# Prediabetes ML Modelling

## Pre-Diabetes Diagnosis Codes

We have used the Livongo codes (file attached below) and added additional conditions. Using IC9 coding as ICD 10 coding is not available in the data source.

The

Main groups:

1. Testing for glucose tolerance
2. Gestational diabetes
3. Overweight/Obesity
4. Family history of diabetes
5. Hypertension
6. Clinical conditions with increased risk of prediabetes
7. Testing for glucose tolerance

|  |  |
| --- | --- |
| 79021 | Impaired fasting glucose |
| 79022 | Impaired glucose tolerance test (oral) |
| 79029 | Other abnormal glucose |
| 7902 | Abnormal glucose |

1. Gestational diabetes

We removed any codes that reflected a mother with pre-existing diabetes

|  |  |
| --- | --- |
| V1221 | Hx of gestational diabetes |
| 64800 | DIABETES IN PREG-UNSPEC |
| 64801 | DIABETES-DELIVERED |
| 64802 | DIABETES-DELIVERED W P/P |
| 64803 | DIABETES-ANTEPARTUM |
| 64804 | DIABETES-POSTPARTUM |
| 6488 | Abnormal glucose tolerance |

1. Overweight/Obesity

|  |  |
| --- | --- |
| V8522-25 | Overweight category |
| V85.30-39 | Obese category |
| V85.41-45 | Severe Obesity category |

1. Family history of diabetes

|  |  |
| --- | --- |
| V180 | Family history of diabetes |

1. Hypertension

Some codes related to hypertension

|  |  |
| --- | --- |
| 4010 | MALIGNANT HYPERTENSION |
| 4011 | BENIGN HYPERTENSION |
| 4019 | HYPERTENSION NOS |
| 402.xx | HYPERTENSIVE HEART DISEASE- malignant, benign, unspecified |
|  |  |

1. Clinical conditions with increased risk of prediabetes

|  |  |
| --- | --- |
| V69.0 | Lack of physical exercise |
| 287.03 | Obesity hypoventilation syndrome |
| 253 | Acromegaly |
| 255 | Cushing’s syndrome |
| 2513 | Postpancreatectomy (complete) (partial) |

## Data Source Used:

Schema (Database)- edh analytics solutions db (QC)

Table 1- udp person idmapping (This table is used to get Person Internal Id)

Table 2- participant integrated (This table is used to get other demographics columns)

Table 3- compass claim raw person identifiers (This table is used to get Patient Key/Person Key)

Table 4- compass claims raw med claim rollup (This table is used to get start date and end date for)

Table 5- compass claims raw diagnosis (This is used to get the diagnosis code)

Table 6- compass claims rollup 2019/2020/2021 (This table is used to get claims data wrt each year)

Table 7- compass claims raw cpt hcpcs procedure (This table is used to get procedure codes)

Table 8- compass claims raw procedure category (This table is used to get procedure category codes)

Data is pulled using a sql file which is created using above tables. Flags are created using diagnosis code from the excel file shared by Mike for year 2019, 2020 and 2021.

## SQL File 1

**At first,**

1. Testing for glucose tolerance
2. Gestational diabetes
3. Overweight/Obesity
4. Family history of diabetes
5. Hypertension
6. Lack of physical exercise

**The codes from above six variables are used to create the label column and,**

**Below mentioned variable codes for are used to create separate features-**

1. Sleep apnea
2. Cushing syndrome
3. Acromegaly

**Below mentioned variable codes for number 10 and 11 are used to create a single feature**

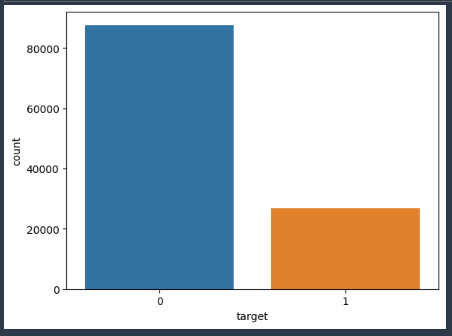
1. Encounter for screening for diabetes mellitus
2. Dietary counselling surveillance

