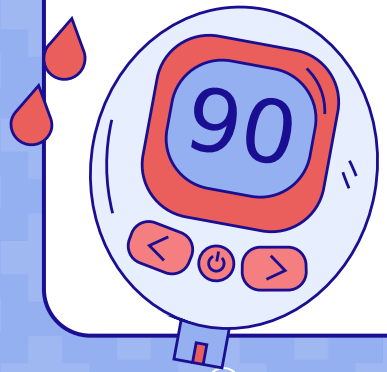
Diabetes Risk Identification And Prediction

A Predictive Model And Pre-screening Tool For Diabetes Risk

by Chisum Lindauer



TYPE 2 DIABETES *FACTS*

UNSEEN

22% of adults undiagnosed

RISK

10% - 40% will have kidney failure

IMPACT

Over 460 Million Have Diabetes

COST

$412 Billion Annually

COMMON

Type 2 makes up 95% of all cases

AVOIDABLE

Up to 58% of cases can be avoided.

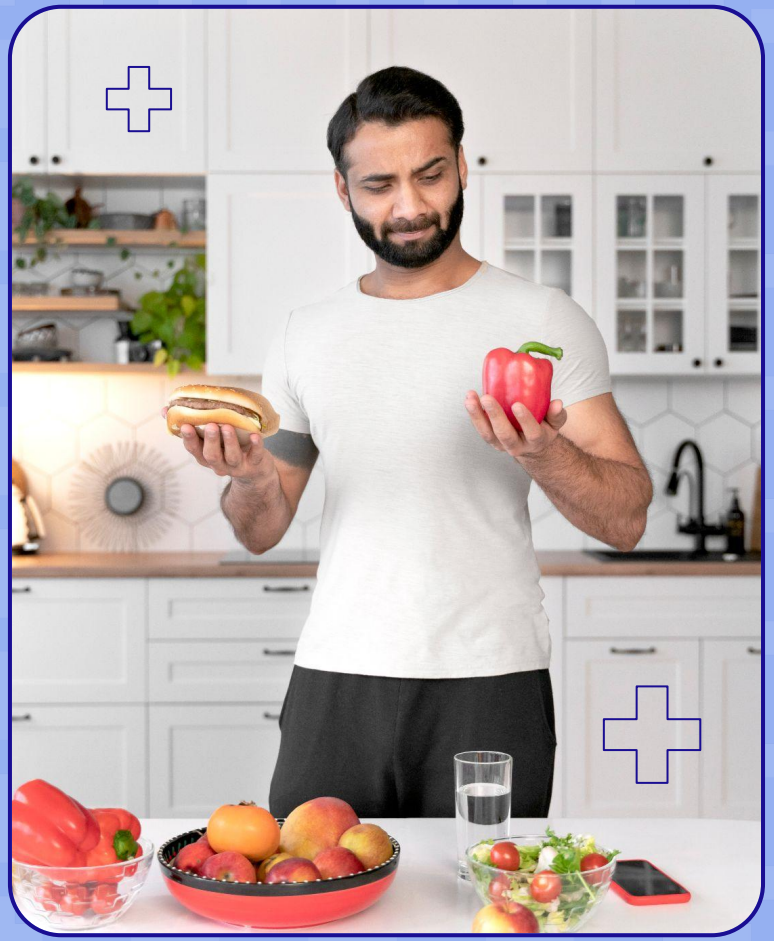
BUSINESS *UNDERSTANDING*

Stakeholders

+ Healthcare Providers
+ You

