Heinz 95-845: Project Proposal

Yalu Tong YALUT/YALUT@ANDREW.CMU.EDU
Heinz College
Carnegie Mellon University
Pittsburgh, PA, United States

Youngjoo Son YSON/YSON@ANDREW.CMU.EDU
School of Architecture
Carnegie Mellon University
Pittsburgh, PA, United States

Sidharth Srikumar SSRIKUMA/SSRIKUMA@ANDREW.CMU.EDU
College of Engineering
Carnegie Mellon University
Pittsburgh, PA, United States

1. Project Details

Your project will involve the use of the machine learning pipeline. This is an opportunity for you to explore some interest you have in an applied domain and the machine learning suitable for the task.

The purpose of the project is to conduct an analysis that is novel in some way. The novelty could be in terms of development of machine learning, the assessment of a wide variety of machine learning algorithms at a focused task, or the application of a single machine learning algorithm that impacts a real societal problem.

A list of exemplary papers are available in the Possible Data Sets slides on Canvas. The examples may be helpful in identifying how you conduct your study and prepare your write-up. Two additional resources for finding a problem domain include: (1) Data is Plural (`https://goo.gl/UgKgLC`) and, (2) the url: `https://github.com/awesomedata/awesome-public-datasets`. We recommend you do not choose a fully pre-processed data set. We do recommend you choose a data set that will fit in memory (or that can be run on your laptop) so that your machine learning process will be manageable.

The proposal for the project is due on **November 7th**. Please use this TeX template in Section 2 and submit on Canvas **a link to a git repository with instructions on access (particularly if it is a private repository)**. You may find the online editors ShareLatex or Overleaf helpful in drafting your TeX file. However, in order to learn git version control (which will help you checkpoint during your project), we require the submission to be in a git repository. The git repository should include at minimum the .tex, .bib, and .pdf file for your proposal.

1.1 Objectives

The objective of this project proposal is to generate a proposal for your course project. It should be concise and describe the following components:

- Construction and description of an analytic framework that motivates the use of machine learning for your task
- Presentation of machine learning techniques appropriate for the task
- Description of the data
- Description of possible limitations of the study
- Description of the likely analysis outcomes and their impact.

1.2 Parameters

The project will be conducted in groups of 2-3.

The project you propose should be different from an existing analysis, including publicly available analyses and analyses from other class projects of yours. It is permissible to perform an analysis in data that warrants a secondary analysis. My guideline here is that the analysis must be greater than 50% new. To get approval for these studies, please describe the existing project and highlight the difference and contribution of this class's project. Provide any relevant documents (proposals, manuscripts, and/or citations). If the project has overlap with work from another course, you must also provide documented approval from the other faculty member/research collaborator(s).

Your team is free to use programming language(s) of your choosing, however, we may only be able to support your endeavors in R.

2. Proposal Details (10 points)

Please provide information for the following fields. Your proposal write-up should be no more than 2 pages.

2.1 What is your proposed analysis? What are the likely outcomes?

Our proposed analysis: Test if retinopathy can be detected by fasting glucose for people who are over 40. The likely outcomes are that the fasting glucose variable would consistently have a major impact on determining retinopathy in different classification models.

2.2 Why is your proposed analysis important?

Early detection of retinopathy is important in that it helps to examine early diabetes. By focusing on the certain age range, we would be able to determine the applicability of fasting plasma glucose in discovering retinopathy depends on the age.

2.3 How will your analysis contribute to existing work?

There are existing studies tried to detect retinopathy by values of fasting plasma glucose (Patel et al. (2017); , Takao et al. (2018)) and fasting plasma glucose showed association with retinopathy. We are to focus on the age of older than 40 years old, so that we could find out the different degree of the association compared to precedent studies. This contributes

to the efficiency of fasting plasma glucose in detecting retinopathy for people who are older than 40.

