THE UNIVERSITY OF CHICAGO

Health Analytics
University of Chicago

Diabetes Prediction

Prepared for: Arnab Bose

Submitted by: Ashish Mahadik

January 2019

Executive Summary

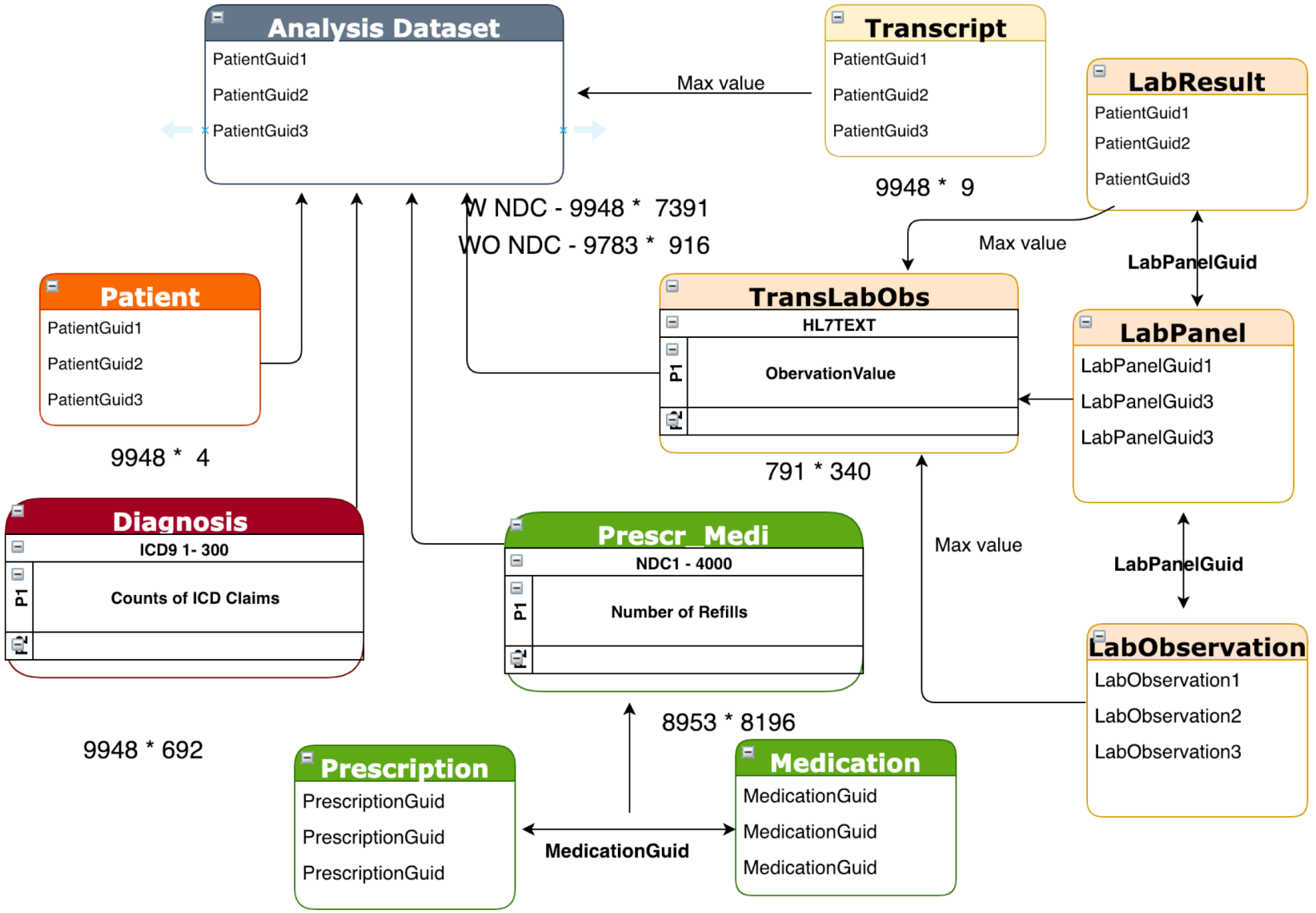
Background & Objective: Goal of this project was to build ML predictive model for predicting diabetic Type 2 condition of a patient from around 9900 patients health EMR records.

Key Process Steps and Findings:

- Analysis dataset was created by merging Lab, Transcript, Diagnosis and Prescription files. Analysis dataset is designed where unique Patient Ids are on row level and Key features at column level.
- Age, BMI, SystolicBP, Weight, DiastolicBP, Temperature, Respiratory rate and Acute condition are important variables to predict diabetic condition.
- Male has higher proportion of diabetic patients but gender does not come as important variable to predict diabetic condition.
- 401 – Hypertension and 272 – lipoid metabolism conditions coexist with diabetic conditions and they are important variables for prediction.
- XGBoost reports lowest brier score of 0.1230 but SVM is better model if we consider False Negative rate as evaluation metric.

