Team 14

Team Alpaca

Roni Fridman - Technion - Data and Science Engineering

Amit Meir - BGU - M.D.

Nati Dadon - JCT - Physics and Electro-optics Engineering

Ilya Vorotyntseb - Brauda - Software Engineering

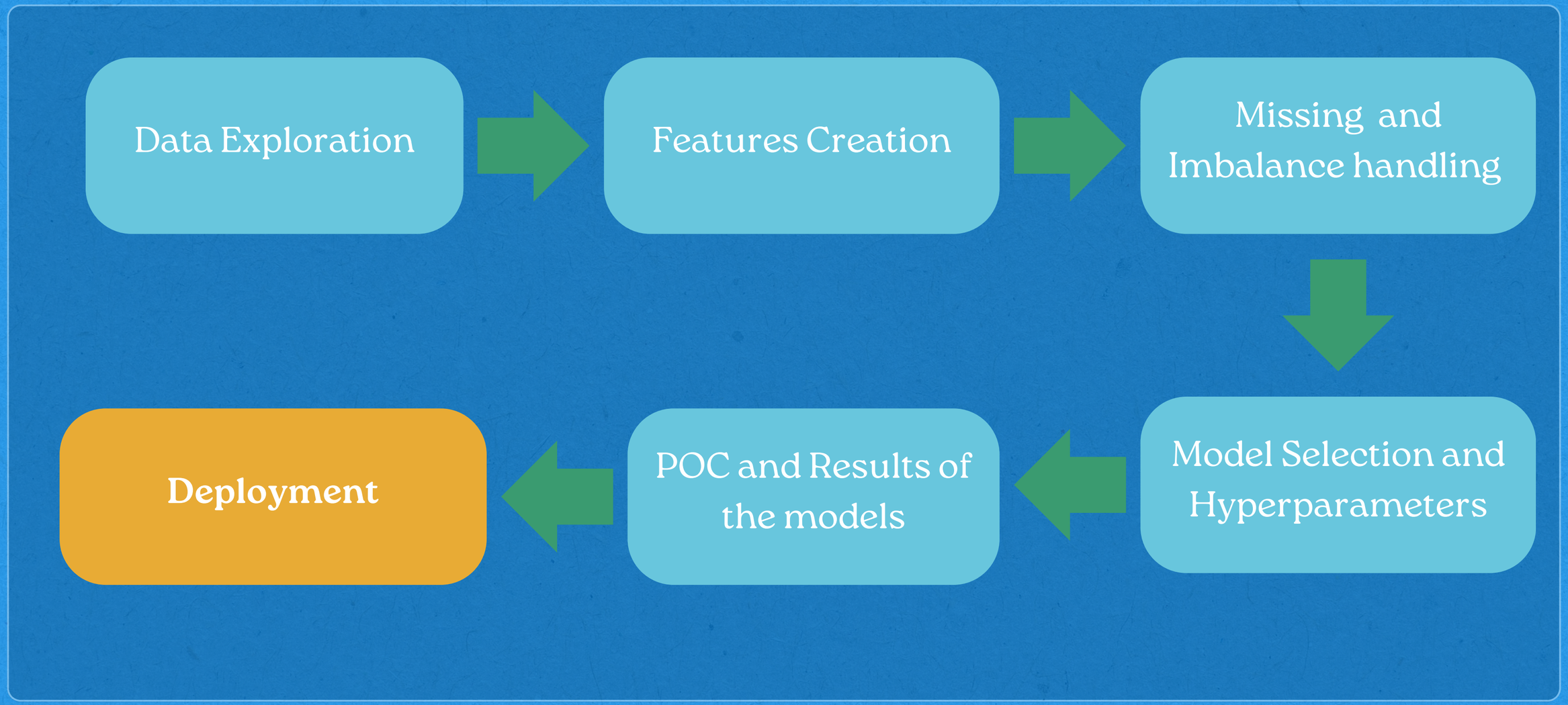
Evgeny Lambord - Technion - Computer Science

Our Data

Our model is based on the “Cardiovascular Study Dataset” Which tries to predict the 10-year risk for CHD complications.