**Below mentioned variable codes for number 13, 14 and 15 are used to create a single feature**

1. Polyuria
2. Nocturia
3. Polydipsia

**Note- The labels are created for year 2019,2020 and 2021 separately and later combined.**

**Event rate is 77% for Class 0 and 23% for class 1.**



**Also, Alight data source do not have ICD-10 diagnosis code. Hence, we have used ICD-9 diagnosis code.**

ICD stands for- The International Classification of Diseases, Tenth Revision, Clinical Modification — more commonly known as ICD-10-CM — is a classification system of diagnosis codes representing conditions and diseases, related health problems, abnormal findings, signs and symptoms, injuries, and external causes of injuries

## Files created-

### Pre-Diabetes-EDA.ipynb:

Exploring Pre-Diabetes data.

**S3 Path**- s3://adl-core-sagemaker-studio/external/artichauhan/Pre-Diabetes/Raw Data/data in parquet format

### Pre-Diabetes-Processing.ipynb:

This file is created to join Pre-Diabetes data and demographic data and remove duplicates.

**S3 Path for Pre-Diabetes parquet files-**

s3://adl-core-sagemaker-studio/external/artichauhan/Pre-Diabetes/Raw Data/data in parquet format/

**Two demographics files used in this notebook.**

**S3 path-** s3://adl-core-sagemaker-studio/external/artichauhan/Pre-Diabetes/Raw Data/Demographics files/

**Two files are-**

Hype\_ML\_demographics\_19\_20\_21\_1csv

Hype\_ML\_demographics\_19\_20\_21\_2.csv

After combining data and removing duplicates, data is then rolled up to single category wherever there are multiple variants of one category. Further, data is split into train and test set. Other transformation is performed such as missing value imputation, outlier treatment separately on the train set and later test set to avoid data leakage. Final train and test sets after cleaning are stored in s3.

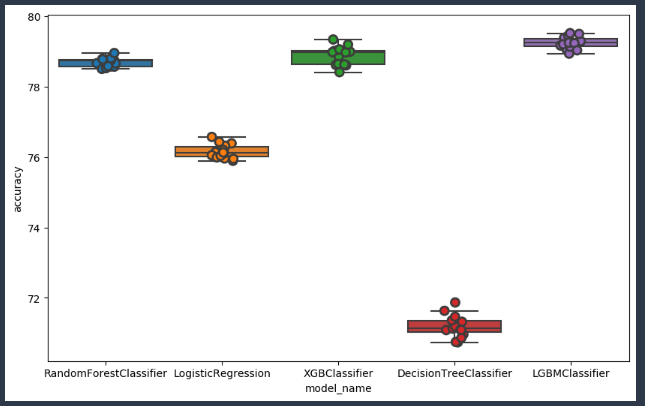
### Pre-Diabetes-Modelling.ipynb:

In this notebook, we will perform below steps

* Load transformed training data and testing data
* Build multiple ML model using training data
* Hyper-parameter tuning of best model to identify best parameters which give good results
* Evaluate model performance on various metrics and graphs such as ROC and PR curve.
* Saving the final (best) ML model in S3 which will be used to make inference on new data
* Interpreting Model using SHAP values

## Modelling based on above SQL file-

## Baseline Model comparison on Un-balanced Dataset



|  |  |  |
| --- | --- | --- |
| **Model Name** | **Mean Accuracy%** | **Std. Dev. Accuracy** |
| Decision Tree Classifier | 71.19 | 0.13 |
| LGBM Classifier | 79.25 | 0.17 |
| Logistic Regression | 76.15 | 0.19 |
| Random Forest Classifier | 78.68 | 0.12 |
| XGB Classifier | 78.88 | 0.25 |

As from above comparison, it can be observed that Mean Accuracy is greater for LGBM, and second-best model is XGBoost.

Further, comparing classification report from LGBM and XGBoost.

