SCREENING FOR CHRONIC KIDNEY DISEASE

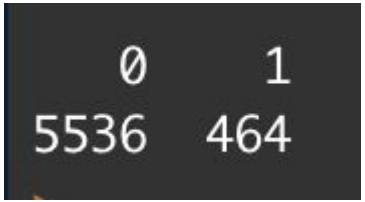
PRATHAMESH BAPAT

INTRODUCTION

Objective To develop an easy to use screening tool that identifies patients who are at higher risk of developing CKD based on various factors.

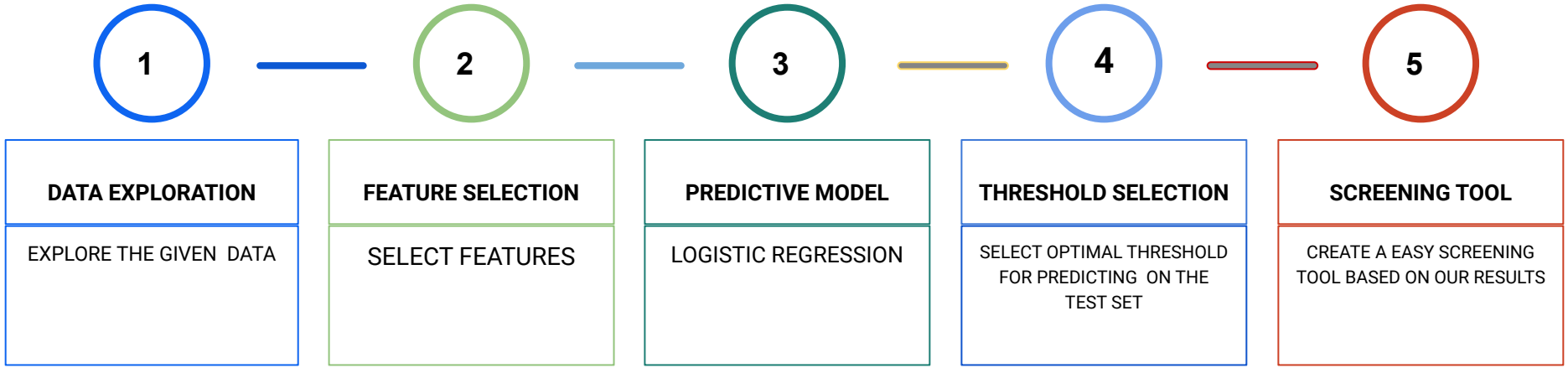
WHAT WE ALREADY KNOW?

Hypertension and Diabetes are one of the main cause of the chronic kidney disease



- According to the data, Age, Hypertension and Diabetes were positively correlated to CKD.
- The nature of the data is imbalanced since only 464 out of 6000 have CKD and might produce biased predictions.
- From literature survey, black and Hispanic are more prone to CKD but in data, they are underrepresented.

