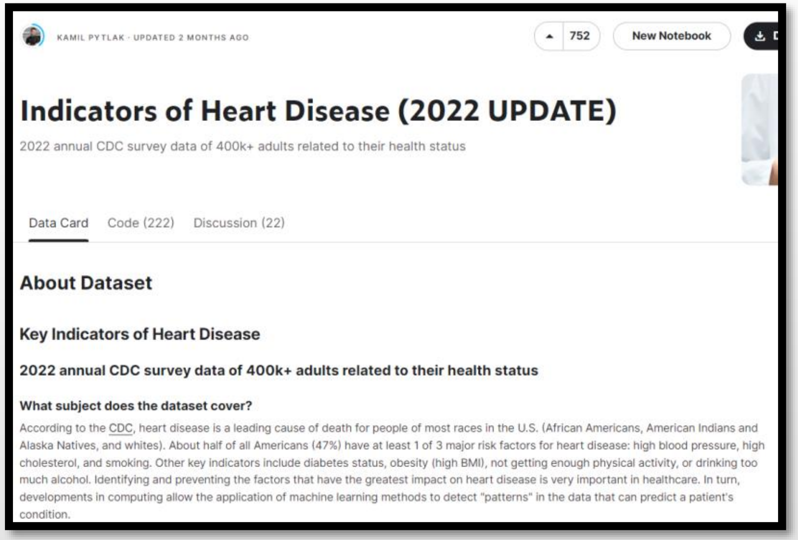
Understanding Heart Disease:

Can You Predict Who Gets It?

OLIVER KISZA, SPENCER AUSLANDER, VIGNESH CHERIATH, BRANDON MATA, FEMI ADEMUWAGUN, CHRIS MANFREDI

Data Retrieval, Multivariate Regression

- Data pulled from Kaggle data sets (data originates from CDC)
- Dataset originally comes from the CDC and is a major part of the Behavioral Risk Factor Surveillance System (BRFSS)
- Initially looked at both data sets (2020, 2022) but went with 2020 due to data simplicity and usage of Heart Disease (Yes/No) as dependent variable
- Initial Data Cleaning was required to use SPSS as our initial data variable understanding
- Regressions were not very powerful but did yield ideas around important variables to use and look through in the population



Data Retrieval, Multivariate Regression

2020 – Heart Disease

ed/Removed[a]

Model	Variables Entered
1	GeneralHealth
2	Stroke
3	Diabetic
4	KidneyDisease
5	DifficultyWalking
	Sex
	ancer
	oking
	Days
	Days
	BMI
	nking
	Sleep
	thma
	vities

Model Summary[p]

Model	R	R Square
1	.254[a]	0.065
2	.307[b]	0.094
3	.328[c]	0.107
4	.340[d]	0.115
5	.348[e]	0.121
6	.357[f]	0.127
7	.363[g]	0.132
8	.367[h]	0.134
9	.368[i]	0.135
10	.369[j]	0.136
11	.370[k]	0.137
12	.370[l]	0.137
13	.370[m]	0.137
14	.371[n]	0.137
15	.371[o]	0.137

a. Predictors: (Constant), GeneralHealth

2022 – High Risk

Variables Entered/Removed[a]

Model	Variables Entered	Variables Re
1	HIVTesting	
2	DifficultyConcentrating	
3	AlcoholDrinkers	
4	HadArthritis	
5	HadDepressiveDisorde	
6	FluVaxLast12	
7	Sex	
	lking	
	ncer	
	rands	
	OfHearing	
	lth	
	Ever	
	ivities	
	essingBathi	
	tack	
	onDifficulty	

Model Summary[x]

Model	R	R Square
1	.130[a]	0.017
2	.151[b]	0.023
3	.171[c]	0.029
4	.182[d]	0.033
5	.191[e]	0.036
6	.197[f]	0.039
7	.201[g]	0.040
8	.204[h]	0.041
9	.205[i]	0.042
10	.206[j]	0.042
11	.207[k]	0.043
12	.207[l]	0.043
13	.208[m]	0.043
14	.208[n]	0.043
15	.209[o]	0.044
16	.209[p]	0.044
17	.209[q]	0.044
18	.210[r]	0.044
19	.210[s]	0.044
20	.210[t]	0.044
21	.210[u]	0.044
22	.210[v]	0.044
23	.210[w]	0.044