Data Transformation

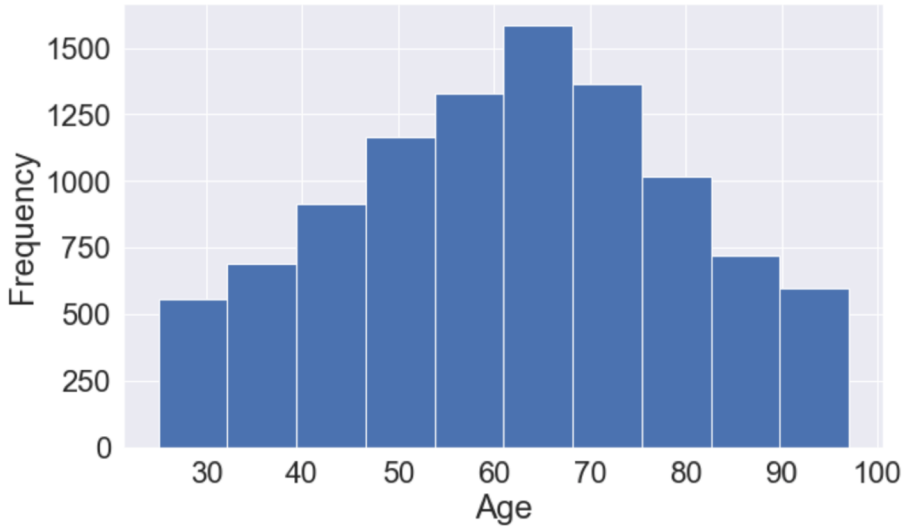
- LabResult , LabPanel and LabObservation datasets are merged together to create TransLabObs dataset with PatientGuid, max observation values at Row level and HL7TEXT as columns.
- Transcript dataset with maximum value from multiple lab measurements was created.
- Substring of ICD9 3 characters was used to create columns of Diagnosis dataset
- NDC code as 8196 columns were created in Prescription_ medication dataset
- All datasets from step 1 to 4 were merged into patient dataset. This created Patient by condition dataset as final analysis dataset.

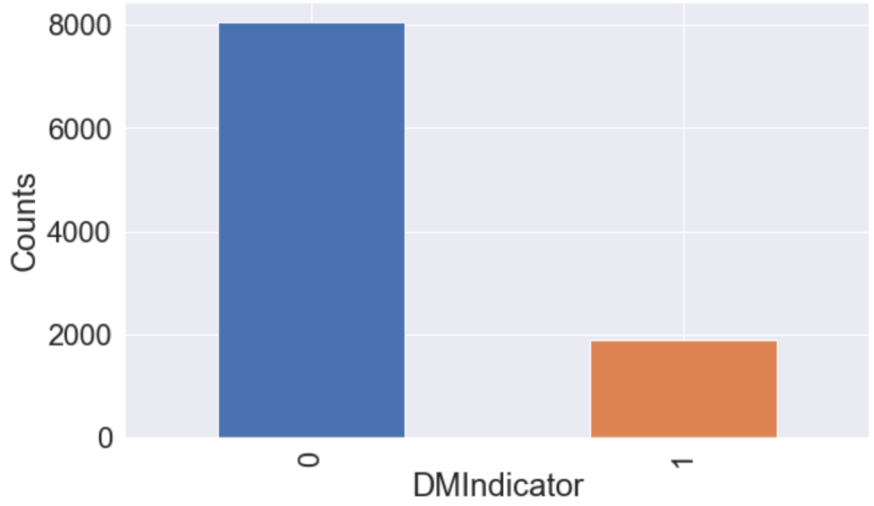


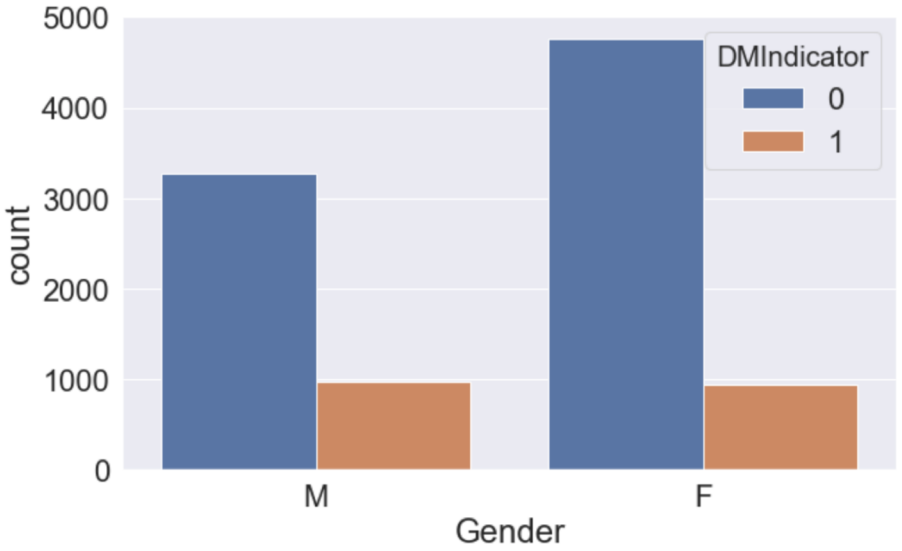
Layout of Analysis Dataset for Predictive Model

