Justin Ku

Capstone Two Final Report

Diabetes Predictor Model

Problem Statement

Diabetes is a serious chronic disease that plagues the modern world. It's a silent killer that causes life-long health complications such as heart damage, kidney failure, amputations, and permanent vision loss. What are the parameters that can predict if someone has a high chance of having or developing Diabetes?

Using a health indicators dataset, a model was created to determine which indicators have a strong relationship with people who have Diabetes. These indicators can be used to help address the areas of concerns for public health or determine if an undiagnosed individual has a predisposition for this serious disease or is unknowingly Diabetic.

Early prevention and diagnosis lead to greater positive health outcomes and can alleviate pressure on our medical system. Thus creating a net positive ripple effect in individuals and the community. Many techniques such as Frequentist Statistics and Supervised Machine Learning were used in this Diabetes Predictor Model.

Data Wrangling

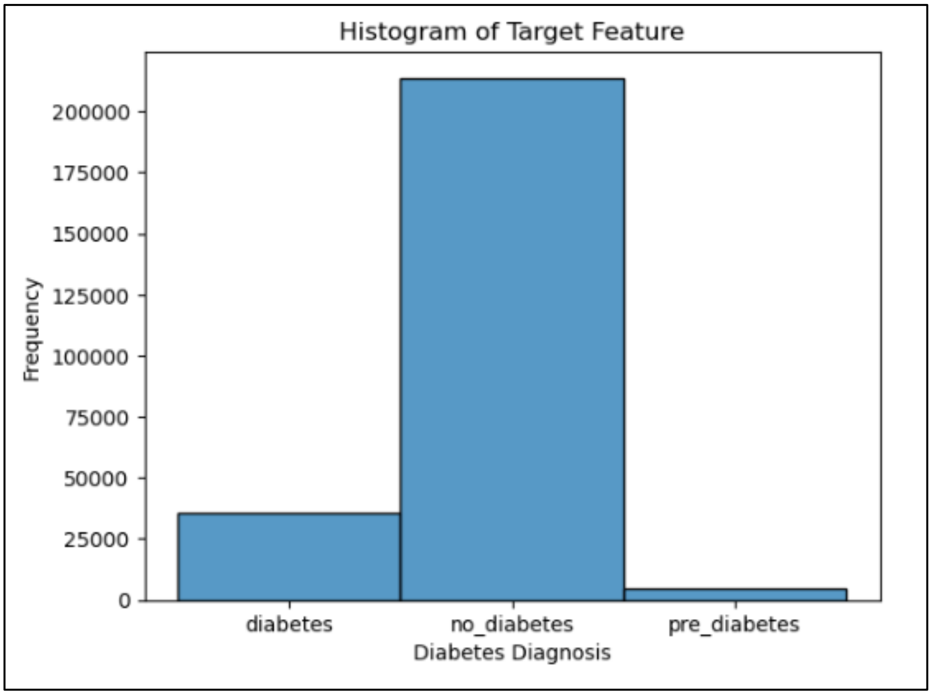
The original Diabetes dataset contained 253,680 rows and 22 columns. Python techniques were used to determine that there were no missing values in the dataset so imputations were not necessary for this dataset. The target feature values were changed to categorical texts for the purposes of analyzing my data.

0.0 was changed to 'no_diabetes'
1.0 was changed to 'pre-diabetes'
2.0 was changed to 'diabetes'

As the original Diabetes dataset was clean and had no data issues, no additional Data Wrangling steps were needed besides changing the name of my target feature column to "DiabetesDiagnosis" for clearer analyzation purposes.