a. Predictors: (Constant), HIVTesting

2022 – Heart Attack

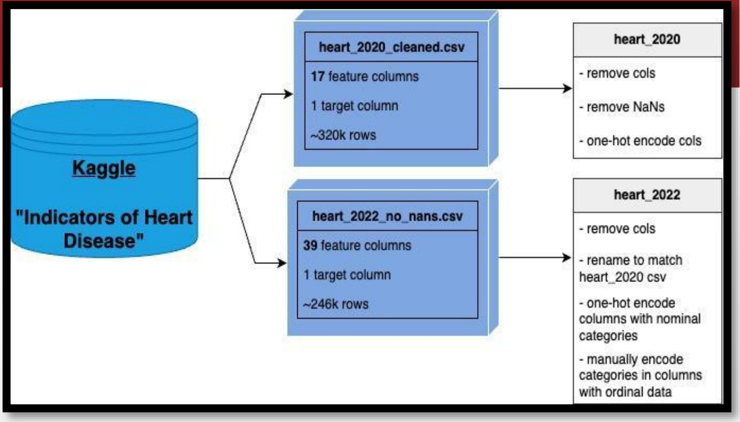
Variables Entered/Remo

Model	Variables Entered
1	HadAngina
2	HadStroke
3	GeneralHealth
4	ChestScan
5	Sex
6	DifficultyWalking
	PD
	Drinkers
	VaxEver
	HardOfHearing
	neyDisease
	lActivities
	VisionDifficulty
	hma
	ritis
	pressiveDisorder
	kLastYear
	ck

Model Summa

Model	R	R Square
1	.455[a]	0.207
2	.470[b]	0.221
3	.477[c]	0.227
4	.481[d]	0.231
5	.483[e]	0.233
6	.485[f]	0.235
7	.486[g]	0.236
8	.486[h]	0.236
9	.487[i]	0.237
10	.487[j]	0.237
11	.488[k]	0.238
12	.488[l]	0.238
13	.488[m]	0.238
14	.488[n]	0.238
15	.488[o]	0.238
16	.488[p]	0.239
17	.488[q]	0.239
18	.488[r]	0.239

Python Data Cleaning



2 Datasets: 2020 And 2022(updated) Annual CDC Survey Data Of 400,000+ Adults

original file (heart_2020_cleaned.csv):

- 17 feature columns (heart disease indicators)
- 1 target column (HeartDisease)
- ~320,000 rows

Cleaned file (2020_cleaned-~310,000 rows), python file on github-FA_data_cleaning_2020.ipynb:

- Renamed columns to match identical columns in heart_2022 with slightly different names(ex: GenHealth:GeneralHealth; PhysicaHealth: PhysicalHealthDays; DiffWalking:DifficultyWalking,etc.)
- Renamed other columns for clarity such as SleepTime:HoursOfSleep
- Removed rows with ambiguous data: Diabetic (Yes, during pregnancy, No, borderline diabetes) to yield binary column
- Created dummy variables ideal for binary categories
- Mapped ordinal variables from least to greatest starting from 0 (GeneralHealth and AgeCategory)
- Further analysis with mapped and dummy values
- 2020 Data Cleaning Code

original file (heart_2022_no_nans.csv):

- 39 feature columns (heart disease indicators)
- 1 target column (“HadHeartAttack”)
- ~246,000 rows

Cleaned file (cleaned_df), python file on github- data_cleaning_2022.ipynb:

- Removed columns (“State”, “HadDiabetes”, etc)
- Renamed other columns for clarity such as “SleepTime”: “HoursOfSleep”
- Removed NaN
- Created dummy variables ideal for binary categories (“Sex”, “PhysicalActivities”, “HadHeartAttack”, “HadAngina”, etc)
- Mapped ordinal variables from least to greatest starting from 0 (“GeneralHealth”, “AgeCategory”, “LastCheckupTime”, “RemovedTeeth”, “SmokerStatus”, and “ECigaretteUsage”)
- 2022 Data Cleaning Code

Analysis On Python Using Mapped/Dummy Variables Using Linear Regression Model

	Heart Disease Indicators	r-value
0	GeneralHealth	-0.24471116023984052
1	AgeCategory	0.23425262301744743
2	BMI	0.052666342464889424
3	PhysicalHealthDays	0.17167977114662242
4	MentalHealthDays	0.028234899261603143
5	HoursOfSleep	0.009221458092622849
6	Smoking	0.10811849761589694

