HEART DISEASE PREDICTION

Jackson Song, Maclain Prom,
Magdalena Trzupek, Charlie Liu

Introduction and Motivation

-In the healthcare field, predicting the chances of developing heart disease over the next ten years is crucial, as it helps people and doctors take steps to prevent heart problems before they occur.

-A doctor who focuses on preventing heart problems would find a predictive model for heart disease very useful, as it would help them decide which patients need extra attention and which treatments are most appropriate.

-Our primary goal was to create a model that can accurately predict people at high risk of heart disease.

-Our secondary goal was to yield interpretive insights about the nature of relationships between our variables.

- Dataset found on Kaggle
- 16 columns
- 4,240 total rows
- categorical variables already written in 0/1 format

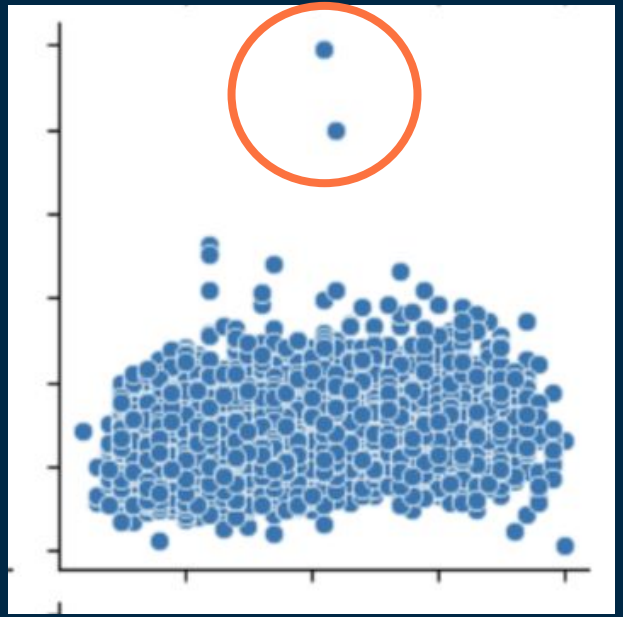
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0
21	1	41	2	0	0	0	0	0	0	195	139	88	26.88	85	65	0
22	0	42	2	1	30	0	0	0	0	190	108	70.5	21.59	72	85	0
23	0	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0
24	0	52	1	0	0	0	0	0	0	234	148	78	34.17	70	113	0
25	0	52	3	1	20	0	0	0	0	215	132	82	25.11	71	75	0
26	1	44	2	1	30	0	0	1	0	270	137.5	90	21.96	75	83	0
27	1	47	4	1	20	0	0	0	0	294	102	68	24.18	62	66	1
28	0	60	1	0	0	0	0	0	0	260	110	72.5	26.59	65	NA	0
29	1	35	2	1	20	0	0	1	0	225	132	91	26.09	73	83	0
30	0	61	3	0	0	0	0	1	0	272	182	121	32.8	85	65	1
31	0	60	1	0	0	0	0	0	0	247	130	88	30.36	72	74	0
32	1	36	4	1	35	0	0	0	0	295	102	68	28.15	60	63	0
33	1	43	4	1	43	0	0	0	0	226	115	85.5	27.57	75	75	0
34	0	59	1	0	0	0	0	1	0	209	150	85	20.77	90	88	1
35	1	61	NA	1	5	0	0	0	0	175	134	82.5	18.59	72	75	1
36	1	54	1	1	20	0	0	1	0	214	147	74	24.71	96	87	0
37	1	37	2	0	0	0	0	1	0	225	124.5	92.5	38.53	95	83	0
38	1	56	NA	0	0	0	0	0	0	257	153.5	102	28.09	72	75	0
39	1	52	1	0	0	0	0	1	1	178	160	98	40.11	75	225	0
40	0	42	1	1	1	0	0	1	0	233	153	101	28.93	60	90	0
41	1	36	3	0	0	0	0	0	0	180	111	73	27.78	71	80	0
42	0	43	2	1	10	0	0	0	0	243	116.5	80	26.87	68	78	0
43	0	41	2	1	1	0	0	0	0	237	122	78	23.28	75	74	0
44	0	52	1	0	0	1	0	1	0	NA	148	92	25.09	70	NA	1
45	1	54	2	0	0	0	0	0	0	195	132	83.5	26.21	75	100	0

