# Evaluation of Different Imputation and Machine Learning Techniques for Prediction of Imbalance Depression Data

CAROLINE SKLAVER

GWU DATA SCIENCE CAPSTONE

07/09/2020

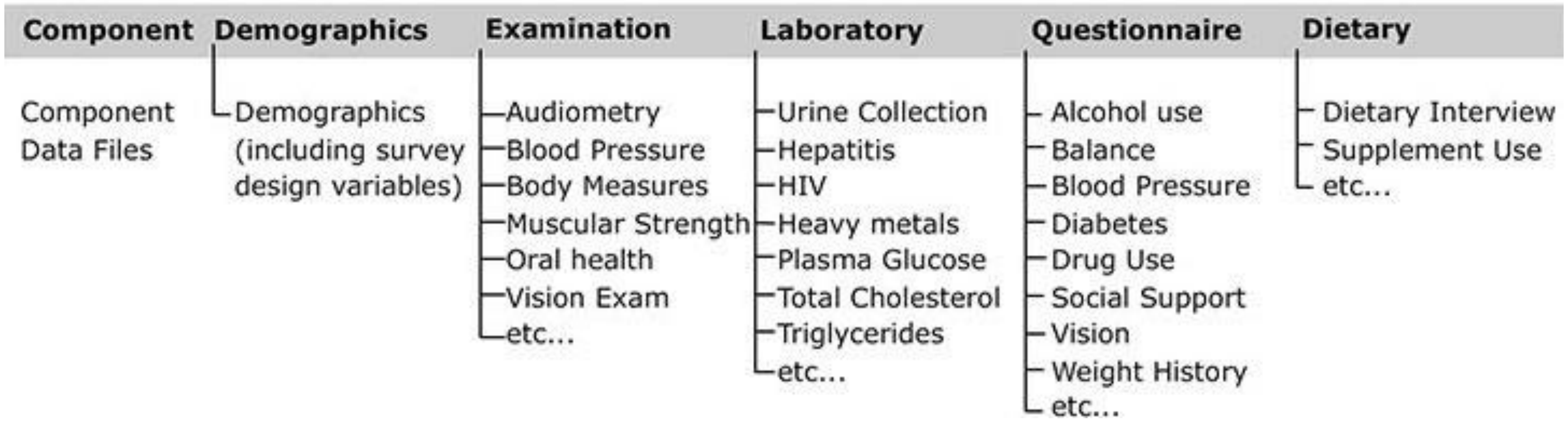
# Outline

- Background & Data
- Problem Statement
- Literature Review
- Data Pre-processing
  - Imputing missing values
- Baseline models
- Imbalanced data models
  - Resampling
  - ANN
- Next steps

# Background

- 1 in every 6 adults in the US will experience depression during their lifetime, which is about 16 million adults each year
- Depression is characterized as a long-term affect in mood that can take form in a variety of symptoms such as feeling sad, loss of interest in activities, changes in appetite and sleeping patterns, increased fatigue, feeling worthless, difficulty concentrating, or thoughts of suicide
- Depression frequently goes undiagnosed and untreated, contributing to the overall burden of disease

# National Health & Nutrition Examination Survey (NHANES)




- Sample of the civilian, non-institutionalized US population amounting to about 5,000 participants each two-year period
- Patient Health Questionnaire nine-item (PHQ-9) depression screening instrument assesses the frequency of depression symptoms over the past 2 weeks
  - Response categories include "not at all," "several days," "more than half the days," and "nearly every day" given a point ranging from 0 to 3. A total score of 10 or higher marked as ‘depressed’
- 31,357 complete depression observations between 2005-2016
- **7.28%** of the observations are classified as "depressed"

# Problem Statement

Depression is a widely undiagnosed mental condition that greatly impacts the burden of disease. In this project, I aim to use the NHANES data from 2005-1016 to build a model that can predict depression among adults in the US. This project requires data imputation techniques to handle missing values and classification models that can perform on this imbalanced dataset.