	Heart Disease Indicators	r-value
0	Diabetic	0.18696346633644118
1	PhysicalActivities	-0.10089330845766817
2	Asthma	0.04108739544718814
3	SkinCancer	0.09320241203489989
4	AlchoholDrinking	-0.032941641681676424
5	Stroke	0.19801184782142203
6	DifficultyWalking	0.20292116027668886
7	KidneyDisease	0.14542892385601908

GeneralHealth	Map	GeneralHealth
Excellent		4
Very Good		3
Good		2
Fair		1
Poor		0

LastCheckupTime	Map	LastCheckupTime
Within past year (anytime less than 12 months ago)		0
Within past 2 years (1 year but less than 2 years ago)		1
Within past 5 years (2 years but less than 5 years ago)		2
5 or more years ago		3

RemovedTeeth	Map	RemovedTeeth
None of them		0
1 to 5		1
6 or more, but not all		2
All		3

SmokerStatus	Map	SmokerStatus
Never smoked		0
Former smoker		1
Current smoker - now smokes some days		2
Current smoker - now smokes every day		3

ECigaretteUsage	Map	ECigaretteUsage
Never used e-cigarettes in my entire life		0
Not at all (right now)		1
Use them some days		2
Use them every day		3

AgeCategory	Map	AgeCategory
Age 18 to 24		0
Age 25 to 29		1
Age 30 to 34		2
Age 35 to 39		3
Age 40 to 44		4
Age 45 to 49		5
Age 50 to 54		6
Age 55 to 59		7
Age 60 to 64		8
Age 65 to 69		9
Age 70 to 74		10
Age 75 to 79		11
Age 80 or older		12

	Heart Disease Indicators	r-value
1	GeneralHealth	-0.244711
2	AgeCategory	0.234253
3	DifficultyWalking	0.202921
4	Stroke	0.198012
5	Diabetic	0.186963
	PhysicalHealthDays	0.171680
	KidneyDisease	0.145429
	Smoking	0.108118
	PhysicalActivities	-0.100893
	SkinCancer	0.093202
	BMI	0.052666
	Asthma	0.041087
	AlchoholDrinking	-0.032942
	MentalHealthDays	0.028235
	HoursOfSleep	0.009221

Spark Ingestion, Data Understanding, Exporting Model

- Reading our cleaned data file from GitHub

```
import pandas as pd

heartdata2020 = 'https://raw.githubusercontent.com/oliverkisza/Final-Project-Team-1/main/Resources/2020_cleaned.csv'

pd_df = pd.read_csv(heartdata2020)
heart2020_df = spark.createDataFrame(pd_df)

heart2020_df.limit(5).show()
```

```
+------------+-----+-------+--------------+------+------------------+----------------+---------------+------+-----------
|HeartDisease|  BMI|Smoking|AlcoholDrinking|Stroke|PhysicalHealthDays|MentalHealthDays|DifficultyWalking|   Sex|AgeCategory|
+------------+-----+-------+--------------+------+------------------+----------------+---------------+------+-----------
|          No| 16.6|    Yes|            No|    No|               3.0|            30.0|             No|Female|      55-59|
|          No|20.34|     No|            No|   Yes|               0.0|             0.0|             No|Female|80 or older|
|          No|26.58|    Yes|            No|    No|              20.0|            30.0|             No|  Male|      65-69|
|          No|24.21|     No|            No|    No|               0.0|             0.0|             No|Female|      75-79|
|          No|23.71|     No|            No|    No|              28.0|             0.0|            Yes|Female|      40-44|
+------------+-----+-------+--------------+------+------------------+----------------+---------------+------+-----------
```