Dataset

Data Cleaning Decisions

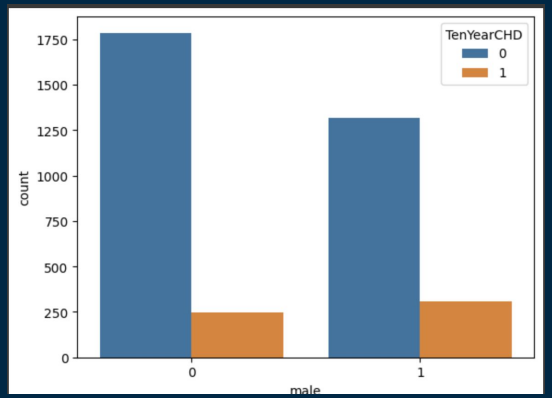
- Age, male, education, currentSmoker, totChol to predict tenYearCHD
- 582 explicit missing values that were dropped
- No implicit missing values detected
- 52 significant outliers in totChol (total cholesterol) column that we decided to remove
- 3,657 rows left after cleaning

—>



Visualizations of variables with association

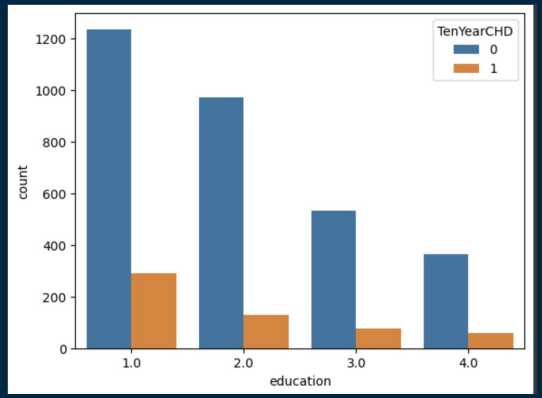
Gender and TenYearCHD



-medium relationship

-men slightly more likely to be diagnosed

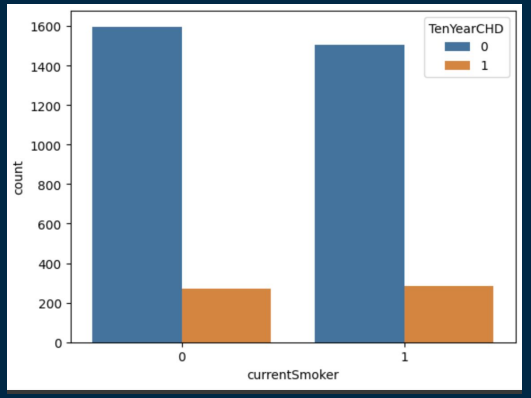
Education and TenYearCHD



-high relationship

-The higher the education, the smaller the likelihood of diagnosis

