Diabetes Risk Prediction

Using BRFSS 2024 Survey Data

Regis University
MSDS Practicum- 1
Pavan Kalyan Kamasaani

Research Question

Is it possible to screen diabetes with no blood tests and just survey questions

- There are approximately 37 million diabetics in. about 1 out of 5 are not even aware of it.
- Catching it early can stop serious problems like kidney damage, blindness, and nerve damage.
- Goal: I built a model that aims to catch a possible number of diabetics using only behavioral and demographic data from the survey.

Dataset Overview

BRFSS 2024 — the biggest health survey in the world, which is conducted by CDC anually.

457,670	14	14.89%	22
People surveyed	Original features	Have diabetes	Features after engineering

Demographic: Age, Sex, Race, Education, Income

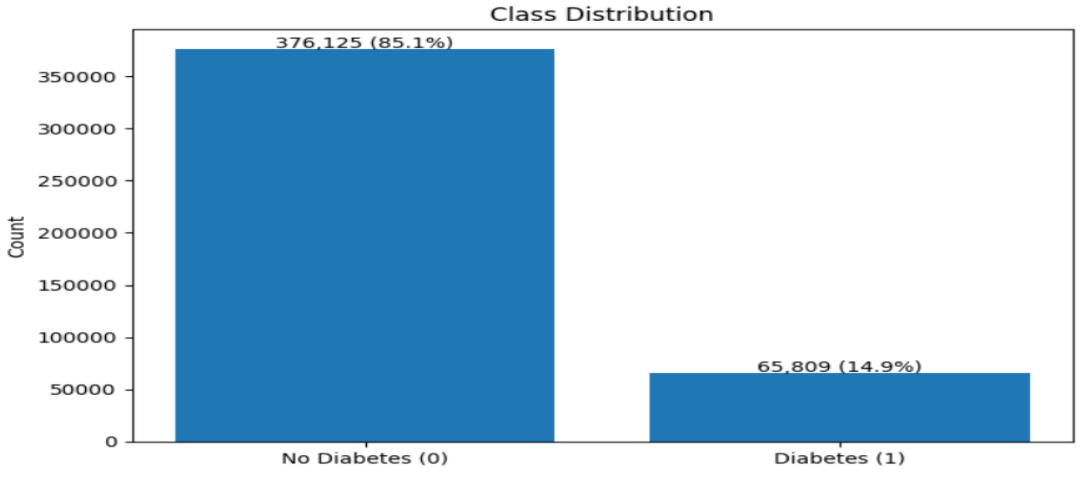
Health Behavior: Smoking, Exercise, Checkup, Medical Cost

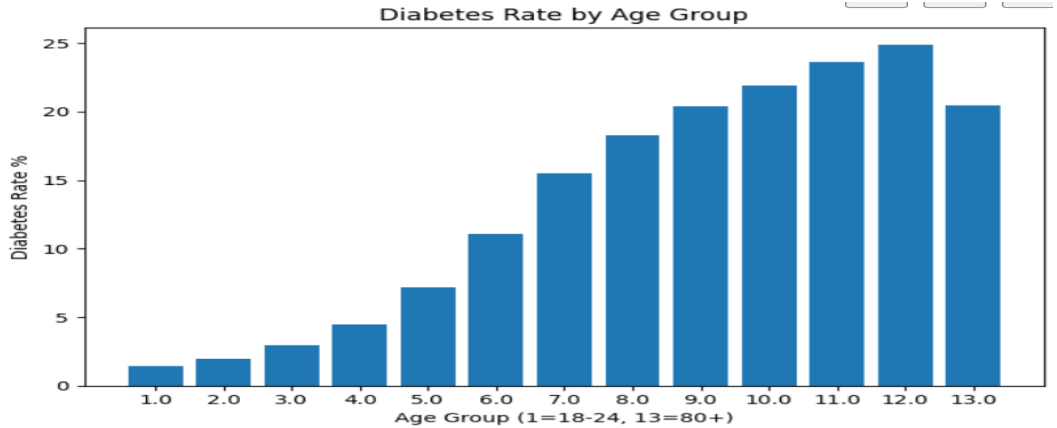
Clinical: BMI, General Health, Physical Health, Mental Health, Difficulty Walking