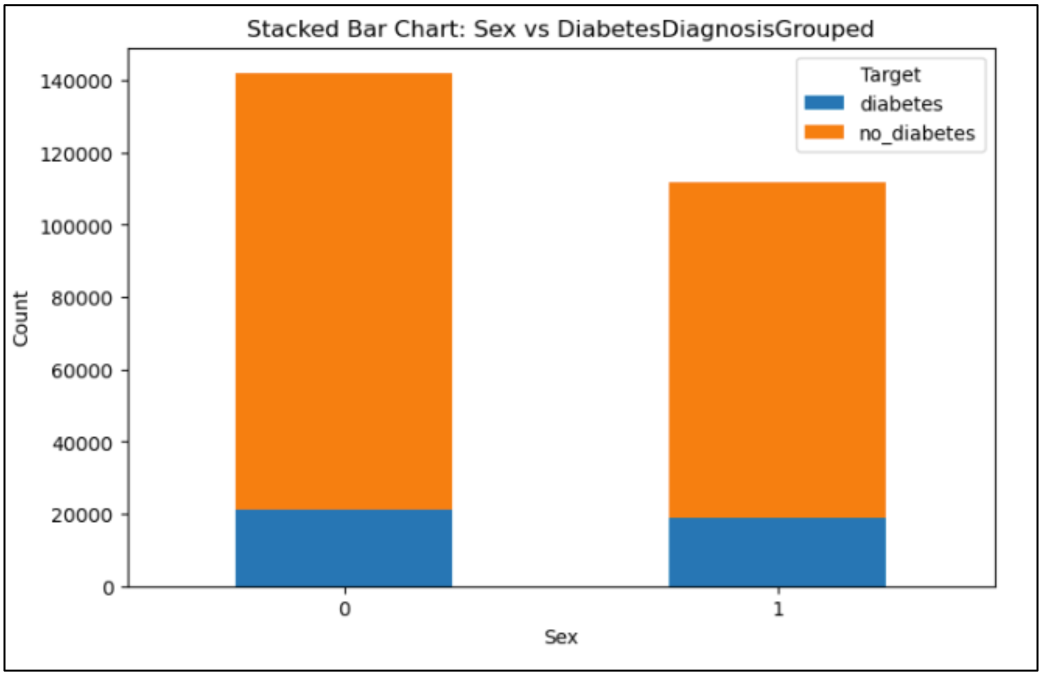
Exploratory Data Analysis

During the EDA step, each feature of the Diabetes dataset was examined and compared with other features. Since the target feature of Diabetes Diagnosis has a larger quantity of 'no_diabetes' diagnosis, the 'pre_diabetes' and 'diabetes' categories were combined.

