ClosedLoop.ai

Six Mistakes You Can Avoid in Healthcare Data Science

Contents

“Smart people learn from their mistakes. But the real sharp ones learn from the mistakes of others.”

– Brandon Mull, Fablehaven

The cover of the May 5, 2017 Economist made a bold assertion: the world's most valuable resource was no longer oil, but data. The rush for this new raw material is being fueled by artificial intelligence (AI), which has proven itself to be a capable disruptor and engine for innovation across nearly every industry, particularly advertising, transportation, and financial services.

While AI's impact continues to grow, healthcare is behind when it comes to realizing the benefits of these advanced technologies. In part, it is because healthcare is a digital laggard. Despite medicine's sophisticated devices and complex treatments, healthcare has been slow to adopt digital technologies for workflow automation or health service innovation. Exhibit A is the patient check-in experience; it is essentially the same as it was 30 years ago.

It is also because healthcare confronts unique challenges that other industries do not face, or at least not to the same degree. First is the nature of the decisions being made. Healthcare decisions include extremely sensitive information, require timely action, and can have life or death consequences. Each informational element brings unique demands and creates a high hurdle to the universal use of AI. Healthcare also has some of the strictest rules for data privacy. Patient privacy is important but the rules protecting it make accessing data quite cumbersome an effect that is compounded when multiple data sources are involved and when the cadence of data availability is not in lock step. Finally, even if electronic health records data were readily available, the majority of factors that determine a person's health (and thus the data needed to model it) lay outside the walls of the clinic.

The progress of AI in healthcare is also held in check by the required combination of skills needed to practice healthcare data science. There is already a shortage of data scientists; the estimated shortfall of trained professionals is in the hundreds of thousands. And healthcare data scientists need to know not only the coding, statistics, and machine learning skills needed to do modeling, but must also master subject-specific topics for medical coding and health measures (e.g. diagnosis codes, procedure codes, medications, lab results, quality measures, outcomes, etc). The highly challenging nature of healthcare, combined with the fact that data science is itself a young field, means that mistakes are not uncommon in healthcare data science. The goal here is simple – to highlight some of the more common errors that healthcare data scientists make and offer suggestions for ways to avoid them.

“The goal here is simple – to highlight some of the more common errors that healthcare data scientists make and offer suggestions for ways to avoid them.”

Not Considering How the Model Will be Used

One of the data scientist's most important responsibilities is to work with stakeholders so that modeling efforts are focused where they drive better outcomes and positioned so they achieve what they can achieve.

A common stumbling block is that models are built without considering the context in which they will be deployed. This 'context' has implications well before deployment; it also affects how a model is designed and trained. Take for example a model that predicts whether a patient will be readmitted to a hospital within 30 days of having been discharged. In addition to the patient, several other stakeholders want to prevent this outcome, including the hospital, the payer, and the primary care provider. Despite their shared goal, each stakeholder would need a different model to achieve it. This is because the context of each situation would be quite different.

To illustrate, consider a readmission model. One version might be used by a hospital's care team while planning for a patient to be discharged. In this setting, clinical data is often updated several times each day. The models would ideally incorporate any new clinical data, along with actions taken by the care team plus any baseline risk factors extracted from a patient's medical history. Such a model could be run multiple times per day so the training data would need shaped to match this tempo.

A different readmission model might be used by an Accountable Care Organization (ACO). Their model would not likely have the same kinds of data, particularly a patient's vitals over the course of an inpatient stay. Some ACOs have clinical data from office visits while others predominantly use claims, which can take weeks-to-months before being available. ACOs can partially bridge this gap with 'ADT records' (i.e. files with information about the admissions, discharges and transfers of their patients). While these records lack comprehensive clinical details, they send more timely signals than claims and usually include diagnostic information. Training data would need to reflect these distinct latencies.