# Literature Review

- Linear **associations with depression**:
  - Dietary practices including caffeine consumption (Iranpour & Sabour, 2019)
  - Physical activity levels, metabolic syndrome, and body mass index (Liu, Ozodiegwu, Yu, Hess, & Bie, 2017)
  - Chronic conditions such as asthma, kidney disease (Patel, Patel, & Baptist, 2017), and osteoporosis (Cizza, Primma, & Csako, 2009)
- Jerez et al. 2020 used **imputation methods** based on statistical techniques (mean, hot-deck and multiple imputation) and machine learning techniques (multi-layer perceptron (MLP), self-organization maps (SOM) and k-nearest neighbor (KNN))
  - Machine learning imputation performed best for ANN classification
- **Shallow convolutional networks** to predict Coronary Heart Disease with NHANES data
- **Support vector machines and ensemble modeling** to predict risk of Diabetes with NHANES data
- **Repeated random sub-sampling** & Random Forests to predict the risk of eight chronic diseases (Khalilia M., 2011)

# Data Preprocessing

- Data was downloaded and converted to CSV files using python script found on GitHub
- Remaining preprocessing including importing files, extracting and renaming columns, and merging data for each year was done in Python Jupyter Notebook
- The target column of depression was calculated using the sum of the Patient Health Questionnaire (PHQ-9) nine-item survey ranging from 0 to 27, with a score of 10 or higher marked as 'depressed'. Only data with complete PHQ-9 questionnaires were used in this project
- Features of this data frame were chosen based off prior associations found in the literature (Appendix ,Table 1)
  - Examples include demographics such as age, income, education level, and race/ethnicity
  - physical features such as BMI and cholesterol
  - Health status indicators including chronic diseases, alcoholism, and physical activity levels
- 31,357 observations with 55 columns including respondent sequence number, and survey year

# Data Imputation

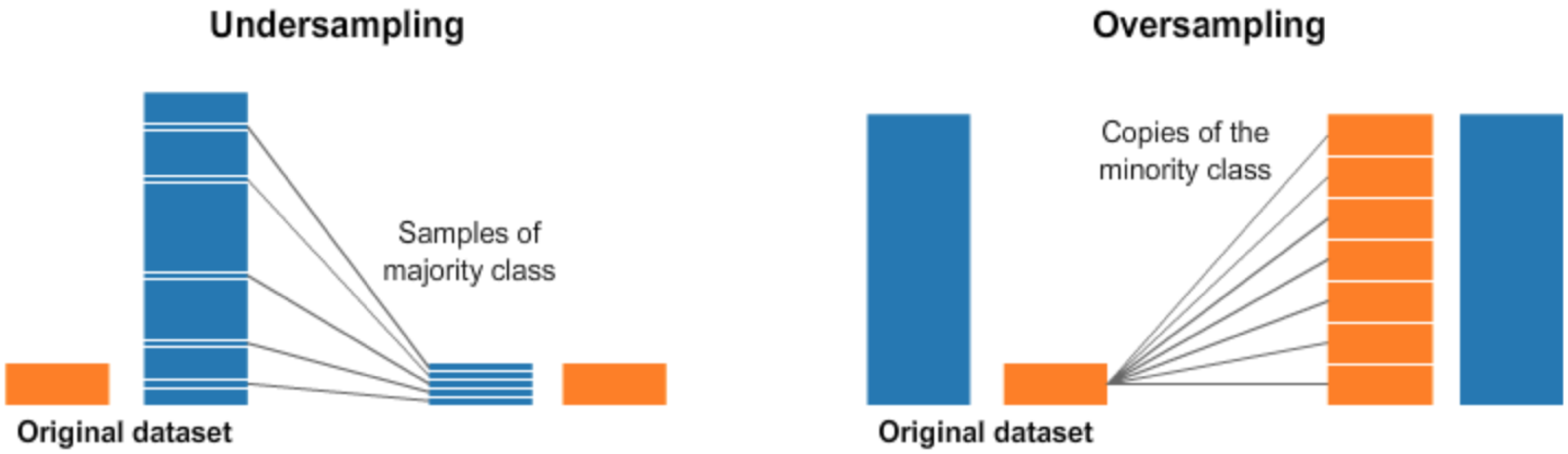
- Statistical methods:
  - Simple mean, median, and mode missing value imputation
  - Implemented after dropping columns with a proportion of missing values exceeding a given input threshold
- Machine Learning methods:
  - Aim to use available information within the dataset to estimate and substitute the missing values
  - First, separated the data frame into complete data and incomplete data
  - Complete data (excluding the ultimate target variable ‘depressed’) was used to predict values of the next-most-complete column. These predicted values were then imputed into that column and used with the complete data to predict the missing values of the subsequent column
  - Thus this method progressively predicts missing values with the imputed data
  - Used both Multi-layer perceptron (MLP) k-nearest neighbors (KNN)