2.4 Describe the data. If applicable, please also define Y outcome(s), U treatment, V covariates, and W population.

National Health and Nutrition Examination Survey (NHANES) dataset, which contains examination, demographics, dietary, laboratory, questionnaire, and limited access dataset were used. First, we found our outcome variable from the examination dataset, which indicated the presence of retinopathy. Second, fasting plasma glucose variable (U) was found in laboratory dataset. For covariates (V), since eye health can influence our outcome, we plan to include lots of variables from Vision and Ophthalmology dataset. Also, some variables in demographic and dietary dataset are considered as listed below. The targeted population was people older than 40 years old.

Y: presence or absence of retinopathy (OPDDARMA,OPDSARMA), U: Fasting glucose (mg/dL, LBXGLU), V: gender (DMDHRGND), age (DMDHRAGE), weight (WTLAI8YR), Dietary Supplement Use 30 Day (Energy, kcal, DSQIKCAL; total sugars, gm, DSQISUGR, protein), some variables in experimental data of Vision and Ophthalmology (like OPXFDT), W: Age, older than 40.

2.5 What evaluation measures are appropriate for the analysis? Which measures will you use?

The evaluation measures for the models that we use will be ROC curve. We will use AUC to see the accuracy and also balance the recall and specificity to make the false negative rate smaller since it has higher cost.

2.6 What study design, pre-processing, and machine learning methods do you intend to use? Justify that the analysis is of appropriate size for a course project.

Study design: We will analyze participants who are over 40 years old from 2005-2008 NHANES survery. Retinopathy result was obtained from retinal imaging in ophthalmology assessment. We assumed participants who had retinopathy for either eye as having retinopathy.

For the pre-processing we plan to do some data cleaning and data imputation work for the missing data, we intend to use multiple imputation or missing data elimination depends on the type of the missing. Moreover, since the retinopathy may happen in one eye or both eyes so that it has two variables to present it in the data set, we may combine them together to create a new indicator.

We intend to use multiple classification algorithms such as (logistic regression, random forest, etc) to predict the retinopathy and compare the model performances.

The data set we use is in the year of 2005-2008 and we narrow down the population to people who are over 40 in the survey, the size of our analysis is not too big for a course project.

2.7 What are possible limitations of the study?

Since we are using the NHANES data set, limitations of the study largely come down to limitations in the data set. Firstly, data in the dietary data set are estimates based on the patients food and beverage consumption in the 30 days period prior to the interview. Thus, accuracy of the data about total consumption is dependant on the reliability of the self-reportage by participants in the survey. Since we are using second hand data, assessing the reliability of this data becomes difficult.

Another point of limitation is the amount of data available for us to make the models. While we hope to combine data collected over multiple data collection cycles (NHANES surveys are typically carried out every 2 years), the surveys that contained the Ophthalmology test are few. Given that our study focuses only on people 40 years or older, the amount of available data can mean that the uncertainty in estimates for our model parameters (which will be reported along with our results) is potentially significant, and further data would have to be collected in future iterations of the study in order to reduce the uncertainty.

Finally, since the data has not been collected in a specifically designed setting, there is uncertainty about the orthogonality of the data. This might mean that there are some confounded effects, or mistaken attribution of causation.

2.8 Who will use your analytic pipeline? In one or two sentences, describe an example of its use.

The doctors and researchers can use the pipeline to detect the presence of retinopathy for a patient over 40 which might be useful to examine early diabetes.

References

Y. R. Patel, M. S. Kirkman, Hannon T. S. Considine, R. V., and K. J. Mather. Retinopathy predicts progression of fasting plasma glucose: An early diabetes intervention program (edip) analysis. In *Journal of diabetes and its complications, 31(3), 605-610*, 2017.

T. Takao, K. Inoue, M. Suka, H. Yanagisawa, and Y. Iwamoto. Optimal cutoff values of fasting plasma glucose (fpg) variability for detecting retinopathy and the threshold of fpg levels for predicting the risk of retinopathy in type 2 diabetes: A longitudinal study over 27 years. In *Diabetes research and clinical practice, 140, 228-235*, 2018.