Model Health and Performance Monitoring

Ensure Your Predictive Models are Running Optimally with Model Health Monitoring

NLP LOGIX
AI REALIZED

a white paper

Introduction

As more and more businesses rely on predictive models within their day-to-day operations, it is imperative to ensure their models are sustainable and working as they were originally intended. Model Health monitoring is an essential, but often overlooked, process to identify signals that a model may be degrading before it becomes completely compromised. This paper will review how monitoring model health is different than system monitoring, key measures to monitor, and the benefits of implementing model health monitoring.

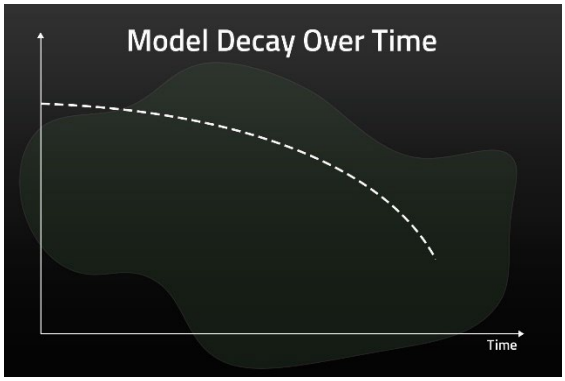
Proactively Detect Threats to Model Health

Causes of Model Performance Decay

A core tenant of modeling is the assumption that future data will look like and behave similarly to that of the past. While traditionally a solid assumption, there are many things that can occur after model deployment that can impact the data a model is dependent upon. These can be due to the issues with the training dataset, changes in the target behaviors or population of interest over time or as a result of an unanticipated event, or software system or process changes. Each of these can impact the data being processed and cause model performance to decay if not detected, which makes monitoring model health so important.

The data used to train the model is important as it is the basis for determining how each attribute influences the behavior of interest. While using a representative sample of records for training the model is the best practice, this may not always be available, or it may not be known if the data matches the greater population in which the model will be applied. If the training data was based on a subset of the population (for instance, limited to select groups, geographic areas or business units), the subset may not behave the same way as the total population, leading to potentially inaccurate predictions when applied to those not in the subset. Similarly, introducing data from a new population (e.g., company gains a new customer base or new geography) to an already deployed model may not produce optimal results for this new group, as the model was not exposed to any key differences or behaviors from this population during training. The model creators and end users of the deployed model may not have had visibility to whether the training sample was representative of the current population or if the current population has shifted from the original subset.

Even if the population remains stable, their behaviors can change over time, potentially impacting the ability of a model to accurately predict the target behaviors. External factors like changes in funding, policies, emerging trends, and technological advances can impact behaviors in unanticipated ways and change what drives key behaviors. For example, a model implemented by a health insurance company may be impacted by a newly established tax credit for certain individuals who purchase health insurance. After this credit availability, individuals who were previously identified as the least likely to purchase health insurance may now be much more likely to buy. The introduction of this new policy could change the motivating factors for purchasing health insurance, compromising the model accuracy. Monitoring how the model is performing and attributes of key behaviors can help to identify situations in which behaviors are changing and the model is at risk of no longer performing as intended.



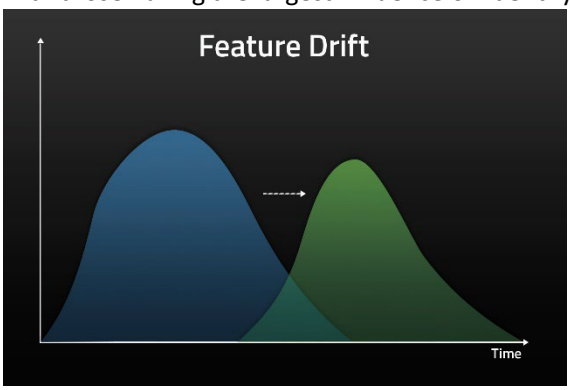
Unanticipated events, like the COVID-19 pandemic, can have a more immediate impact on model performance. Any event that impacts a sizable portion of the population or target behavior of interest can render a prior model useless. As we know with COVID-19 consumer purchasing behavior rapidly shifted to focus on essential needs, home improvements, and online shopping. The shift in consumer purchasing behavior could not be anticipated and the driving factors that were previously used to predict consumer spending or purchasing behaviors were likely diminished. While COVID-19 had an immediate and substantial shift in behavior, there could be other, less extreme events that affect the population of interest and impact their target behaviors.