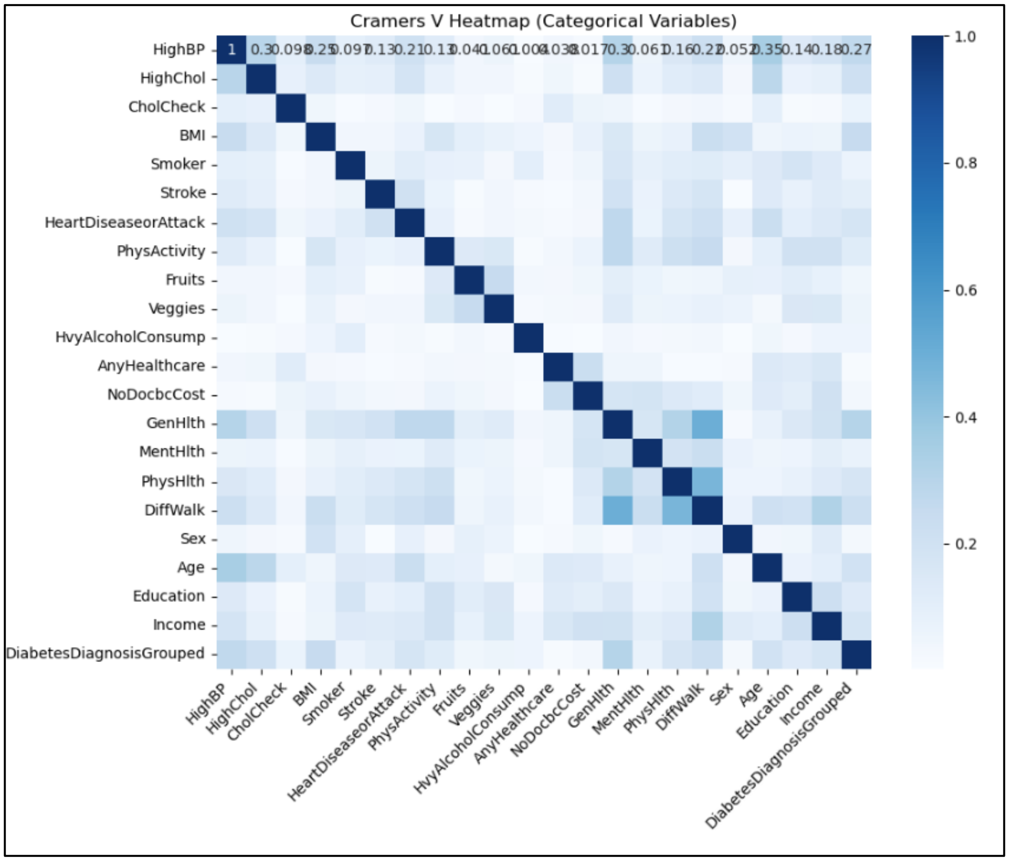


As most of the features within this dataset were categorical features, the continuous features such as 'Menthlth' (Mental Health) and 'PhysHlth' (Physical Health) were turned into categorical features using buckets to represent weeks. Both of these features originally had values between 0-30 to represent days. During the EDA process, most features were converted to integer data types to prepare for the modelling process.

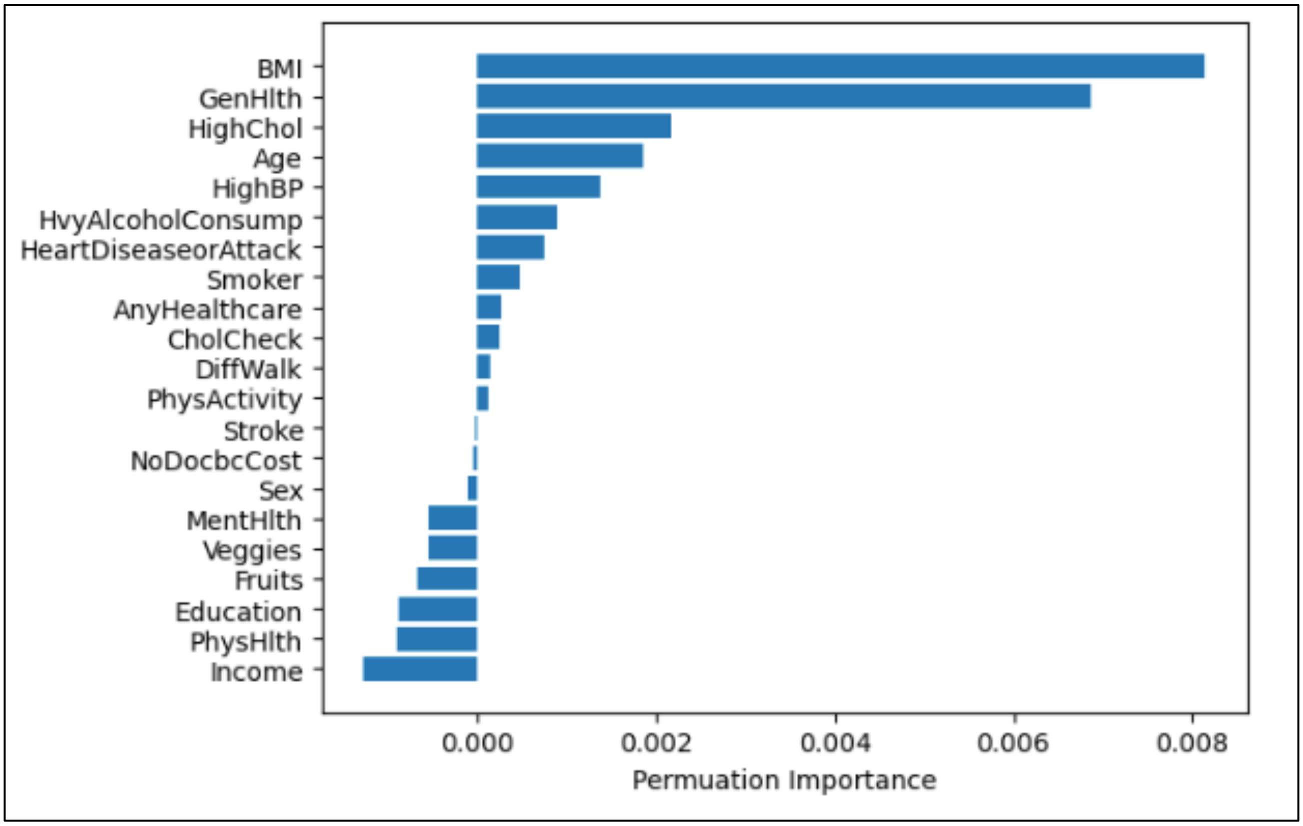
Each feature was then compared to the Diabetes Diagnosis target feature to visually determine if there were any significant relationships to be had using stacked bar charts and a boxplot.