***Note- All the metrices are based on Test Set.***

### Classification Report for XGBoost

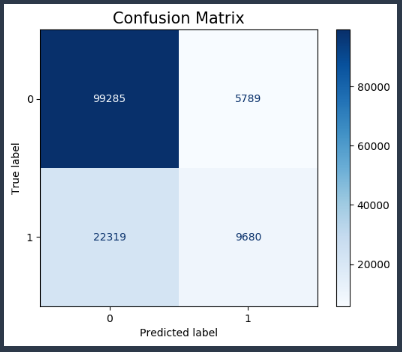
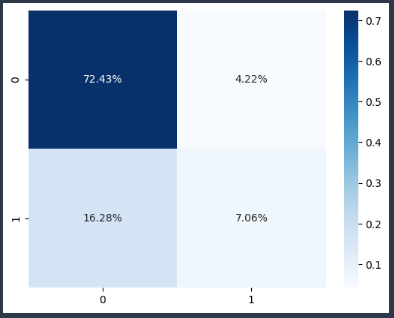
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.82 | 0.95 | 0.88 |
| **1** | 0.63 | 0.30 | 0.41 |
| **Macro Average** | 0.72 | 0.62 | 0.64 |
| **Weighted Average** | 0.77 | 0.80 | 0.77 |
| **Accuracy** | | | 0.80 |

### Classification Report for LGBM

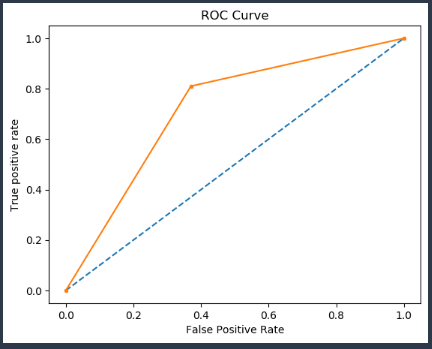
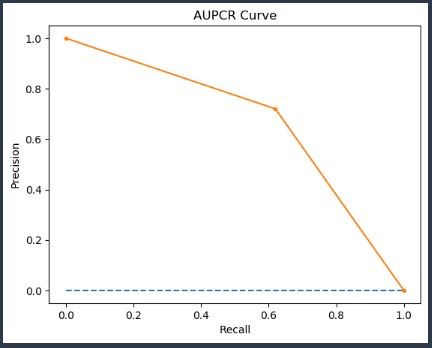
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.81 | 0.95 | 0.88 |
| **1** | 0.64 | 0.28 | 0.39 |
| **Macro Average** | 0.72 | 0.61 | 0.63 |
| **Weighted Average** | 0.77 | 0.79 | 0.76 |
| **Accuracy** | | | 0.79 |

After comparing classification report for XGBoost and LGBM, It can be observed that Recall and F1 score is better for XGBoost. Hence, we will work with XGBoost model and check other metrices.

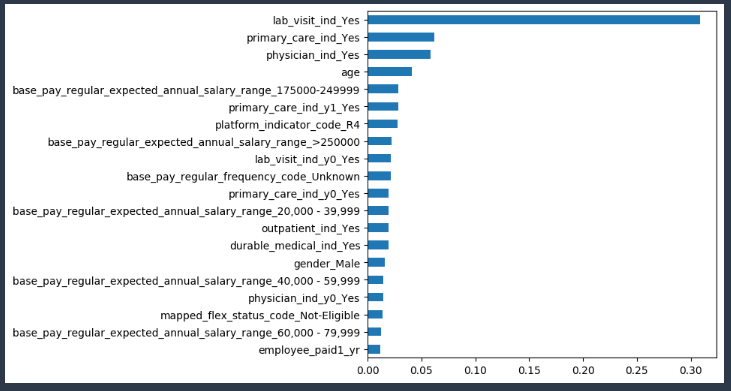
### Confusion Matrix for XGBoost

## ROC and Precision-Recall Curve XGBoost

### Features Importance from XGBoost Model with default parameters



## Baseline Model comparison on Balanced Dataset after SMOTE

* *SMOTE – Synthetic Minority Oversampling Technique.*
* Implemented an oversampling technique (SMOTE) to balance the dataset between the two target classes on XGB Classifier and LGBM Classifier as they gave the best accuracy on unbalanced dataset.
* Ran different models on the balanced dataset and checked model accuracy,
* Mean accuracy is based on running K =15 folds cross validation.
* XGBoost gives us the highest accuracy (80.32%), LGBM a close second of (80.17%)

