Diabetes Action Canada: Advanced Analytics Data Workshop

July 2019

From the working files of Conrad Pow and Tao Chen

Diabetes Action Canada, in collaboration with the Fields Institute Centre for Quantitative Analysis and Modelling (Fields-CQAM), and the Vector Institute for Artificial Intelligence held a two-day data workshop June 17th and 18th 2019. Students, post docs and researchers applied advanced analytics to a de-identified dataset comprised of electronic medical records (EMRs) of patients living with diabetes.

The exercise provided much needed insights into the feasibility of using advanced mathematic modelling and artificial intelligence on Canadian EMR data in a high-performance computing environment.

The analytic environment was housed at the Centre for Advanced Computing (CAC) at Queen's University in partnership with Indoc Research. Two environments were created to offer flexible and scalable solutions to meet the needs of both streams of analysis.

The hardware environment for "Environment One" was a 6-core, 16GB memory and 1TB storage Windows 10 server. The software environment consisted of Python 3 and R. The users used Remote Desktop Protocol to access the server. Commonly used machine learning and visualization packages in Python and R were preloaded; these included numpy, scipy, sklearn, matplotlib, ggplot2 and glmnet. Tensorflow and PyTorch were also installed to enable the use of deep learning models.

The hardware environment for "Environment Two" was a 32-core, 192GB memory and 1TB storage CentOS Linux server. Each server also consisted of an NVIDIA Tesla V100 32GB card to speed up the development of deep learning models. The users used Secure Shell (SSH) to access the server. The suite of installed software was the same as Environment One.

Environment One: Mathematic Modelling/Machine Learning

This Environment was used to predict the response to a specific class of medication (SGLT2 Inhibitors) for diabetic patients using machine learning techniques.

The data were first divided into training and testing datasets. Different machine learning algorithms were then applied to the training dataset; these created models for predicting the response of patients to the medication class. This model was then validated on the testing dataset.

The methods that were used are: Random Forest, Naive Bayes, Support Vector Machine (SVM) and XGboost. Random forest and XGBoost outperform with 89% accuracy to predict the response of patients to the medication. The results of each method are shown in Figure 1 below:

Numbers	Modeling Algorithms	Accuracy	Recall	Precision
1	Naïve Bayes	52%	73%	53%
3	SVM	86%	89%	86%
5	**Random Forest**	**89%**	**96%**	**85%**
6	XGboost	**89%**	**89%**	**90%**

Figure 1: Comparing the results of all algorithms

The results show that it is possible to predict patient responses to this category of medication with high accuracy based on their health records. We can replicate this method so the prediction model can be expanded to all medication classes

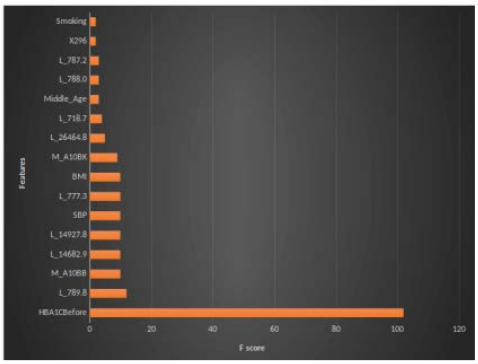


Figure 2: Feature Importance based on XGBoost

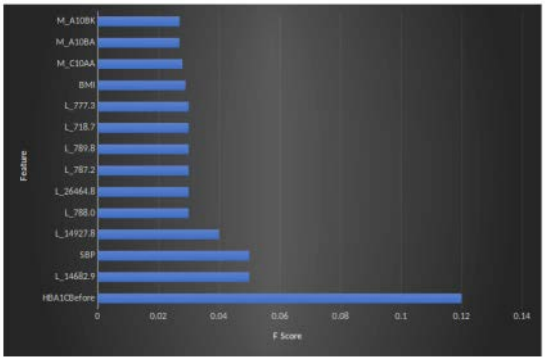


Figure 3: Feature Importance based on Tree Classifier

Figure 2 and Figure 3 show that the four most important features for predicting responses to the medication are HbA1c, serum creatinine, triglyceride levels, and systolic blood pressure. These features, as well as the other less important features, could be used to help physicians decide whether a medication is likely to be effective at lowering A1c given patient characteristics. Better information on likely treatment effectiveness could also reduce the financial burden for the patient and on the healthcare system.

