Data Wrangling and Management

Kelly Im, Brian Wang, Alvin Zou

Datasets

- For the purpose of this project, we will be using the CHIS datasets across multiple years.
- Data is available for the years 2001, 2003, 2005, 2007, 2009, 2011 - 2020.
- However, for the years 2001 and 2011, the data is missing a key variable denoting the type of diabetes survey participants have (if they have diabetes). Which is why we have decided to exclude the two years from our study.

- Additionally, we are considering using the NHIS dataset, but that will be for an additional modeling attempt after we have completed the model with the CHIS dataset.

Inconsistent Variables

- Problem: Across the different years, there are hundreds of variables that are present in some years and not in others.
- Solution: Include all columns in a combined dataset and fill in missing values with NA and combine those that are the same, but named differently.

- We have also considered using interpolation to fill in these missing values, but are unsure if this would be our best option considering how much missing data there is.
- Having a lot of missing data may lead to reduced statistical power and biases.
- However, using interpolation may lead to inaccurate results.

- Question: What would you consider to be the best approach to the issue of missing data?

Class Imbalance

- 1% Type 1 Diabetes
- 9% Type 2 Diabetes
- 90% Inapplicable/No diabetes
- Undersampling and oversampling may have issues
  - Potential overfitting when oversampling
  - Loss of useful data when undersampling
  - Increased learning and processing time when oversampling
- Consider using Cost-Sensitive Learning