CASE STUDY

Using a Narrow Data Set to Accurately Predict Diabetes

hc1.com Model Predicts Diabetes Risk with 77% Accuracy –
Without Reliance on Historical Indicators Such as Glucose Results

BACKGROUND

A world where machines know more than people do about health and human behavior is becoming our reality. Every day, machines help researchers and caregivers better understand patient populations – whether it be those who may become sick, relapse, or even face a transplant rejection. New correlations and indicators are ever evolving and play a vital role in both preventing and managing disease.

While the possibilities are exciting, the daily pressure that healthcare organizations face to improve the quality of care while also reducing costs can be overwhelming. The endless stream of data produced and stored by healthcare entities can be extremely difficult to compile into actionable insight without the help of machine learning.

The data collected by caregivers must be processed and analyzed in order to be utilized effectively. However, the amount of manual work and resources required to sift through all of that data is something few hospitals and healthcare systems have available. By using machine learning, these healthcare providers can create highly personalized patient experiences that cost significantly less. Machines can quickly analyze large sets of data and then apply relationship modeling in a predictive way and in near real time.

Our team at hc1 recently developed and completed a healthcare-focused machine learning case study by utilizing our proprietary data lake and patent-pending data refinery to support deep learning.

The hc1 case study sought to answer the following questions:

1. Would it be possible to produce highly accurate predictive results using a narrow set of healthcare information?
2. More specifically, would it be possible to predict diabetes risk based solely on unrelated historical test results?

Result: By leveraging 5 million results in the hc1 data lake (which houses over 3.8 billion lab test results) and applying machine learning technology, the team was able to accurately predict diabetes risk with an 77% accuracy rate.

CASE STUDY DESIGN

The case study was used to infer diabetes risk based on historical lab work for a sample population and was designed as follows:

1. Only lab test results were allowed as variables.
2. The model used a combination of tests highly linked to diabetes to define the outcome variable. As a result, these tests were excluded from all predictive data sets.
3. The model was created internally based on outcomes that are used to determine diabetes without using a diagnosis code.
4. The machine learning model predicted the probability for diabetes detection based on past lab results unrelated to glucose or diabetes.

DATA PROPOSED

- At least 3 years of a data sample of de-identified patients who had their sugar levels tested at least once.
- For those patients, the study only used lab results from the above – no other variables.
- No geographical or demographic information was allowed.

CONTROL

The study used outcome results defined above to assess the accuracy of the predictions as to who developed diabetes.

FINDINGS

2.5 million records were uploaded into the hc1 environment. That data was automatically cleaned, conditioned and categorized. Then, the outcome variable was defined using a set of parameters for typical diabetes testing.

hc1 created a statistical model using any other non-diabetes tests performed on those same patients as predictive variables. The random forest regression method for learning was then applied. Using 500,000 records, the model, through machine learning, learned the typical profiles and trending for diabetic, normal and pre-diabetic patients. The resulting model was then applied to the rest of the population of 1.9 million records to predict each patient's categorization. Using the known patient results for diabetic testing, the model prediction was compared to the actual result to produce the following probabilities:

Prediction Percentage Correct

- **77% - diabetic**
- **89% - normal**
- **57.8% - pre-diabetic**

New Correlations

The hc1 machine learning model highlighted the following variables as highly significant in the prediction of a diabetic outcome.

- **Globulin**
- **Creatinine**
- **Hemoglobin**
- **Triglycerides**
- **Chloride**

Interestingly, the model highlighted the importance of chloride levels, historically considered a less relevant indicator as related to diabetes, but a possible indicator of kidney disease complications. Several new diabetic testing panels have begun including Chloride testing as part of their Diabetic panels. Other scientific studies have echoed this sentiment relating chloride levels back to kidney disease complications.

FINAL RESULT

The study used the outcome results defined above to assess the accuracy of the predictions as to who developed diabetes.

NEW POSSIBILITIES

How Healthcare Can Use Machine Learning to Solve Other Challenges

As confirmed by the diabetes case study, hc1 offers organizations a way to effectively predict the health of a specific population. These results foster actions related to:

- Where further patient engagement is necessary
- Relevant follow-up testing for patients
- Whether preventative measures or patient counseling is necessary

Using machine learning capabilities, healthcare organizations are able to address important initiatives such as:

- **Population Health Management** – Using similar models, hc1 could allow for predictions around categorization of care for homeless populations, heart disease, chronic kidney disease, as well as sepsis using blood culture, UTI, or ammonia testing. The model can be set up to retrain itself as the population changes and rerun on a preset basis.
- **Resource Management** – Healthcare providers are increasingly moving to a quality-based care paradigm that demands higher service delivery levels while lowering costs. The best, and possibly only, way to accomplish this at scale is through implementing machine learning and artificial intelligence solutions into their larger strategies.

By leveraging machine learning technology, health systems are able to greatly reduce medical treatment costs, optimize organizational processes and improve healthcare delivery. These benefits also apply to large medical and state governments looking to influence health policy changes in their areas.

Ultimately, machine learning and artificial intelligence position the healthcare industry to increase efficiencies, save money, and save lives.