### Classification Report on XGBoost after SMOTE

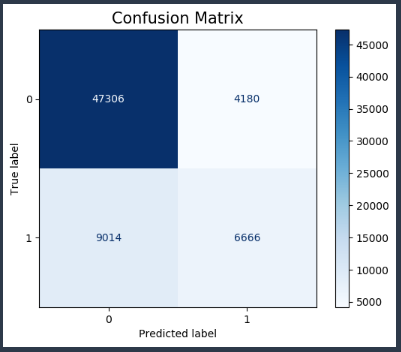
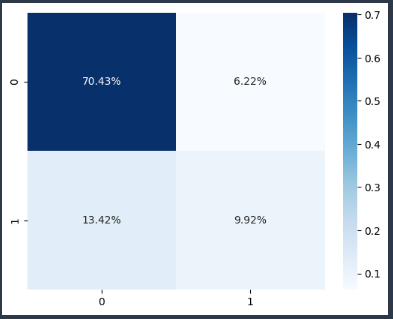
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.83 | 0.90 | 0.86 |
| **1** | 0.54 | 0.37 | 0.44 |
| **Macro Average** | 0.68 | 0.64 | 0.65 |
| **Weighted Average** | 0.76 | 0.78 | 0.76 |
| **Accuracy** | | | 0.78 |

### Classification Report on LGBM after SMOTE

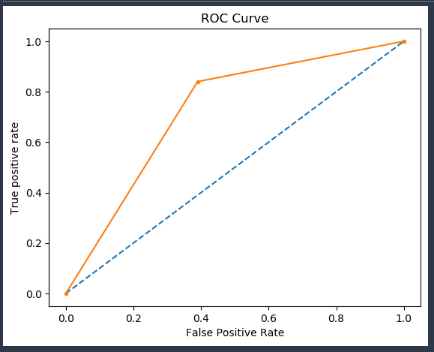
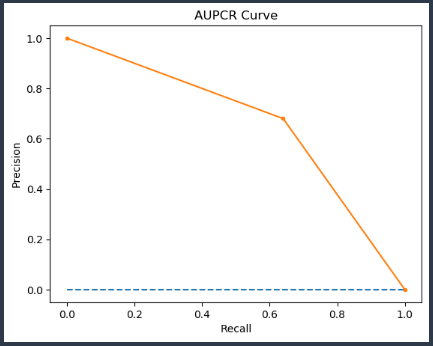
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.82 | 0.90 | 0.86 |
| **1** | 0.53 | 0.35 | 0.42 |
| **Macro Average** | 0.67 | 0.63 | 0.64 |
| **Weighted Average** | 0.75 | 0.77 | 0.76 |
| **Accuracy** | | | 0.77 |

As it can be observed from above comparison that XGBoost is slightly better than LGBM based on Recall and F1 score. Further, will check matrices for XGBoost.

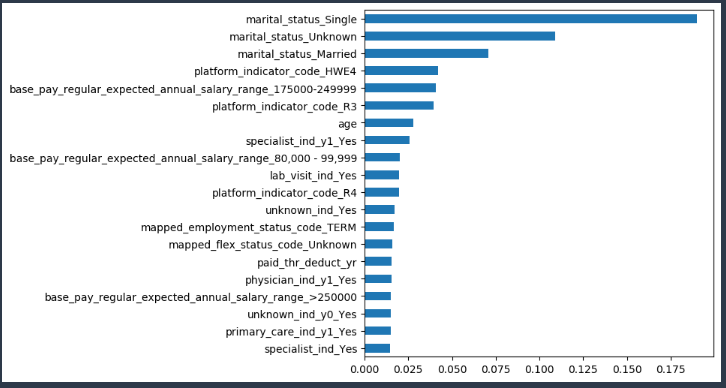
### Confusion Matrix for XGBoost after SMOTE

### ROC and Precision-Recall Curve for XGBoost

### Features Importance from XGBoost Model with default parameters after SMOTE



## Hyper-Parameter Tuning

On the balanced data, XGBoost was the model with the highest accuracy.