It is also important to consider what decisions might be made with the outputs from these models. This is because models are often built to be used with specific interventions. In the discharge planning example, the model may help care teams decide whether to keep the patient in the hospital for an additional day. For the ACO, this decision would not likely be possible since in most cases, the patient would already have been discharged. Instead, the ACO's intervention strategy might be to dispatch a home health nurse to the patient's home to follow up and ensure a good transition of care. For the data scientist, the key is being able to quickly build and modify models that accurately reflect the context as more becomes known and decisions are made about how it will be used.

The ClosedLoop platform is ideally suited to this. For example, Its automated feature engineering calculates features with respect to reference times, not fixed moments in time. This means that the effort needed to take an existing set of features and synthesize time series is minimal. In addition, feature sets are highly customizable (e.g. filtering based on admission type is as simple as looking up diagnosis codes). This makes it easy to build models for specific interventions and reduces barriers to deployment, which is key to promoting adoption.

Not Anticipating Deployment During Model Creation

Some models can be difficult to operationalize after they are created. To avoid this, data scientists, despite often having scientific backgrounds, would benefit from adopting an engineering mentality.

One of the main roadblocks is computational complexity. Perhaps the most famous example is the original 'Netflix Challenge'. Netflix sponsored a competition to create better recommendations wherein the winning team was awarded $1 million. Unfortunately, the winning model was too computationally complex to deploy. Complexity can be a serious and sometimes fatal barrier to deployment and healthcare is no exception; its data requires significant preprocessing to be made suitable for modeling. If you leverage GPUs or spark clusters to create the model, but those resources are not available in production, the model will never get used.

Data and featuring engineering, in addition to having its own complexity, can also bring deployment challenges. In healthcare, it requires heavy amounts of 'time windowing' and data scientists often use hard-coded date windows and boundaries to get their models up and running. But the use of fixed dates needs to end when moving models into production and if such windows exist for multiple features, the likelihood of mistakes increases dramatically. Another issue linked to deployment is instrumentation. Data scientists do a great deal of performance measurement during training but often stop monitoring performance once models are in production. This is a critical mistake because the underlying conditions that affect a model's performance can change.

There is an assumption in modeling – that observations are independent and identically distributed (IID) – and it is critical for a model to be valid. As time progresses, however, this assumption may become invalid. Measuring model performance and data drift is as important after deployment as it is during training to ensuring the model is performing consistently and as expected.

The ClosedLoop platform was built with deployment in mind. While having a deployment mindset can help, having access to the right tools makes this much easier. The ClosedLoop platform leverages cloud computing, so resources scale as computational needs grow; its proprietary language, CL Expressions, makes dynamic time windowing effortless; and its automated ML Ops monitor every step of production, including data uploads, model runs, and reporting.

> **"Measuring model performance and data drift is as important after deployment as it is during training to ensuring the model is performing consistently and as expected."**

Allowing Data Leakage

Data leakage is one of the biggest problems in data science. It happens when data from outside the training set is used to build the model. This allows the model to learn using data it should not know or have access to and can lead to overly optimistic if not outright invalid models.

Most data scientists know this and understand how to navigate the important details of the train-validation-test process. Yet healthcare data science models often undergo a performance degradation when they are deployed, a tell-tale sign of data leakage. The most common reasons are due to the additional handling that healthcare data requires. A major contributor is that healthcare data science is innately time-ordered, and time series data brings several complications that time agnostic data does not. Even common tasks like computing population level statistics become exacting when using time-ordered data because the population being measured is changing over time.

A person's health is a dynamic and shifting characteristic that is best represented by using multiple data points over time. Healthcare's time-ordered data often includes multiple years of information about a person which can open significant opportunities when handled correctly. But it also requires that a model's features and outcomes be calculated dynamically and across multiple moments in time. And while this permits data scientists to create more data for everyone in the population, it significantly increases the risk for data leakage when handled incorrectly.

