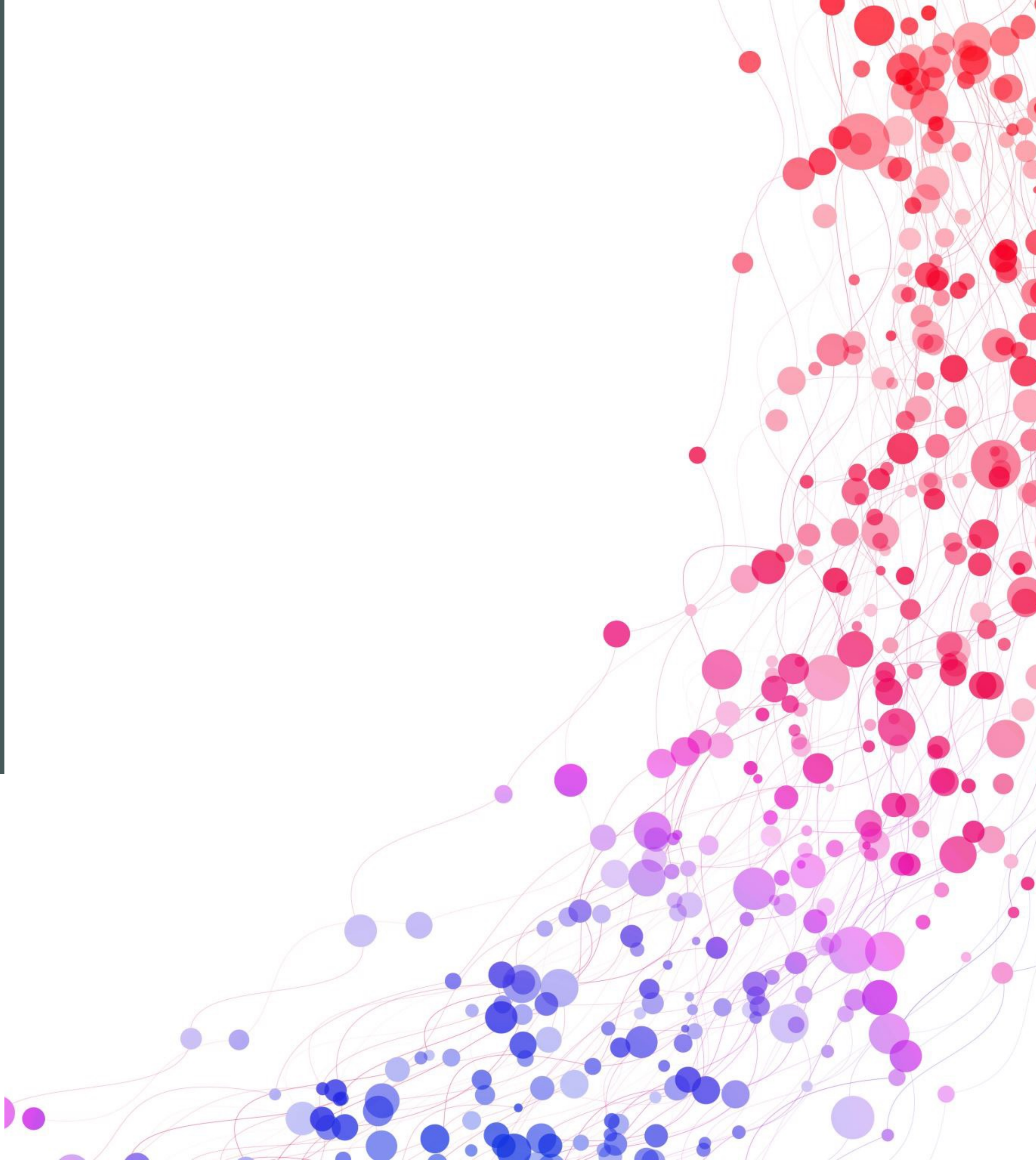
SCREENING TOOL FOR CHRONIC KIDNEY DISEASE

Anagha Srivaths



OBJECTIVE

- Develop an easy-to-use screening tool that helps identify patients at high risk for chronic kidney disease

WHAT WE ALREADY KNOW

- CDC and NHCS collects data from nationwide surveys of US adults
- Hypertension and Diabetes are identified as major causes for CKD
- Age is also an important factor