Although sparse in both rows and column, we managed to provide great results!

Our Model


Data Exploration
Features Creation
Missing and Imbalance handling
Model Selection and Hyperparameters
POC and Results of the models
Deployment

Feature Creation

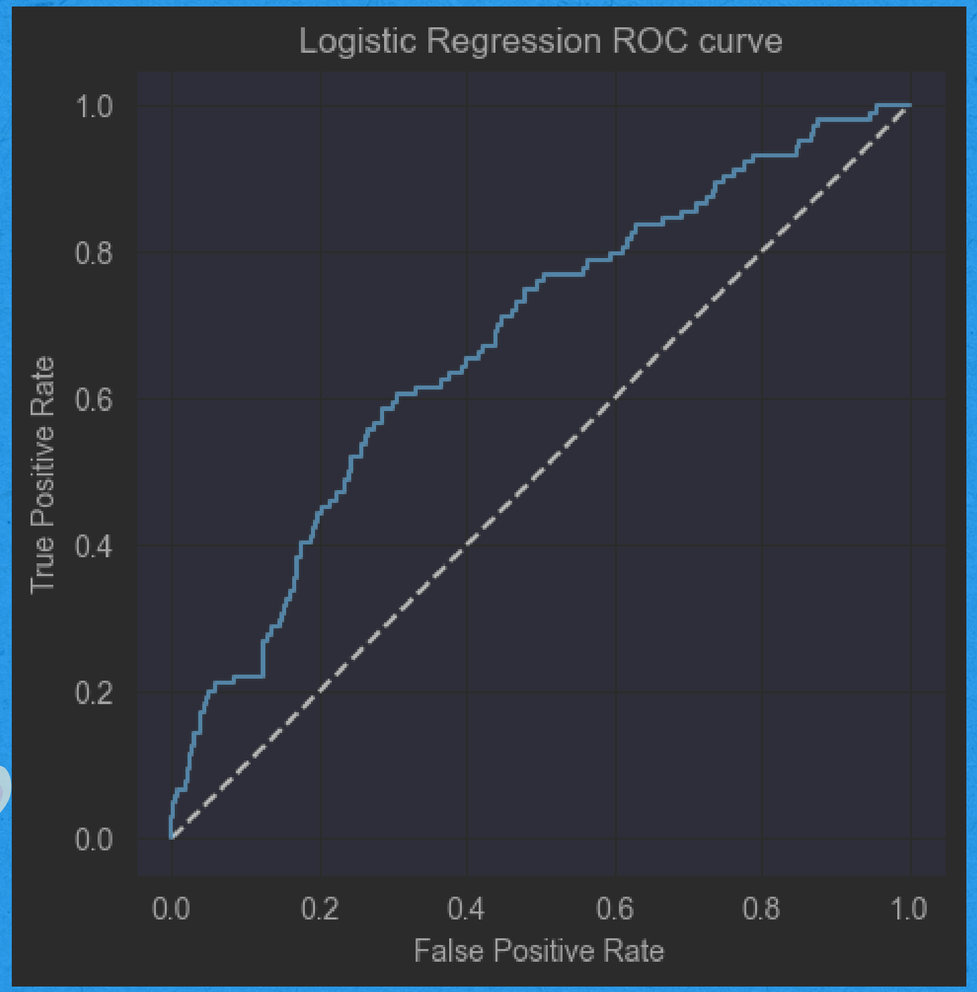
After the exploring the data, we created these columns:

- Packs of cigaretts based on the amount in the “cigsPerDay“ column
- is the user in a state of Hypertension basen on the systolic and diastolic BP.
- glucose_level: categorical cutoffs for the glucose column
- diabetes_score2 : a simplified version on the well known metric for diabetes detection.

