Diagnosis of Diabetes

Dream Team

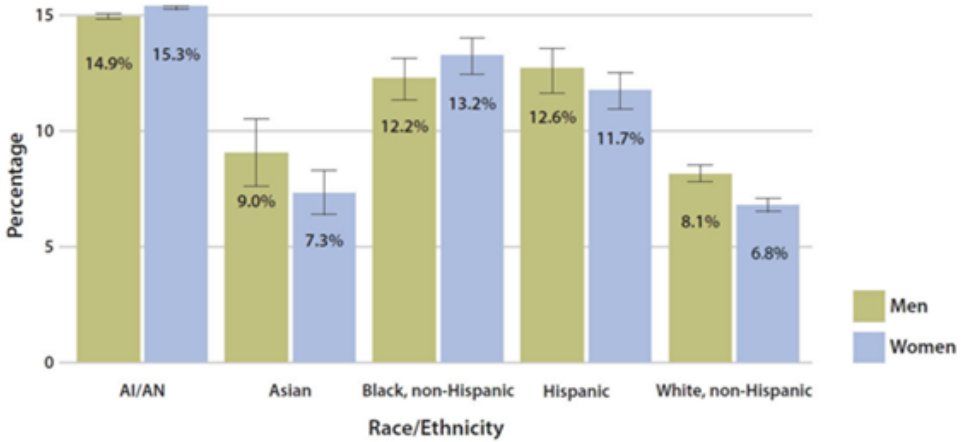
Team: Dream Team

Member: Ran Dou, Muhammad Furqan Shaikh, Tianyi Zhou, Mduduzi Langwenya, Siyan Lin, Qimo Li

CONTENT

I. INTRODUCTION: DIABETES IN THE U.S.

- More than 100 million U.S. adults are diabetic or are pre-diabetic
- 25% of adults living with diabetes are not aware of their condition
- Onset of diabetes increases with age
    - Adults aged 18-44: **4%**
    - Adults aged 54-64%: **17%**
    - 65 and older: **25%**
- Native Americans have a higher predilection for Diabetes



Source: cdc.gov

Project Motivation

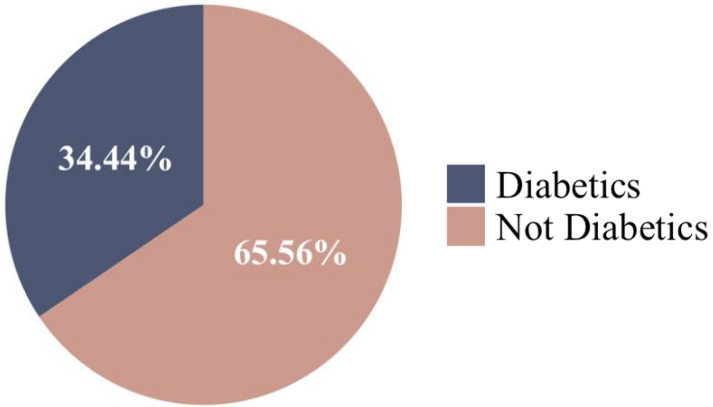
- Medical analysis is a new trend
- Can we predict Diseases like Diabetes?
- identify at-risk individuals and intervene
- Reduce unnecessary tests and save money
- **Caveat:** limited applicability of model to Pima Indian Dataset