PatientGuid	Transcript Variables (~10)	Lab Variables (~300) –HL7 Text	ICD9 Diagnosis Code(~ 600)	NDC Codes(~8000)	DMIndicator
PAtient1	Max Observation Values	Max Observation Values	Number of occurrence of ICD9 code	Number of refills for NDC	
Patient2					
Patient3					

Visual Analysis of Input Variables

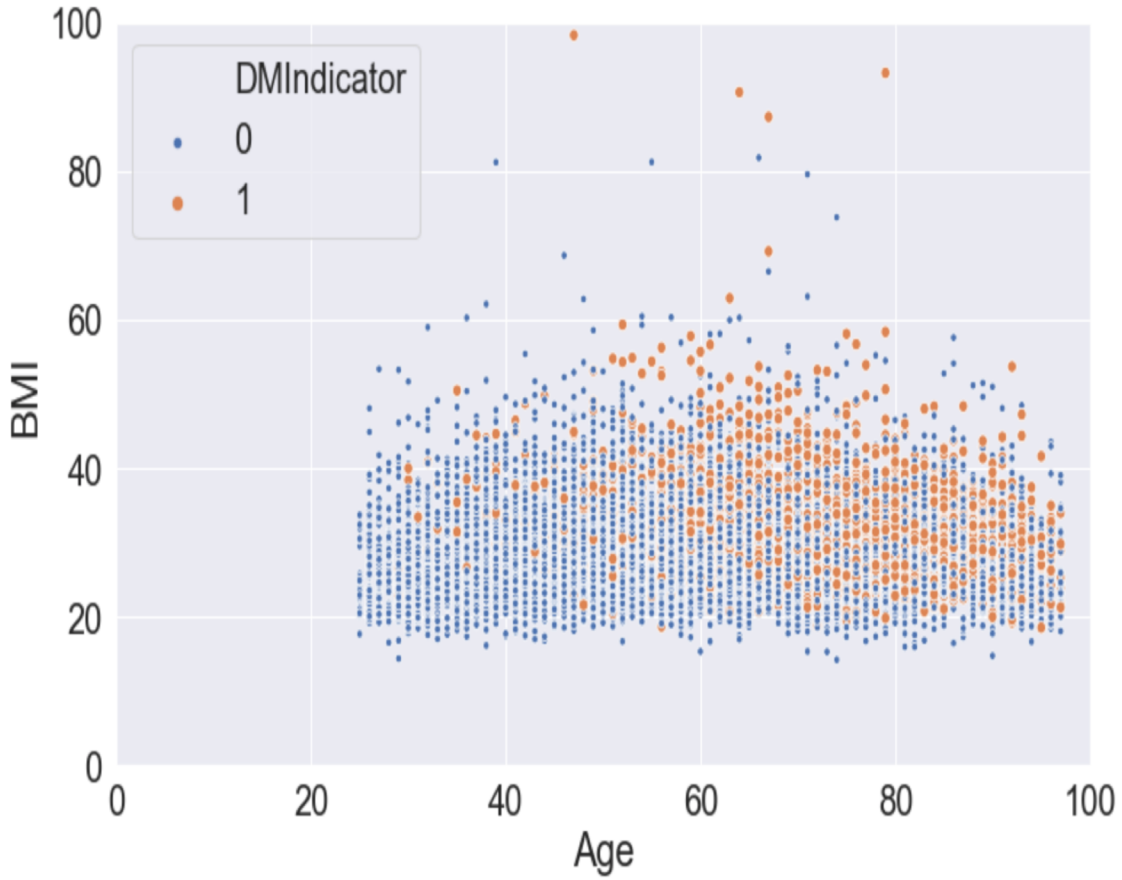


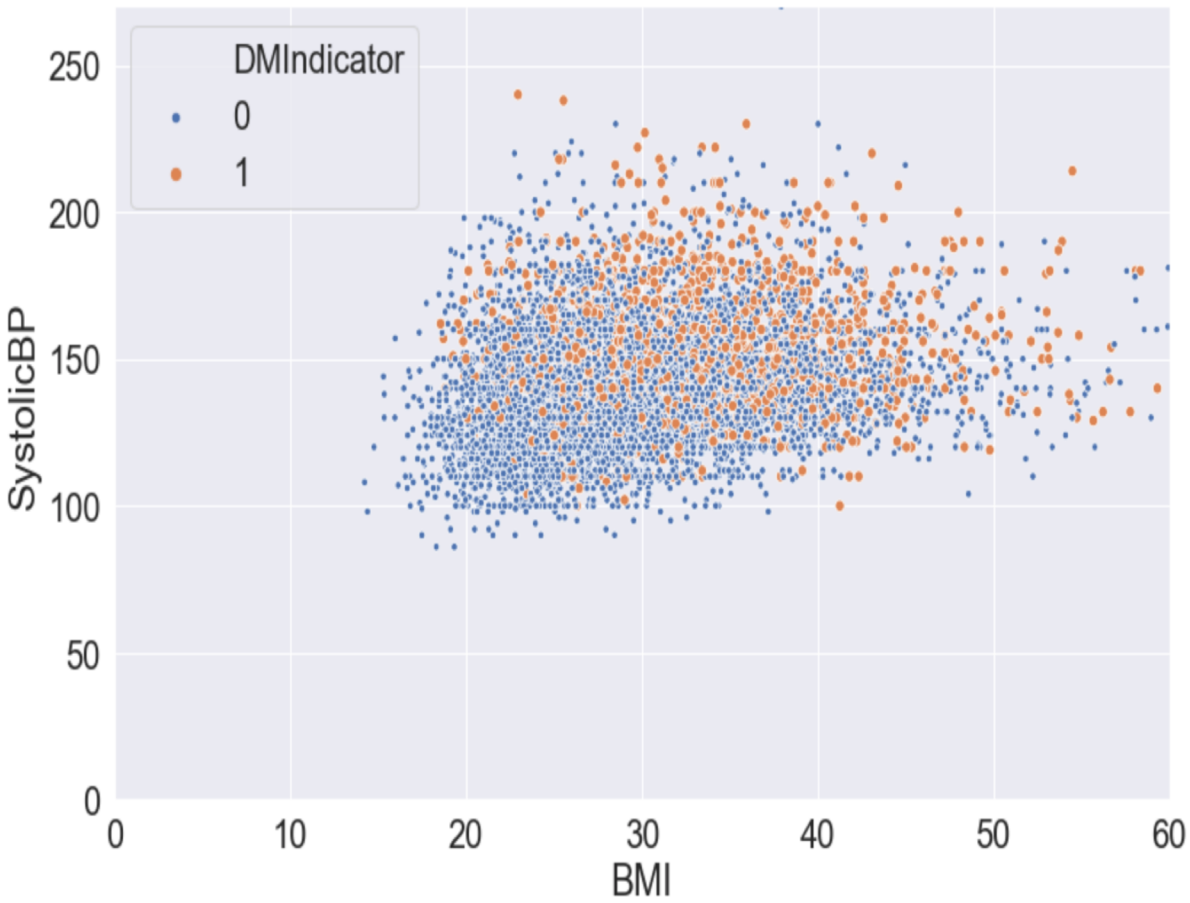




- Average age from sample data is around 61 years
- Only 18% of samples have positive diabetes condition
- Male patient has higher probability of being diabetic