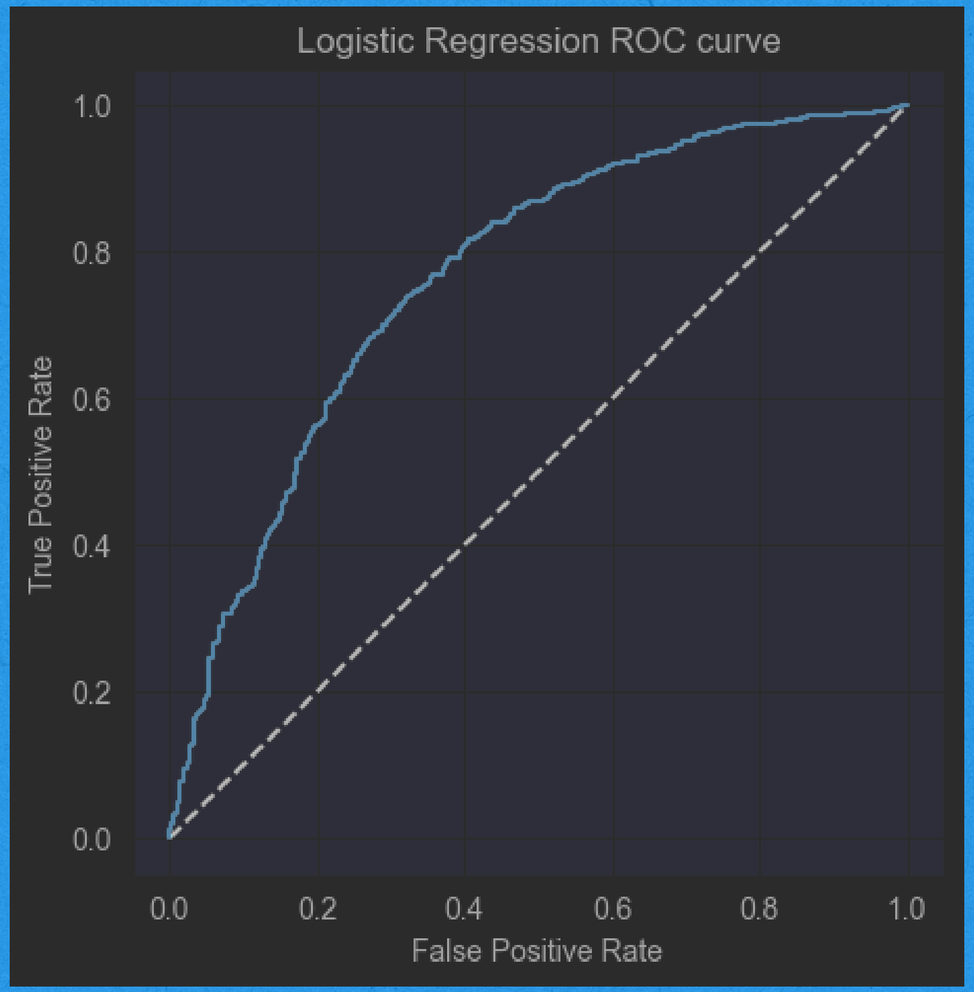
Handling Missings and Imbalance

- We have tried using a KNN imputator, but found out the bias in the data is too big, and the predictions become unrelayable. We decided the use the standart method: Median for continuous values, and Mode for categorical values.
- The binary target varible contained only 25% points from the ‘1’ category. Thus, we oversampled from the category using the SMOTE algorithm, which created a much more stable data.

Before oversampling



After oversampling



POC and Results of the models

The models we chose to use are both the Random Forest and Gradient Boosting Tree models.
Tree based models are a good fit in the world of medicine, since they provide great explainability between the models engineers and the doctors using them.