Credit: www.dreamstime.com

II. DATA PREPARATION

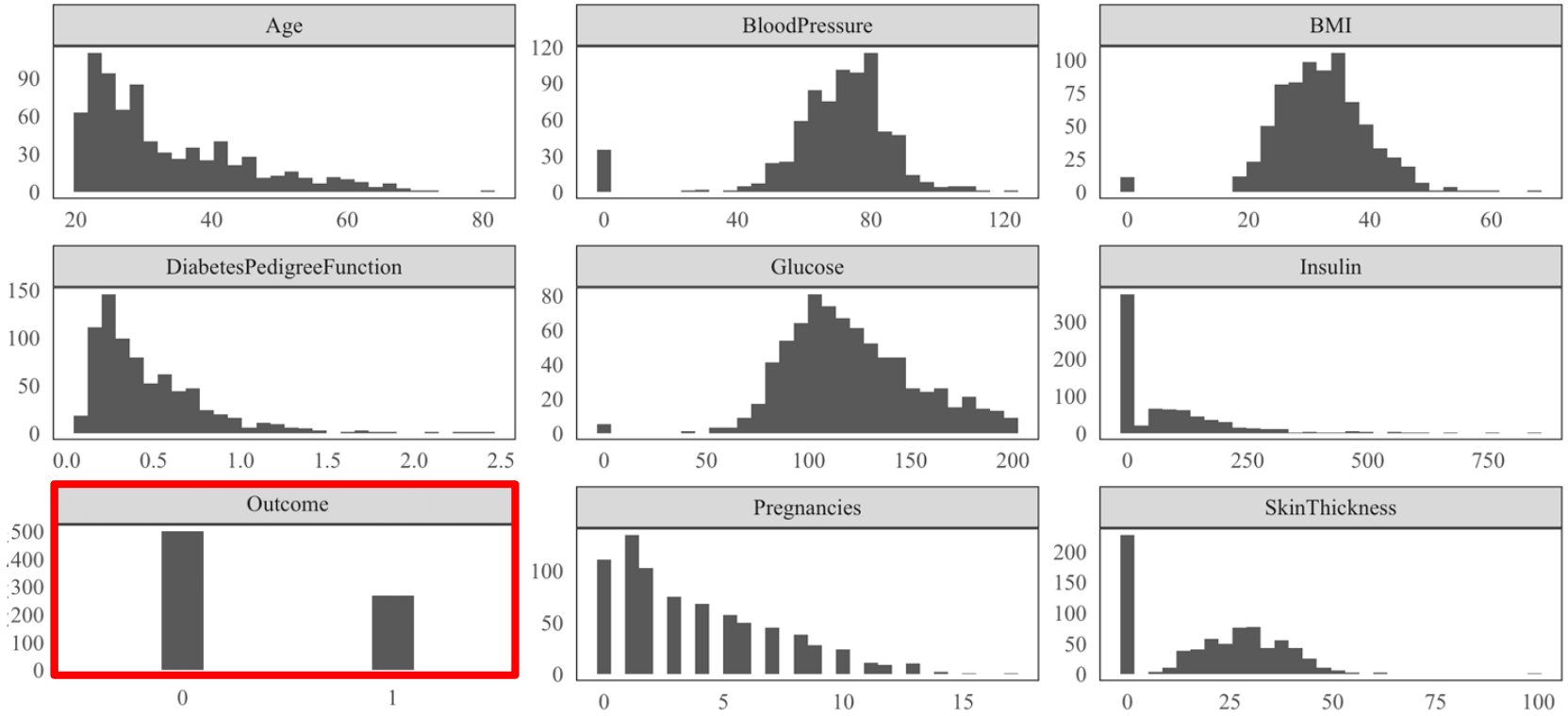
Data Source

National Institute of Diabetes and Digestive and Kidney Diseases



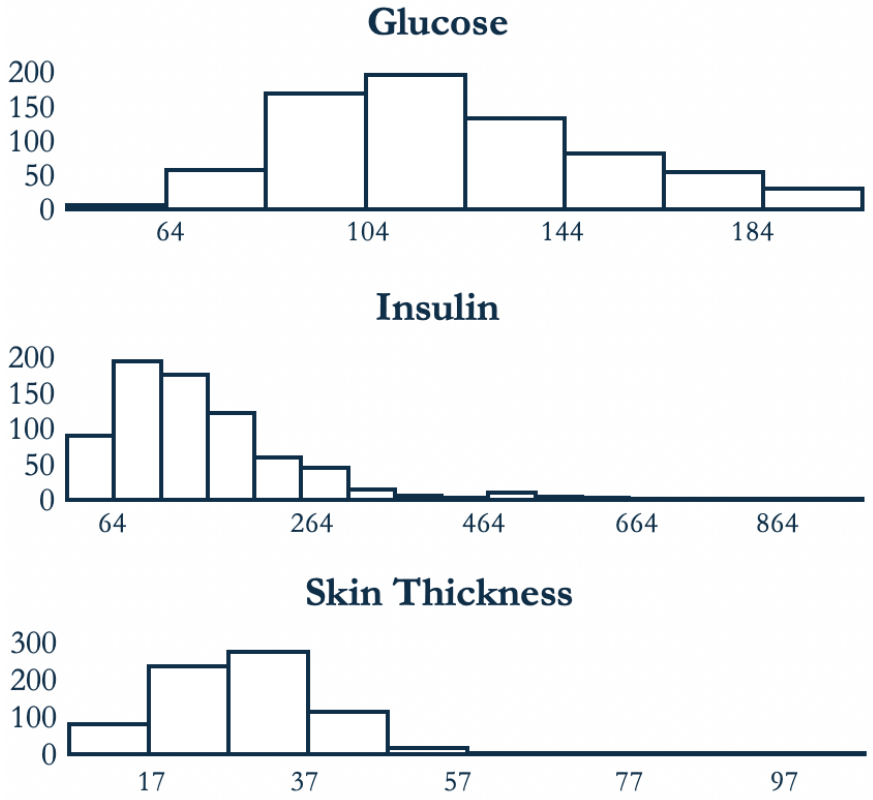
Definition of variables

1. Glucose
2. BMI
3. Diabetes Pedigree Function
4. Age
5. Insulin
6. Skin Thickness
7. Pregnancies
8. Blood Pressure

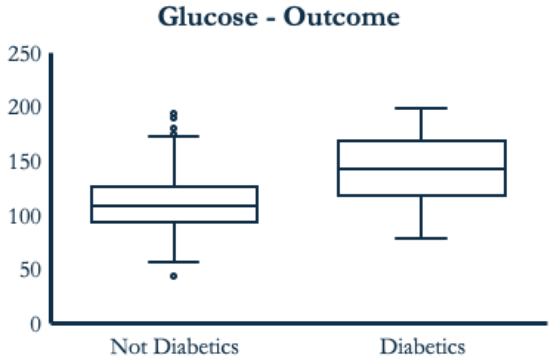


Features Overview

- Glucose, BMI, and Blood Pressure had missing data
- Insulin and Skin Thickness had many zero values
- Stepwise regressions for imputing missing values
  - Insulin ~ Glucose + BMI
  - Skin Thickness ~ BMI