Process

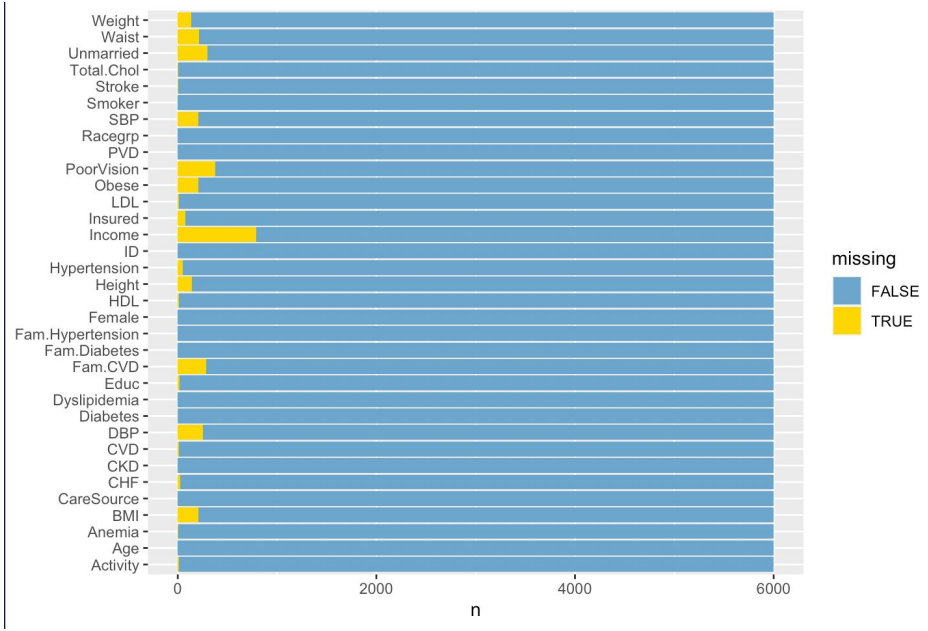

1
2
3
4
5
DATA EXPLORATION
EXPLORE THE GIVEN DATA
FEATURE SELECTION
SELECT FEATURES
PREDICTIVE MODEL
LOGISTIC REGRESSION
THRESHOLD SELECTION
SELECT OPTIMAL THRESHOLD FOR PREDICTING ON THE TEST SET
SCREENING TOOL
CREATE A EASY SCREENING TOOL BASED ON OUR RESULTS

FEATURE SELECTION

Racegrp	CKD=0	CKD=1	Total
Black	1001	77	**1078**
Hispa	1688	70	**1758**
Other	184	6	**190**
White	2663	311	**2974**
Total	**5536**	**464**	**6000**
	Chi-square		71.7

CareSource	CKD=0	CKD=1	Total
Vlinhth	1169	100	**1269**
Dr./HMO	3156	326	**3482**
Noplace	911	14	**925**
Other	298	24	**322**
Total	**5534**	**464**	**5998**
	Chi-Square		63.2

	CKD=0			CKD=1			
	Average	Std Dev	Count	Average	Std Dev	Count	T-stat
Age	47.15	17.90	5536	73.05	11.71	464	-43.56
Weight	79.17	19.60	5432	77.74	19.25	435	1.49
Height	167.25	10.12	5433	165.29	10.41	428	3.77
BMI	28.24	6.22	5377	28.35	5.98	417	-0.36
Waist	96.54	15.24	5365	100.10	14.44	420	-4.85
SBP	124.27	20.14	5352	141.47	25.28	442	-13.94
DBP	71.86	12.24	5318	67.73	14.28	430	5.83
HDL	51.97	15.79	5529	50.08	16.18	463	2.41
LDL	151.85	42.46	5529	157.20	44.02	463	-2.52
Total Chol	203.82	42.04	5531	207.28	44.98	463	-1.60
Activity	2.06	0.82	5530	1.69	0.67	462	11.11