We removed the "fake" data we got from oversampling, and checked the accuracy, AUC score and ROC curve of these models, and these are the results:

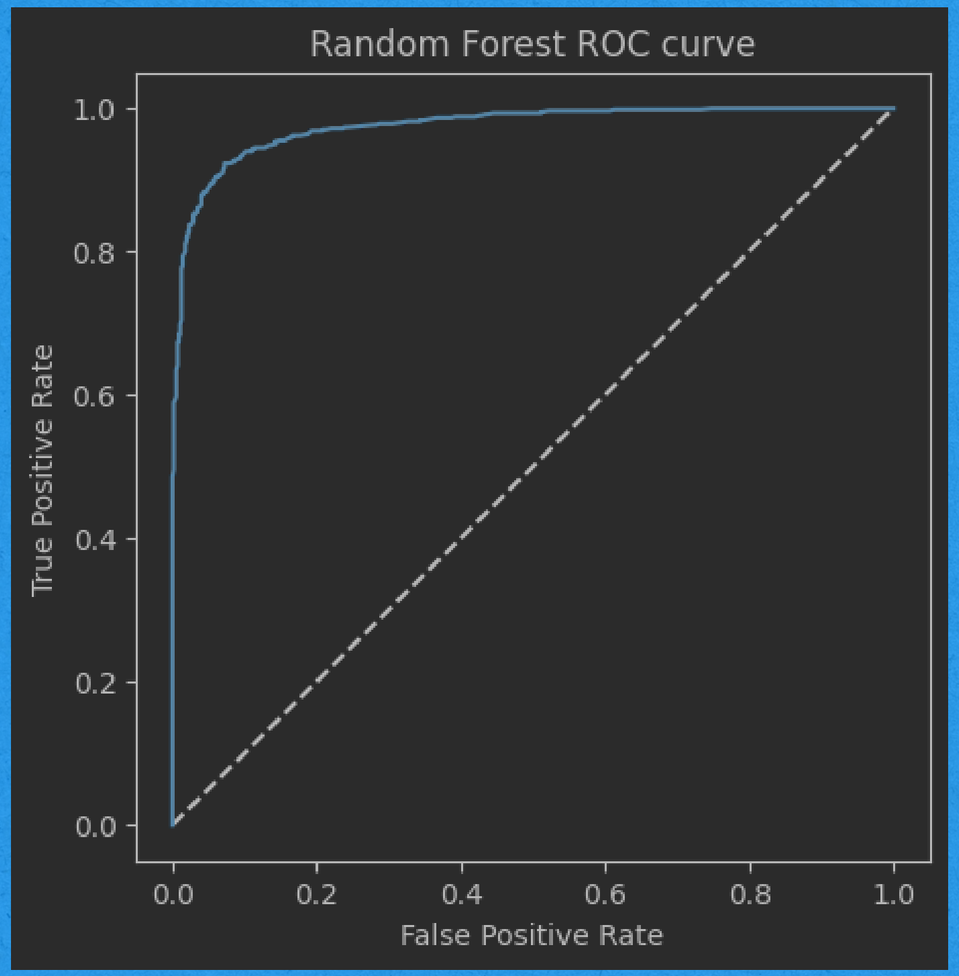
All data (Including SMOTE):