After doing initial comparison of the health indicators with the 'diabetes' and no_diabetes' categories, a correlation heatmap using Cramer's V score was created to further explore the relationships between the indicators and the Diabetes diagnosis.



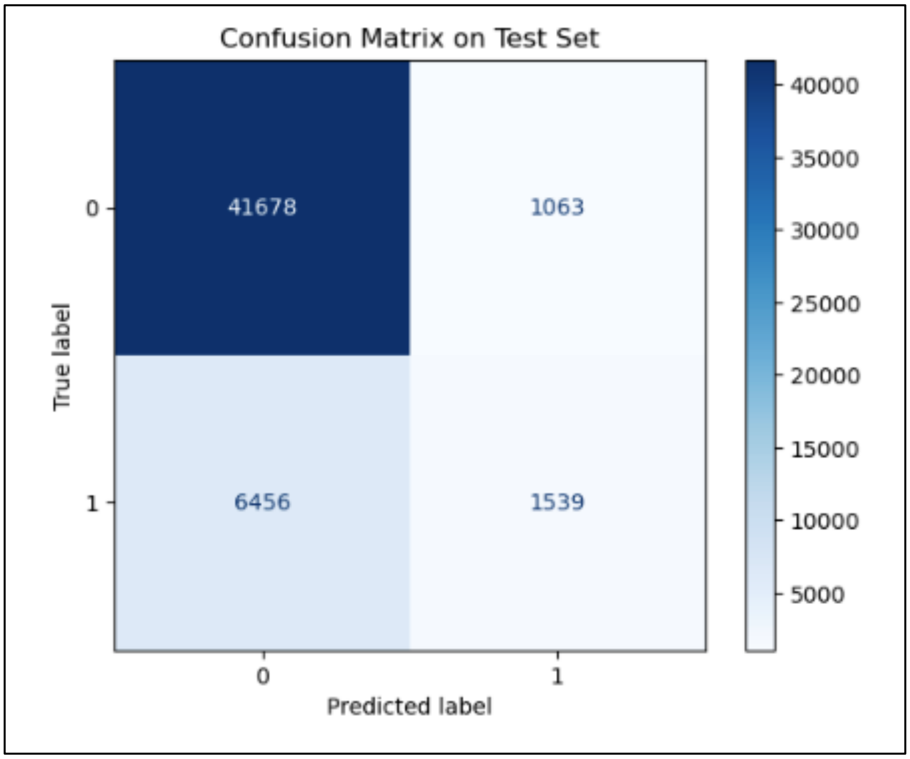
For the last step of the Exploratory Data Analysis portion of the Data Science process, a Random Forest model was instantiated to perform Feature Importance. To overcome the drawbacks of default Feature Importance computed with mean impurity decrease, a second Permutation based Feature Importance was calculated. The findings determine that BMI, General Health, and High Blood Cholesterol are the indicators that best predict a Diabetes diagnosis.