For example, if data for a particular individual are allowed within both training and test data sets, the model's performance will be artificially inflated. Complicating the risk of leakage is the problem of healthcare data lag, which is particularly true of administrative data (e.g. medical and pharmacy claims). Data is lagged when it becomes available long after the service it represents has been provided (e.g. a hospital admission or physician visit). The extent of the lag varies by type of service but can be weeks or months. In addition, claims data can be unreliable when it initially becomes available and is known for being corrected or altered.

In this instance, data scientists must fight their instinct to use as much data as possible, since it will reliably lead to differences in testing and deployment performance. Instead, the question should be "What would the data look like at the time the prediction is being made." It should be an 'as of' view of the past. Exploratory data analysis (EDA) should be performed to understand what the data lags look like and how to construct a reliable source of information.

Even for a data scientist who is experienced with all these issues, it is far too easy for some of the myriad details to get missed during the modeling process, especially if they are being addressed manually. The ClosedLoop platform deals with all these issues. It automatically partitions individuals into training and testing populations before generating time sequences. All features are generated dynamically using CL Expressions, and time lags can be used with a simple modification to any expression.

Inadvertently Introducing Bias

Researchers recently found that a widely used algorithm sold by the healthcare analytics company Optum dramatically underestimated the health needs of the sickest black patients and assigned healthier white patients the same risk score as it did black patients who were less healthy.

This happened even though race had been specifically excluded as a variable. At issue was that the risk scores were used to prioritize resources for patient outreach and support, a disparity which would almost certainly exacerbate the already existing health disparities experienced by African Americans. This bias in risk scores happened because Optum's algorithm, which is not unlike many other risk algorithms used in healthcare, uses total healthcare costs as the proxy to define a patient's risk. In other words, the algorithm was not directly predicting a person's future health or healthcare needs; it was predicting their future healthcare costs which would then be used to identify who would benefit from additional health services.

On the one hand, using healthcare costs as a proxy for risk is understandable. Sicker patients often use more healthcare resources than those who are less sick, and cost data from claims has been one of the few types that has been consistently defined, structured, and available. Nevertheless, using cost as a proxy frequently gives rise to bias. This is because African American patients have lower healthcare costs. Data shows that African American patients generally use fewer healthcare services than white patients and, when they are used, services are often reimbursed at lower levels, particularly under Medicaid coverage. This source of bias can be greatly reduced by using outcomes other than total costs when designing algorithms. Alternatives would be to predict an adverse health event (e.g. a hospital admission) or a decline in health (e.g. disease progression). When researchers use these alternatives, they found the new algorithm had far less bias. The problem was not the algorithm per se; it lay in what the algorithm was being asked to do.

Focusing on events and other health outcomes has other benefits as well. A key value of models often flows from understanding why the model made a particular prediction. It is well understood that high costs correlate with poor health, but such knowledge is rarely actionable on its own. Rather than seeing high costs as the health outcome, it is more useful to see them as the consequence of an adverse event or poor health outcome which has a specific explanation (e.g. hospital admission, urgent dialysis start, kidney disease progression, etc.). With the right tools, data scientists can leverage data (e.g. diagnoses, procedures, medications, lab results, social determinants, patient reported outcomes, etc.) that allow these explanations to rise to the surface and make the models significantly more actionable.

Bias can creep in for other reasons too including framing the wrong problem, training with biased data, being introduced through feature selection, or using an inadequate definition of fairness. Data scientists must always measure the degree of bias in their models and work with stakeholders to identify its sources and reduce its impact. A standard metric used to evaluate fairness is called disparate impact. It measures the proportion of a group that receives a positive benefit and compares proportions between unprivileged and privileged groups. The common benchmark is a "four-fifths rule"; if an unprivileged group receives a positive benefit less than 80% of their proportion compared to the privileged group, it is deemed a disparate impact violation.

