Awareness to Diabetes

Che-Yu Liu

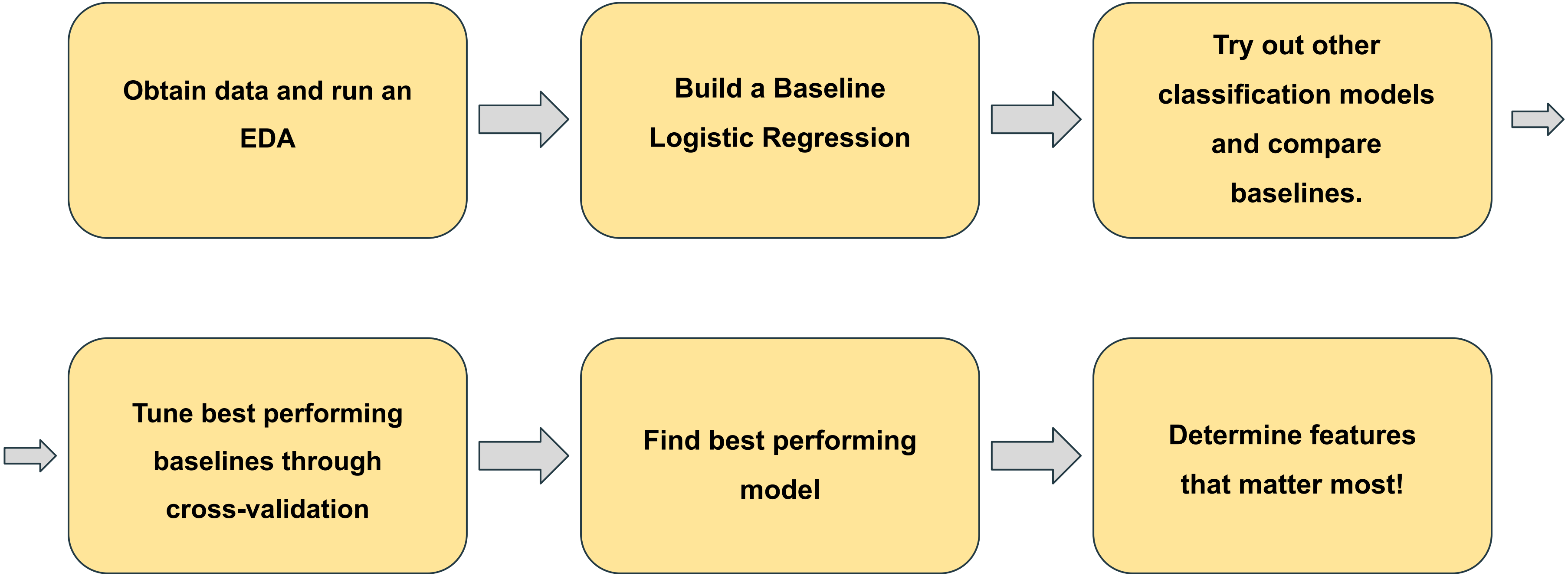
Introduction

- As of 2022, over 11% of the US population has been diagnosed with diabetes and around 90-95% of those people have type 2 diabetes.
- Apple would like to work with the CDC to create a campaign to bring awareness to diabetes indicators.

Data

- The data was obtained Kaggle dataset that uses a CDC survey responses from a Diabetes Health Indicators dataset.
  - Asked questions like:
    - Gen Health questions
    - Demographic questions
    - Socioeconomic questions
- ~70000 observations balanced data having 50% of observations being positive for diabetes.

Methodology



Obtain data and run an EDA

Build a Baseline Logistic Regression

Try out other classification models and compare baselines.

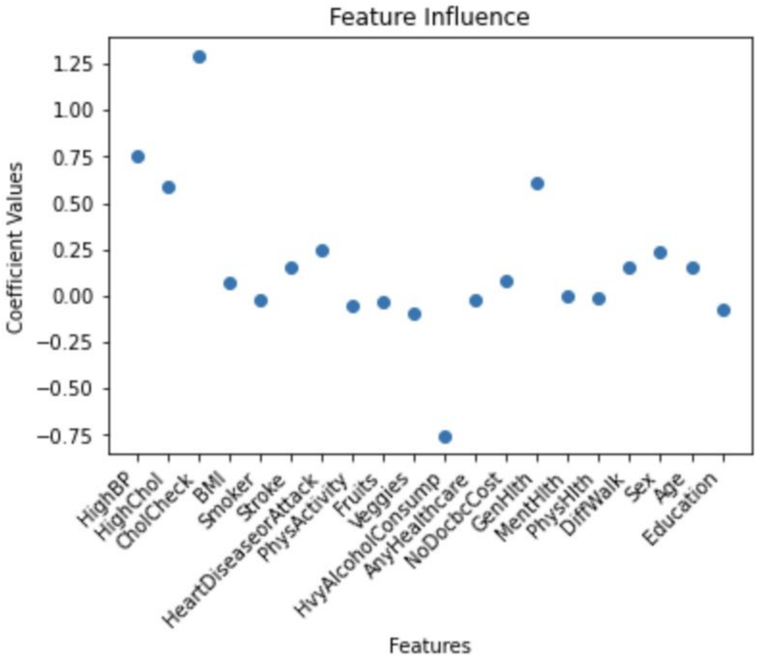
Tune best performing baselines through cross-validation

Find best performing model

Determine features that matter most!