Visual Analysis of Input Variables

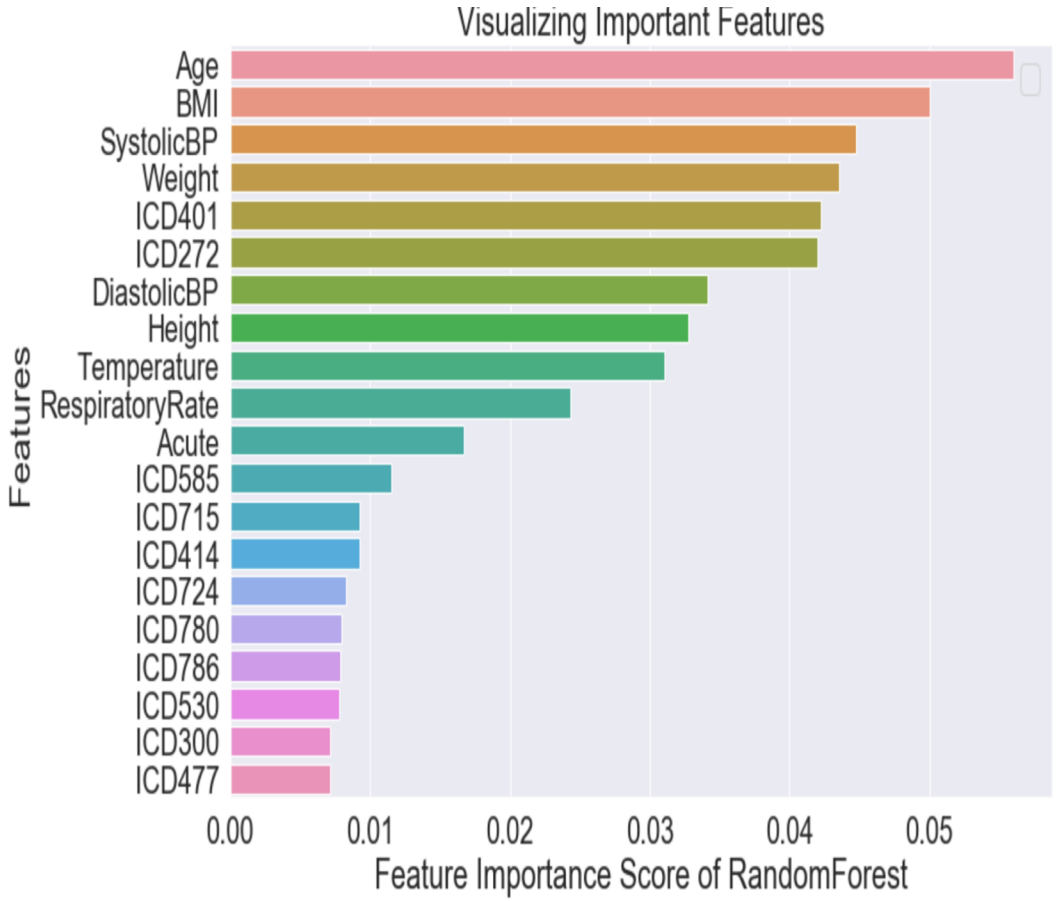




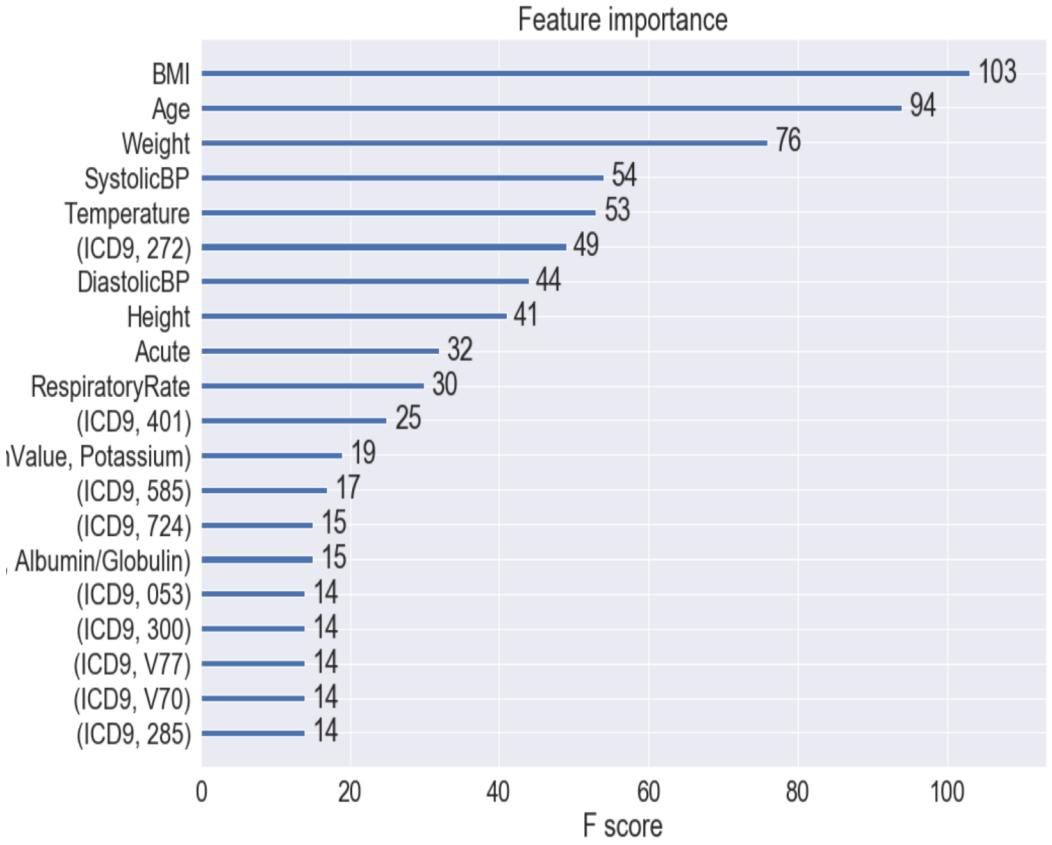
- **Patients with age higher than 60 tend to be diabetic**
- **Patients with higher BMI and higher Systolic blood pressure tend to be diabetic**