# Baseline Model Results

| | Model F1-micro Score | | | | | |
|---|---|---|---|---|---|---|
| Imputation Strategy | MLP | XGBoost | Random Forest | Decision Tree | Logistic Regression | Naive Bayes |
| KNN | 0.9266 | 0.9272 | 0.9251 | 0.8723 | 0.7442 | 0.8048 |
| MLP | 0.9265 | **0.927367** | 0.9252 | 0.8716 | 0.7437 | 0.8029 |
| Mean - Drop >75% NA | 0.925373 | 0.927366 | 0.925214 | 0.871516 | 0.742914 | 0.80327 |
| Mean - Drop >50% NA | 0.9262 | 0.9268 | 0.9251 | 0.8734 | 0.7429 | 0.8022 |
| Median - Drop >75% NA | 0.9259 | 0.9271 | 0.9249 | 0.8712 | 0.7445 | 0.8038 |

| MLP Imputation XGBoost Confusion Matrix | | |
|---|---|---|
| | Positive | Negative |
| Positive | 31 | 439 |
| Negative | 29 | 7773 |

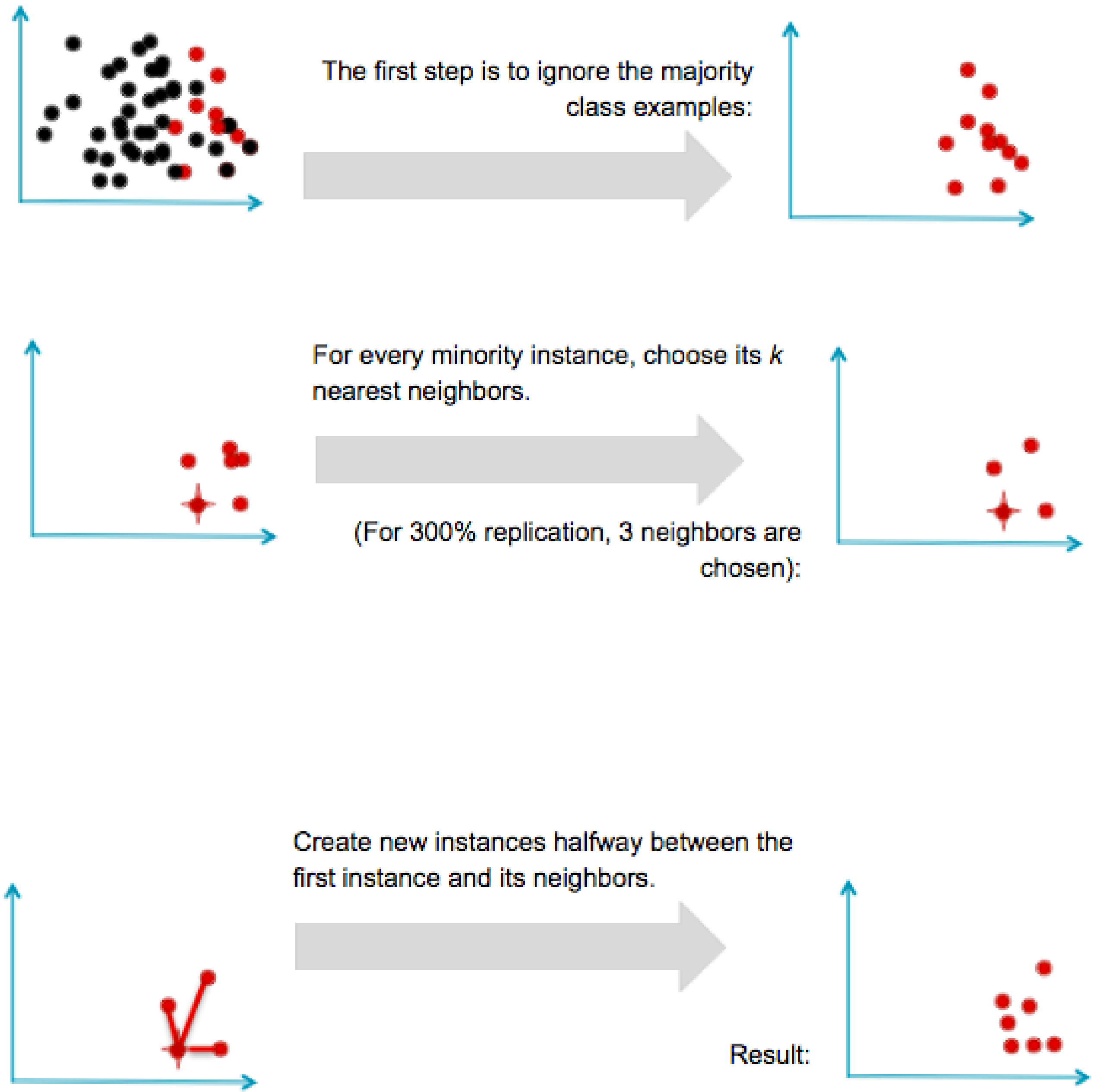
- The problem with imbalanced data, model performance is not representative of model predctions
  - High number of false negatives