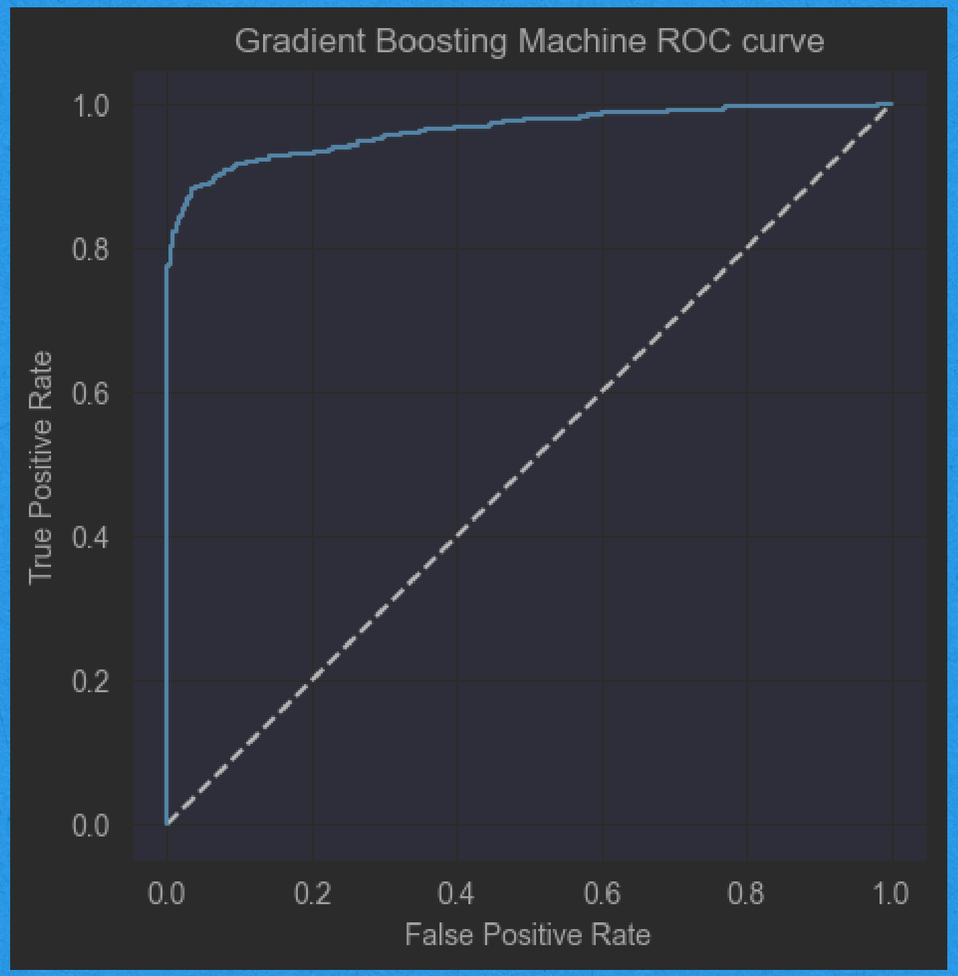
- Random Forest: accuracy on test = 0.9175 AUC = 0.9721
- Gradient Boosting Machine: accuracy on test = 0.9157 AUC = 0.9696

Clean data (Excluding SMOTE):

- Random Forest: **accuracy on test = 0.886 AUC = 0.8985**
- Gradient Boosting Machine: **accuracy on test = 0.8725 AUC = 0.8552**