Comparison of Key Variables of Two Models

Feature Importance from Random Forest



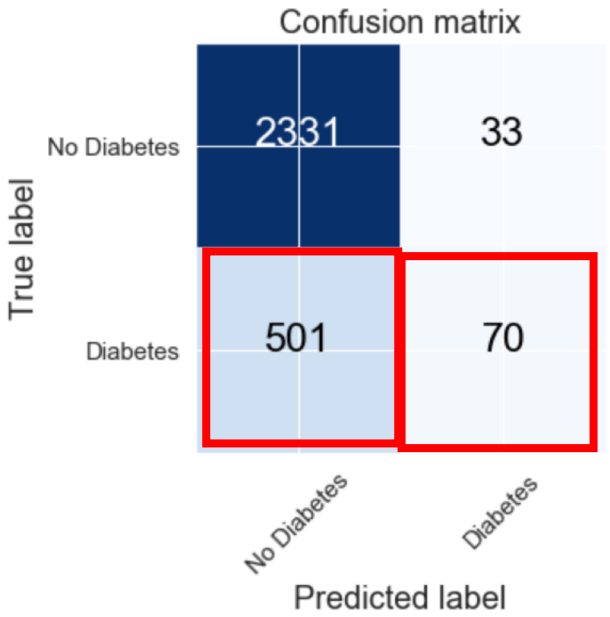
Feature Importance from XGBoost



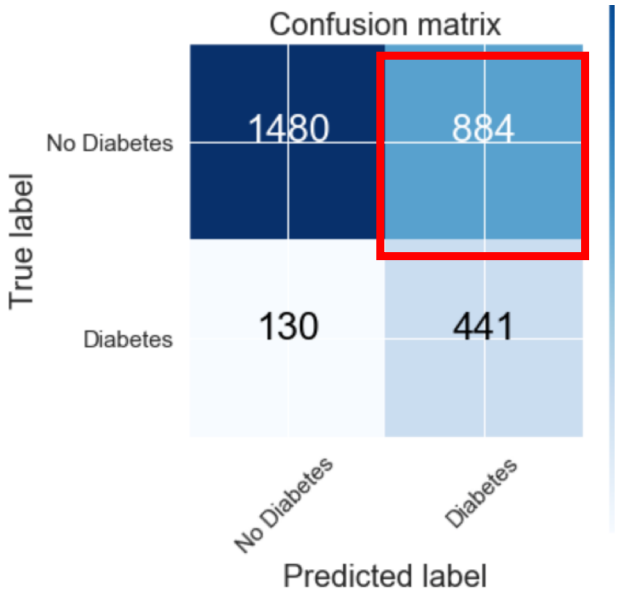
- **401 – Hypertension and 272 – lipoid metabolism conditions coexist with diabetic conditions**
- **Age, BMI, SystolicBP, Weight, DiastolicBP, Temperature, Respiratory rate and Acute condition are important variables to predict diabetic condition.**