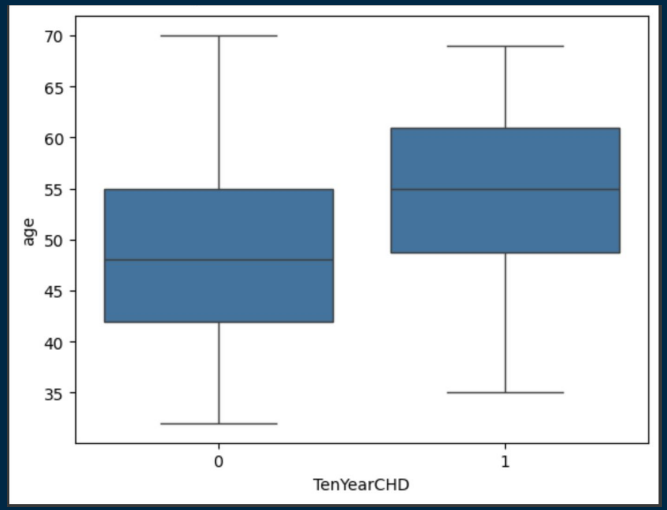
currentSmoker and TenYearCHD



-weak relationship

- Individuals who did and did not smoke will not really impact the likelihood of getting a diagnosis.

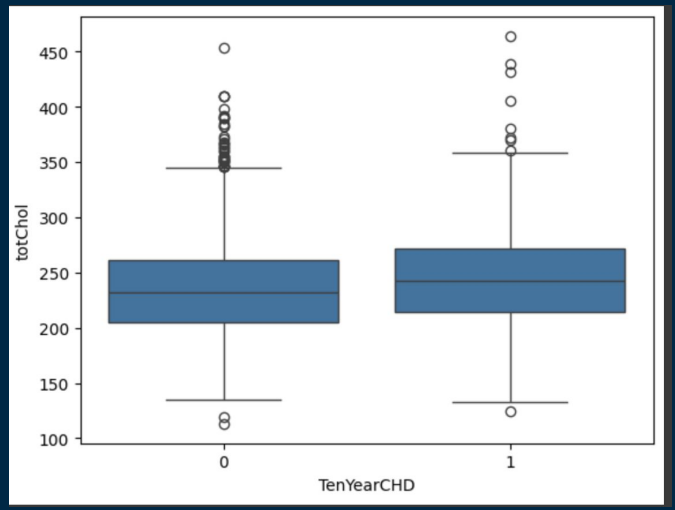
Age and TenYearCHD



-medium relationship

-older individuals more likely to be diagnosed

TotChol and TenYearCHD

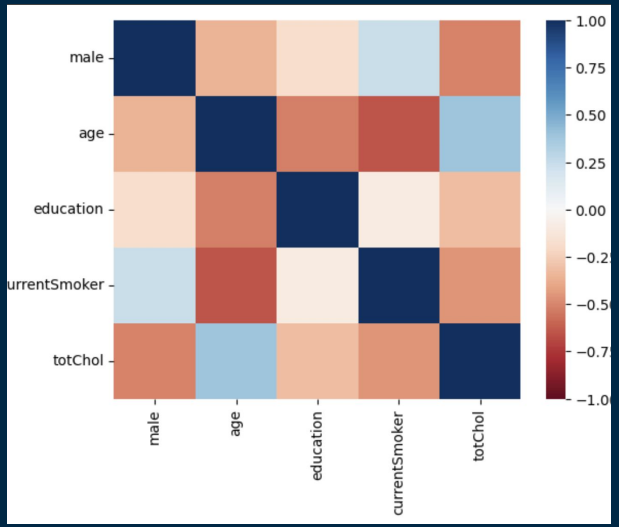


-medium relationship

- When you have high cholesterol, it can be the cause of heart disease.

Relationships between Explanatory Variable Pairs

-There are no pairs of explanatory variables that show strong associations that could lead to multicollinearity. The highest correlation is male and currentSmoker, which is 0.2.

