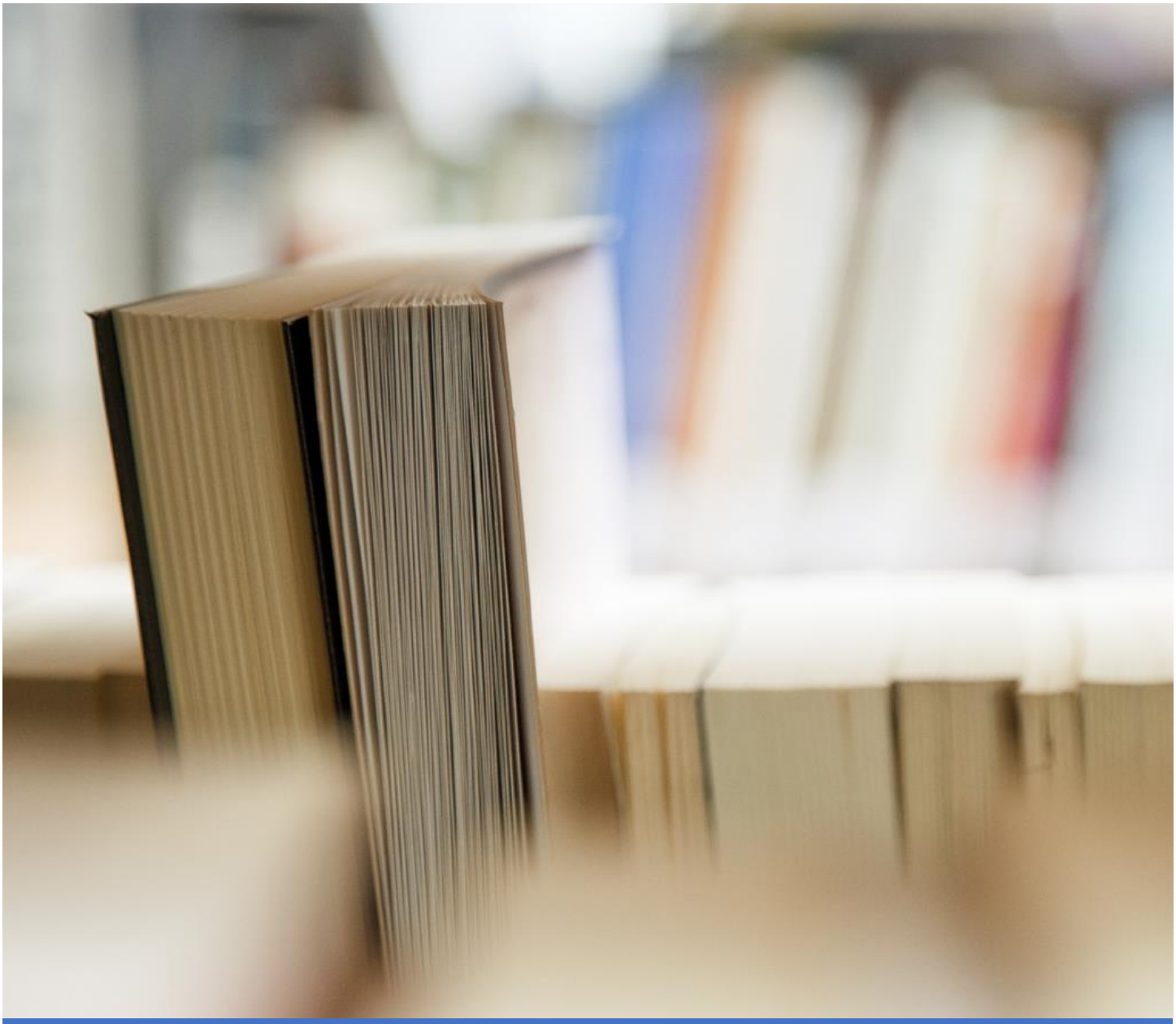

Capstone Proposal

Machine Learning Engineer Nanodegree

Ish Gupta

Ish.gupta@outlook.com

Contents

domain background

The National Longitudinal Mortality Study (NLMS) is a national, longitudinal, mortality study sponsored by the parts of the National Institutes of Health, and Center for Disease Control and Prevention and the U.S. Census Bureau for the purpose of studying the effects of differentials in demographic and socio-economic characteristics on mortality.

The main NLMS consists of a database developed for the purpose of studying the effects of demographic and socio-economic characteristics on differentials in U.S. mortality rates. It consists of U.S. Census Bureau data from Current Population Surveys (CPS) and a subset of the 1980 Census combined with death certificate information to identify mortality status and cause of death. The study currently has approximately 3.8 million records with over 550,000 identified mortality cases. The content of the socio-economic variables available offers researchers the potential to answer questions on mortality differentials for a variety of key socio-economic and demographic subgroups not covered as extensively in other databases.

Mortality information is obtained from death certificates available for deceased persons through the National Center for Health Statistics. Standard demographic and socio-economic variables such as education, income, and employment, as well as information collected from death certificates including cause of death are available for analyses.

problem statement

The survey consists of the reasons of death for the people who lost their life, and were included in the Survey. The reasons are related to Health, accidents, and intended loss of life, i.e., Suicide.

The data can be analysed and studied to see if there can be any measures that the government can take to reduce Suicidal attempts, by differentiating the deaths caused by intended action, vs accidents.

The datasets and inputs

The 11-year follow up consists of a subset of the 39 NLMS cohorts included in the full NLMS that can be followed prospectively for 11 years. The content of each record on the file includes demographic and socioeconomic variables combined with a mortality outcome, if there is one. To prevent disclosure, all of the records have been concatenated into a single file and the temporal dimension has been altered. In lieu of identifying the CPS year and starting point of mortality follow-up for each file, all of the records in have been assigned an imaginary starting point conceptually identified as April 1, 1990. These records are then tracked forward for 11 years to observe whether person in the file has died. This approach results in a maximum of 4018 days of follow up for this cohort.

For those who have died, the underlying cause of death and follow-up time until death have been provided. For those not deceased by the end of 4018 days follow-up period, the follow-up time provided is the full observation length, 4018 days or 11 years. In the construction of data, it was assumed that these surveys, collected from throughout the 1980s and 1990s, would adequately reflect the U.S. non-institutionalized population on April 1, 1990. Under this assumption, the separate CPS samples have been combined and can be viewed as one large sample taken on that date.
The data attributes are explained here.

solution statement

- Segregate accidental and suicidal cases using Classfication algorithms.
- analyze them to look for any possible reasons which could potentially be the reasons for suicide.

benchmark model

- assess usage and performance of xgboost for the Classification problem using accuracy against validation dataset, and confusion matrix
- compare the performance of xgboost with neural networks (using pytorch) to see how well each one performs in predicting the potential Suicides

evaluation metrics

- prediction accuracy against test/validation dataset
- prediction accuracy in terms of confusion matrix

project design

1. Beginning with extracting the cases of accidental and suicidal deaths, data will be cleaned for removing unwanted attributes, fixing missing values.
2. Using stratified sampling to sample the data in train and testing set to have similar proportions of the cause of deaths
3. Build an xgboost model, and assess the performance
4. Perform hyper-parameter tuning, and see if it improves the performance
5. Use Neural networks to classify similarly, and assess the performance