Problem Solved

Helping pre-screen people for Diabetes Risk



DATA *UNDERSTANDING*

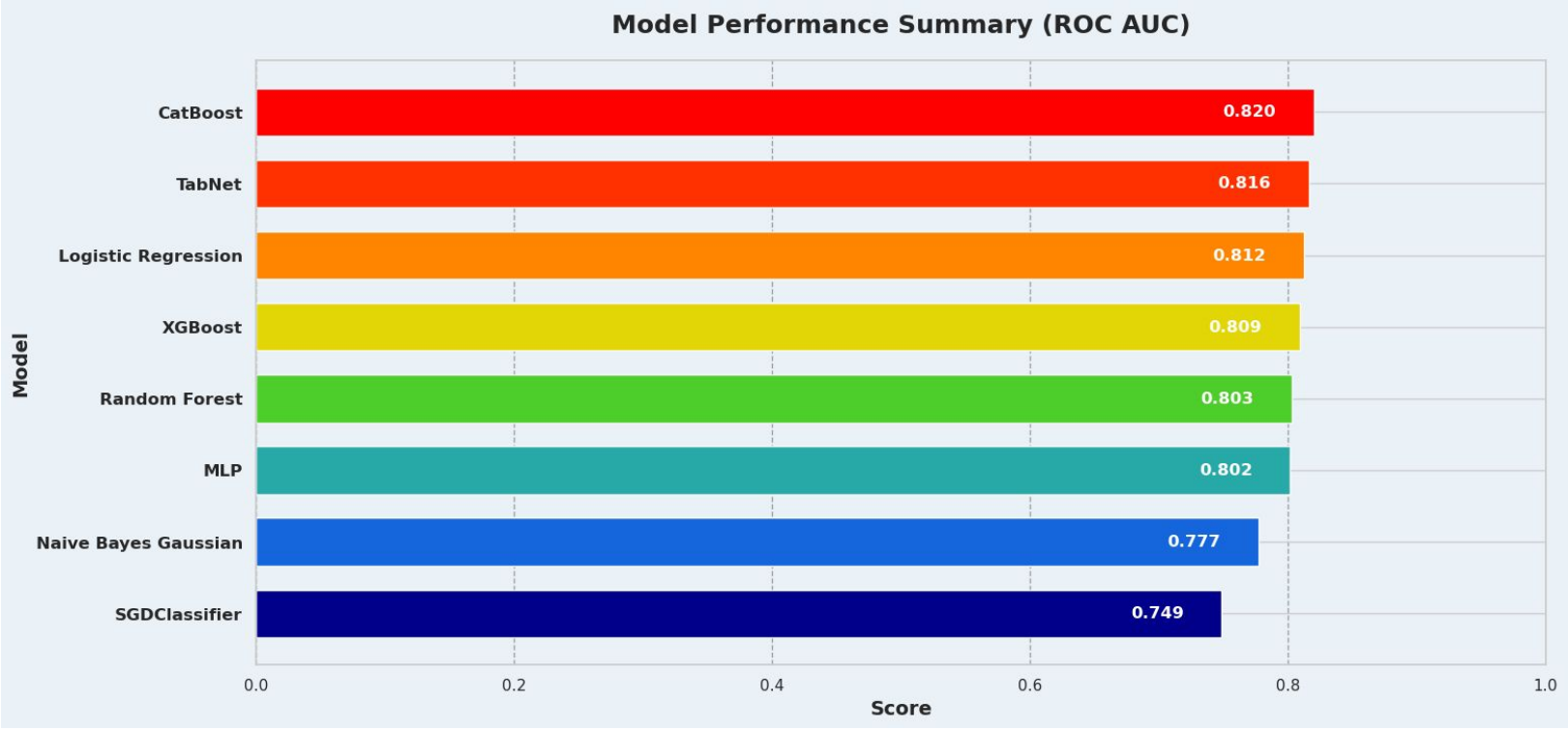
CDC Diabetes Health Indicators from the UCI Machine Learning Repository

Respondents	220,000
Health Indicators	23
Data Contains	Respondent answers to survey questions and if they have diabetes
Download Via	UCI Machine Learning Repository or ucimlrepo in python
Data Used To	Train models to predict diabetes based on health indicators
Robustness	A large dataset is more reliable and the CDC is a trusted source

