Diabetes Prediction

Lucas Wilkerson
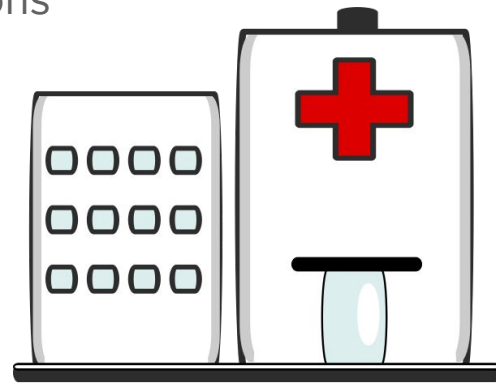
Summary

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

Multiple model iterations performed

- **Goal ➜ Find the best model to predict Diabetes risk/ presence**

Insight generated will be used to make actionable recommendations

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
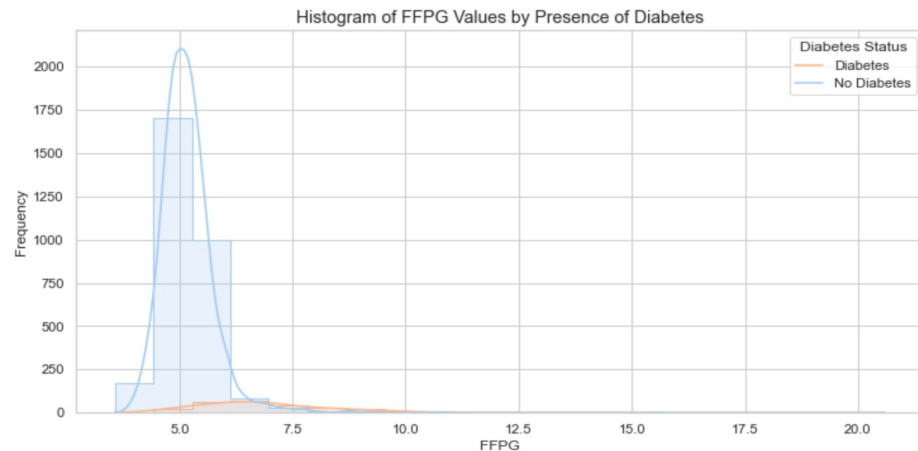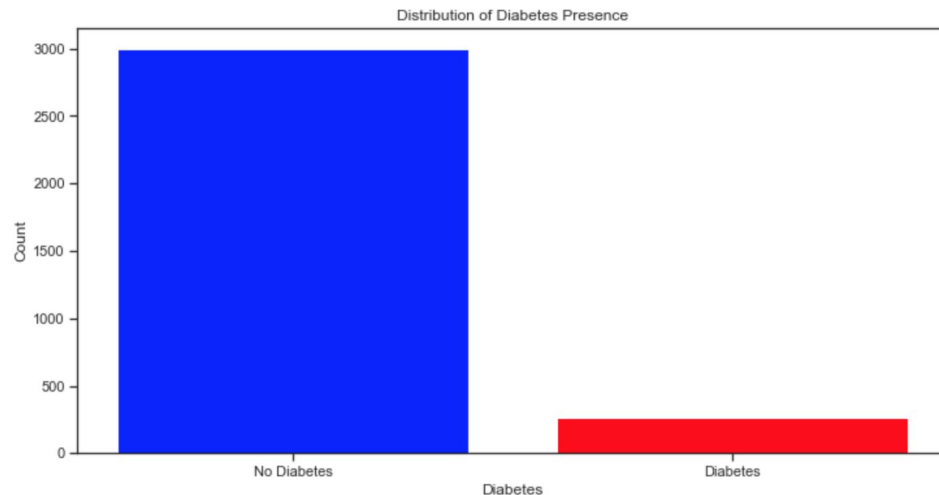- Health Interventions

Data/Methods

EDA:

- 3260 Rows
- 18 Columns
- Highest Correlation: FFPG, FPG, Age

Modeling:

- Multiple models
- Best Model:
    - Tuned Baseline Logistic Regression

Distribution of Diabetes Presence

Histogram of FFPG Values by Presence of Diabetes

Model Performance: Tuned Baseline Model

Evaluation Metrics:

- Accuracy Score: 0.9042944785276074

- Precision Score: 0.45081967213114754

- **F1 Score:** 0.5851063829787234

- **Recall Score:** 0.8333333333333334

Feature Importance: Best Model

Most Important Features:

- **FFPG: Final Fasting Plasma Glucose**

- Age

- FPG: Fasting Plasma Glucose

Feature Importances for Tuned Baseline Model
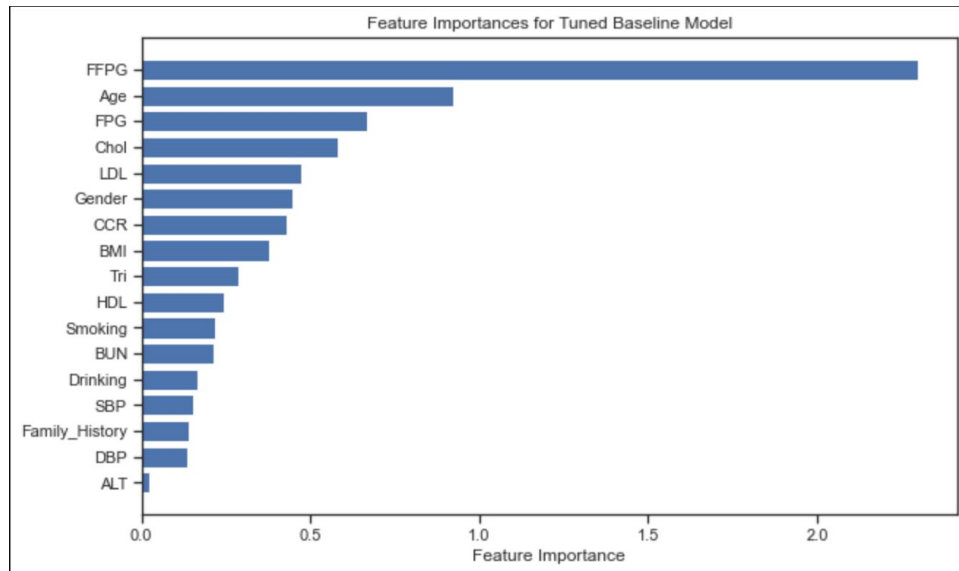
Conclusions/Recommendations

Early Detection: Practitioners should use this as an early detection/ screening tool

Timely Treatment: Practitioners can implement early health interventions

Health Metrics: Practitioners should monitor 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!

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