SCREENING TOOL FOR CKD

THE DATASET

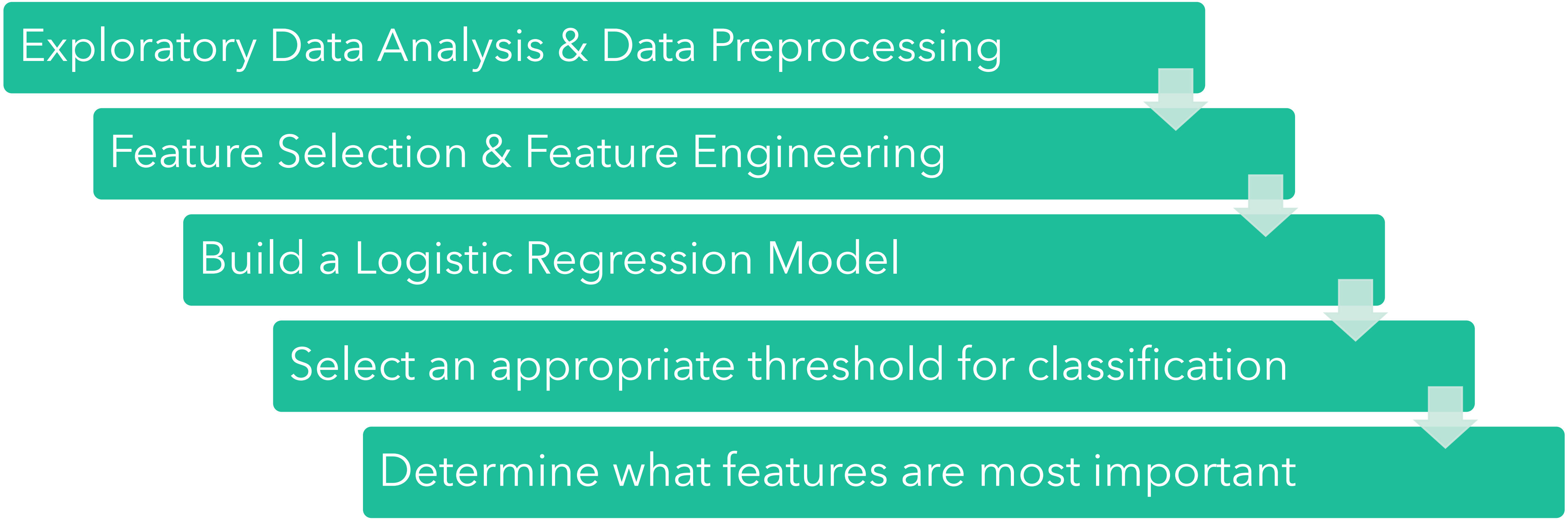
Dataset consists of 8819 adults : 6000 records for training data and 2819 records for test data

10 continuous and 23 categorical variables

The target variable CKD is imbalanced : only 464 out of 6000 reportedly have CKD

Blacks and Hispanics are more prone to CKD but are under-represented in the dataset

PROCESS



Exploratory Data Analysis & Data Preprocessing

Feature Selection & Feature Engineering

Build a Logistic Regression Model

Select an appropriate threshold for classification

Determine what features are most important

DATA PREPROCESSING

Removed 2819 rows with missing values for CKD and stored separately as test dataset

6000 rows remain – stored as training dataset

Removed rows with missing values in any column, rather than imputing values

Fixed the datatypes of the variables

FEATURE SELECTION

1. Remove highly correlated variables using Variance Inflation Factor Tests. This addresses the problem of multicollinearity

2. Two sample t-tests between numerical predictor variables and categorical target variable to include only strong and significant predictors

3. Chi-Square tests between categorical predictor variables and target variable to include only strong and significant predictors