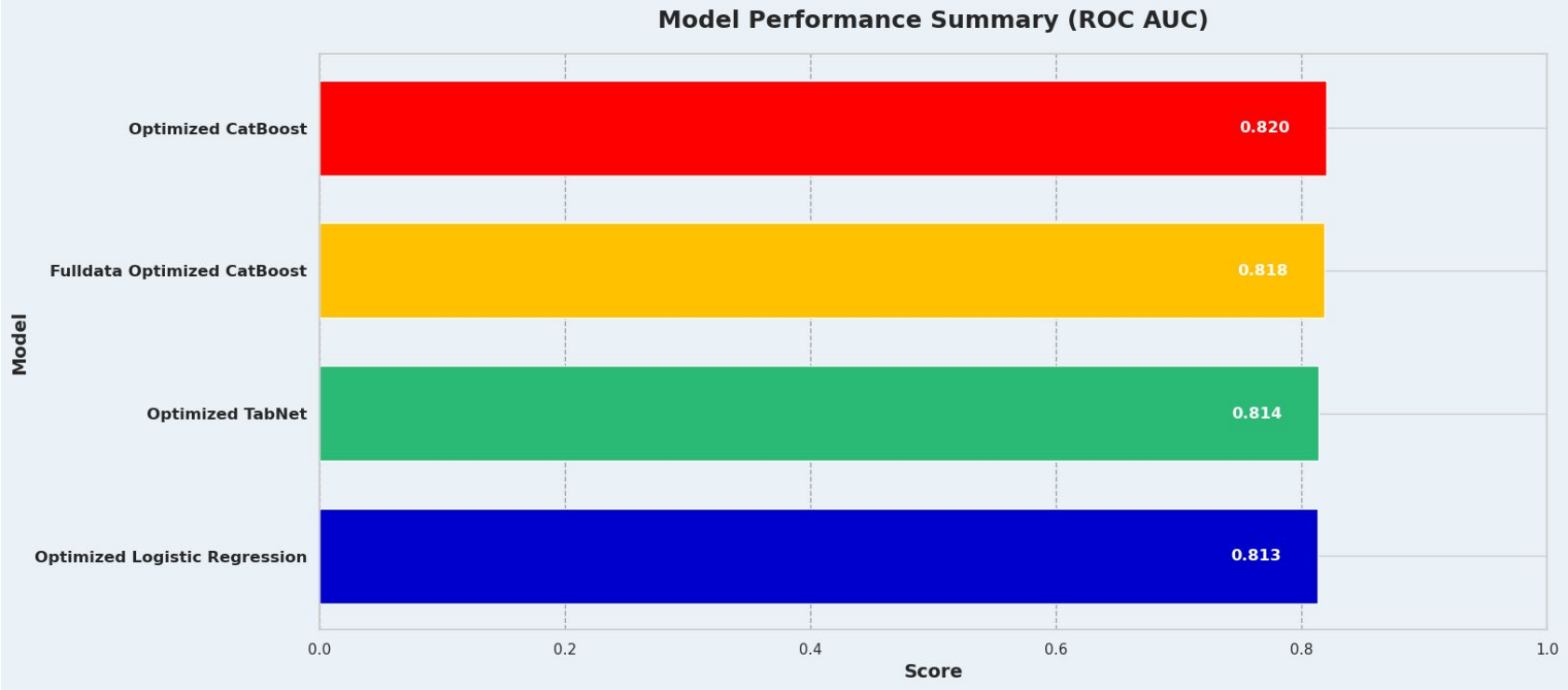
DATA PREPARATION

- Thoroughly prepared the data to ensure readiness for modeling
- Reduced memory usage by 83.6%
- Removed over 20,000 duplicates
- Evaluated features and created features based on custom reports
- Balanced the data
- Split data into training, testing, and holdout sets for reliable evaluation