But the disparate impact measure is completely unsuited to healthcare situations. First, it only accounts for one type of modeling error – the false positive. This is woefully inadequate in situations where errors in the other direction – false negatives – cost dearly; individuals who could benefit from an intervention do not get it. The measure also fails to adjust for instances where the alarm rate for the reference group is too low, a problem which is not solved by the arbitrary benchmark.

The ClosedLoop platform has built-in capabilities for helping to combat bias. ClosedLoop developed a new method for measuring fairness called Group Benefit Equality (GBE). It explicitly addresses the shortcomings of standard fairness metrics when used in healthcare. GBE measures the rate at which a particular event is predicted to occur within a subgroup compared to the rate at which it actually occurs. GBE is easily explained, has transparent procedures, and uses clearly defined thresholds to assess when models are biased.

Data scientists should think critically about potential sources of boas and take seriously any differences they observed. And while no single metric is a silver bullet against algorithmic bias, the ClosedLoop platform includes a metric that is always available, completely transparent, easy to explain, and suited to healthcare.

Failing to Distinguish Between Missing Data and No Data

When working with data that represent a person's conditions and treatments, data scientists must realize they will face situations where there is no data: sometimes it is because there is none to be had; at other times, data exists but is 'missing'.

As a rule, data is created whenever a person interacts with the healthcare system. For example, a person generates data when they go to the doctor. There will be data extracted from the bills sent to the insurance company and, because of the increased use of electronic medical records (EMRs), a good chance there will also be a detailed record with clinical data (e.g. symptoms, chief complaints, lab results, vital signs, prescribed medications, etc.). Unfortunately, this rule is not universal. It is possible to have no data on a patient, despite their interaction with the healthcare system. It is also possible to have some data about a patient but not all of it, including having different data for otherwise identical patients.

These situations happen for multiple reasons including insurance coverage gaps, inadequate data integration capabilities, dissimilar EMR systems, or different coding practices. Consider a healthy adult who mainly has an annual physical as compared to someone in poor health with little-to-no insurance coverage. Both will generate small amounts of data despite significant differences in their health. Accounting for coverage gaps is important; they are correlated with other factors that drive poor health outcomes and failing to do so could exacerbate existing health disparities.

As another example, consider a patient with asthma and their use of rescue inhalers. Rescue inhalers are not medications one hopes will be needed on a frequent basis. To gain insights into how often this is happening, data scientists can examine the number of times those medications are filled in the previous year. Three such fills might be considered reasonable if the patient had pharmacy coverage for the entire period. But, if the patient had just established coverage or had intermittent eligibility, this same number might suggest a materially different usage pattern. For the data scientist, one approach is to average the number of fills over the number of coverage eligibility months.

This type of normalization will make high utilizers more comparable. It can also be useful to segment populations based on such characteristics. In clinical settings, it is important for data scientists to avoid assuming that data missingness is random or because of imperfections in the data collection process. In some instances, missing data can be an indicator that a key process has not occurred, and so leaving Null values in place could be useful. At other times, data scientists may impute missing values. Data scientists may find missingness patterns that correlate with adverse events. Such patterns, while real, can be specific to a particular clinical context and data capture system; they may not generalize to other institutions or settings. In such situations, data scientists should take care to train models with an institution's specific data.

The ClosedLoop platform makes handling these situations straight forward. Feature creation is geared toward time windowing, and the platform automatically tracks and versions populations. This makes it easy to create segments of intermittent coverage and to generate time sequences of eligibility months based on the provided intervals. One simply counts the months within the relevant range and provides that count as the normalization constant for feature creation. The platform is also robust to missing values. Null values can be left as is or, in instances where it is important to fill in missing data, use conditional logic that allow the application of various rules (e.g. if a reading is dropped in a time series, the average of reading over a fixed time window can be substituted).

“In clinical settings, it is important for data scientists to avoid assuming that data missingness is random or because of imperfections in the data collection process. In some instances, missing data can be an indicator that a key process has not occurred, and so leaving Null values in place could be useful.”