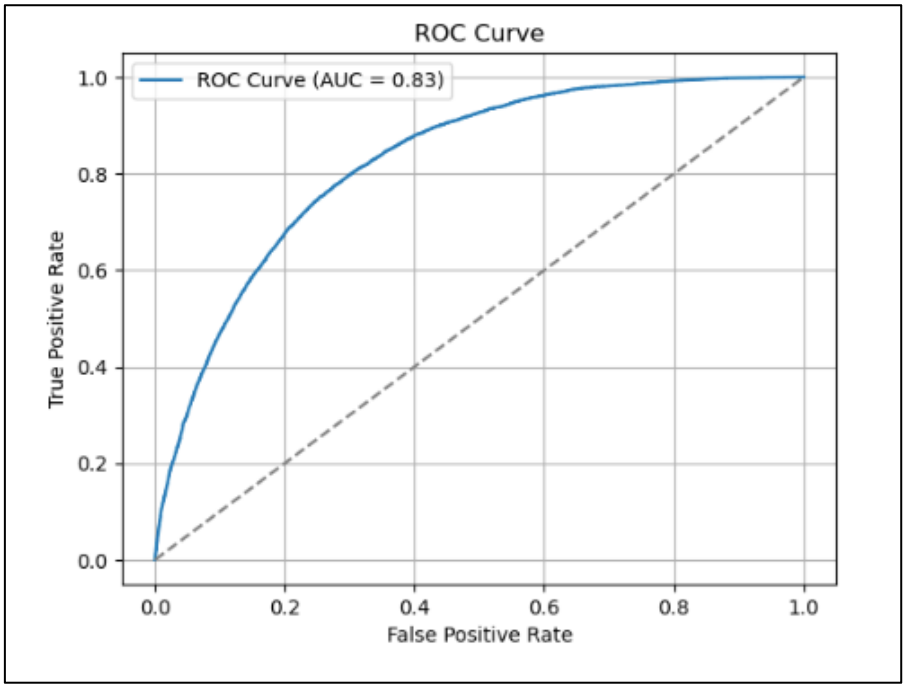


Model Selection

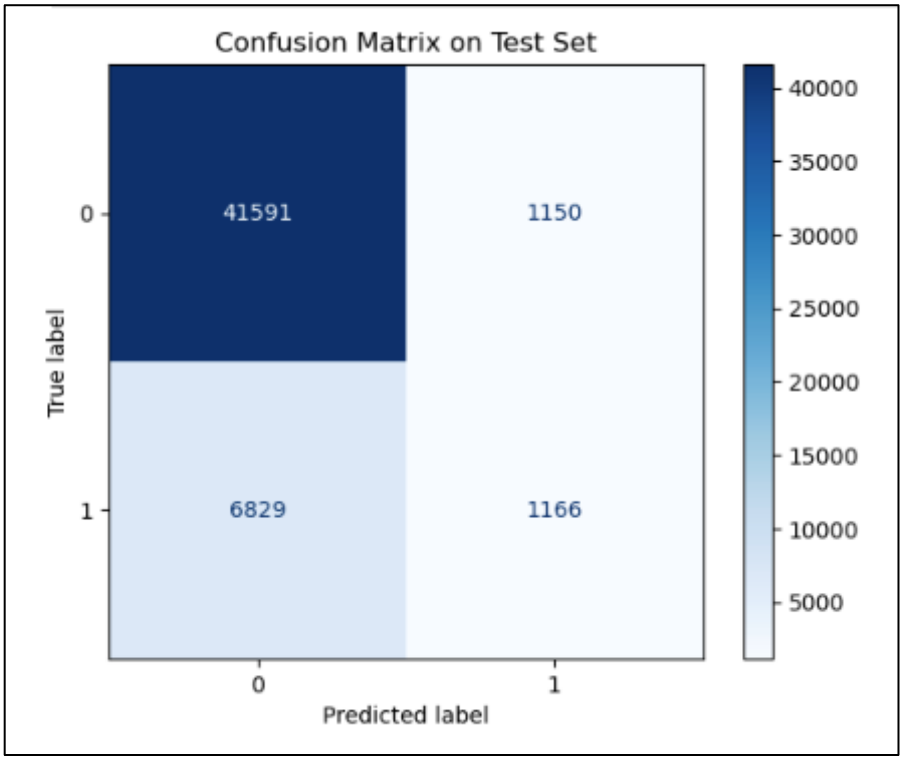
The Model Selection process was used to determine the best model that could train the best on the split training set and perform consistently on the testing set. These were the different models used: XGBoost model, KNN (K-Nearest Neighbors) model, and a SVM (Support Vector Machine) model. A RandomizedSearchCV was used for all three models with ‘Accuracy’ scoring to determine the best hyperparameters.