BASELINE *MODELS*

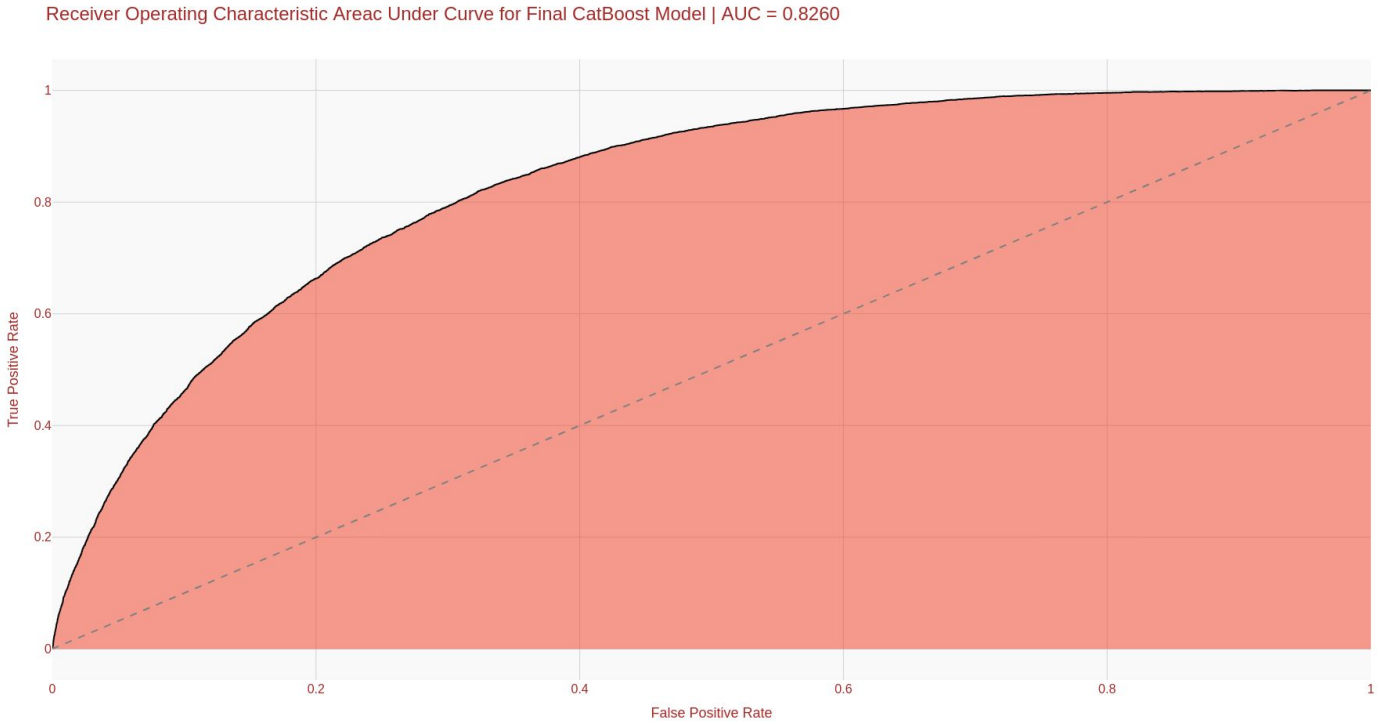


HYPERTUNED *MODELS*

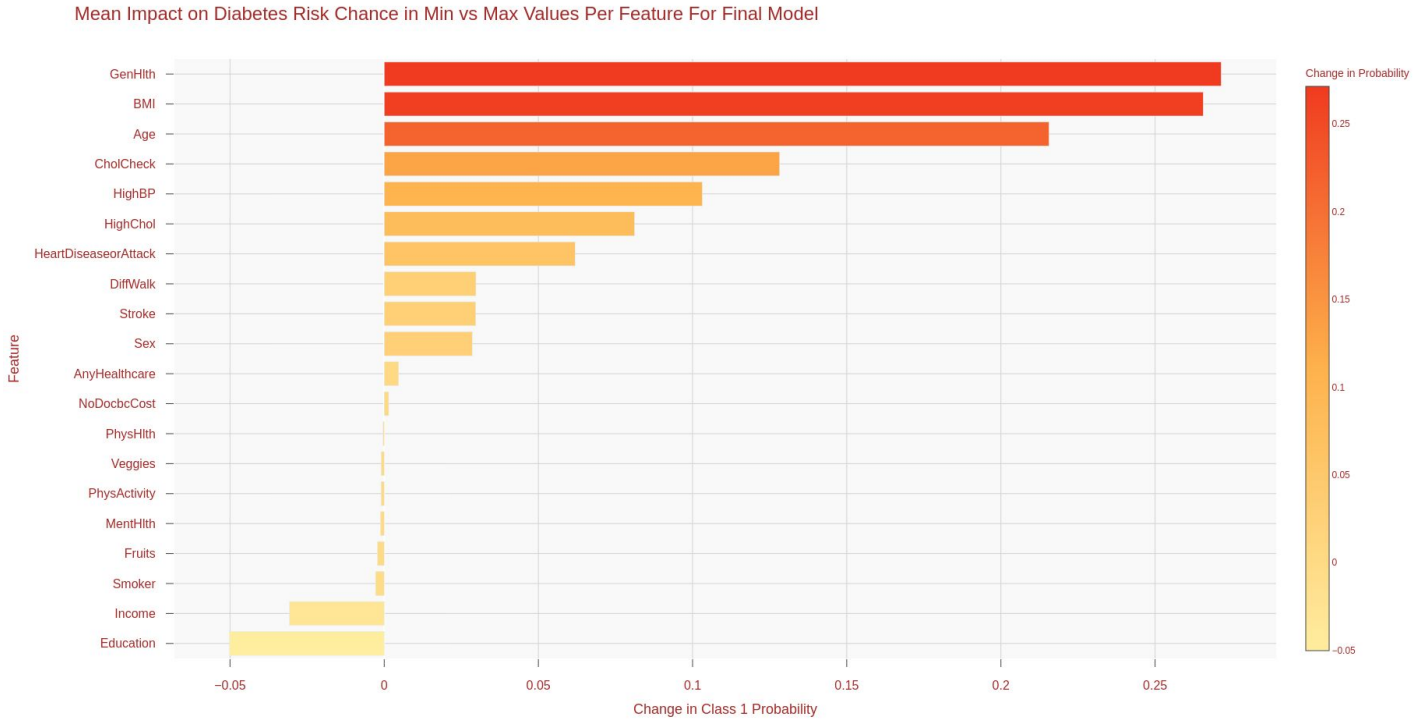


COMBINING *FINETUNING*

FINAL *MODEL*


Receiver Operating Characteristic Areac Under Curve for Final CatBoost Model | AUC = 0.8260
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1