Baseline Logistic Regression

- Logistic Regression
  - Recall Train: 0.769
  - Recall Validation: 0.765

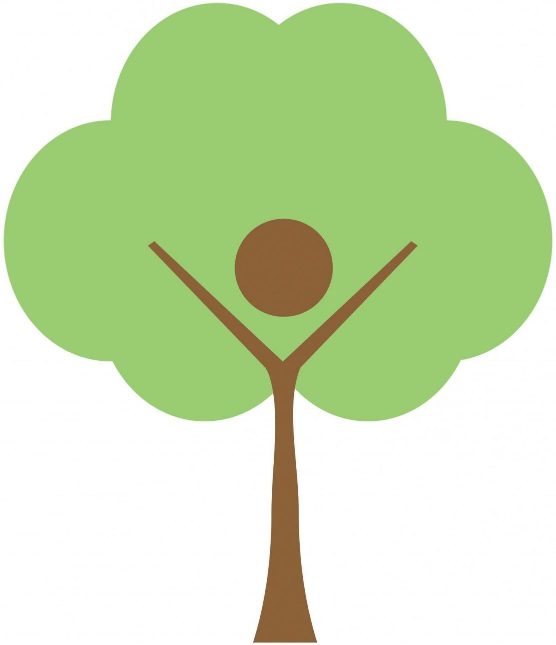


Other Baseline Models

- Random Forest Classifier
  - **Recall Train: 0.988**
  - **Recall Validation: 0.769**
- Gradient Boosting Classifier
  - Recall Train: 0.988
  - Recall Validation: 0.789
- XGBoost Classifier
  - **Recall Train: 0.831**
  - **Recall Validation: 0.789**
- Naive Bayes
  - Recall Train: 0.689
  - Recall Validation: 0.691

Decision Tree Tuning

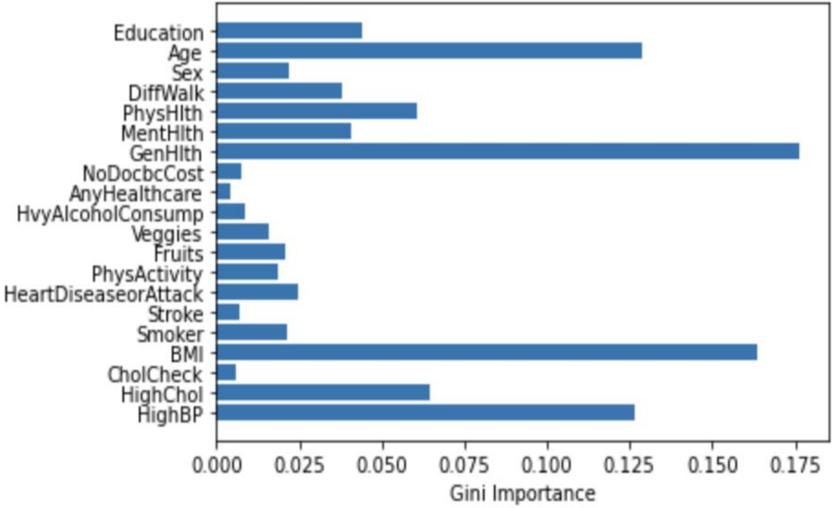
- Random Forest and Gradient Boosting performed similarly.
  - Random Forest after tuning:
    - **0.769 → 0.793**
  - Recall Test Score: **0.796**
- XGBoost showed promise
  - Recall Score after tuning:
    - **0.789 → 0.792**
  - Recall Test Score: **0.789**



Feature Importance

- From the Random Forest Classifier feature importance function most important features were:
  - Age
  - General Health
  - BMI
  - High Cholesterol
  - High Blood Pressure

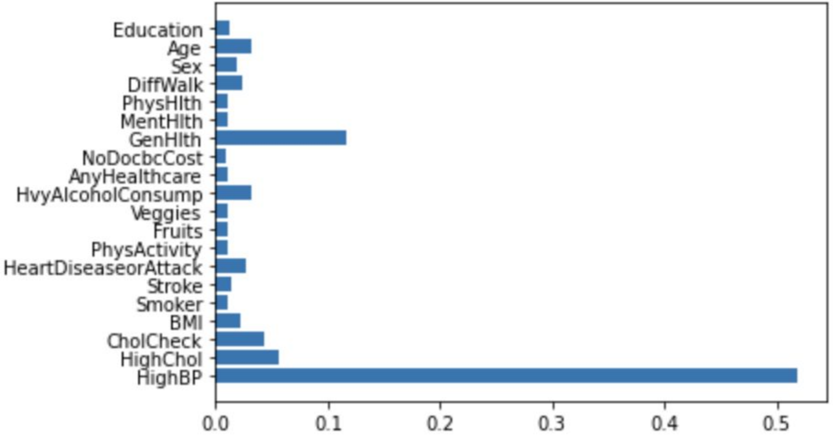
Influence on Diabetes



Feature Importance

- From the XGBoost feature importance function most important features were:
  - General Health
  - High Cholesterol
  - High Blood Pressure

Influence on Diabetes



Conclusion & Future Direction

- The best performing model at predicting diabetes is:
  - Random Forest Classifier
- Try out other ensembling methods
- Combine other diabetes indicators datasets.