Comparisons of Confusion Matrices

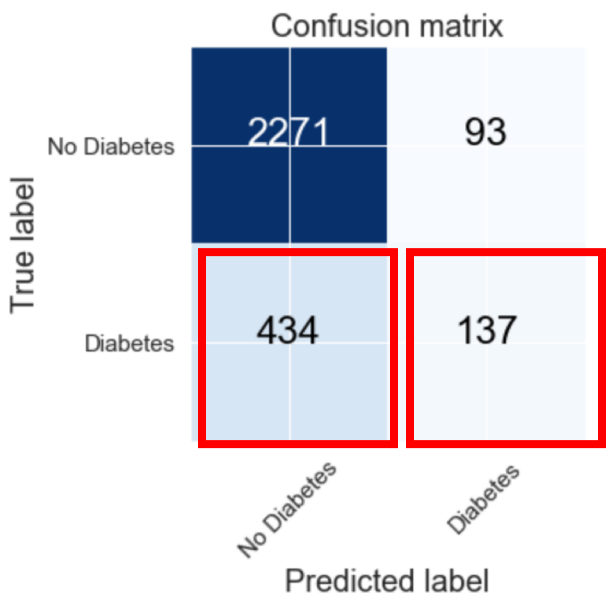
Random Forest



SVM



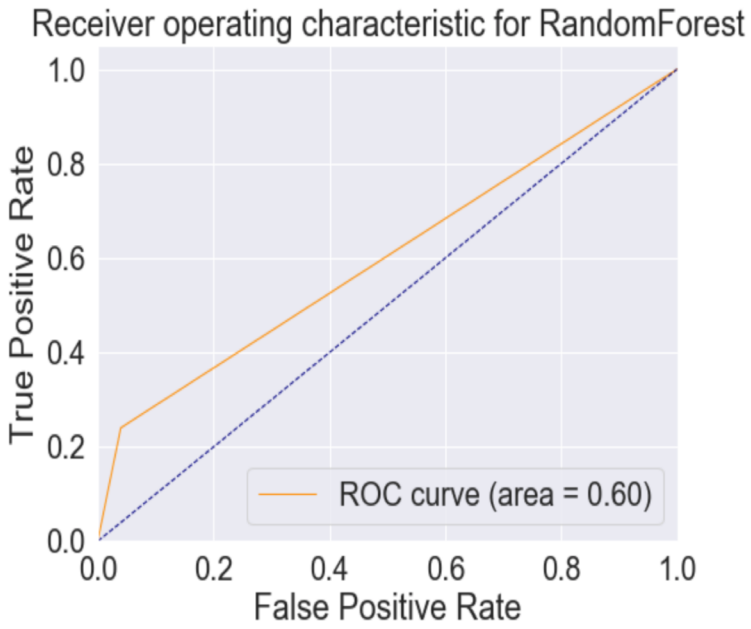
XGBoost with SMOTE

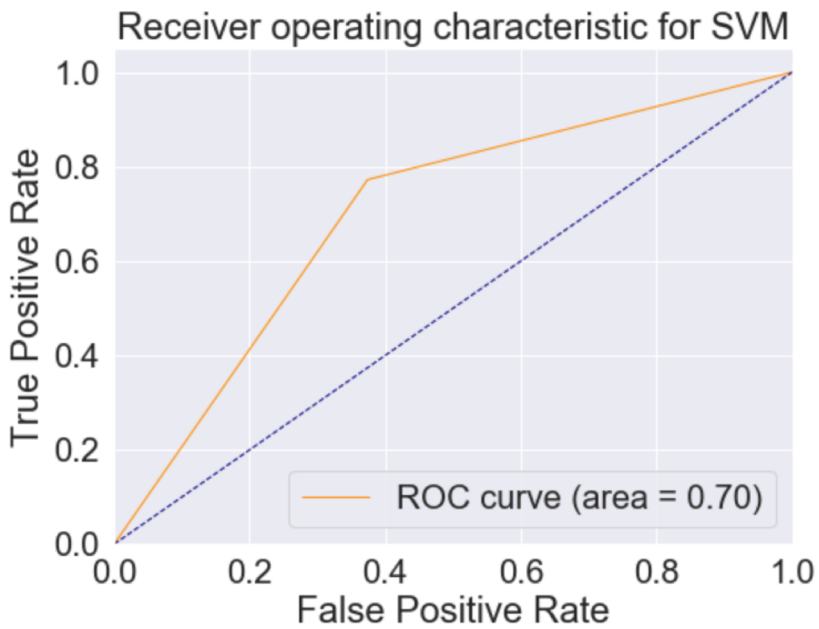


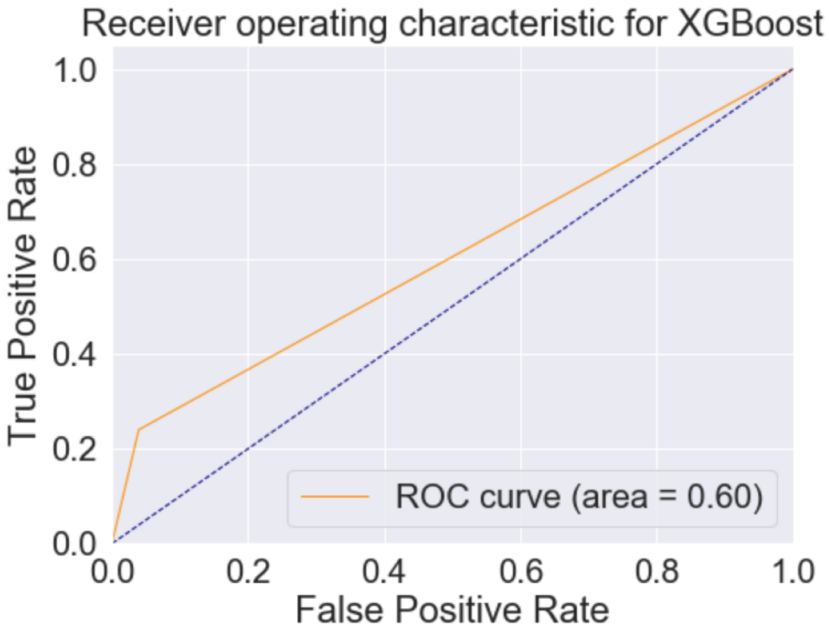
***Red boxes highlight poor performance of models

- **Random Forest and XGBoost does good job in predicting true negative - “No Diabetes” condition but does relatively bad in predicting true positive - “Diabetes” condition.**
- **SVM is doing good in predicting True positive of “Diabetes” condition.**