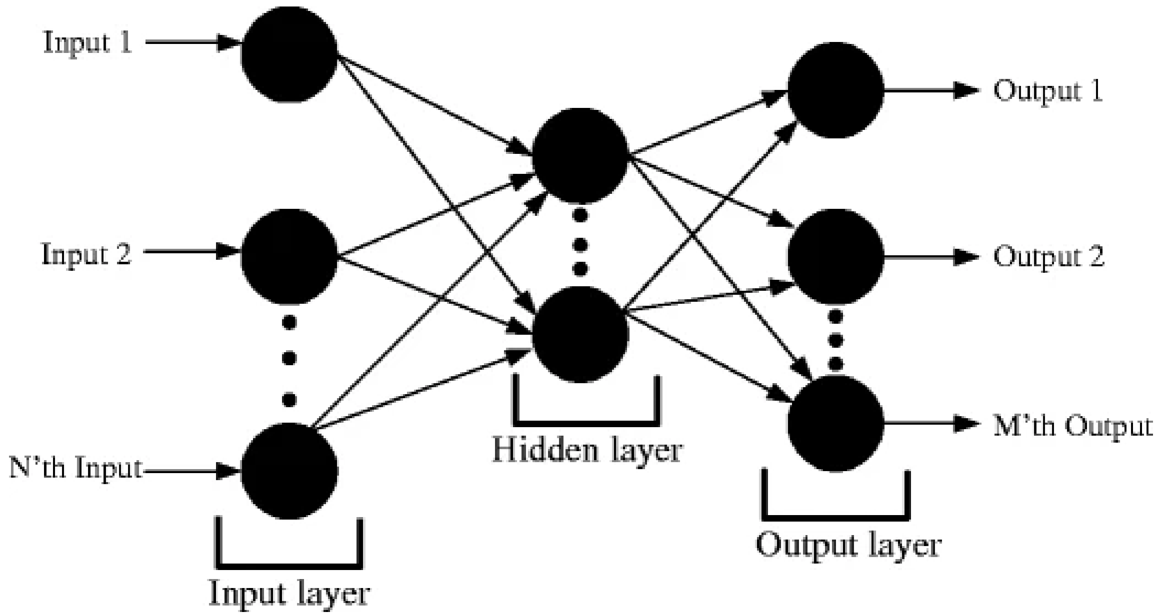



# Resampling Techniques

# Resampling Cont.

## Synthetic Minority Oversampling Technique (SMOTE)






# Keras ANN

- Sequential model of 2 fully connected dense layers
- Trained on resampled data, tested on original testing set