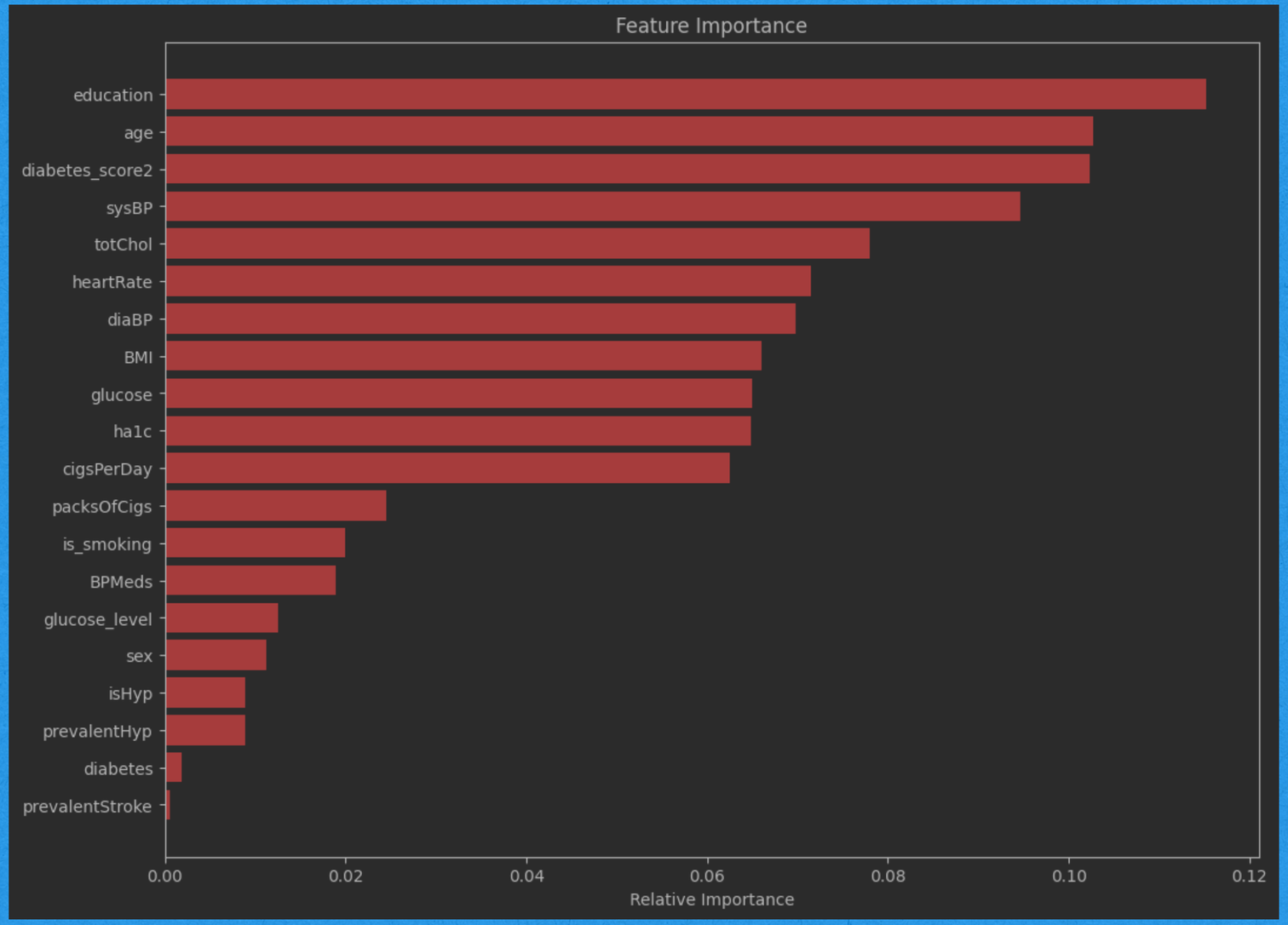
POC and Results of the models





Importance of Features

These are the importance of the features accoring to the Random Forest classifier.
We can see the contribution of the diabetes_score2 column added!



But we didn’t stop there

We wanted to use our model to actively help other diabetic paitients. Using the models we developed, we wanted to provide paitients with a pragmatic solution for the unfortunate classification.

We want to answer the question :
“What is th minimal path to avoid the cardiovascular complication?”

D.O.C

Diabetic Optimal Counselor

What is a Counterfactual Example?

A counterfactual explanation (CE) is the minimal change to a datapoint such that after the change, the trained model will flip its prediction.

Age = 67	BMI = 34	Gender = M	...	Stroke

Age = 67	BMI = 25	Gender = M	...	No Stroke

Main problems of CE's

Purely theoretical (impractical) explanations:

Impossible implementation:

Age = 67	BMI = 34	Gender = M	has Head = Yes		Headache
Age = 67	BMI = 34	Gender = M	has Head = No		No Headache

Unrealistic implementation:

Age = 67	BMI = 34	Gender = M		Stroke
Age = 67	BMI = 31.123569988745	Gender = M		No Stroke

Optimization

The algorithm we use for optimization, can find a robust, minimal and optimal change to measured values. Using those, we can develop a plan for the user to better his situation through improving his habits and lifestyle!

The optimizer, developed in a noval article we used, is capble of solving all problems mentioned using various models like Decision Trees, Random Forest, XGBoost and even simple Neural Networks!