VIF TESTS

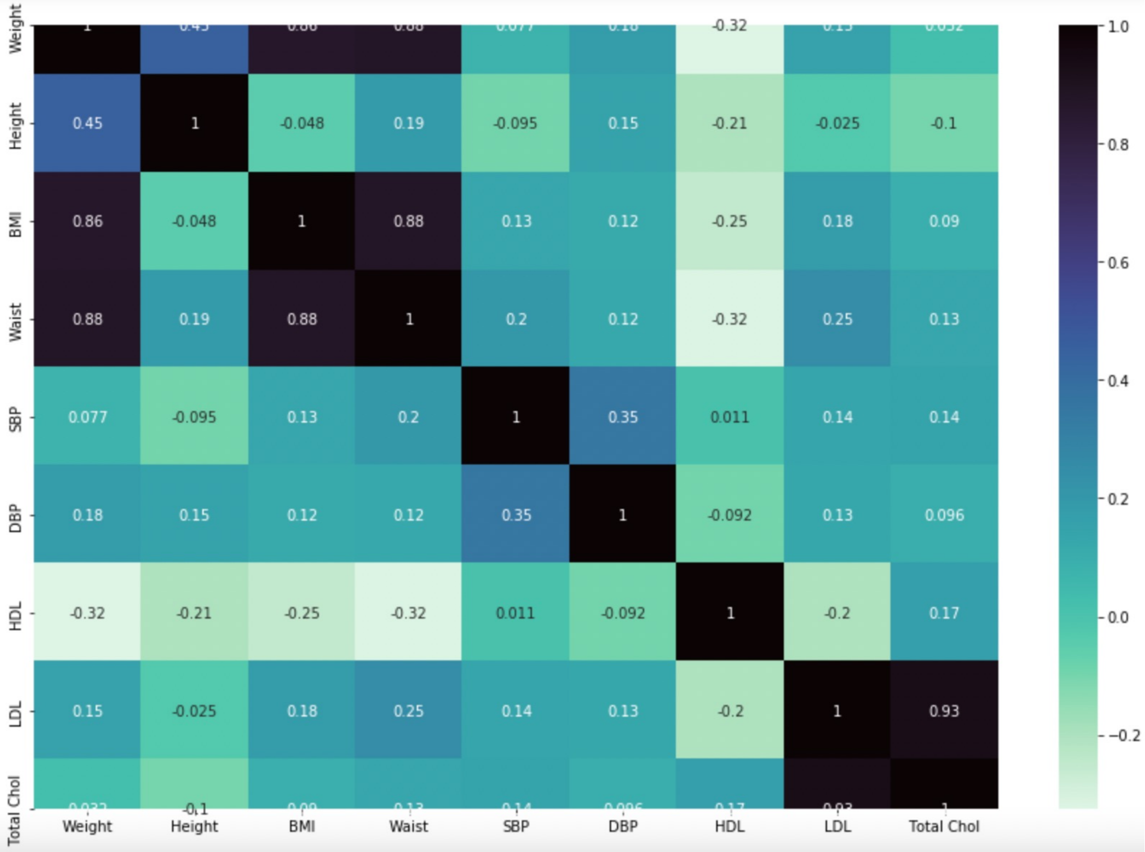
VARIABLE	VIF
Age	2.263513
Female	2.428272
Educ	1.249053
Unmarried	1.174469
Income	1.292704
Insured	1.328344
Weight	93.847686
Height	24.843140
BMI	73.446612

VARIABLE	VIF
Obese	2.589737
Waist	8.297568
SBP	2.174986
DBP	1.314585
HDL	100.000000
LDL	100.000000
Total Chol	100.000000
Dyslipidemia	1.166641
PVD	1.083044

VARIABLE	VIF
Activity	1.078867
PoorVision	1.055844
Smoker	1.097945
Hypertension	1.808456
Fam Hypertension	2.635881
Diabetes	1.205459
Fam Diabetes	1.106771
Stroke	1.785756
CVD	2.026512

VIF TESTS

VARIABLE	VIF
Fam CVD	2.707840
CHF	1.157335
Anemia	1.032044
Racegrp_black	1.289021
Racegrp_hispa	1.497455
Racegrp_other	1.066870
CareSource_	1.019016
CareSource_clinic	1.132133
CareSource_noplace	1.368662
CareSource_other	1.092397



VIF TESTS

VARIABLE	P-VALUE
Hypertension & Fam Hypertension	0.00020
Diabetes and Fam Diabetes	4.1091e-52

Removed BMI, Height and Waist – High Correlation with Weight

Removed LDL - High Correlation with Total Cholesterol

Removed Fam Diabetes & retained Diabetes

Removed Fam Hypertension & retained Hypertension

t - tests

Two sample t-tests between continuous predictor variables and target variables to retain only statistically significant variables