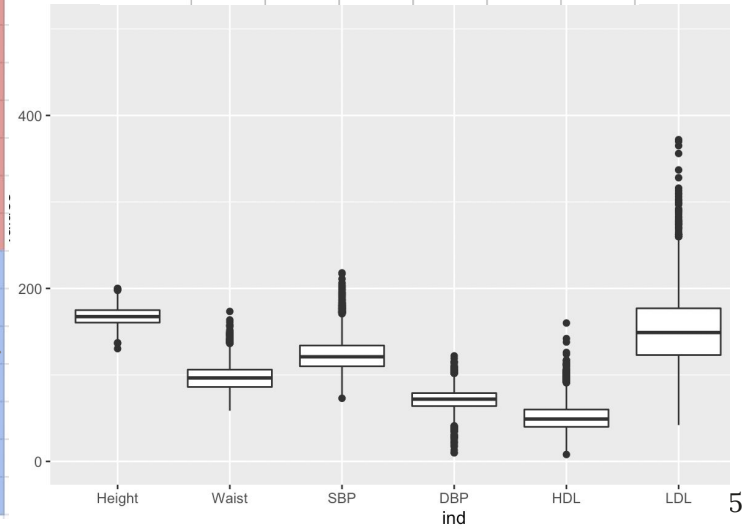


	Variable=0			*Variable=1*			
Variable	**CKD=0**	**CKD=1**	**%1s**	**CKD=0**	**CKD=1**	**%1s**	**Chi-square**
Female	2655	210	7.3%*	2881	254	8.1%	1.3
Educ	3064	308	9.1%	2458	155	5.9%	21.1
Unmarried	3335	227	6.4%	1926	211	9.9%	23.1
Income	2723	293	9.7%	2088	104	4.7%	44.5
Insured	1137	17	1.5%	4329	439	9.2%	78.2
Obese	3708	281	7.0%	1669	136	7.5%	0.4
Dyslipidemia	4951	414	7.7%	585	50	7.9%	0.0
PVD	5379	395	6.8%	157	69	30.5%	171.1
Poor Vision	4932	355	6.7%	277	60	17.8%	57.0
Smoker	3902	273	6.5%	1634	191	10.5%	27.4
Hypertension	3476	97	2.7%	2007	367	15.5%	322.0
Fam Hypertension	4231	388	8.4%	1305	76	5.5%	12.5
Diabetes	4998	334	6.3%	537	130	19.5%	145.3
Fam Diabetes	3829	307	7.4%	1707	157	8.4%	1.8
Stroke	5403	404	7.0%	128	59	31.6%	153.7
CVD	5249	348	6.2%	275	115	29.5%	276.7
Fam CVD	3469	306	8.1%	1824	118	6.1%	7.7
CHF	5401	404	7.0%	113	56	33.1%	158.3
Anemia	5434	442	7.5%	99	22	18.2%	18.9

FEATURE ENGINEERING

TARGET	CKD	
CATEGORICAL	Racegrp	CREATE SEPERATE FEATURES FOR EACH LEVEL
	Educ	
	CareSource	CREATE SEPERATE FEATURE FOR EACH LEVEL
	PVD	
	Activity	ORDER THE FACTORS
	PoorVision	
	Smoker	
	Hypertension	FACTOR EVERYTHING ELSE
	Fam.Hypertension	
	Diabetes	
	Stroke	
	CVD	
	Farm.CVD	
	CHF	
	Anemia	
NUMERICAL	Height	NORMALIZE EVERYTHING ELSE
	Waist	
	SBP	BOX PLOTS
	DBP	
	HDL	
	LDL	
	AGE	CREATE 4 BUCKETS

	Height	Waist	SBP	DBP	HDL	LDL
Height	1	0.19	-0.1	0.15	-0.21	-0.02
Waist	0.19	1	0.2	0.12	-0.32	0.25
SBP	-0.1	0.2	1	0.35	0.01	0.14
DBP	0.15	0.12	0.35	1	-0.09	0.13
HDL	-0.21	-0.32	0.01	-0.09	1	-0.2
LDL	-0.02	0.25	0.14	0.13	-0.2	1



LOGISTIC REGRESSION

```
# Get Train and Validation Set
train = ch_train_final[1:3308,]
val = ch_train_final[3308:4135,]
```