| Keras ANN, 50 epochs | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **MLP Imputed** | | | | | | | | | |
| *Resampling Technique* | Loss | Accuracy | Precision | Recall | AUC | True Positive | True Negative | False Positive | False Negative |
| Undersampling | 0.51497 | 0.76563 | 0.18227 | 0.66067 | 0.78504 | 294 | 4508 | 1319 | 151 |
| Oversampling | 0.51242 | 0.79050 | 0.15734 | 0.74242 | 0.79460 | 343 | 3973 | 1837 | 119 |
| SMOTE | 0.69220 | 0.88584 | 0.26755 | 0.29936 | 0.68599 | 141 | 5415 | 386 | 330 |
| | | | | | | | | | |
| **KNN imputed** | | | | | | | | | |
| *Resampling Technique* | Loss | Accuracy | Precision | Recall | AUC | True Positive | True Negative | False Positive | False Negative |
| Undersampling | 0.55518 | 0.72720 | 0.17623 | 0.72068 | 0.80520 | 338 | 4223 | 1580 | 131 |
| Oversampling | 0.56160 | 0.78603 | 0.18175 | 0.56793 | 0.73750 | 255 | 4675 | 1148 | 194 |
| SMOTE | 0.68333 | 0.87022 | 0.21265 | 0.24846 | 0.66293 | 121 | 5337 | 448 | 366 |
| | | | | | | | | | |
| **Mean/Mode imputed** | | | | | | | | | |
| *Resampling Technique* | Loss | Accuracy | Precision | Recall | AUC | True Positive | True Negative | False Positive | False Negative |
| Undersampling | 0.55616 | 0.70982 | 0.17282 | 0.75424 | 0.80886 | 356 | 4096 | 1704 | 116 |
| Oversampling | 0.52157 | 0.80660 | 0.20382 | 0.54584 | 0.74403 | 256 | 4803 | 1000 | 213 |
| SMOTE | 0.56145 | 0.87293 | 0.23706 | 0.29461 | 0.67449 | 142 | 5333 | 457 | 340 |

# Preliminary Results

# Next Steps

- Refine resampling techniques
- Build more sophisticated neural network (e.g. weights, layers, neurons, cross-validation)
- Try resampling with other ML algorithms
- More literature review of imbalanced data

# Appendix

*Table 1. Chosen Features*

*Gender*
*Age*
*Race/Ethnicity?*
*What is the highest grade or level of school you have completed or the highest degree you have received?*
*Marital status?*
*Annual household income?*
*Total number of people in the Household?*
*Ever have 4/5 or more drinks every day?*
*Doctor told you have diabetes or prediabetes?*
*In general, how healthy is your overall diet?*
*How many of those meals did you get from a fast-food or pizza place in the past week?*
*Have serious difficulty hearing?*
*Have serious difficulty seeing?*
*Ever used cocaine/heroin/methamphetamine?*
*Covered by health insurance?*
*Ever told you had weak/failing kidneys?*
*Ever been told you have asthma?*
*Doctor ever said you had arthritis?*
*Ever told had congestive heart failure?*
*Ever told you had coronary heart disease?*
*Ever told you had heart attack?*
*Ever told you had a stroke?*
*Ever told you had thyroid problem?*
*Ever told you had chronic bronchitis?*
*Ever told you had any liver condition?*
*Ever told you had COPD?*
*Ever told you had cancer or malignancy?*
*Broken or fractured a hip?*
*Ever told had osteoporosis/brittle bones?*
*Vigorous recreational activities?*
*Moderate recreational activities?*
*Minutes sedentary activity?*
*Past week number of days cardiovascular exercise?*
*Past week number of days strengthened muscles?*
*Number of sex partners/lifetime?*
*Describe sexual orientation.*
*Ever told by doctor have sleep disorder?*
*Smoked at least 100 cigarettes in life?*
*Caffeine Intake (mg/day)*
*Systolic: Blood pressure (mm/Hg)*
*Diastolic: Blood pressure (mm/Hg)*
*Body Mass Index (kg/m**2)*
*Waist Circumference (cm)*
*Total Bone Mineral Density (g/cm^2)*
*Direct HDL-Cholesterol (mg/dL)*
*LDL-cholesterol (mg/dL)*
*Total Cholesterol (mg/dL)*
*Triglyceride (mg/dL)*
*Glycohemoglobin (%)*
*Herpes Simplex Virus Type 2*
*HIV antibody test result*