Engineered : age_bmi interaction, bmi_health combined, health_exercise_risk, and 5 more

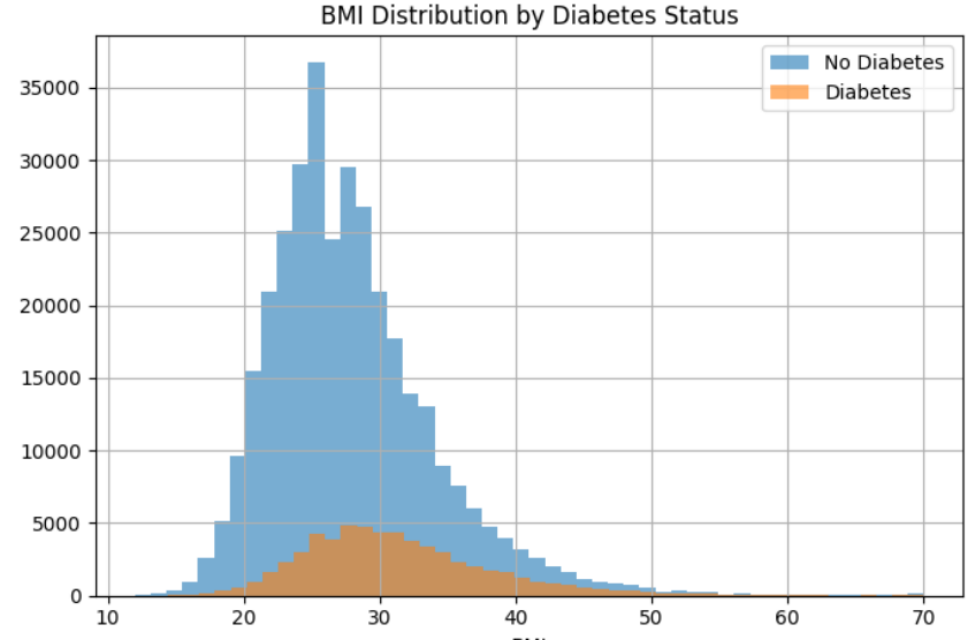
Exploratory Data Analysis

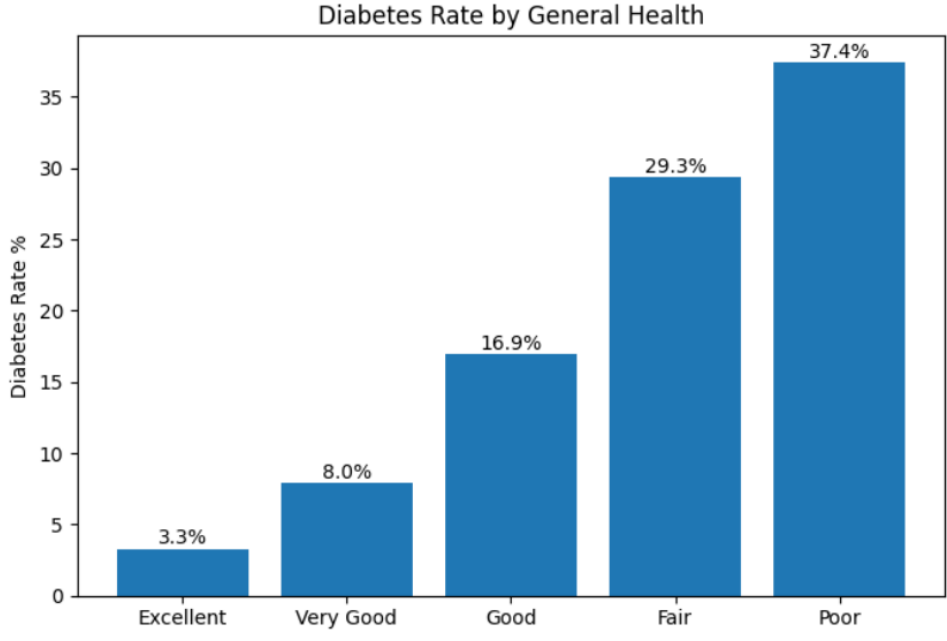
- 85% of people don't have diabetes, only 15% does. so accuracy is not a correct metric here
- Diabetes rate is at 2% in the youngest group to 25% for older group
- These patterns told us age and BMI would be the most important features to focus on