Environment Two: Environment two: Applying Artificial Intelligence using Unsupervised Machine Learning

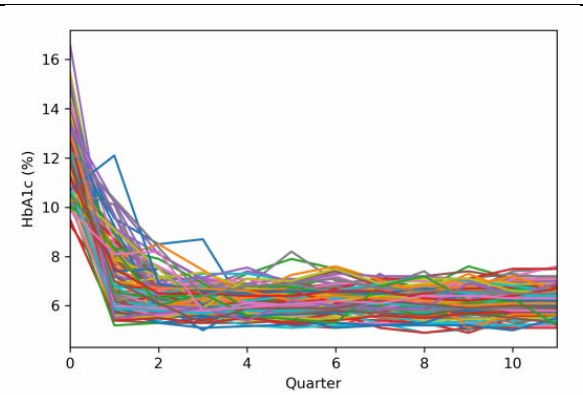
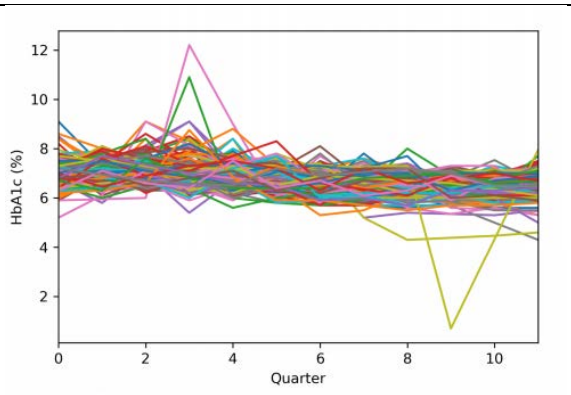
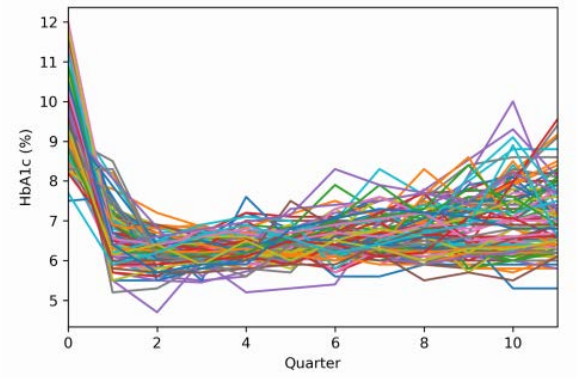
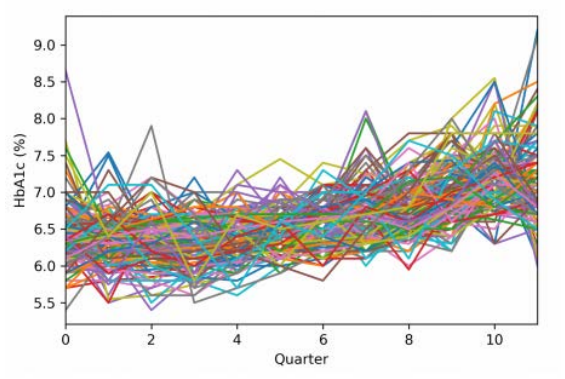
Data Scientists aimed to find out if common hemoglobin A1C trajectories can be identified using unsupervised learning techniques for adults living with diabetes.

This analysis used unsupervised learning using hierarchical clustering, five clusters and SciPy.

Characteristics of included patients:

	No A1c	Only 1 A1c	More than 1 A1c
Patient characteristics			
# of patients	21812	6527	81660
Female	10688	3329	39197
Age	57(±20)	56(±19)	60(±14.6)
BMI	31 (±6.95)	30 (±6.77)	31 (±6.31)

After applying unsupervised machine learning, 5 main clusters were identified. These clusters were not pre-defined and instead were identified through hierarchical clustering (a type of unsupervised learning). These included: (1) Started with higher HbA1c and ended with lower HbA1c (bad start, good finish), (2) Started with lower HbA1c and ended with lower HbA1c (good start, good finish), (3) Started with higher HbA1c and ended with higher HbA1c (bad start, bad finish), (4) Started with lower HbA1c and ended with higher HbA1c (good start, bad finish), (5) remaining population.

 **(1) Bad start, Good Finish**	 **(2) Good Start, Good Finish**
Age 65; 46% female; BMI = 32 Female provider = 40% Metformin use = 83%	Age 55; 36% female; BMI = 32 Female Provider = 33% Metformin use = 91%
 **(3) Bad Start, Bad Finish**	 **(4) Good Start, Bad Finish**
Age 55; 24% female; BMI = 32 Female provider = 34% Metformin use = 90%	Age 64; 41% female; BMI = 32 Female provider = 35% Metformin use = 90%

Cluster 5 contained the remaining patients, summarized into: Age 63, 46% female, BMI 32, Female provider 40%, Metformin 79%

The results of this analysis provide a proof of concept that unsupervised machine learning can help identify clusters of patients based on their hemoglobin A1C trajectory. Thereafter, a deep-dive can be performed to identify unique patient or provider-level characteristics within each cluster. Identifying these characteristics can help identify what trajectory future patients will fall into and allow tailoring of their management to optimize the treatment of their diabetes

Challenges / Lessons Learned:

During the start of the event, users commented that it was difficult to log in and access the restricted environment. OpenVPN proved to burdensome at first but became more straightforward once users gained familiarity. Some issues were due to operating systems in individual machines, such as Windows, Mac OS, Linux and Ubuntu.

We are working with CAC/Indoc to develop an intuitive step by step instruction of OpenVPN installation and secure server login in Linux, Windows and Mac OS.