FEATURE *IMPACT*



MODEL *DEPLOYMENT*

PRIVATE

On device and never sends or stores any data

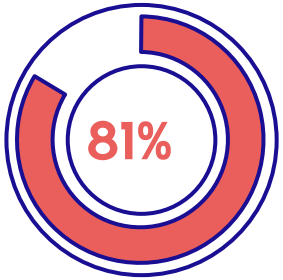
WEBPAGE

Works on any device and is mobile friendly

PREDICT

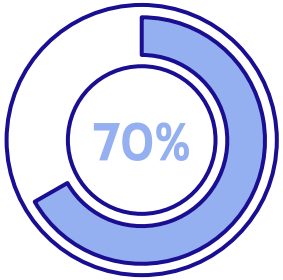
If risk is over 50% it recommends further screening

PROJECT *CONCLUSIONS*



IDENTIFIES

Identifies 81% of at risk cases on unseen data.



ACCURATE

70% of overall predictions are accurate

PROVIDERS

Prevent more diabetes cases

PATIENTS

Easy way to estimate and understand risks of diabetes

MODEL *IMPROVEMENTS*

INDICATORS

Explore additional health indicators

NOVEL

Explore new approaches with current data

TEST

Combine new indicators and approaches

MORE DATA

Find other datasets and repeat the process

NEXT *STEPS*

FEEDBACK

Create feedback system

LIVE TESTING

Gather feedback and test

EXPAND

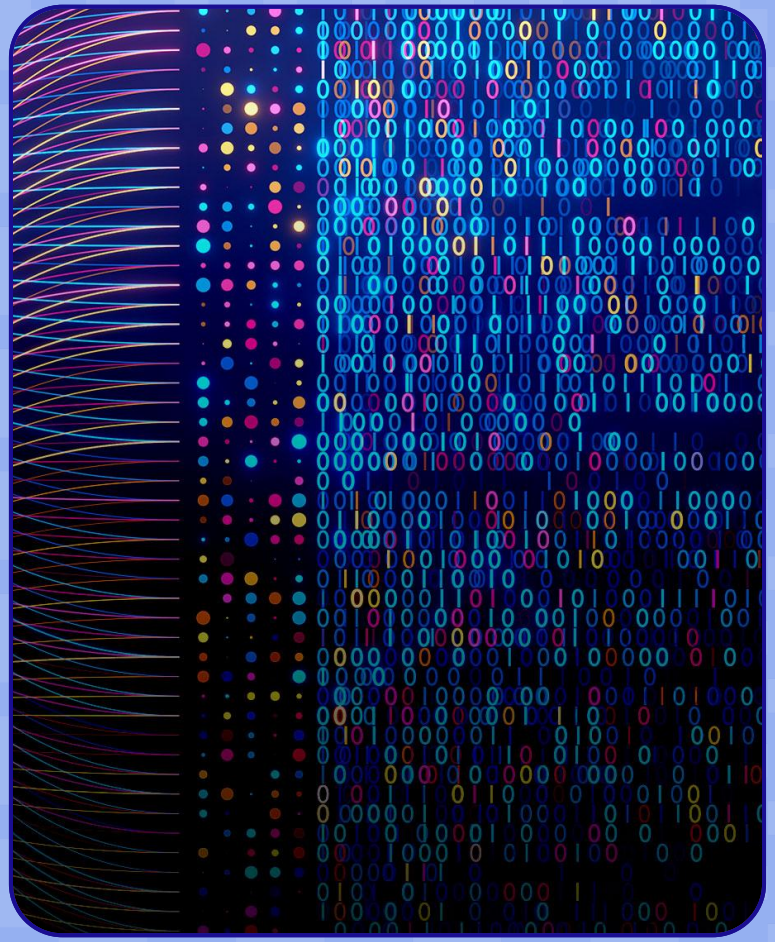
Find additional stakeholders and users

ITERATE

Iterate on the model and tool feedback

IMPACT

Identify those at risk for diabetes on a large scale



MANY THANKS!

Do you have any questions?

chisum@atrixtech.com
linkedin.com/in/chisum-lindauer/
github.com/ds-minerva