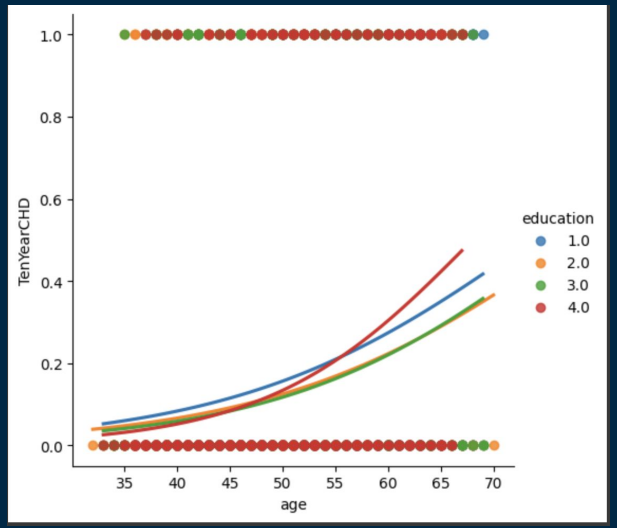


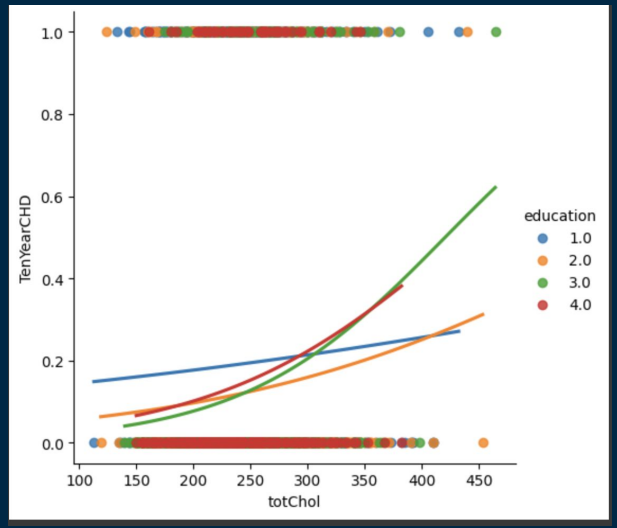

```
clean[['male', 'age', 'education', 'currentSmoker', 'totChol']].corr()
```

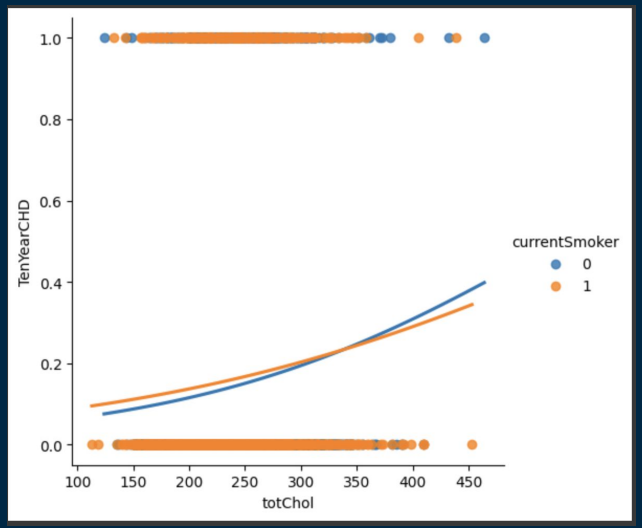
	male	age	education	currentSmoker	totChol
male	1.000000	-0.024345	0.017736	0.205925	-0.068960
age	-0.024345	1.000000	-0.159502	-0.210724	0.270128
education	0.017736	-0.159502	1.000000	0.025259	-0.013693
currentSmoker	0.205925	-0.210724	0.025259	1.000000	-0.049296
totChol	-0.068960	0.270128	-0.013693	-0.049296	1.000000

Interaction Effects

-Education and age, education and totChol, and currentSmoker and totChol seem to have interaction with TenYearCHD.







Model Data Preprocessing

1. 0/1 Response Variable
2. Features Matrix and Target Array
3. Scaling the explanatory variable
4. Indicator variables by using the get dummies
5. Concat the numerical and indicator dataset as feature.