Memory issues appeared when analysts tried to join data. This was due to researchers using pure python code; running the analysis in SQL would vastly improve the run time. As an example, a researcher tried to remove 5 digits off a medication variable. The code took over 6 hours to run (overnight). The command was cancelled, and re-run with a limited cohort of 10k patients: it then took 28 minutes. If done in SQL, the task would have been completed in 4 minutes.

Before the workshop, we realized that the complete repository database format could be overwhelming for naive R and Python programs. In the workshop, we provided the data in two formats: in a database format and using the HL7 Fast Healthcare Interoperability Resources (FHIR) format. An advantage of FHIR is that each patient's data resides in a single file as a single block. It allows a program to process patient data in sequence; processing can be stopped as needed. The disadvantage is the data format is less generally known and requires a bit more programming. The database format, on the contrary, requires that data for all patients be processed concurrently. This could be appropriate for small numbers of patients but requires careful programming and tools such as SQL server in order to process data for larger sets of patients (several thousands and more). During the workshop, we did notice that the database format was problematic for Fields participants. As the complete repository data is rather large, it requires advanced programming skills to transform the database format data into a format that is suitable for analysis. Vector participants seemed more comfortable with the FHIR format and were able to more readily convert the data into an analysis-ready format.

It was clear that participants would have benefited from a training dataset; this would be a flat file with data on about 1000 patients with all values available so the data could be interpreted and understood.

To address these issues, we plan to 1) develop an analysis-ready format for data, 2) create a sample database consisting of data on 1000 patients, and 3) install SQL server in the secure environment. In the analysis-ready format, each row is one patient, and the columns are diagnoses, medication usage, lab results, # of visits and others in a certain time frame. This analysis-ready format allows quick analysis that helps the user to gain a good understanding of the data. A sample database of 1000 patients will allow users to efficiently explore the database, quickly test ideas and gain confidence; it will decrease the more tedious aspects of early stage data processing. The availability of SQL Server will provide a very useful and scalable data processing tool. There is a chance the analysis-ready format may bias the types of analyses that could be done by imposing our methods of data processing for users; this may restrain novel uses and analyses of data. We consider the risk as being minimum as the users also have access to the sample database; the examination of the sample database helps to understand the derivation of the analysis-ready format from the database and inspire creative idea formulation.

General discussion during the day included what "Time 0" was for these patients. Was is the date of inception for diabetes or the first visit date? Additionally, laboratory test dates were rounded to the nearest visit data; there was discussion on whether this should be done or whether the actual date of the test should remain. The repository does not provide a definite answer to these questions. We can, however, improve the value of these variables for analysis by providing more information through improvements in the data dictionary; we can improve efficiency by using the analysis-ready format. Rounding lab tests to the nearest visit make the time relationship among observation, diagnosis and treatment explicit, which could help discover patterns among them. At the same time, such rounding introduces strong assumptions and may lead to undesired or artificial results.

Investigators asked whether it was possible to plot postal codes on google maps. One researcher wanted to examine the longitude and latitude effects on diabetes management. A downloadable PCCF tool from Statistics Canada allows mapping patient FSA/postal code to longitude and latitude coordinates, however, Google maps requires internet access to generate the map and connect to Google map APIs and plot data. This limitation was established by DAC to protect the data and ensure the environment is "locked down". We may be able to apply some leniency in a project by project basis with the express approval of the Research Governing Committee. This still needs to be discussed and reviewed from a privacy point of view. We could also examine the feasibility of creating a shared drive between Windows and Linux within the environment to allow multiple groups to collaborate.

Investigators had some difficulties with variability in data entry habits for health data in EMRs. For example, modified ICD9-ON coding is used in Ontario for health conditions, while ICD9 is used in other provinces (250.01 vs 2501). Additionally, primary care physicians generally use ICD9 codes while hospitals use ICD10 codes. It was also requested that categorize medications; it would be helpful to have this done prior to analysis and to have clear definitions available: for example, 1 would represent SGLT2s, 2 would equal GLP1s, 3 for Insulin, etc.

We plan to add a glossary section in the data dictionary to explain the many abbreviations and discuss the caveats. We also intend to add some initial drug classes (metformin, sulfonylurea, DPP4, GLP-1, SGLT-2, insulin, and other glucose lowering medication) in the data dictionary and the analysis-ready format.

Conclusion

The workshop provided an opportunity to link mathematicians from Fields and AI researchers from Vector to diabetes researchers at the Banting Institute in order to collaborate in a meaningful way to provide new insights into the prevention, detection and treatment of diabetes.

There is still some work that needs to be done to make the data machine-learning ready. The data would benefit from being augmented with outcomes data. Since outputs are limited by available data, we will now focus on improving these data. We will refine our data cleaning methods and improve the completeness and integrity of our data.

We can conclude that the National Diabetes Repository in combination with the HPC environment at the CAC can provide a secure analytics platform that supports for AI projects. There was synergy and value in the collaboration between clinicians, mathematicians, health and AI researchers and data scientists. In just under 12 hours, we were able to demonstrate results from advanced analytics.

In conclusion, we demonstrated that we can provide a rich data resource to Canadian researchers, allowing them to use advanced analytics for projects that bring value to patients living with diabetes.