Demonstration

D.O.C - Vision

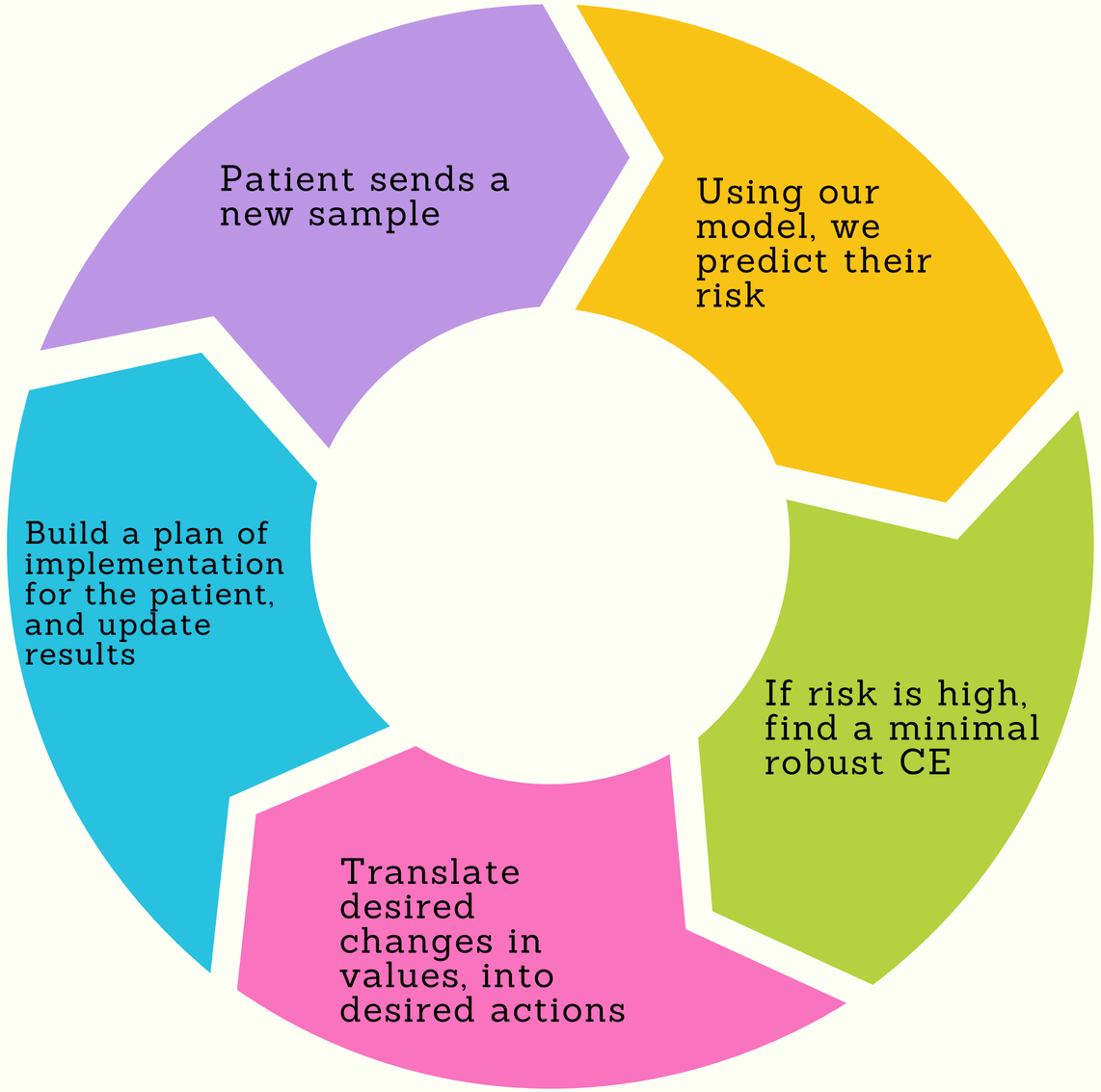
NEW SAMPLE

PREDICTION

OPTIMIZER

TRANSLATE

BUILD PLAN



Refrences

“Finding Regions of Counterfactual Explanations via Robust Optimization”, 2023, Maranago et. al.

“SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe”, 2023, L. Rydén et al.

Thank You!

From Team Alpaca