Bayesian Optimization (BO) – An approach that identifies hyperparameters best suited to generate a model with the highest accuracy possible.

BO implemented on XGBoost.

Accuracy for XGBoost increased from 78% to 79% but recall for 1 call decreased from 35% to 31%.

As we want to predict more accurately the pre-diabetic employees. We choose the default parameter model to be the best model on balanced train dataset.

**Now, I thought of changing some variables in the dataset and after discussion with Kyle and letting him know the plan he agreed to change the dataset and work on that. Below is the second SQL file created.**

## SQL File 2

**Here in second SQL file, basically changed the variables used for creating target column and individually used every other variable.**

1. Testing for glucose tolerance
2. Gestational diabetes

**The codes from above two variables are used to create the label column and,**

**Below mentioned variable codes for are used to create separate features-**

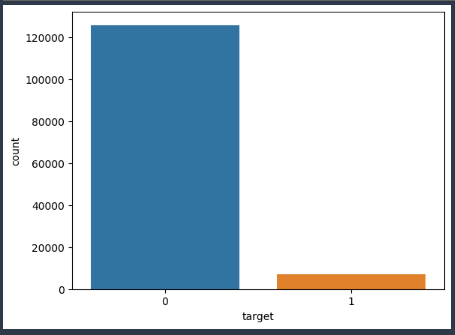
1. Overweight/Obesity
2. Family history of diabetes
3. Hypertension
4. Lack of physical exercise
5. Sleep apnea
6. Cushing syndrome
7. Acromegaly
8. Encounter for screening for diabetes mellitus
9. Dietary counselling surveillance
10. Polyuria
11. Nocturia
12. Polydipsia
13. Preventive visits
14. Blood glucose test
15. Ovarian disorder
16. Heart disorder

**Note- The labels are created for year 2019,2020 and 2021 separately and later combined.**

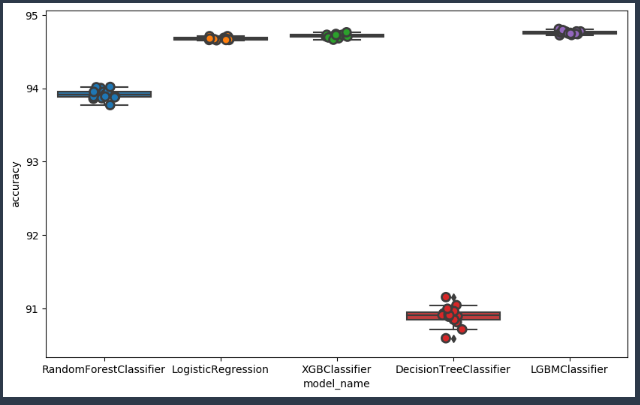
Some Other variables used are-

|  |
| --- |
| 1. Tot billed amt |
| 1. Avg billed amt |
| 1. Std billed amt |
| 1. Max billed amt |
| 1. Employer paid |
| 1. Employee paid1 |
| 1. Employee paid2 |
| 1. Paid thr deduct |
| 1. Specialist ind |
| 1. Primary care ind |
| 1. Physician ind |
| 1. Unknown ind |
| 1. Durable medical ind |