Data Cleaning



↑ *Glucose – Outcome*

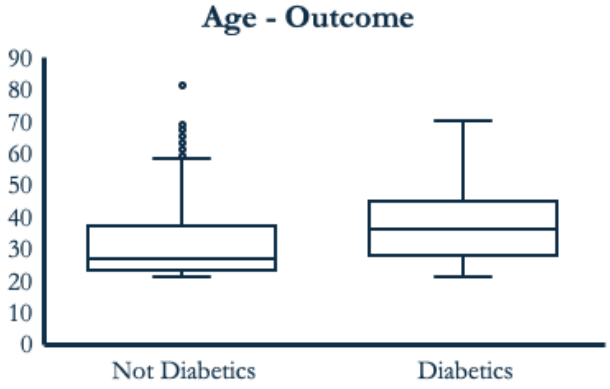
Means: *111.06 – 142.61*

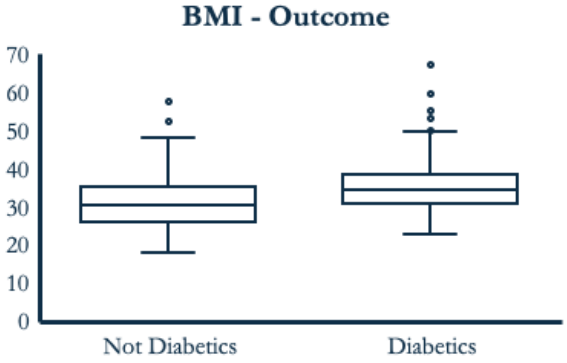
Higher glucose level

↓ *Age – Outcome*

Means: *31.26 – 37.34*

Seniors are at more risk





↑ *BMI – Outcome*

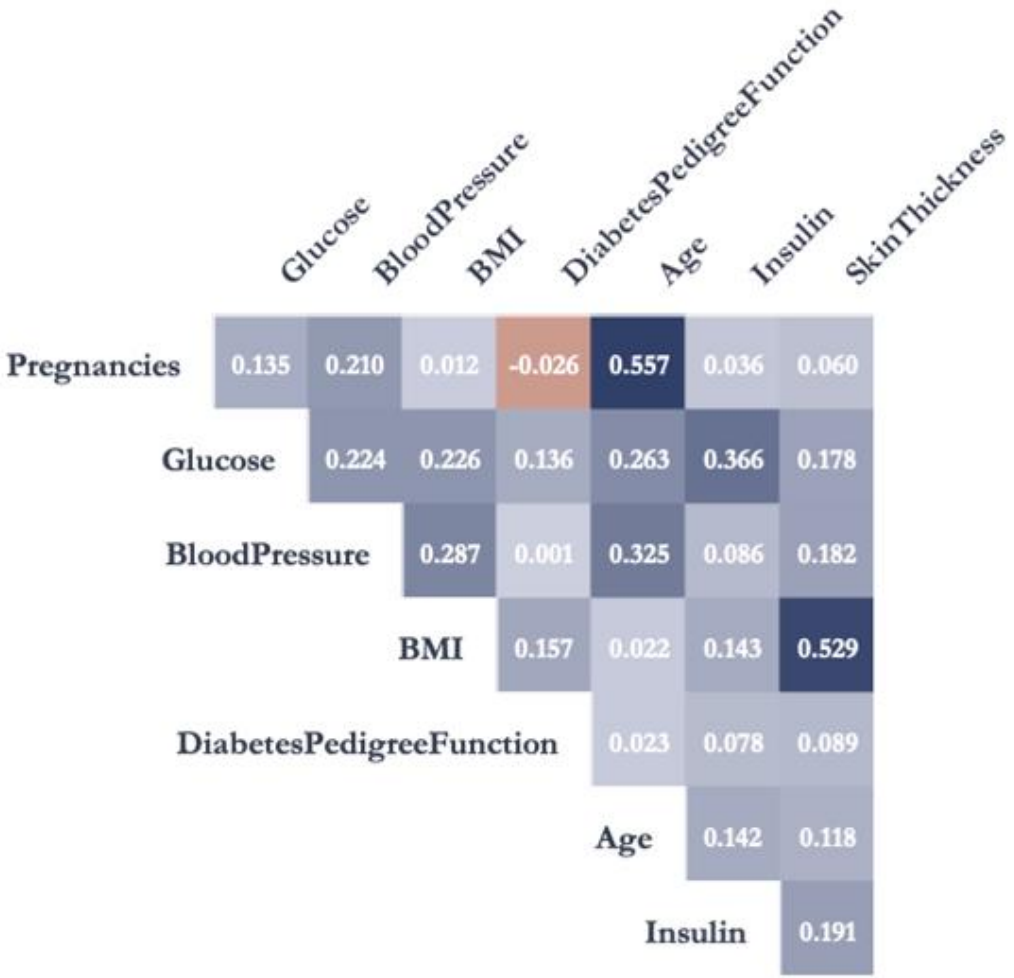
Means: *30.95 – 35.31*

Fat → Diabetes

III. DESCRIPTIVE STATISTICS & VISUALIZATION

Correlation Matrix

- Non-Multicollinearity
- Pregnancies - Age
- BMI - Skin Thickness
- Pregnancies - DPF
- Glucose - insulin
- Pregnancy's influence