Not Predicting Impactable Risk

A common assignment for data scientists is to build a model that predicts individuals who are high risk. Such models are then often used to target care management programs that deliver specific health services designed to reduce risk and mitigate or prevent future adverse events.

The problem is that intervening with the highest risk people does not necessarily lead to the greatest impact. For example, most models would predict a 97-year-old obese diabetic as having a higher risk of hospitalization than a 37-year-old with the same conditions. Yet few experts would prioritize the 97-year-old for an intervention that emphasized weight loss and lifestyle changes. On the other hand, if the model were predicting the risk of readmissions and the intervention was delivering home health and care giver support, experts might see the intervention as more beneficial to the 97-year-old. Such stark contrasts are rare in the real world, but the point is that impactability is not solely a function of risk; it emerges from the interaction of the person and the program.

Healthcare organizations want to do more than predict high risk. They want to identify where their program will have an impact and predict which individuals are likely to be the most impacted by it. Data scientists need to be vigilant as they answer these questions because they place distinct demands on the kinds of models that need built and the data that will be required. From a modeling perspective, while the ideal approach would use models built on a causal chain, this field of machine learning is still maturing. In lieu of causal modeling, other machine learning approaches can be used, especially if explainability is emphasized and data scientists engage subject matter experts to assess the results. If data scientists opt for a 'black box' approach to modeling, such collaboration becomes intractable and can lead to poor uptake from health teams.

> **“Data scientists need to be vigilant as they answer these questions because they place distinct demands on the kinds of models that need built and the data that will be required.”**

One feature of the ClosedLoop platform is the ability to tailor the population used for specific models. This helps avoid including known high-risk patients (e.g. cancer patients who have care management intrinsic within their treatment plans). It also allows models to be aimed at populations who are the target for the intervention (e.g. pediatric asthma). Another feature is the platform's focus on explainability.

Using its 'Contributing Factors' technology, ClosedLoop's platform surfaces the specific factors that best explain each prediction, which are calculated at both population-specific and patient-specific levels. Finally, the platform is built to support the rapid iteration and collaboration needed to produce the models capable of achieving a meaningful impact. This capability is crucial; models are sensitive to the different variables they are trained upon, and they will learn the unique Contributing Factors when the data is changed.

Conclusions

Healthcare data scientists must confront a host of challenges that do not exist in other industries. The fact that many data scientists come to healthcare from non-healthcare backgrounds means they will not be familiar with the subtle-yet-vital details waiting for them. Using all-purpose tools makes avoiding them effortful even for data scientists that are skilled in the profession's best practices.

The ClosedLoop platform is built for healthcare data science. Its entire purpose is to unlock the potential of data science to fuel meaningful change in healthcare. By providing a tool that automatically handles many of the procedural details, data scientists can focus their critical thinking on ways to best leverage data to solve real world problems.

About the Author

Joseph Gartner is a PhD physicist and data scientist with nearly 18 years of professional coding experience and a passion for unlocking the potential of machine learning and AI. At ClosedLoop, he serves as Director of Data Science, applying his expertise to solve some of the most challenging problems in the healthcare industry. He was instrumental in the development of ClosedLoop's C-19 Index, an open source, AI-based predictive model that identifies individuals with a heightened vulnerability to severe complications from COVID-19.

Prior to joining ClosedLoop, Joseph served as Lead Instructor and Principal Data Scientist at Galvanize, a technology education community that specializes in providing immersive data science and software engineering courses. Throughout his career, Joseph has remained at the forefront of scientific progress and the development of innovative technologies. As a particle physicist, he was on shift when the first collisions occurred at CERN's Large Hadron Collider, and he contributed to DARPA's XDATA project, an initiative dedicated to developing the tools necessary to efficiently analyze and disseminate massive volumes of data to inform mission-oriented decisions. Joseph graduated from the University of Florida with a PhD in particle physics. He currently lives in Austin and enjoys learning Brazilian Jiu Jitsu in his free time.