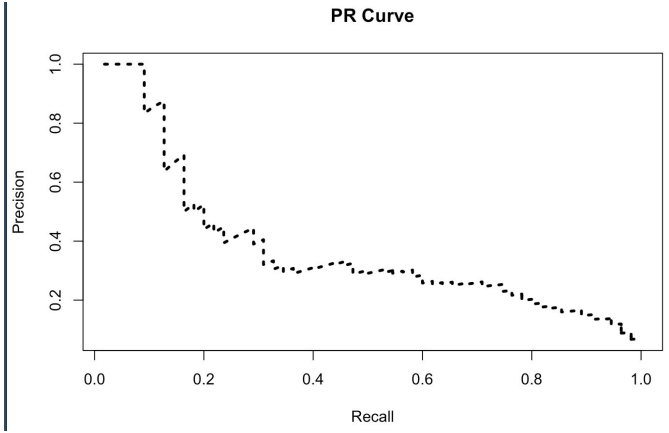
IMPORTANT FEATURES

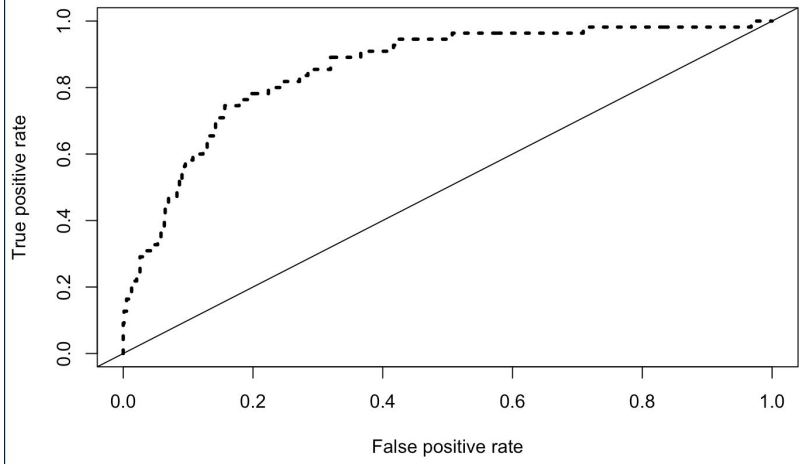
AGE
PVD
HYPERTENSION
DIABETES
CVD

Positive change in Log Odds for these features when 'yes' respect to 'No'

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.038926   2.187443  -4.132 3.59e-05 ***
Height           0.014185   0.010051   1.411  0.15817
Waist            0.003290   0.006527   0.504  0.61417
SBP              0.000118   0.004557   0.026  0.97934
DBP             -0.008249   0.006829  -1.208  0.22707
HDL             -0.007131   0.006003  -1.188  0.23484
LDL              0.002874   0.001993   1.442  0.14926
Activity2       -0.155066   0.181445  -0.855  0.39276
Activity3       -0.605910   0.320950  -1.888  0.05904 .
Activity4       -1.622040   1.029744  -1.575  0.11521
agegroup         1.189989   0.136139   8.741  < 2e-16 ***
Educ1           -0.164875   0.187561  -0.879  0.37938
Unmarried1       0.261124   0.183297   1.425  0.15427
Income1         -0.106592   0.201706  -0.528  0.59718
Insured1         0.401112   0.398470   1.007  0.31411
PVD1             0.563184   0.259491   2.170  0.02998 *
PoorVision1      0.120599   0.258029   0.467  0.64022
Smoker1          0.030848   0.171583   0.180  0.85732
Hypertension1    0.923285   0.221803   4.163 3.15e-05 ***
Fam.Hypertension1 -0.149682  0.346807  -0.432  0.66603
Diabetes1        0.644011   0.199995   3.220  0.00128 **
Stroke1          0.187980   0.383731   0.490  0.62422
CVD1             0.746711   0.288660   2.587  0.00969 **
Fam.CVD1         0.043865   0.308173   0.142  0.88681
CHF1             0.060903   0.345705   0.176  0.86016
Anemia1          0.909321   0.635895   1.430  0.15272
RaceGrp_black1   0.505923   0.781486   0.647  0.51738
RaceGrp_white1   0.747863   0.763347   0.980  0.32723
RaceGrp_hispa1  -0.255118   0.785089  -0.325  0.74522
CareSrc_clinic1 -0.402877   0.395879  -1.018  0.30883
CareSrc_noplace1 -1.118675  0.596190  -1.876  0.06060 .
CareSrc_DrHMO1  -0.463328   0.368796  -1.256  0.20900
```

ROC – PR CURVE – OPTIMAL THRESHOLD SELECTION






```
#Optimal Threshold with FN cost 10 times FP cost
cost.perf <- performance(pred_roc,"cost",cost.fp = 1, cost.fn = 10)
pred_roc@cutoffs[[1]][which.min(cost.perf@y.values[[1]])]
```

0.086

OPTIMAL THRESHOLD

ACCURACY	0.84
RECALL	**0.73**
PRECISION	0.24
F SCORE	0.37

ACTUALS

	pred 0	pred 1
0	652	121
1	15	40

CONFUSION MATRIX – TEST SET

SCREENING TOOL[1]

1) What's your Age?

20-40	40-60	60-80	80+
+1	+2	+3	+4

2) Do you have any stroke or Peripheral vascular disease?

EITHER ONE	BOTH
+1	+2

3) Do you have diabetes or any family diabetes history or hypertension?

EITHER ONE	EITHER 2	ALL THREE
+1	+3	+4

IF YOU SCORED MORE THAN 5 THEN *TEST FOR CKD* AND CONSULT YOUR DOCTOR

THANKS!

Any Questions?

Reference for screening tool - [1] https://github.com/kcngnn/Chronic-Kidney-Disease-Prediction/blob/master/Final_Report.pdf