IV. MODELS COMPARISON

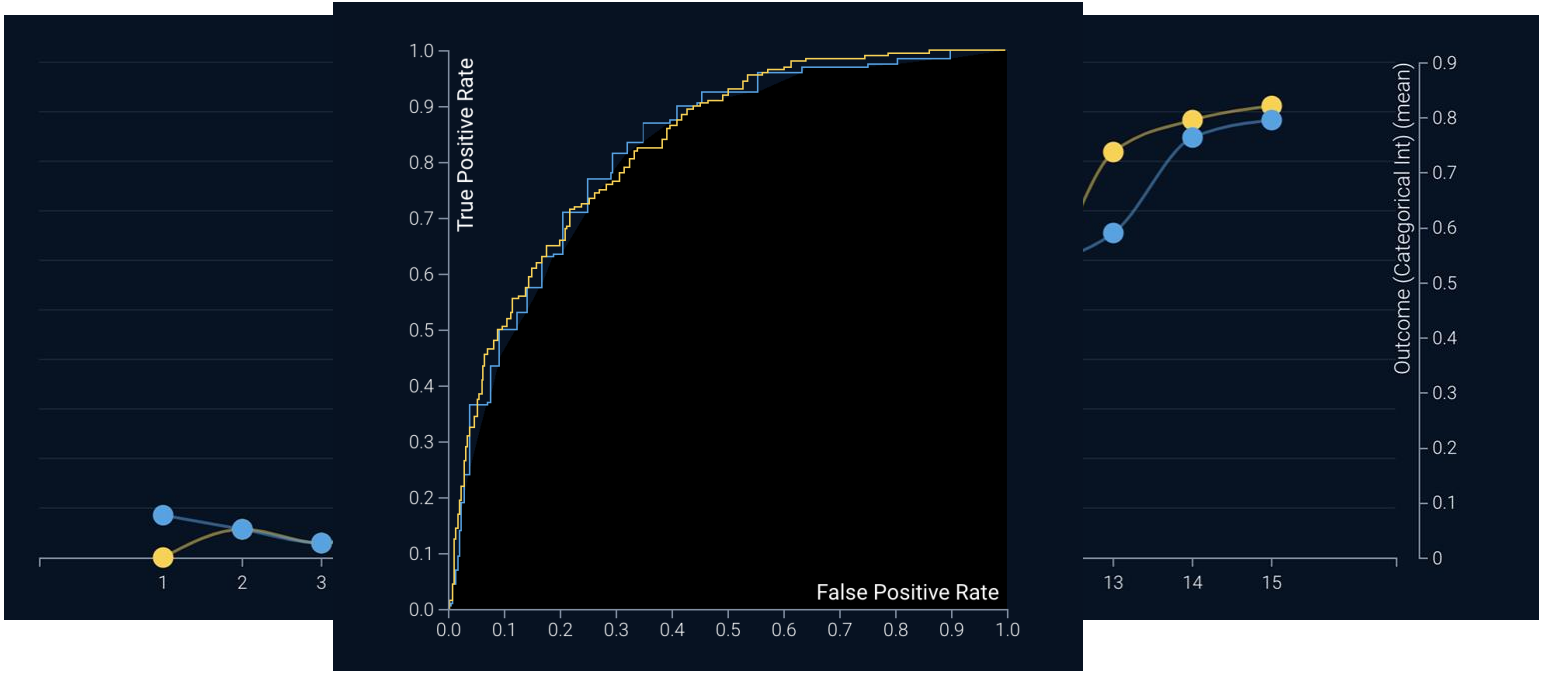
Method	Accuracy	Sensitivity	Specificity
Random forest	0.82	0.73	0.87
Auto-tuned K-Nearest Neighbors Classifier	0.79	0.81	0.78
Logistic Regression	0.79	0.78	0.80 (0.81)
Decision Tree Classifier	0.75	0.71	0.77
Gradient Boosted Trees Classifier	0.66	0.97	0.49
PLS Blender	0.80	0.77	0.82
LGBM Blender	0.84	0.77	0.87

- Blender models are a mixture of RF, k-NN and Logistic Regression models.
- Accuracy: **LGBM Blender**
- Interpretability **K-NN and Logistic regression**