VARIABLE	P-VALUE
Age	2.841272e-113
Weight	8.060397e-01
SBP	9.505868e-40
DBP	2.818226e-03
Total Chol	1.492170e-01
HDL	3.800099e-03

Eliminated variables with p-values less than 0.05

Dropped Weight and Total Chol as they were not statistically significant

CHI SQUARE TESTS

Chi Square tests between categorical predictor variables and target variables to retain only statistically significant variables

VARIABLE	P-VALUE
Female	5.155155e-01
Racegrp	8.446885e-10
Educ	4.421879e-05
Unmarried	1.434536e-03
Income	3.562931e-08

VARIABLE	P-VALUE
CareSource	1.511427e-06
Insured	3.471815e-11
Obese	1.683670e-01
Dyslipidemia	1.000000e+00
PVD	3.039083e-23

VARIABLE	P-VALUE
Activity	1.241635e-09
PoorVision	4.104805e-10
Smoker	8.361968e-04
Hypertension	9.625535e-47
Diabetes	3.662216e-23

CHI SQUARE TESTS

Chi Square tests between categorical predictor variables and target variables to retain only statistically significant variables

VARIABLE	P-VALUE
Stroke	3.480925e-21
CVD	3.695777e-36
CHF	1.021426e-12
Anemia	2.471501e-01
CKD	0.000000e+00

Eliminated variables with p-values less than 0.05

Dropped Female, Obese, Dyslipidaemia & Anemia as they were not statistically significant

FINAL VARIABLES SELECTED FOR MODELLING

Age Racegrp Educ Unmarried Income CareSource

Insured SBP DBP PVD Activity PoorVision

Smoker Hypertension Diabetes Stroke CHF

PREDICTIVE MODELLING – LOGISTIC REGRESSION

- Target Variable – CKD → Categorical variable with 2 levels :
  - 0 indicating the absence of CKD
  - 1 indicating the presence of CKD
- Train – Validation set split in the ratio 80 : 20
- Normalised all continuous variables
- Created Dummy variables for all categorical variables

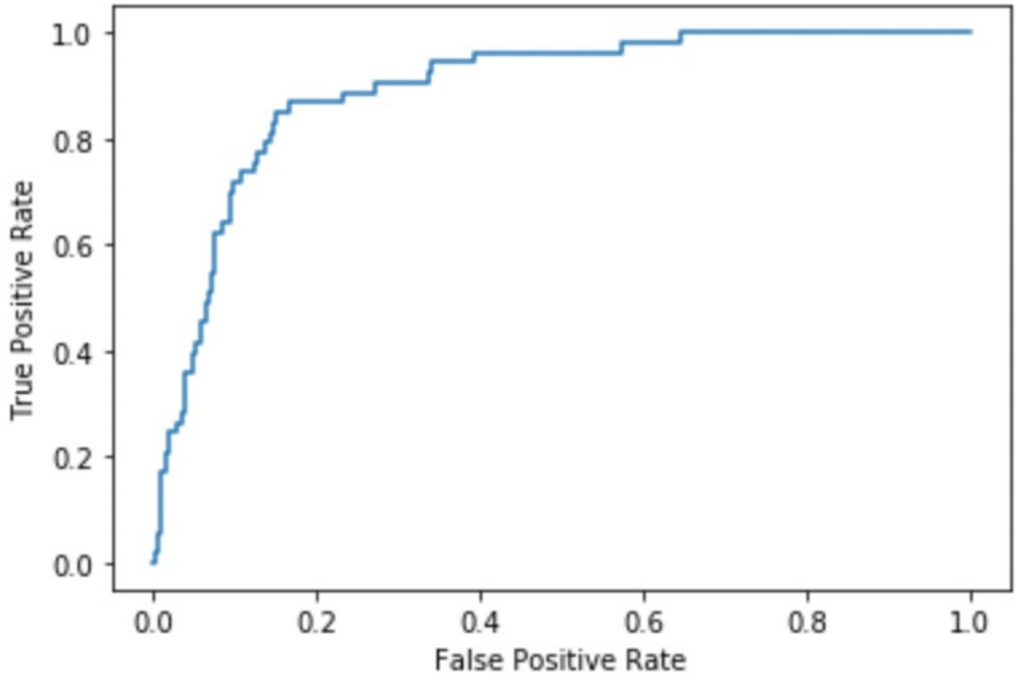
OPTIMAL THRESHOLD SELECTION

A logistic regression assumes equal probability (50/50) of belonging to either target class