Features

	age	totChol	male	currentSmoker	education_2.0	education_3.0	education_4.0
0	-1.232347	-0.955631	1	0	0	0	1
1	-0.414774	0.303332	0	0	1	0	0
2	-0.181182	0.188880	1	1	0	0	0
3	1.337166	-0.268924	0	1	0	1	0
4	-0.414774	1.104489	0	1	0	1	0
...	...	...	...	...	...	...	...
3652	0.052410	1.745416	1	1	0	0	0
3653	0.169206	-0.680948	1	1	0	1	0
3654	0.286002	0.738246	0	0	1	0	0
3655	-1.115551	-1.184533	1	0	0	1	0
3656	-1.232347	-0.932741	0	1	0	1	0

Target

	TenYearCHD
0	0
1	0
2	0
3	1
4	0
...	...
3652	1
3653	0
3654	0
3655	0
3656	0

Model Selection Technique

- Model selection technique used was forward selection algorithm

- The variables added in order were age, followed by male, followed by currentSmoker, followed by totChol, followed by education(2.0), followed by education(3.0), finally followed by education(4.0)

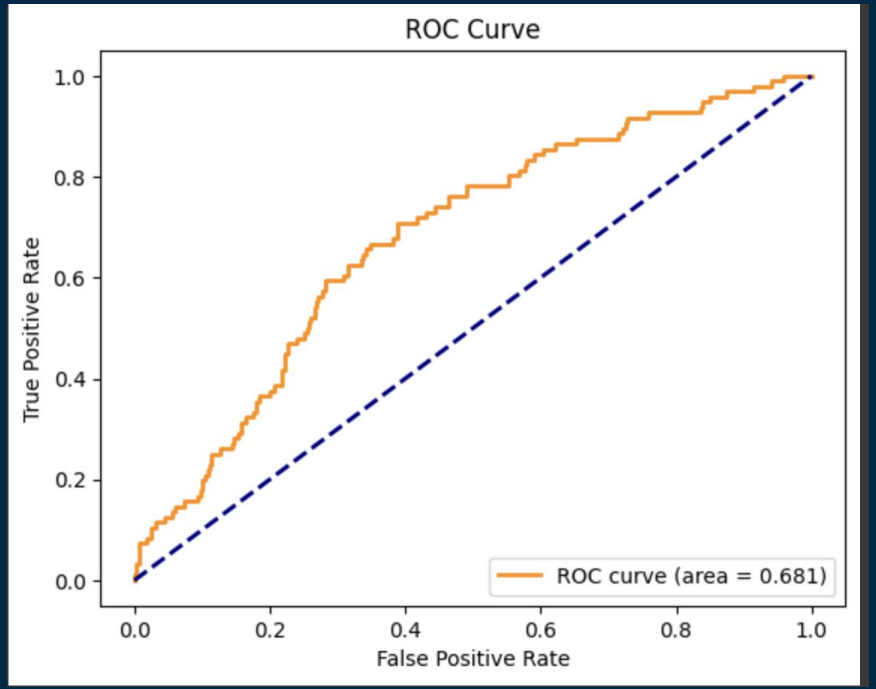
Best Model

-The best model ended up being the full model.

- as the original model is the best model the test AUC score is the same (0.681)