IV. MODELS COMPARISON

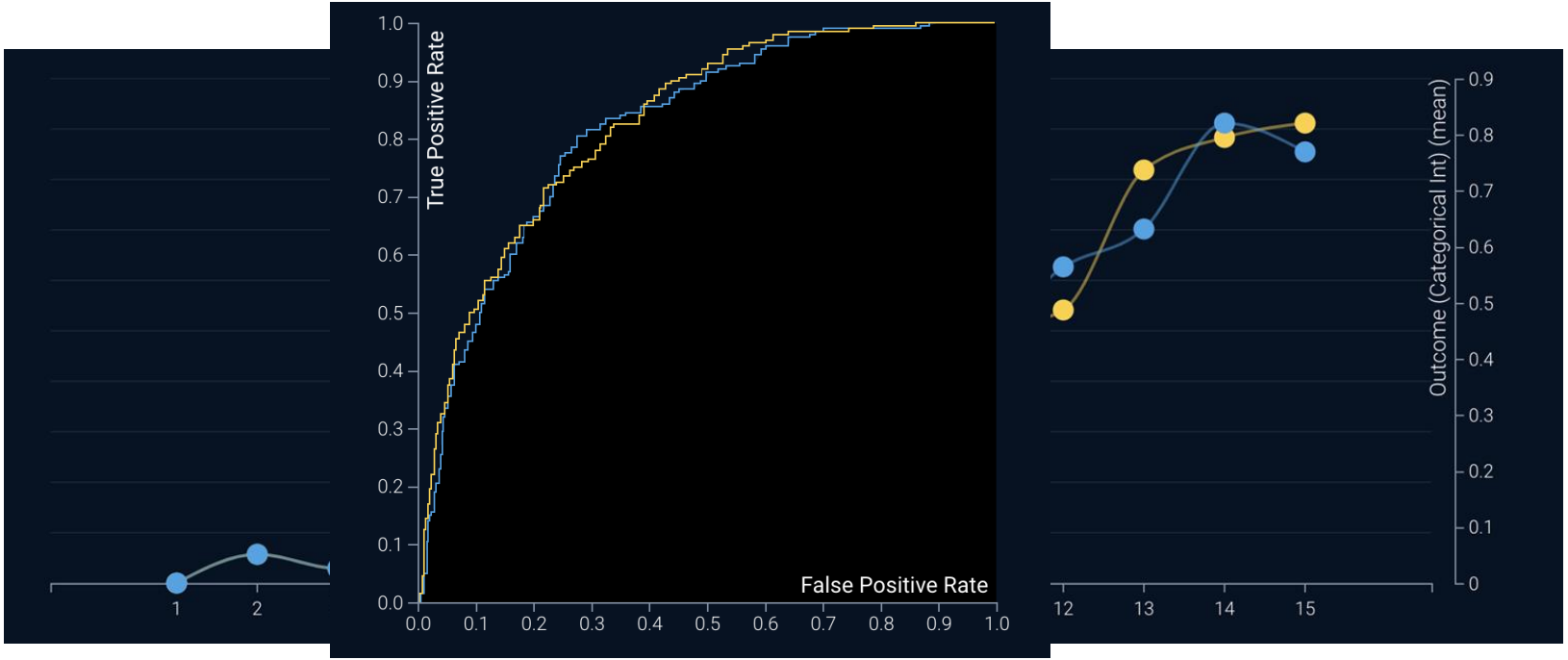
Logistic Regression and **k-NN**

ROC Curve



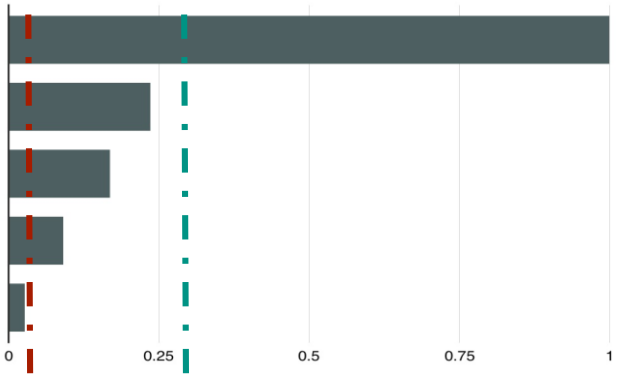
IV. MODELS COMPARISON

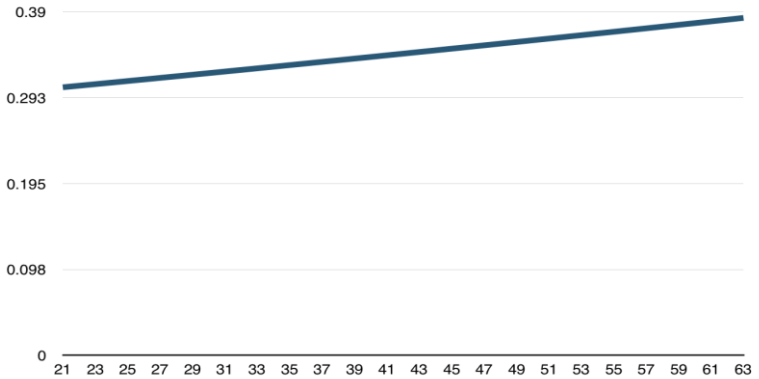
Logistic Regression and **LGBM Blender**