With an imbalance in the target variable, this approach would be invalid

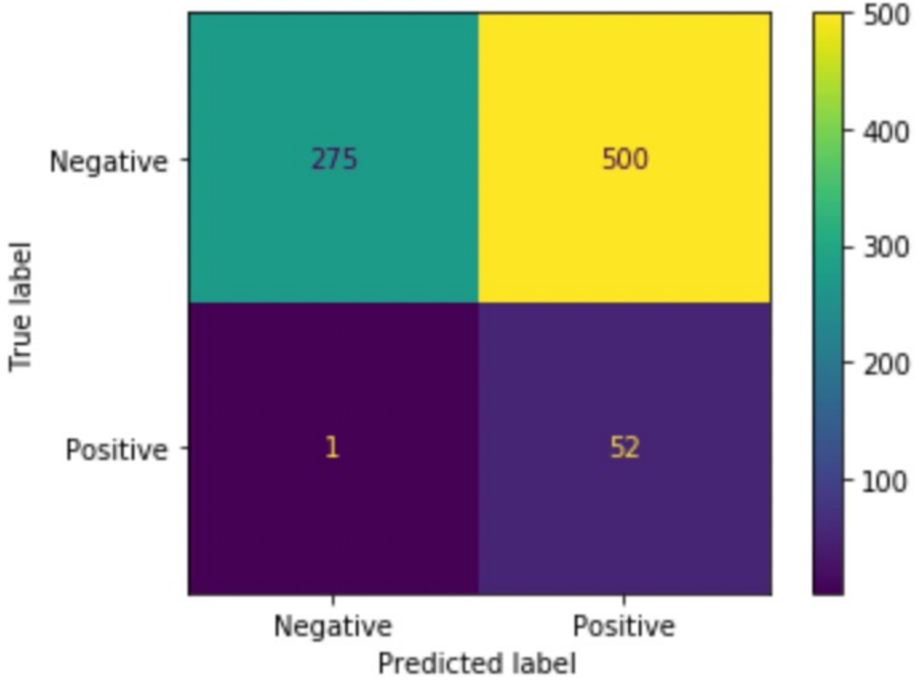
Thus, an optimal threshold is selected using the ROC curve

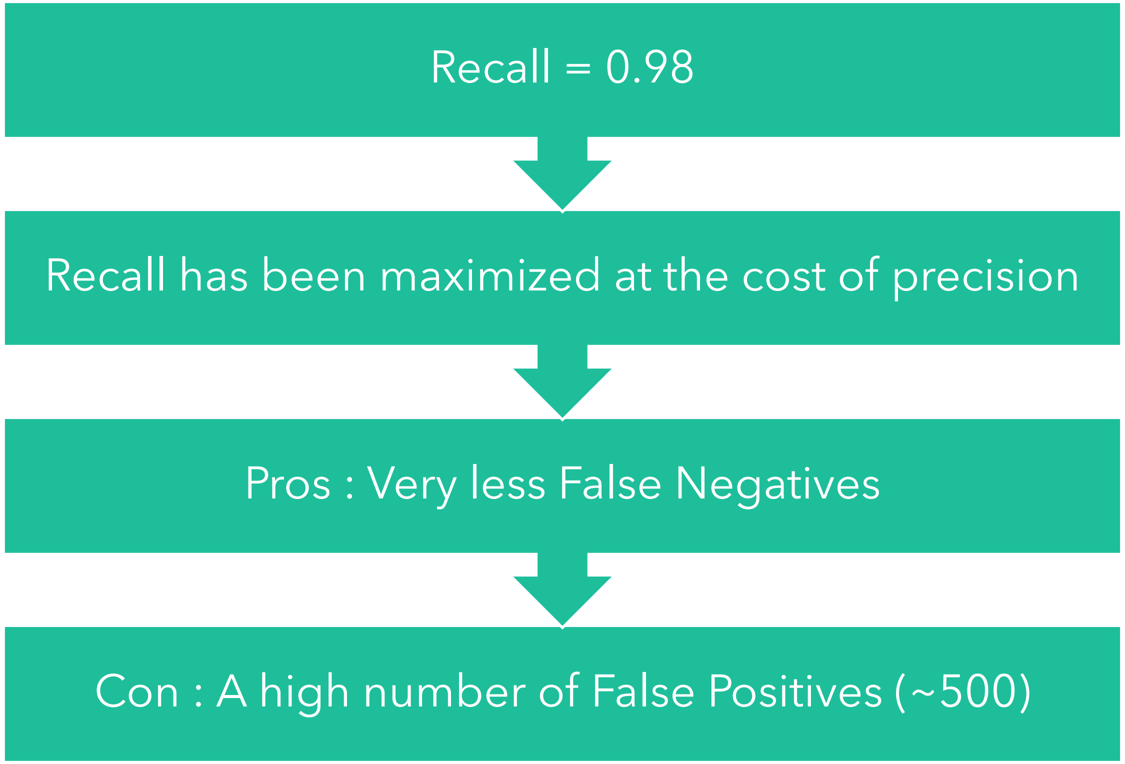
For this particular application : **Cost of FN > Cost of FP** → Choose a threshold to maximise ***recall (TPR)***