Exploratory Data Analysis





- Diabetic people usually have higher BMI. but there's also a lot of overlap with healthy people
- The general health of 3.3% diabetes says "Excellent" 37.4% says "Poor". This is around 10x difference
- BMI alone isn't enough. so we combined it with age and health status, which takes to feature engineering

Feature Engineering

created new features that will help the model find patterns it couldn't see with the original data itself.

#1

bmi_health_combined

BMI times general health rating gives high BMI + poor health as high risk

#2

age_bmi_interaction

Age times BMI says a BMI of 32 at age 25 is very different when compared to age 65

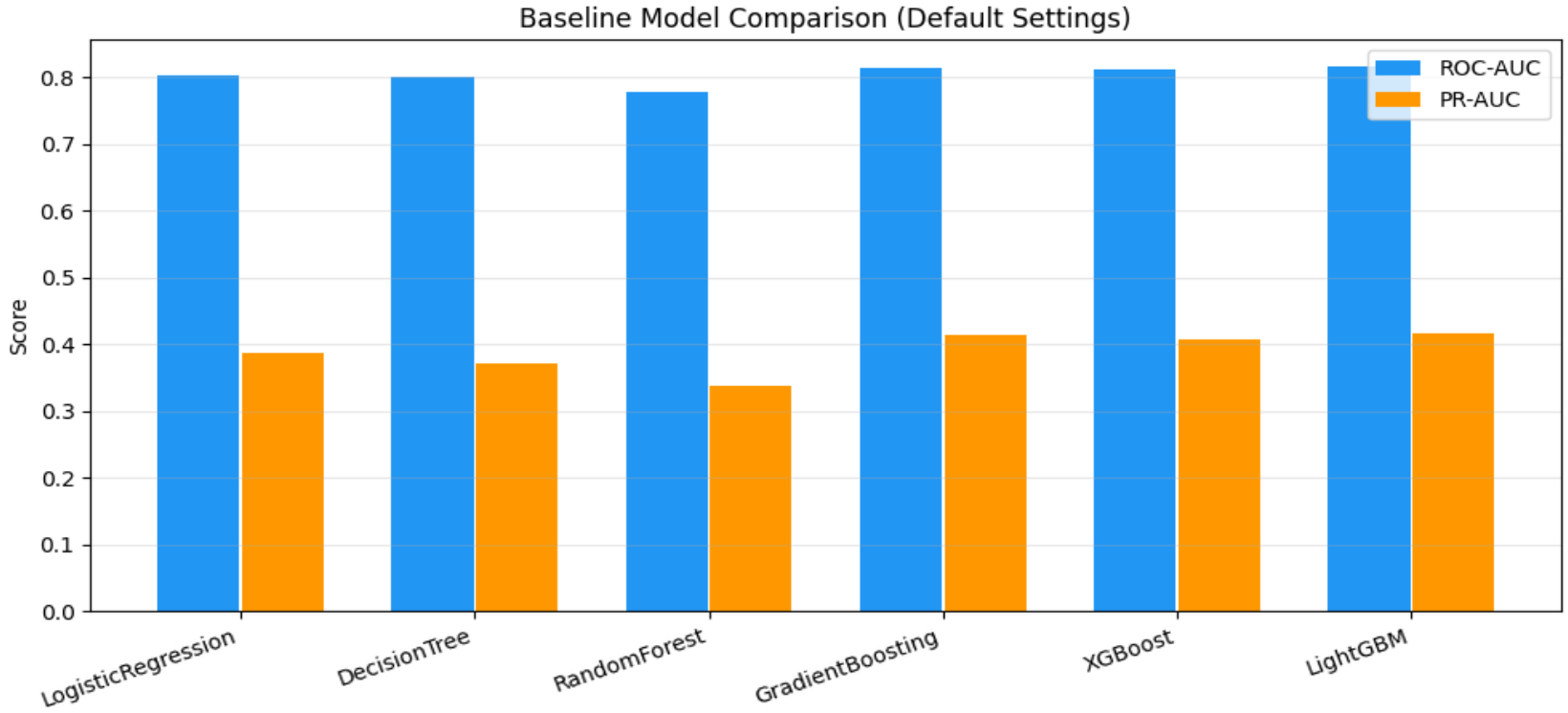
#3

health_exercise_risk

Health rating times exercise — poor health AND no exercise compounds risk

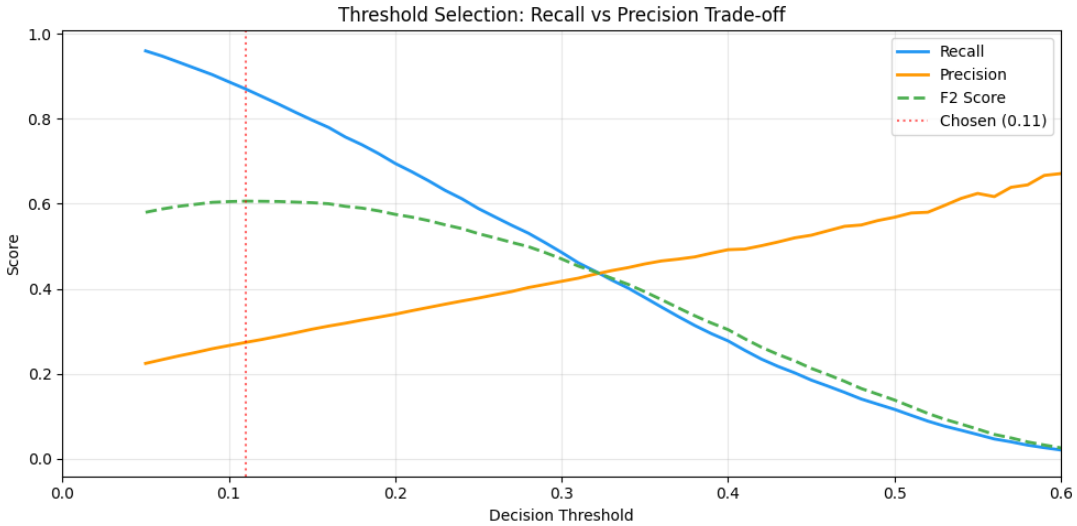
Model Comparison

tested 6 models at default settings. 3 gradient boosting methods came out on top.