- Data querying to get a better understanding of the data

```
a2020q1 = """
SELECT
  HeartDisease,
  COUNT(*) AS TOTAL,
  ROUND(COUNT(CASE WHEN Smoking = 'Yes' THEN Smoking END) / TOTAL * 100,2) AS PERCENT_SMOKING,
  ROUND(COUNT(CASE WHEN AlcoholDrinking = 'Yes' THEN AlcoholDrinking END) / TOTAL * 100,2) AS PERCENT_DRINKERS,
  ROUND(COUNT(CASE WHEN Stroke = 'Yes' THEN Stroke END) / TOTAL * 100,2) AS PERCENT_STROKE,
  ROUND(COUNT(CASE WHEN Diabetic = 'Yes' THEN Diabetic END) / TOTAL * 100,2) AS PERCENT_DIABETIC,
  ROUND(COUNT(CASE WHEN Asthma = 'Yes' THEN Asthma END) / TOTAL * 100,2) AS PERCENT_ASTHMA,
  ROUND(COUNT(CASE WHEN KidneyDisease = 'Yes' THEN KidneyDisease END) / TOTAL * 100,2) AS PERCENT_KIDNEY_DISEASE,
  ROUND(COUNT(CASE WHEN SkinCancer = 'Yes' THEN SkinCancer END) / TOTAL * 100,2) AS PERCENT_SKIN_CANCER
FROM heart20
Group by HeartDisease
ORDER BY HeartDisease DESC
"""
spark.sql(a2020q1).show()
```

```
+------------+------+---------------+----------------+--------------+----------------+--------------+----------------------
|HeartDisease| TOTAL|PERCENT_SMOKING|PERCENT_DRINKERS|PERCENT_STROKE|PERCENT_DIABETIC|PERCENT_ASTHMA|PERCENT_KIDNEY_DISEASE|
+------------+------+---------------+----------------+--------------+----------------+--------------+----------------------
|         Yes| 26584|          58.66|            4.16|         16.07|           33.69|         17.89|                 12.65|
|          No|286430|          39.58|            7.09|           2.6|           11.12|          12.9|                  2.84|
+------------+------+---------------+----------------+--------------+----------------+--------------+----------------------
```

- Exporting our data as a csv for modeling

```
[11] heart2020_df.toPandas().to_csv('2020_cleaned.csv')
```

Initial Model Creation, Running Into Issues With Precision and Prediction

```
new_df = df.drop(['AgeCategory', 'SleepHours', 'HeightInMeters', "WeightInKilograms", 'BMI','H
                 'CovidPos_No','PneumoVaxEver_Yes','HadAngina_Yes', 'HadAngina_No'], axis=1)
```

- First model run with 2022 data
- “Had HeartAttack” as our target
- Removed columns that were not binary

features, for better optimization

- Classification reports presented a poor precision

percentage for predicting Yes for Heart Attack

```
y = df['HadHeartAttack_Yes']
X = df.drop(columns=['HadHeartAttack_Yes', 'HadHeartAttack_No'])
X_train, X_test, y_train, y_test = train_test_split(X,y,random_state = 42, stratify=y)
```

```
# Create and save the testing classification report
testing_report = classification_report(y_test, testing_predictions)

# Print the testing classification report
print(testing_report)
✓ 0.0s

              precision    recall  f1-score   support

           0       0.95      0.99      0.97     37015
           1       0.59      0.26      0.36      2741

    accuracy                           0.94     39756
   macro avg       0.77      0.62      0.66     39756
weighted avg       0.92      0.94      0.92     39756
```

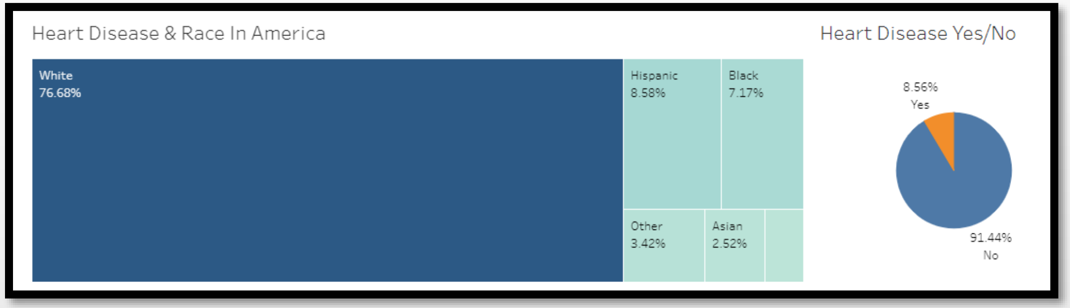
Model Problems, Reasoning For Model Changes