OPTIMAL THRESHOLD = 0.0053

MODEL PERFORMANCE





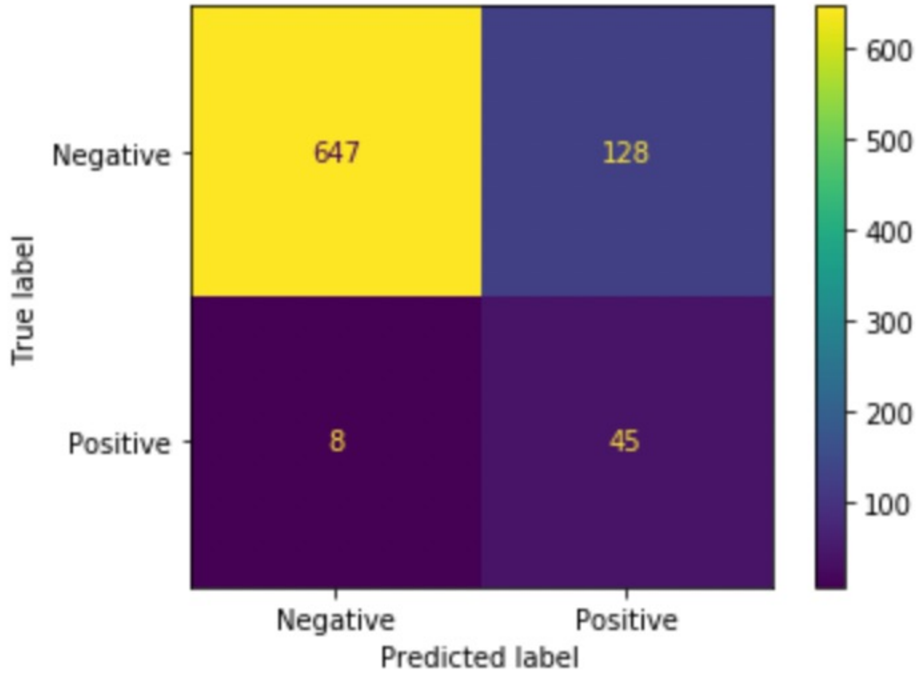
Recall = 0.98

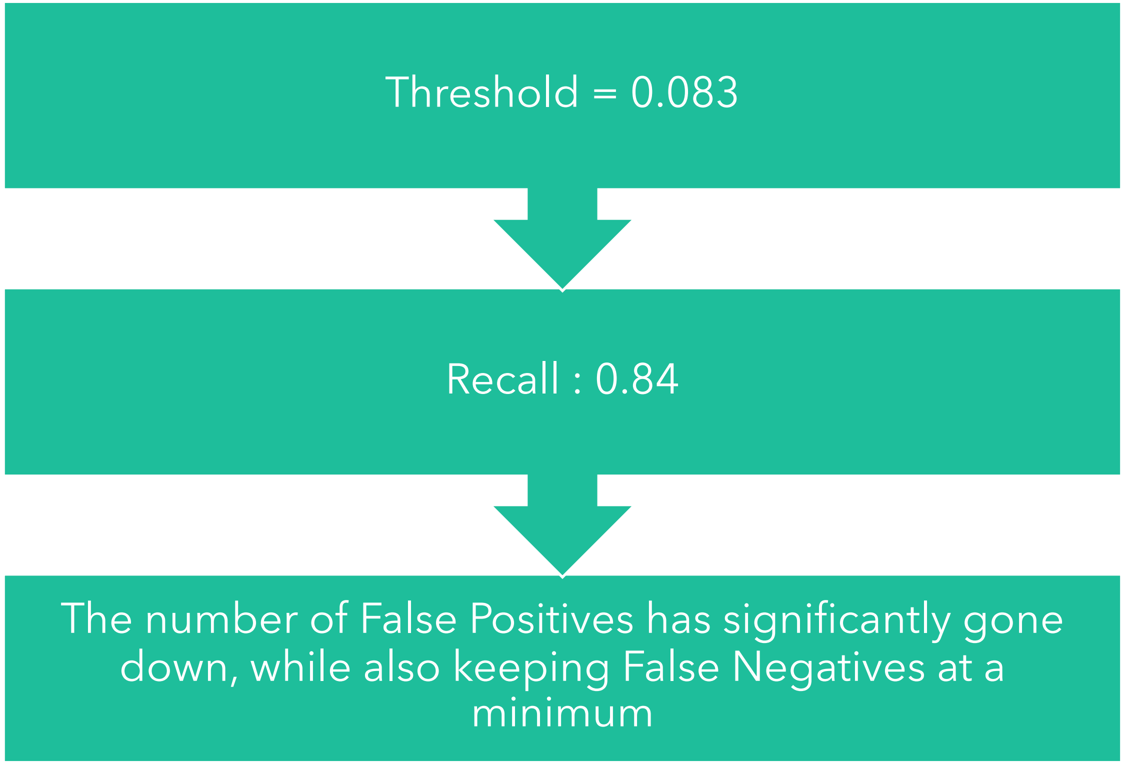
Recall has been maximized at the cost of precision

Pros : Very less False Negatives

Con : A high number of False Positives (~500)

ANOTHER THRESHOLD : BALANCE BETWEEN PRECISION & RECALL





Threshold = 0.083

Recall : 0.84

The number of False Positives has significantly gone down, while also keeping False Negatives at a minimum

FEATURE IMPORTANCE : WHAT CAN BE USED TO SCREEN PEOPLE?

VARIABLE	COEFF	P>|Z|	ODDS RATIO	VARIABLE	COEFF	P>|Z|	ODDS RATIO
Age	0.0867	0.002	1.090569	**Income_1**	-0.1396	0.466	0.869706
SBP	-0.0060	0.189	0.994018	**CareSource_clinic**	-0.0527	0.797	0.948665
DBP	0.0018	0.793	1.001802	**CareSource_noplace**	-0.6563	0.103	0.518767