Threshold Optimization

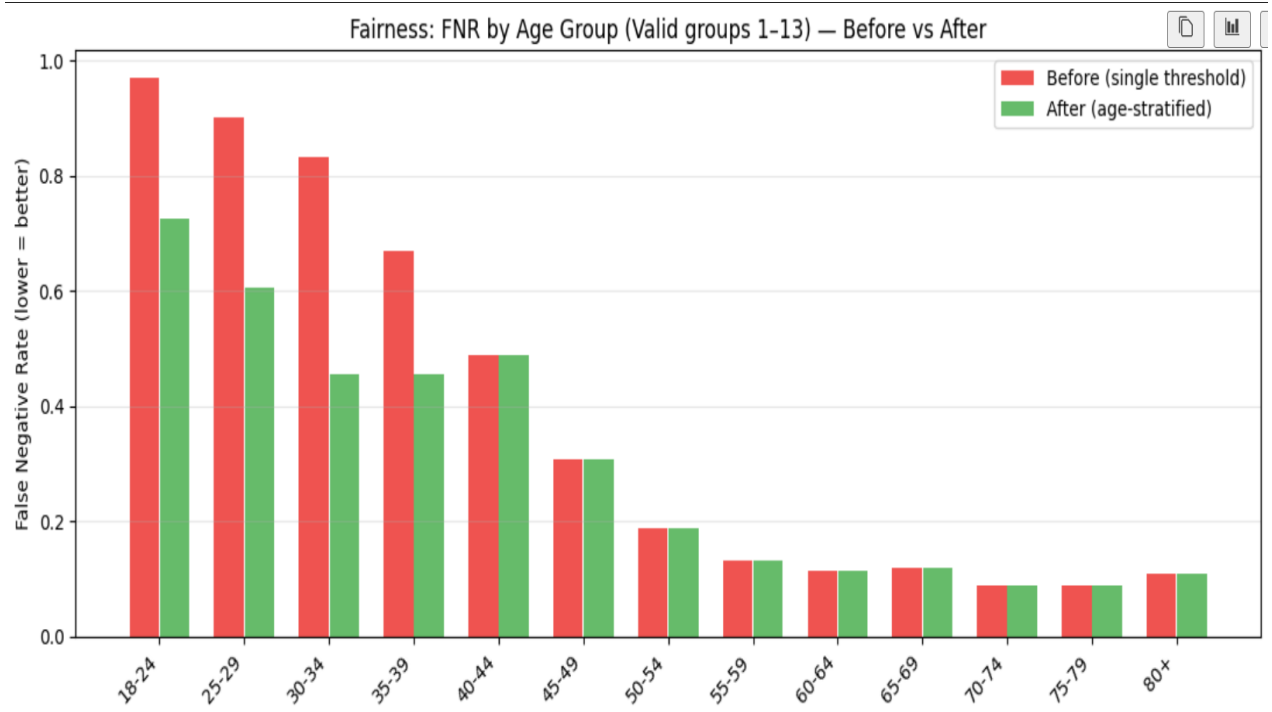
- At the default of 0.50 cutoff, the model missed 88% of diabetics.
- so I lowered it to 0.13 using the F2 score, which cares more about catching sick people rather than avoiding the false alarms.



