Capstone Project Proposal

Domain Background

Overweight or obesity has become a common problem nowadays. It is estimated 40% of adults are overweighed[1]. It will be interesting to quantify these factors that have impact to overweight. Some research has shown that obesity is not a homogeneous condition, it maybe the case for overweight as well[2]. Hopefully the outcome of the analysis will help to bring up some effective solutions to prevent us from overweighting[3].

Another motivation of coming up with this proposal is due to personal reasons. Losing weight has been my life-long project. However it has not been very successful so far. It is common sense that the solution is to eat less and exercise more. I want to use a more data-driven approach to make my weight loss journey a bit more scientific and efficient.

Problem Statement

Whether one is overweight or not, is actually really straightforward to measure. Instead of predicting overweight, this project is trying to predict whether one will increase their weight or not within a year, given the lifestyle they currently have.

It is a **classification problem**. Target is defined as 1 if weight increase is larger than 5 pounds over a year. Inputs are personal demographic, dietary and laboratory data, which are detailed in the next session.

I also want to explore what are the key contributors to weight increase? To be more specific, what are the factors that will help with reduce risk of weight gain? What are the factors that lead to weight gain?

Dataset & Inputs

The dataset I'm going to use is National Health and Nutrition Examination Survey (NHANES) dataset from 2013- 2014. It is available in Kaggle[4] and originally produced by CDC[5].

My overall dataset will be aggregated from 5 datasets, including:

1. Demographics dataset: it includes demographic feature like age, education etc. Most importantly, it also has sample weight[6]
2. Examinations dataset, including Blood pressure, body measures etc.
3. Dietary data, including dietary interview about total nutrient intakes
4. Laboratory dataset, which includes information like Cholesterol – HDL, Insulin etc.
5. Questionnaire dataset, including weight history, alcohol usage etc.

Final dataset has 9813 observations with 1816 features; out of which 1657 are 1 and 8156 ar e 0. With sample weight applied, the weight gain rate is 20%. Hence it is an i**mbalanced data set**. I will split the dataset into training, validation and testing dataset by 60, 20, and 20 resp ectively, with stratifying the data using the target label.

The dataset was carried out using machine-learning techniques for predicting different target as compared to our objective here, for example predicting sleep disorder[7], type 2 diabetics [8]. It gives us confidence that this dataset could be used for this research.

---

[1] https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5986075/
[3] https://www.cdc.gov/nchs/products/databriefs/db340.htm
[4] https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey
[5] Centers for Disease Control and Prevention
[6] **Sample Weights:** The 2-year sample weights (**WTINT2YR,** WTMEC2YR) should be used for all NHANES 2013-2014 analyses. Detailed instructions for combining datasets from previous NHANES cycles are provided in the NHANES Analytic Guidelines[6]. Note that medical dataset is not used here hence another sample weight (WTMEC2YR) will not be used.

Solution Statement

Information from the overall dataset will be used as inputs to predict how likely an individual will gain weight in the next one-year. Target variable is aggregated using weight variable from Questionnaire dataset on difference between WHD020 (How much do you weigh without clothes or shoes?) and WHD050 (How much did you weight a year ago?), if the difference is larger than 5, then target will be 1, otherwise 0.
Our solution is mainly focusing on using supervised learning, including Logistic regression, Gradient Boosting Tree and its derivatives (XGboost, LGB, CatBoost) and Neural Nets.
Top 10 variables from each model will be used as predictive factors to assess their impact to weight gain.

Benchmark Model

Logistic regression model will be built as benchmark. Its strength lies on solving classification problems and very single features can be explained easily. It works well when features have linear relationship with target variable and are independent between each other.
The weakness of Logistic Regression is the fact it is not able to capture non-linearity and can perform poorly when there is high correlation between features. It also does not work too well with imbalanced dataset.
The advantage and disadvantages of Logistic Regression makes it a good candidate as benchmark model.

Evaluation Metrics

There are few metrics that will be used to evaluate mode performance. It includes:

- **AUC**: area under the curve. It is important and most popular metrics for checking classification model's performance. It tells us how well the model is able to discriminate between classes (overweight and not overweight), where 1 to be the best.
- **F1 score:** F1 score is the harmonic average of precision and recall, 1 to be the best and 0 to the worst. F1 score will be used here instead of predictive accuracy. The reason is predictive accuracy is not a useful metrics for imbalanced dataset[9].