**Also, as a smaller number of codes are used in creating variable, event rate decreased from 23% to 5% for 1 class.**



## Baseline Model comparison on Un-balanced Dataset

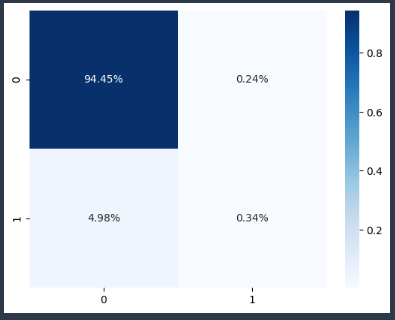
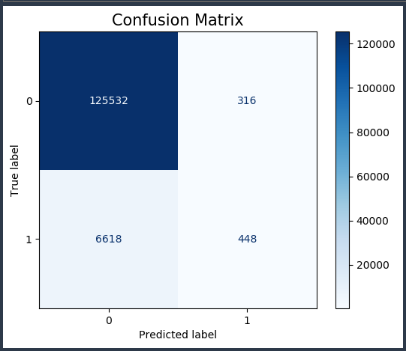


|  |  |  |
| --- | --- | --- |
| **Model Name** | **Mean Accuracy%** | **Std. Dev. Accuracy** |
| Decision Tree Classifier | 90.89 | 0.13 |
| Logistic Regression | 94.67 | 0.02 |
| Random Forest Classifier | 93.91 | 0.01 |
| XGB Classifier | 94.71 | 0.06 |
| LGBM | 94.75 | 0.02 |

It is observed from above comparison that XGBoost is slightly better than other algorithms as per cross-validation mean accuracy.

After comparing Recall and F1 score for individual algorithms over test set, it was observed that XGBoost is the model that gives best Recall out of all. Hence, it was again chosen as the best model and other matrices were identified.

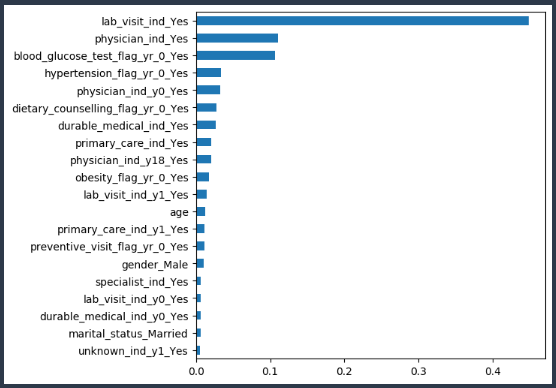
### Confusion Matrix for XGBoost



### Classification Report for XGBoost

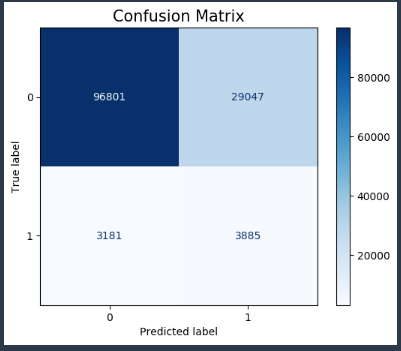
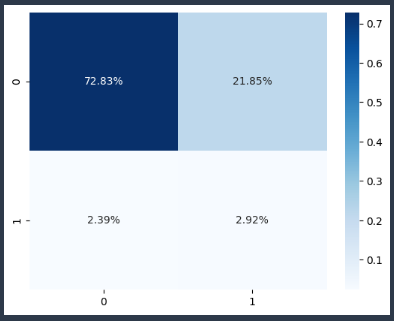
|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.95 | 1.00 | 0.97 |
| **1** | 0.59 | 0.06 | 0.11 |
| **Macro Average** | 0.77 | 0.53 | 0.54 |
| **Weighted Average** | 0.93 | 0.95 | 0.93 |
| **Accuracy** | | | 0.95 |

### Features Importance from XGBoost Model with default parameters



## XGBoost modelling output on balanced dataset after SMOTE

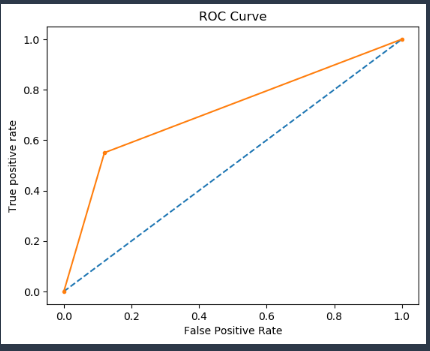
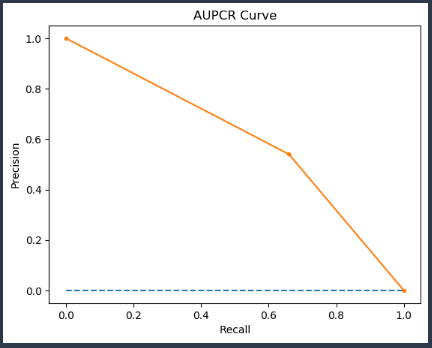
### Confusion Matrix