Comparison of ROC Curves



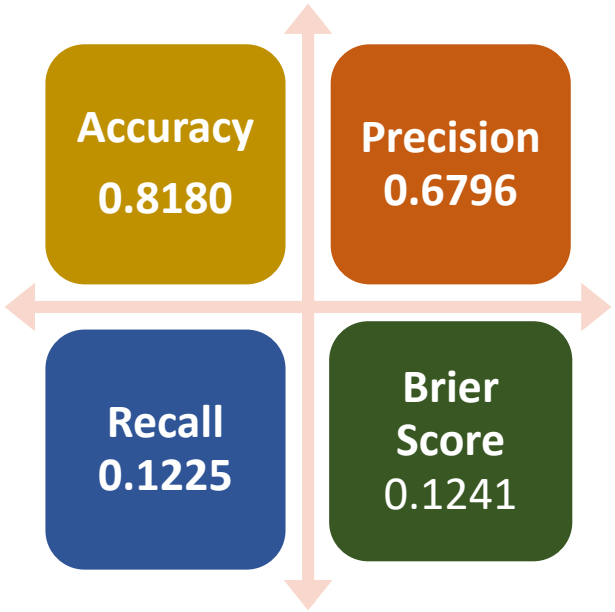




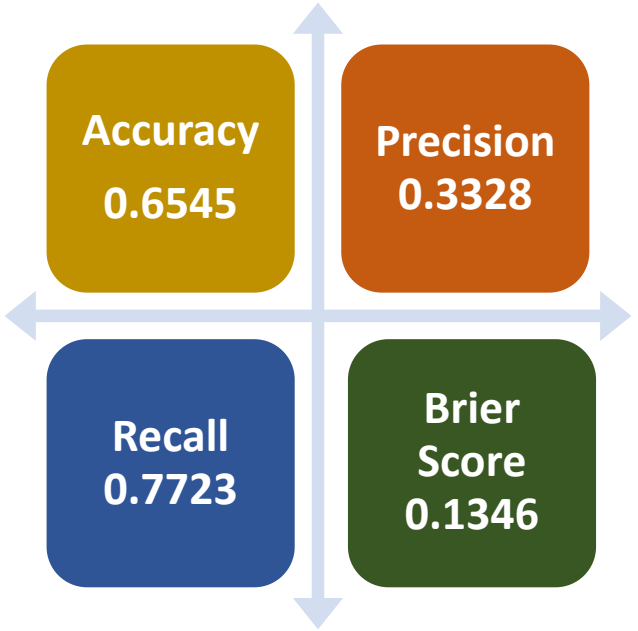
- **ROC Curve of SVM suggest better outcome compare to Random forest and XGBoost.**
- **All Models are using 0.5 default classification threshold.**

Comparison of Key Output Metrics

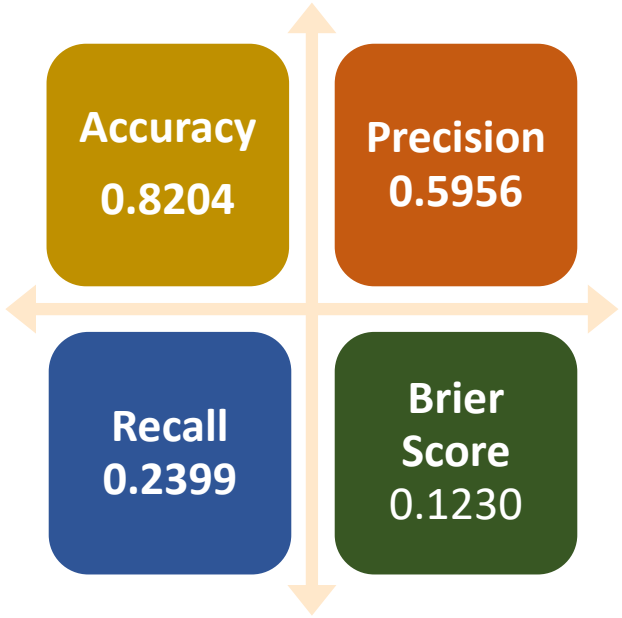
Metrics from Random Forest



Metrics From SVM



Metrics From XGBoost without SMOTE



- **Brier Score of Xgboost without SMOTE – 0.1230 is better model compared to other two models.**
- **Accuracy of Xgboost – 0.8204 is better compared to other two models.**
- **Precision of Randomforest is better compared to other two models.**

Future Improvements

- Important but missing lab results for detecting diabetic condition can be added. E.g- Glucose or A1C, Skin thickness measurements
- Variable for higher Classification of NDC codes such as therapeutic group can be added.
- Better imputation of Missing value can improve model results.