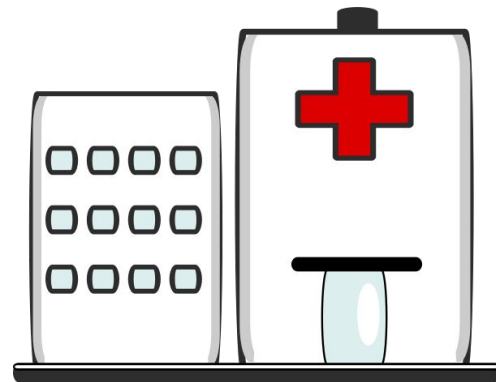# Diabetes Prediction

Lucas Wilkerson

Summary

Data concerning Diabetes and various health markers was obtained from a Kaggle dataset

Multiple model iterations performed

- **Goal → Build a model to predict Diabetes risk/ presence**

Insight generated will be used to give actionable recommendations for the stakeholder (healthcare practice)

Outline

- Business Problem/ Stakeholder
- Data/Methods
- Modeling
- Conclusions/Recommendations

Business Problem/ Stakeholder

A healthcare practice group wants to improve patient outcomes by:

- Early Diabetes Detection
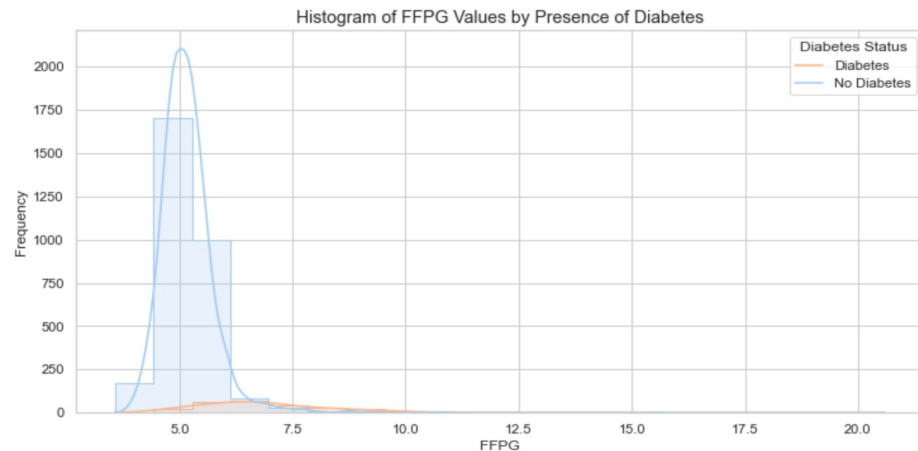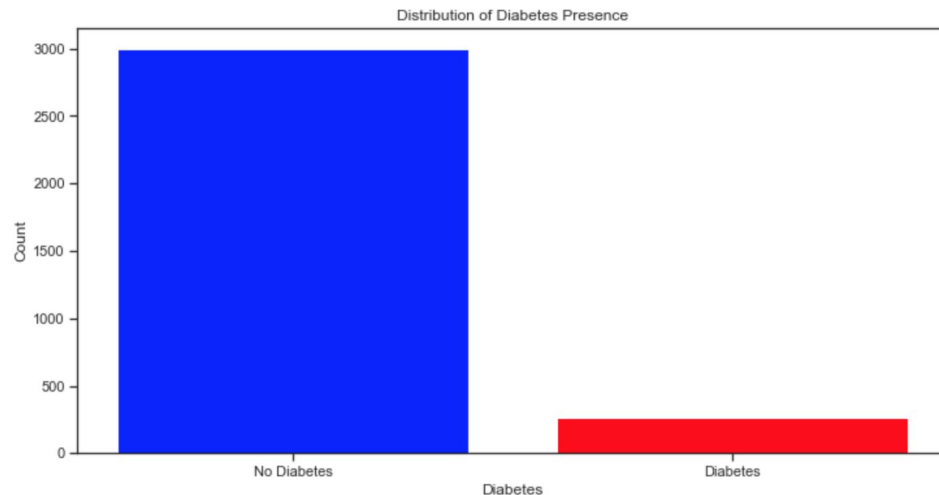- Decreasing risk of Diabetes
- Health Interventions

publicdomainvectors.org

Data/Methods

EDA:

- 3260 Rows
- 18 Columns

Highest Correlated Features:

- FFPG
- FPG
- Age

Distribution of Diabetes Presence

Histogram of FFPG Values by Presence of Diabetes

Modeling

Multiple models were tested:

Best performing model:

- **Tuned Logistic Regression Model**
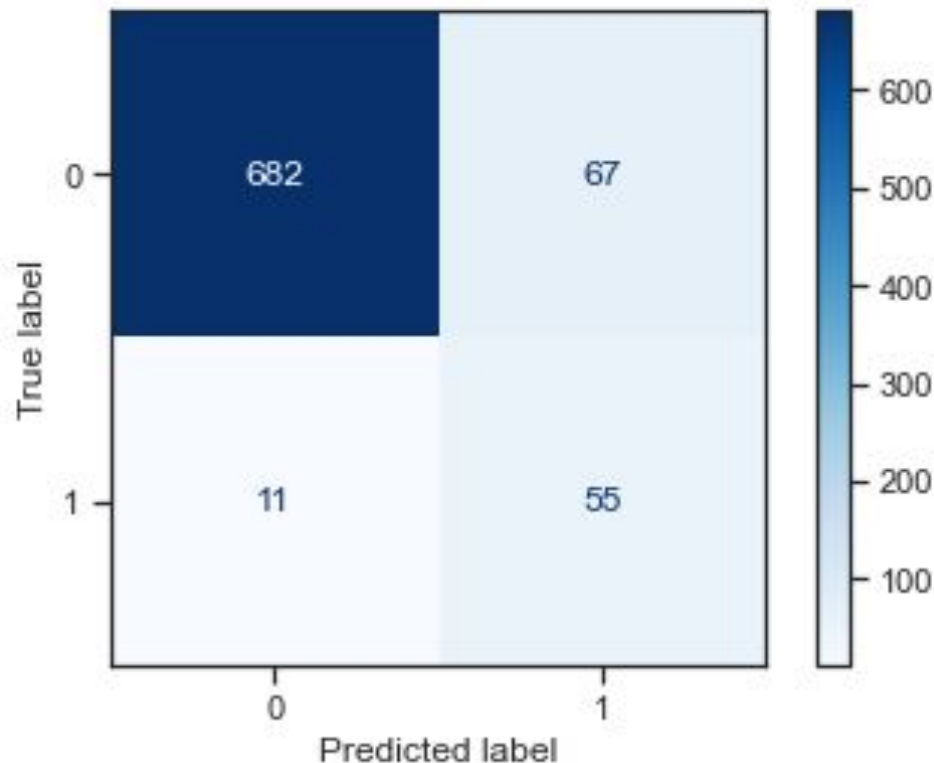
Metrics of Focus:

- **Recall Score**
- **F1 Score**

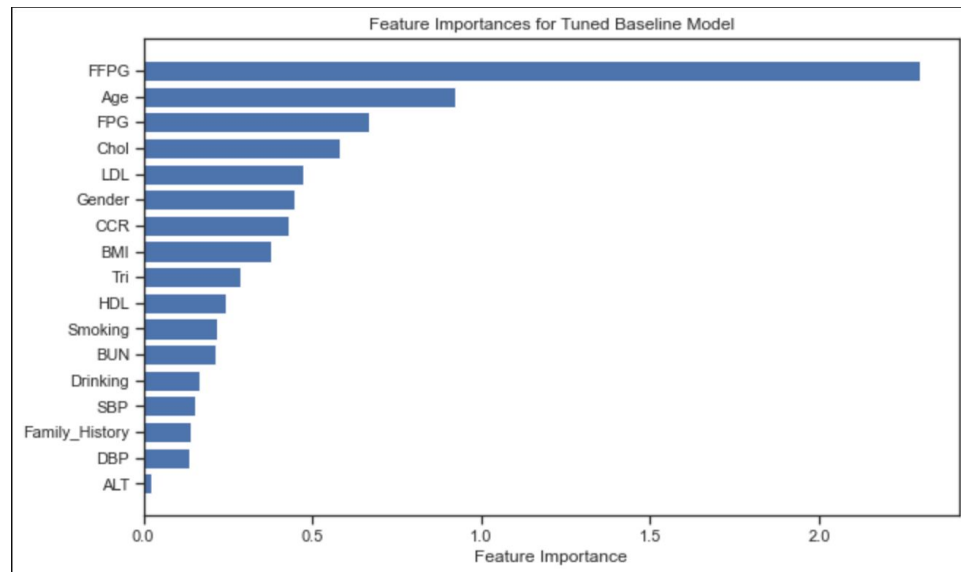# Model Performance: Tuned Baseline Model

Evaluation Metrics:

- Accuracy Score: 0.904

- Precision Score: 0.451

- **F1 Score: 0.585**

- **Recall Score: 0.833**

Feature Importance: Best Model

Most Important Features:

- **FFPG: Final Fasting Plasma Glucose**

- Age

- FPG: Fasting Plasma Glucose

Feature Importances for Tuned Baseline Model

Conclusions/Recommendations

Early Detection:

- Early detection/ screening tool

Timely Treatment:

- Early health interventions/ preventative care

Health Metrics:

- Monitoring of a patient's FFPG and FPG routinely while factoring in age

Limitations

- Higher chance of false positives with higher recall score
- Lower sample size
- Limited feature information from dataset source

Thank You!

Repository Link: phase3_project
Email: ldwilker10@gmail.com
GitHub: @ldwilker10
LinkedIn: https://www.linkedin.com/in/lucasdukewilkerson/