HDL	-0.0146	0.008	0.985506	**CareSource_other**	0.2708	0.453	1.311013
Racegrp_hispa	-1.0693	0.000	0.343249	**Income_1**	-0.1396	0.466	0.869706
Racegrp_other	-0.4799	0.456	0.618845	**Insured_1**	0.0006	0.999	1.000600
Racegrp_white	0.0104	0.963	1.010454	**PVD_1**	0.2470	0.359	1.280179
Educ_1	-0.2520	0.162	0.777245	**Activity_2**	-0.2118	0.231	0.809127
Unmarried_1	0.2203	0.209	1.246451	**Activity_3**	-0.5162	0.081	0.596784

FEATURE IMPORTANCE : WHAT CAN BE USED TO SCREEN PEOPLE?

VARIABLE	COEFF	P>|Z|	ODDS RATIO
Activity_4	1.3483	0.68	0.259681
PoorVision_1	0.1082	0.671	1.114271
Smoker_1	0.1850	0.275	0.831104
Hypertension_1	0.7979	0.0005	2.220872
Diabetes_1	0.4716	0.018	1.602556
Stroke_1	0.1433	0.709	1.154076
CVD_1	0.5252	0.073	1.690797
Fam CVD_1	0.3663	0.051	0.693295
CHF_1	0.0610	0.862	0.940823