XGBoost:

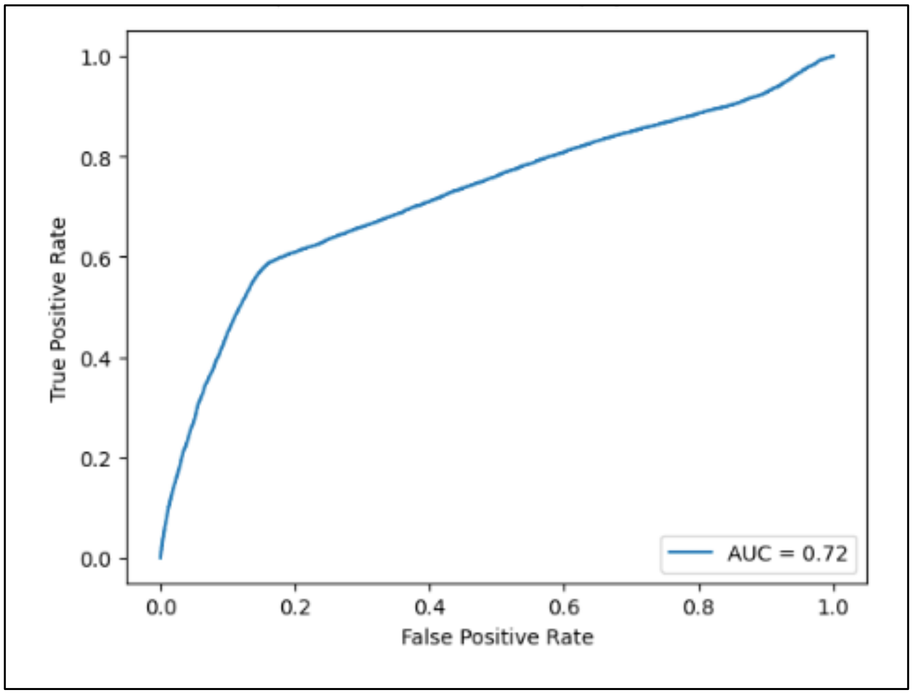

Confusion Matrix on Test Set
41678
1063
6456
1539
0
1
True label
0
1
Predicted label
40000
35000
30000
25000
20000
15000
10000
5000


ROC Curve
ROC Curve (AUC = 0.83)
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

KNN:


Confusion Matrix on Test Set
True label
Predicted label
0
1
0
1
41591
1150
6829
1166
40000
35000
30000
25000
20000
15000
10000
5000

SVM:


True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.72

Takeaways

Out of the box, the XGBoost model performed the best with an AUC of 0.83. An AUC score of a 1 showcases a perfect model and an AUC score or 0.5 is no better than random for reference. The XGBoost model had the highest recall and highest F1 score. In the case of Diabetes diagnosis prediction, having a high recall is important as a missed diagnosis can be severe to a patient's future health outcomes. A Diabetes prediction model should try and catch the most amount of diagnoses. An XGBoost model was also determined to be the best in this scenario as the high F1 score is the harmonic mean of precision and recall. This XGBoost model was able to catch the most amount of Diabetes diagnoses while also being able to predict correct positive cases.

Future Research

For further improvement of this Diabetes Predictor model, more granular health indicators can be used. Some indicators used from this dataset are derived from patient truthfulness. For a more accurate Diabetes model, health indicators that do not rely on patients but utilize concrete health data from blood diagnostics test would best be suited.