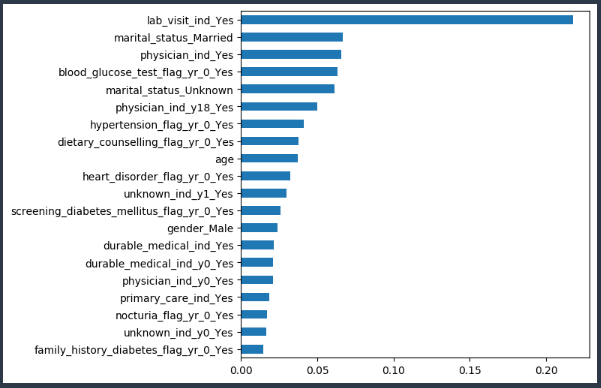
### Classification Report

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** |
| **0** | 0.97 | 0.77 | 0.86 |
| **1** | 0.12 | 0.55 | 0.19 |
| **Macro Average** | 0.54 | 0.66 | 0.53 |
| **Weighted Average** | 0.92 | 0.76 | 0.82 |
| **Accuracy** | | | 0.76 |

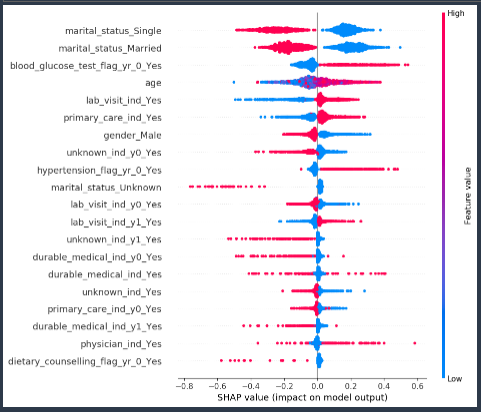
### ROC and Precision-Recall Curve

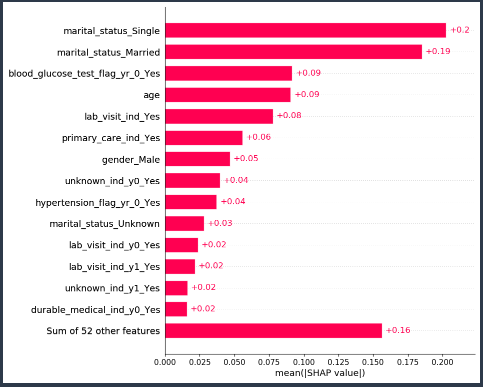
### Features Importance from XGBoost Model with default parameters after SMOTE

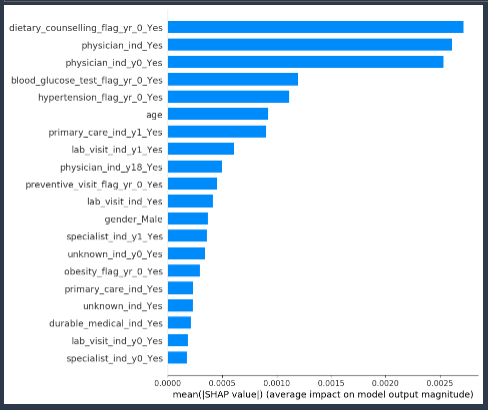


### SHAP Values Summary Plot



### Top 15 Features based on SHAP Mean Values





## Next Step-

* Now, need to create new SQL query to pull data
* Changes that we need to make in SQL are-
* Add Obesity diagnosis code to Pre-Diabetic label, which will increase the event rate for class 1.
* Accordingly, we need to check which other features can be added to create label that will increase class 1 values and further will run the model and evaluate the results.