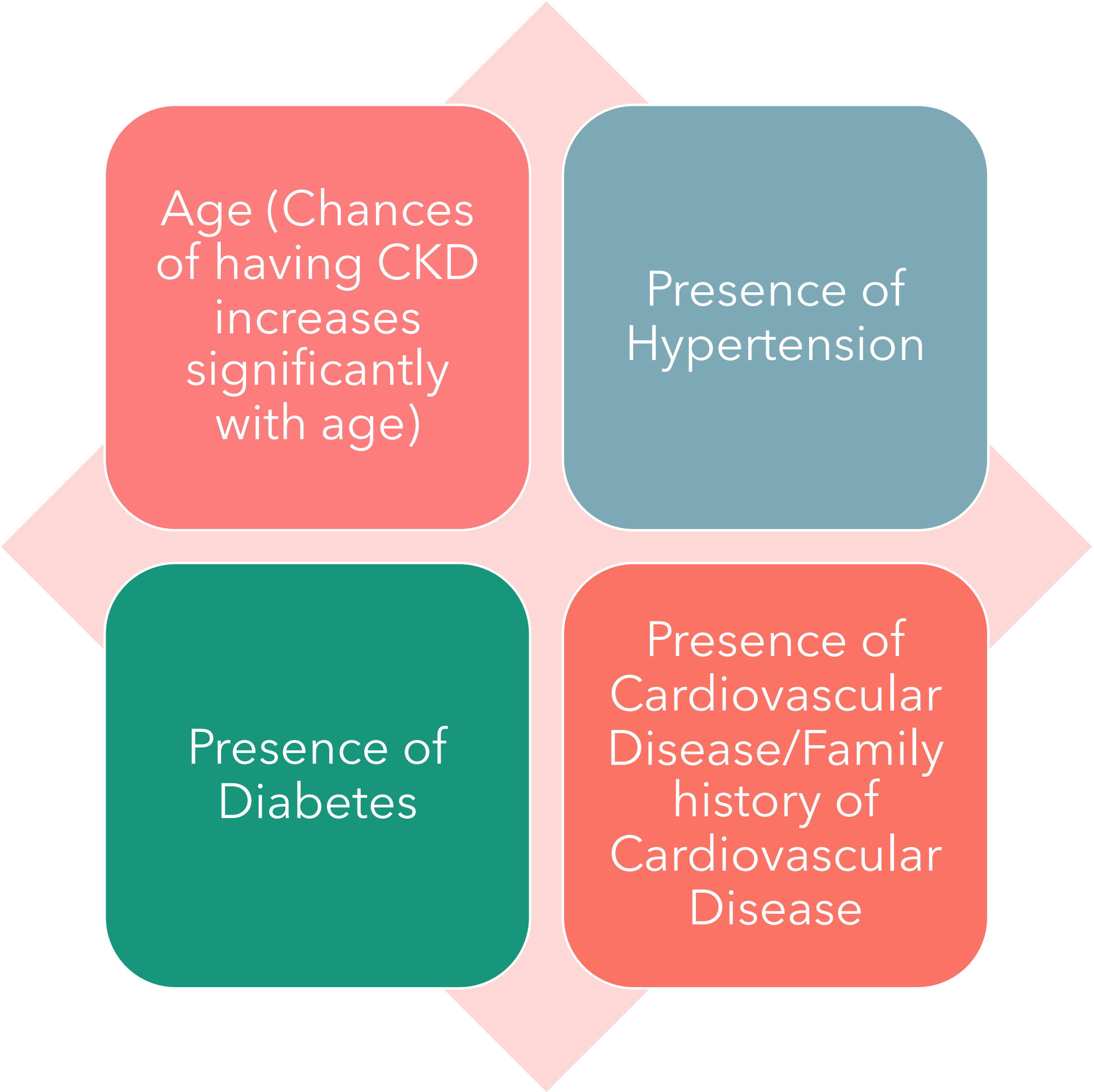
Every unit increase in age increase the probability of having CKD by 9%

Presence of Hypertension increases the probability of having CKD by 122%

Presence of Diabetes increase the probability of having CKD by 60%

Presence of a cardiovascular disease increases the probability of having CKD by 69%

VARIABLES TO BE INCLUDED IN THE SCREENING TOOL



Age (Chances of having CKD increases significantly with age)

Presence of Hypertension

Presence of Diabetes

Presence of Cardiovascular Disease/Family history of Cardiovascular Disease

THANK YOU!