify when those are at odds with their knowledge of the real world, and this is typically the first step in tracking down bugs.

* Informing feature engineering

Feature engineering is usually the most effective way to improve model accuracy. Feature engineering usually involves repeatedly creating new features using transformations of raw data or features that has been previously created. Sometimes users can go through this process using nothing but intuition about the underlying topic. But users will need more direction when having 100s of raw features or when they lack background knowledge about the topic that currently working on.

* Directing future data collection

Users have no control over datasets you download online. But many businesses and organizations using data science have opportunities to expand what types of data they collect. Collecting new types of data can be expensive or inconvenient, so they only want to do this if they know it will be worthwhile. Model-based insights give a good understanding of the value of features you currently have, which will help reasoning about what new values may be most helpful.

* Informing human decision-making

Some decisions are made automatically by models. Amazon doesn't have humans (or elves) scurry to decide what to show them whenever they go to their website. But many important decisions are made by humans. For these decisions, insights can be more valuable than predictions.

* Building trust

Many people won't assume they can trust their model for important decisions without verifying some basic facts. This is a smart precaution given the frequency of data errors. In practice, showing insights that fit their general understanding of the problem will help build trust, even among people with little deep knowledge of data science.