Finally, internal processes responsible for the calculations, capture, or formatting of source data consumed by the model can also unexpectedly change and impact model performance. Upgrading or changing operations, systems or processes can introduce threats to model health that may otherwise go undetected. For instance, migrating to a new CRM system that appends a hyphen to all zip codes, or upgrading to a new call center software that defaults to calculating talk time without hold time, or changing the fields that are required to be populated in an Electronic Medical Records System. In each of these examples, the data the model is dependent on has changed or is provided less consistently, and while the model may still run successfully, there is not a way to detect these shifts without proactively monitoring for these types of incidences.

Leading Indicators of Model Decay

There are several indicators that can be measured and monitored to help detect potential issues that arise leading to model decay. NLP Logix focuses on four core areas of monitoring, including data completeness, feature drift, prediction drift, and prediction performance, to proactively identify potential issues with model health.

In most models there are usually several features that are considered the top drivers of the target behaviors, with those having the largest influence on identifying the behavior of interest. In order to ensure records are scored accurately, it is essential they have all the necessary data provided in a manner the model will be able to recognize. Data completeness analysis ensures that the necessary source data is complete, and the generated features match the expected range and format as the data used for training the model. Models may not recognize unexpected values or value formats and, as a result, handle them improperly. Additionally monitoring the rate of missing data across source data fields helps to ensure the model is creating the most accurate predictions across all records.



In addition to completeness, it is important to also understand if any of the features have changed by testing for feature drift. Tests of feature drift ensure that the distribution of the new data being scored is similar to the data used for training the model. These tests help identify many of the causes of model decay, including when the population changes, or there are errors or issues with the source data. These tests compare the score distributions of all the features against their distributions in the original training data. While some shifts may be expected or driven by seasonal effects, we want to make sure we are actively monitoring the data to be able to track and understand these trends so we can identify performance issues before they become problematic.

Similar to feature drift, we also measure prediction drift to ensure the distribution of generated probabilities or classifications are similar to what was produced with the original model. This is important as a shift can indicate there is potentially a change in the behaviors of interest or in how the features of the model relate to those behaviors. Often in a business processes model predictions are directly translated into decisions and actions, and any unanticipated shifts in probabilities can impact vital processes, potentially targeting the wrong records, or conversely, not identifying critical records.

Finally, the last set of model health tests measure the stability of model performance over time. This includes evaluating whether the model accuracy, precision, recall, and AUC are all in range with the original model. These tests ensure the model is still classifying records in the expected manner by comparing the model classification to an actual result for the target behavior. While these are often highly desired metrics to be tracked, they are only available when the client is able to provide labelled data, which may be difficult. Fortunately, even without labelled data, the other model health indicators are able to help identify signs that model performance may be compromised.

The ongoing health and sustainability of a deployed model is essential to ensuring critical business decisions are made using the most accurate information. By tracking these leading indicators at the onset of deployment and on a regular basis, NLP Logix is able to detect the more subtle changes that can lead to model decay and help to proactively remedy issues before they become disruptive to your business.

Protect Your Investment

The deployment and integration of a predictive model into one's business processes often requires a significant investment from both a time and labor perspective. Currently many companies believe that regular retraining of a model will circumvent threats to model performance and help protect their investment for years to come. However, model retraining, while effective in some circumstances, may not address the underlying issues that are limiting the model's ability to accurately and effectively classify the behavior of interest.

Retraining can help to recalibrate models where the population of interest has shifted, or to include additional population subsets, however, it will not be as effective if data from top features are inconsistently provided, if data changed, or if new data is required to adapt to changing behaviors or events. Without actively monitoring model health metrics for shifts or changes in the source data and model outputs, the causes of model decay may not be apparent, and retraining may not be the appropriate solution to ensure the model is operating as intended.

Active model health monitoring helps to detect these issues before they become problematic and begins to impact key business decisions or operations. With monitoring, NLP Logix will flag any issues that may become problematic so we can work with the client to understand what the root causes are and what the best approach is to handle it. In many cases retraining will be warranted and recommended, however there may be some cases where additional data and/or a full model rebuild may be required to ensure the targeted behaviors are correctly identified.

Summary

With each new model build or rebuild, it is important to protect your investment and ensure the model is performing as it is intended over time. There are many factors that can impact model health that cannot be detected via system monitoring alone. Using model health monitoring, deeper examination of the model inputs and outputs, including completeness of data, feature drift, probability drift, and model performance

attributes, can help to detect threats that lead to model decay. NLP Logix can implement ongoing model health monitoring from day one to proactively detect issues that may lead a model to underperform before it becomes problematic and leads to costly disruptions in business processes and decision-making.

To learn more about model health and performance monitoring and how NLP Logix can help ensure your models are sustainable for years to come, contact us at contact@nlplogix.com or visit NLPLogix.com

About the author:
Mary Sheridan, Ph.D.
Modeling and Analytics Team Lead, NLP Logix

Mary is a research and analytics leader with more than 14 years of experience in transforming data into an action plan to improve business operations, marketing, and product performance.