Fairness — Age-Stratified Thresholds

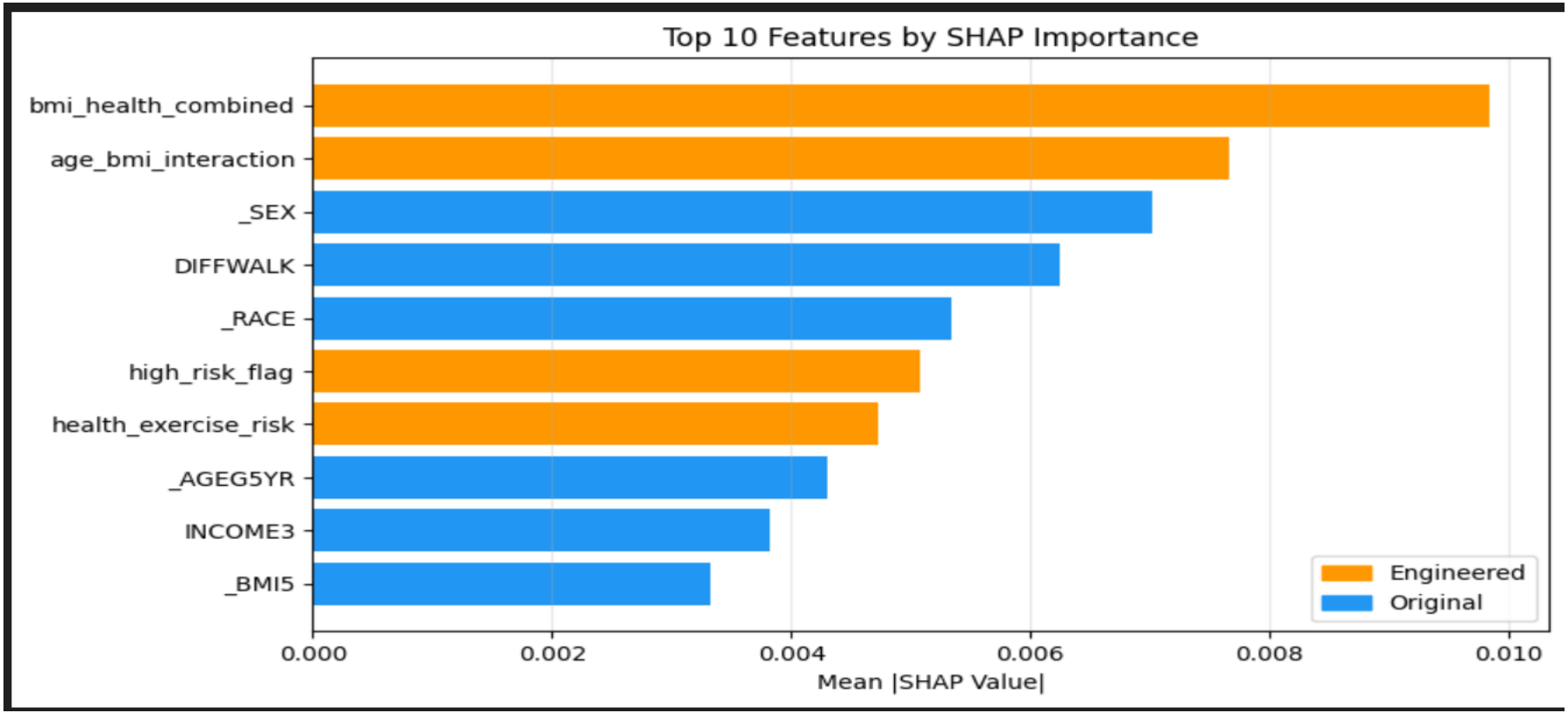
Young adults were being missed at higher rates. We used lower thresholds for younger groups to fix this.

- Young adults were being missed at higher rates.
- We used lower thresholds for younger groups to fix this.



SHAP — What the Model Learned

SHAP tells which features matter most. The orange bars are features I engineered .



Challenges & Solutions

Challenge	How We Solved It
Class imbalance — 85% healthy, 15% diabetic	Used PR-AUC and F2 score instead of accuracy
Default 0.50 threshold missed 88% of diabetics	Checked thresholds, found 0.13 maximizes F2 so recall went from 12% to 84%
Young adults missed at higher rates	Age-stratified thresholds — lower cutoffs for younger groups
Significant amount of income data was missing	Created income missing flag before imputation
Stacking ensemble didn't improve results	kept standalone XGBoost

Final Results

Test set — 66,291 people the model never saw during training

0.844
RECALL

0.406
PR-AUC

0.812
ROC-AUC

0.606
F2 SCORE

Metric	Logistic Regression at 0.50	XGBoost 0.50	Tuned XGBoost at 0.13
ROC-AUC	0.800	0.808	0.812
PR-AUC	0.377	0.396	0.406
Recall	0.120	0.125	0.836
Precision	0.493	0.525	0.288

Limitations & Future Work

Limitations

- The target is self-reported individuals indicate whether they have diabetes, and as such there is noise.
- It's a snapshot— we can only say who has diabetes now, not who will get it
- 3 of 4 flagged people are healthy — fine for screening (just means a blood test), not for diagnosis
- No lab values like blood glucose or A1C which limits how good we can get

Future Work

- Add lab values if available to boost precision
- Test on 2025 BRFSS data to check if results hold up
- Try separate models for each age group instead of just different thresholds
- Look deeper into fairness using other factors
- Build a simple web tool for real-time risk scoring

Key Takeaways

1. Feature engineering helped a lot— the features I built became the top 2 in the entire model
2. Threshold tuning played an important role than model selection for catching diabetics
3. Age stratified thresholds made the model fairer for young adults who were being missed

Thank You

Questions?