STA303H1S/STA1002HS Final Project
Due on 30th August, 2020 11:59 PM EST in Quercus
All relevant work must be shown for credit.

Final Project: The final project is due on **30th August, 2020 11:59 PM EST** and consists of a data analysis on a novel dataset. The deadline will be strictly applied. Under no circumstances can students submit late. Please make sure that you start the submission process early so that your project is graded.

Students will be required to demonstrate their understanding of the methods based on course materials by developing a reasonable regression model using the techniques taught in class. The students will be responsible for choosing the correct methods to apply and providing appropriate justifications defending their choices.

The final project will be done individually, and must be typed and submitted by the stated deadline. The project needs to fulfill the following criteria:

- Font: 12-point font in a style similar to Times New Roman
- Spacing: single-spaced
- The word limit for the final project is **1500**. This excludes the title page, table/figure captions and appendix.
- Maximum 5 tables/figures will be allowed in the project report. The tables and figures should be relevant, and should convey the purpose of the project. All tables and figures should have captions. You may use any combination of tables and figures.
- Up to 3 additional tables/figures but they should only be included if they are relevant to the analysis and are referred to in the main text.
- You must submit the report in a standard file format (e.g., .doc, .docx or a pdf).
- Please submit your `R` codes file. This can be a .r or a .rmd file. No other file format for the codes will be accepted.

In order to pass the course, you must submit the final project.

ACADEMIC INTEGRITY: The University treats cases of plagiarism and cheating very seriously. It is the students' responsibility for knowing the content of the University of Toronto's Code of Behaviour on Academic Matters. All suspected cases of academic dishonesty will be investigated following procedures outlined in the above document. If you have questions or concerns about what constitutes appropriate academic behaviour or appropriate research and citation methods, you are expected to seek out additional information on academic integrity from your instructor or from other institutional resources (see `http://academicintegrity.utoronto.ca/`). Here are a few guidelines regarding academic integrity:

- You may consult class notes/lecture slides during the final project, however sharing or discussing questions or answers with other students is an academic offence.
- Students must complete all assessments individually. Working together is not allowed.

- Paying anyone else to complete your assessments for you is academic misconduct.
- Sharing your answers/work/code with others is academic misconduct.
- Looking up solutions to test/quiz problems online or in textbooks and copying what you find is an academic offence.
- All work that you submit must be your own! You must not copy mathematical derivations, computer output and input, or written answers from anyone or anywhere else. Unacknowledged copying or unauthorized collaboration will lead to severe disciplinary action, beginning with an automatic grade of zero for all involved and escalating from there. Please read the UofT Policy on Cheating and Plagiarism, and don't plagiarize.

Please don't upload this document or the required dataset on any social media platforms, chegg, slideshare or coursehero. Uploading this document to any such website will be treated as a serious academic offense and we will take actions based on University of Toronto's policies regarding plagiarism. We will constantly keep an eye in these websites to find out such incidences.

1 Background

Diabetes is a very prevalent disease that has been linked to a number of lifestyle factors by many clinical studies. Patients with diabetes tend to have worse medical outcomes than similar patients without diabetes and these patients can be quite costly to the healthcare system. However, not all patients with diabetes present the same burden on the healthcare system. It is of interest to identify those patients who are likely to have worse outcomes so that they can be targeted for interventions to improve these outcomes and to reduce costs.

The cost of healthcare cannot always be measured directly, but one of the important measures that are commonly collected in health services research that can act as surrogates for both cost and for poor health outcomes is readmission to hospital within a set number of days of discharge from hospital; readmissions are costly and can occur for several reasons, important ones being (a) inadequate care on the initial stay (perhaps after a discharge that was too soon) and (b) generally poor health of the patient, irrespective of the initial length of stay.

2 The dataset

For your analysis you have the `diabetes` dataset uploaded on Quercus. This dataset is very famous for statistical analysis. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria,

- It is an inpatient encounter (a hospital admission).
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter.

The dataset has 101766 observations from 71518 patients. That is a number of patients had more than one observation. The encounters were stored in the variable `encounter_id` and the patient ID was stored in the variable `patient_nbr`.

2.1 Response

The outcome is the readmission variable named `readmitted`. This variable has three categories:

1. No readmission;
2. A readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate);
3. A readmission in more than 30 days (this one is not so good as well the last one, however, the reason can be the state of the patient).

You can use the variable with its original categories or you can dichotomize it as "no readmission" and "readmission". The choice is yours.

2.2 Predictors/Covariates

There are many covariates in the dataset. The data contains such attributes as race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc. The following variables should not be considered as covariates, `'encounter_id'`, `'patient_nbr'`, `'admission_source_id'`, `'payer_code'`, `'readmitted'` and `'encounter_num'`. The first four variables are some identification variables. The `'readmitted'` is the response and the `'encounter_num'` variable indicates the number of encounter each patient has.

3 Task

Using patient characteristics available from hospital, identify groups of patients who are at different risk of readmission. To answer this question you can use any statistical technique that you learned from the course. However, you need to explain your choice. You should focus on the follwoing aspects:

1. There are many covariates in the dataset. For predicting the probability of a patient being readmitted please select maximum 9-10 covariates. You need to explain why and how you choose the 9-10 covariates for prediction.

2. Since this is a prediction problem you should make one test dataset which you will never use for modelling. Create a test dataset that contains a random selection of 20000 patients. You should not sample from the encounters. you have randomly choose the patients using `'patient_nbr'` variable. You will find the `%in%` code in `R` very useful for this purpose. You should use your student ID as the seed for the sampling.

3. You can fit a GLMM, GLM or GAM (or any other method). But since this is a longitudinal dataset you need to explain what assumptions you need to make to fit a GLM or any other model which assumes independence. If you use GLM then variable selection and prediction becomes very easy, which is not trivial for GLMM. GLMM is, however, the most appropriate analysis technique for this data, but due to the large structure of the data GLMMs may take a long time and may not converge. Thus, you need to properly explain how you choose the modelling technique and also if you fail to perform certain analyses then state that clearly in the limitation section.

4. Make sure to perform exploratory data analysis (basic summary statistics, plots etc.) before moving on to the final modelling.

5. You can do some literature review if that helps.