-threshold of 0.08 is best

-TPR of 0.90625, FPR of 0.72483

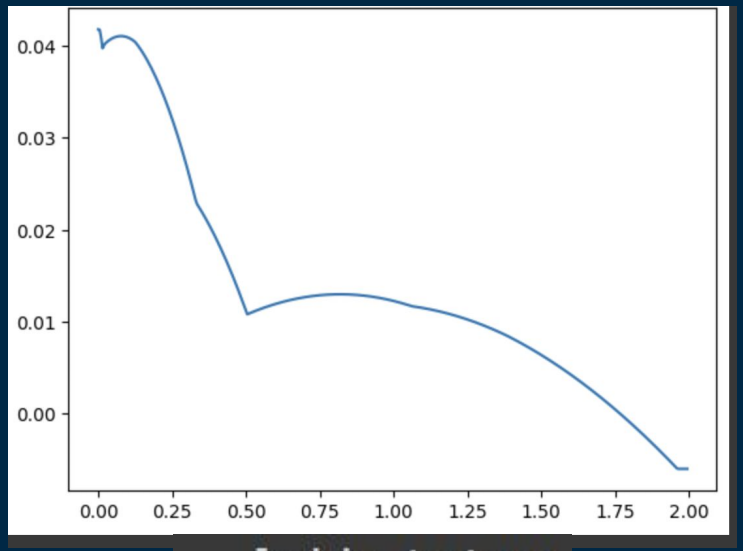


Logistic Regression Model Equation

$$\widehat{\text{TenYearCHD}} = 1 + e^{-(-2.1471 + 0.6707(\text{age}) + 0.1806(\text{totChol}) + 0.5489(\text{male}) + 0.3094(\text{currentSmoker}) - 0.2102(\text{education\_2.0}) - 0.3457(\text{education\_3.0}) - 0.1391(\text{education\_4.0})}$$

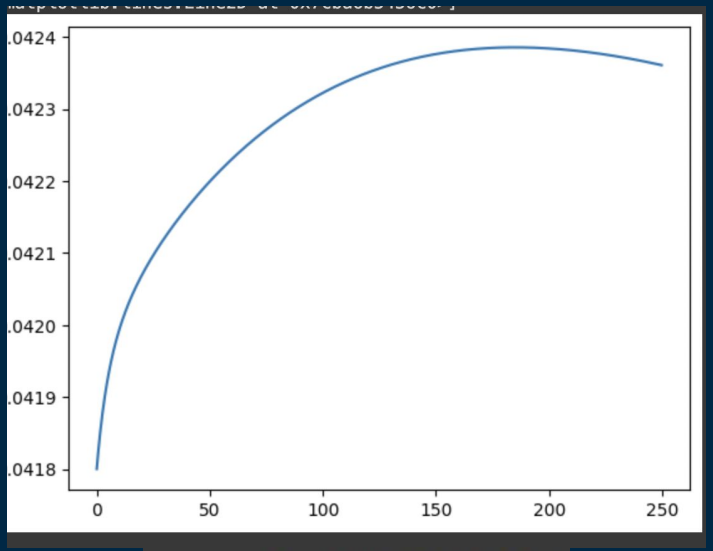
Insights

LASSO



	lambda	test R^2
0	0.0	0.0418

Ridge Regression



	lambda	test R^2
740	185.0	0.042386

	Ridge_model	LASSO_model	Non_regularized_model
male	0.060005	0.061231	0.061231
age	0.007026	0.007008	0.007008
education	-0.006416	-0.006363	-0.006363
currentSmoker	0.016357	0.017184	0.017184
cigsPerDay	0.001773	0.001740	0.001740
BPMeds	0.048989	0.053984	0.053984
prevalentStroke	0.087072	0.141248	0.141248
prevalentHyp	0.026481	0.026863	0.026863
diabetes	0.004361	0.005324	0.005324
totChol	0.000196	0.000197	0.000197
sysBP	0.002480	0.002466	0.002466
diaBP	-0.000535	-0.000561	-0.000561
BMI	0.000642	0.000649	0.000649
heartRate	-0.000705	-0.000687	-0.000687
glucose	0.001317	0.001308	0.001308

Conclusion

-Our AUC of 0.681 is relatively weak, high TPR of 0.90625 but high FPR of 0.724843, so accurate but may deem a lot of people falsely as having a risk of heart disease

-It is imperative that we have accuracy in predicting heart disease over 10 years

-With some adjustments, this model can definitely be improved, but in this context, we need a more accurate model since not accurately predicting heart disease in patients can be life threatening and costly.

-This is currently the highest AUC value we can get from our model. Since we got the model from using all variables, it would not be possible to make it more accurate with what we currently have.

- If we had other variables that can factor into heart disease (ex exercise, diet, age, and stress) we might be able to conceive a better model.