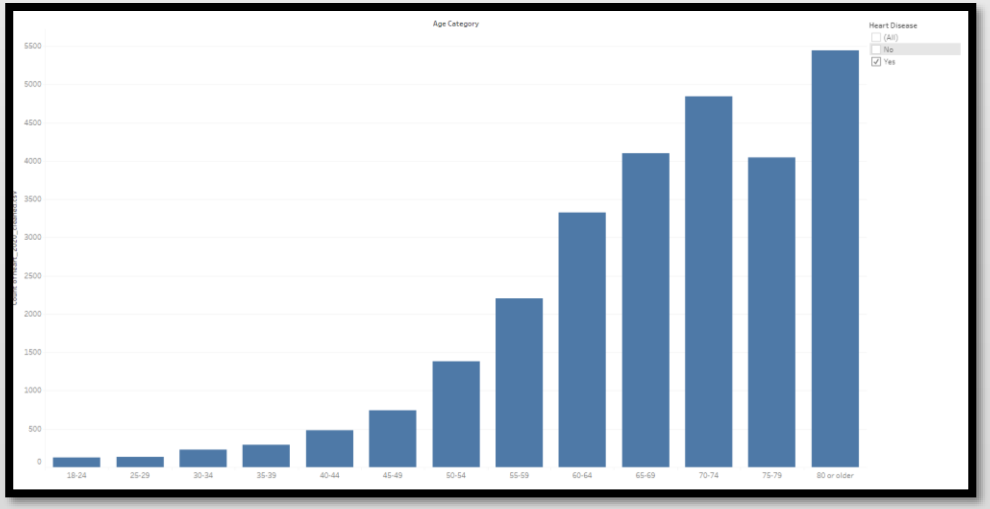
Initial Model Issues

- Strong when predicting who doesn't have heart disease
- Weak when predicting who does!
- Opposite of what you want

Why?

- Data was very skewed (90% didn't have heart disease)
- Model didn't have enough data to identify patients with heart disease





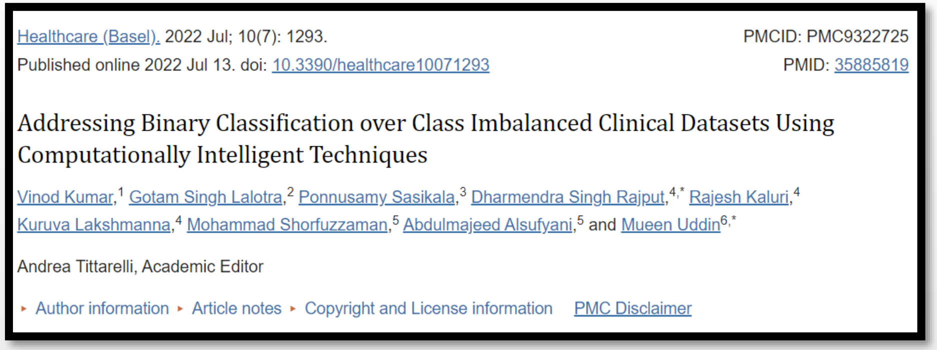
Understanding Issues With Health Care Data

This Is A Common Problem With Heart Disease

- We observed this issue 2020, 2022, other data sources
- Makes sense: most people don't actively have heart disease
- Heart disease seems binary but is likely not

Still A Problem!

- Clearly, our model needs work
- Not good enough to only identify who's healthy, we need to know who's sick
- We iterated through multiple common tactics to refine the model



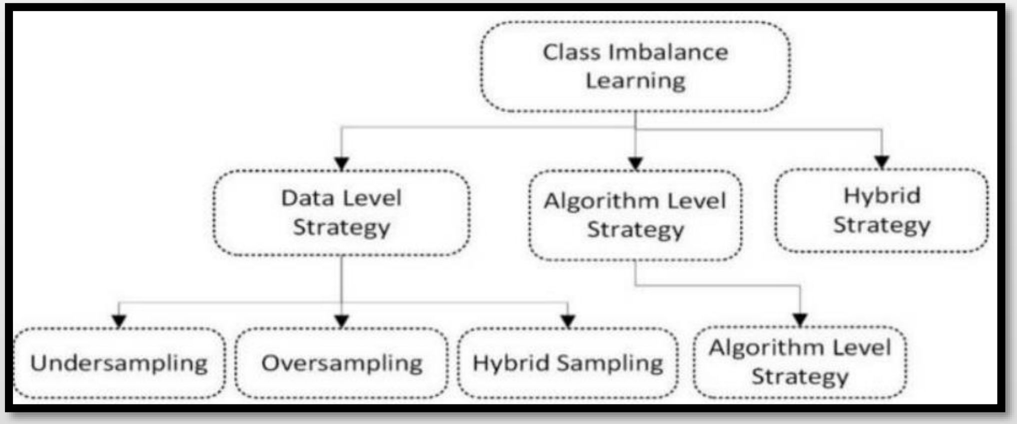
Healthcare (Basel). 2022 Jul; 10(7): 1293. PMCID: PMC9322725
Published online 2022 Jul 13. doi: 10.3390/healthcare10071293 PMID: 35885819