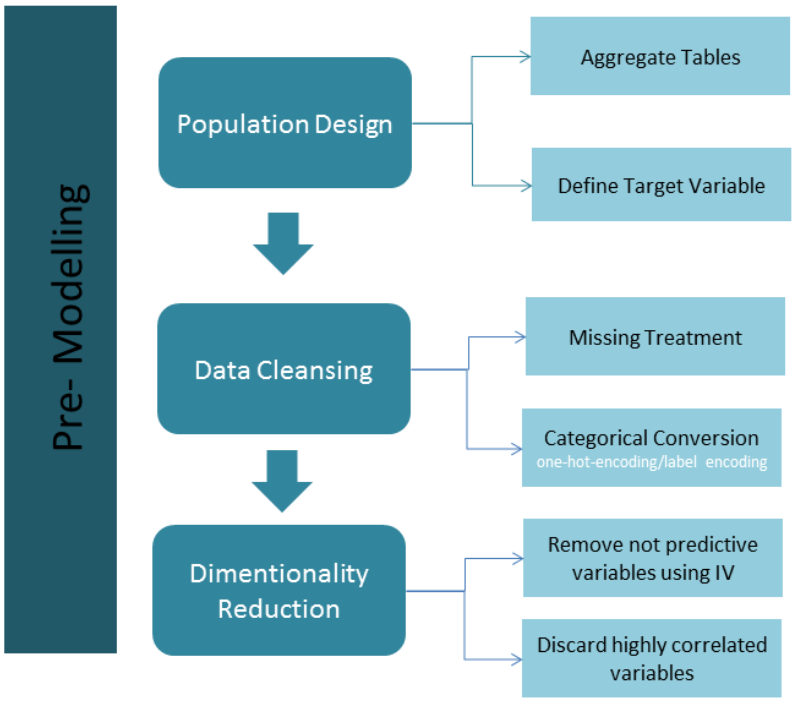
Project Design

Pre-modeling: This phase is critical for modelling. The critical steps are defining target variable (overweight or not), treatment for missing values; converting categorical variables to numerical as most scikit learn machine learning packages can't handle categorical variables. And lastly dimensionality reduction will be done in order to have useful variables as input for modelling.

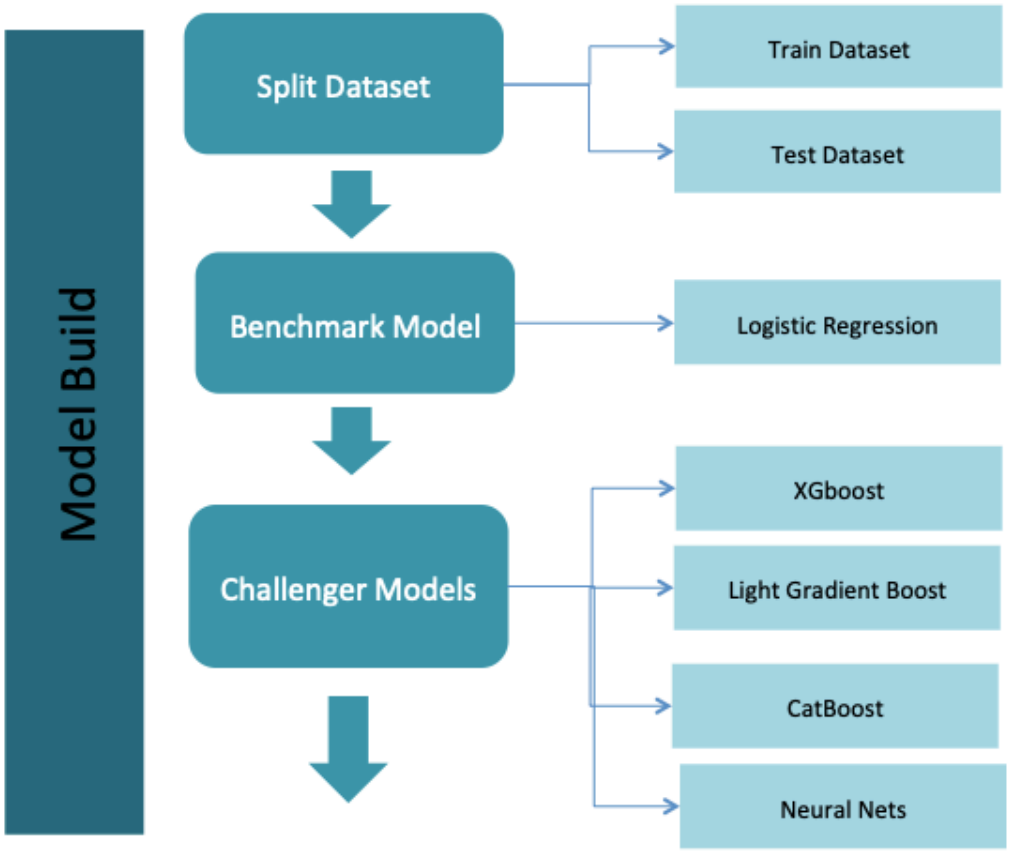
---

[7] http://pitt.edu/~jub69/material/Lec26_ML_exercise.html
[8] https://www.kaggle.com/what0919/diabetes-prediction

[9] https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b



Modelling: I will initially build a benchmark model using logistic regression. Then I will build a couple of challenger models.



Model Evaluation & Results: Models will be compared using AUG and predictive accuracy from confusion matrix. The results will be used to decide the key factors to overweight and

how we can use these results to give suggestions on how to prevent from being overweighed