V. BEST MODEL

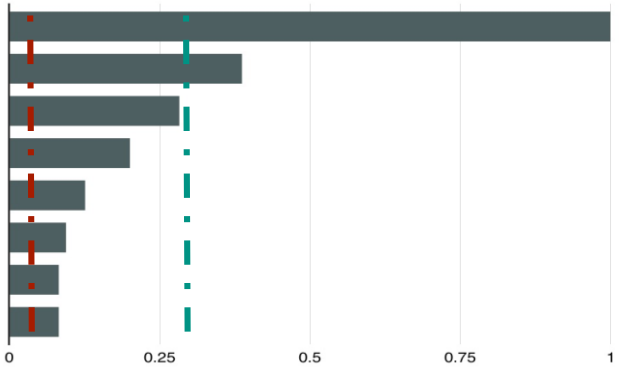
LR

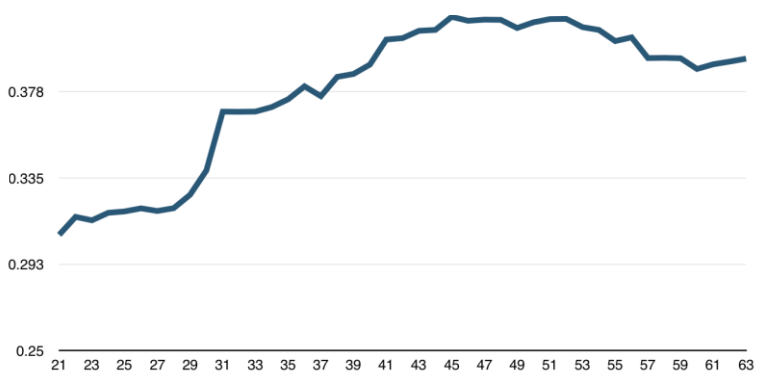




LGBM

Blender

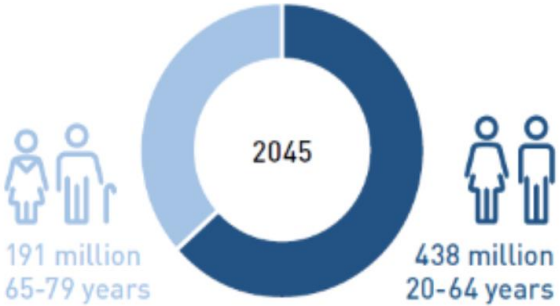


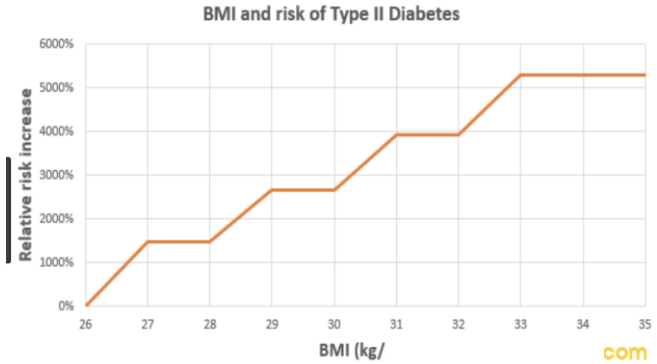


VI. BEST MODEL INTERPRETATION

- **Logistic regression**

Outcome ~ Glucose + BMI+ Pregnancies + Age + Diabetes Pedigree Function





Variable	Beta	Odds
Glucose	0.97	2.64
BMI	0.63	1.88
Pregnancies	0.39	1.48
Pedigree	0.34	1.4
Age	0.15	1.16

While glucose is dominant in diagnosis, combining it with Age, BMI and Pregnancies can reduce misdiagnosis

VII. CONCLUSION AND INSIGHTS

Data Improvement

- Gender, geographically and demographically diverse
- Lipid profile, smoking habits and time to diagnosis
- Other preexisting medical condition
- Information if diabetes in Type 1 or II

Model Improvement

- Using separate models to predict Type I and II
- Use of more indices to reduce dimensionality and mRMR

Q & A

Thank You!

Team: Dream Team

Member: Ran Dou, Muhammad Furqan Shaikh, Tianyi Zhou, Mduduzi Langwenya, Siyan Lin, Qimo Li