Addressing Binary Classification over Class Imbalanced Clinical Datasets Using Computationally Intelligent Techniques

Vinod Kumar,[1] Gotam Singh Lalotra,[2] Ponnusamy Sasikala,[3] Dharmendra Singh Rajput,[4,*] Rajesh Kaluri,[4] Kuruva Lakshmanna,[4] Mohammad Shorfuzzaman,[5] Abdulmajeed Alsufyani,[5] and Mueen Uddin[6,*]

Andrea Tittarelli, Academic Editor

▸ Author information ▸ Article notes ▸ Copyright and License information PMC Disclaimer



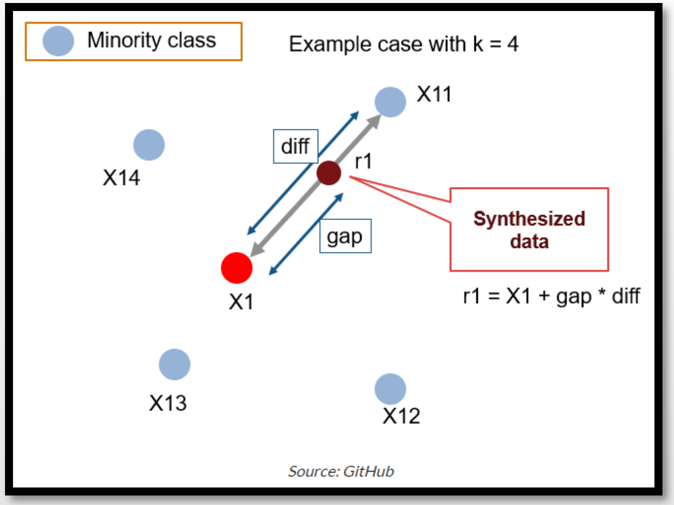
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9322725/

Model Optimization – SMOTE (Synthetic Minority Oversampling Technique)

- There's An Imbalance Ratio Of No Responses For Heart Disease To Yes Responses. Causes The Accuracy Metric To Be Biased And Not Preferable.

- SMOTE Alters The Training Set By Increasing The Number Of Yes Data Points To Match The Volume Of No Data Points.

- Creates Synthetic Samples By Taking A Random Instance Of The Minority Class, Finding Its K-nearest Neighbors And Placing An New Instance At A Random Distance Between.




```
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=78)
```

```
smt = SMOTE()
```

```
counter = Counter(y_train)
print(counter)
```

```
Counter({0.0: 219270, 1.0: 20576})
```

```
X_train_sm, y_train_sm = smt.fit_resample(X_train, y_train)
```

```
counter = Counter(y_train_sm)
print(counter)
```

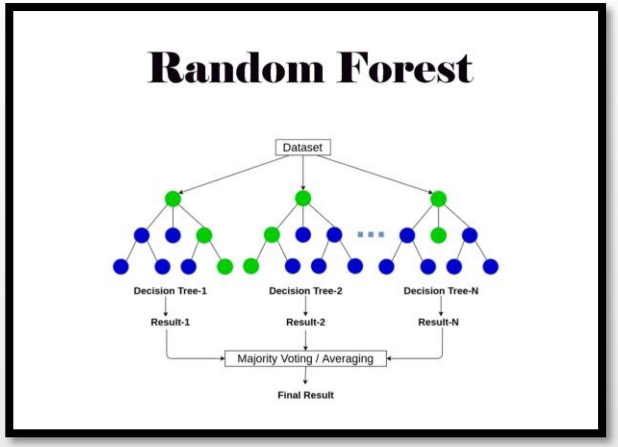
```
Counter({0.0: 219270, 1.0: 219270})
```

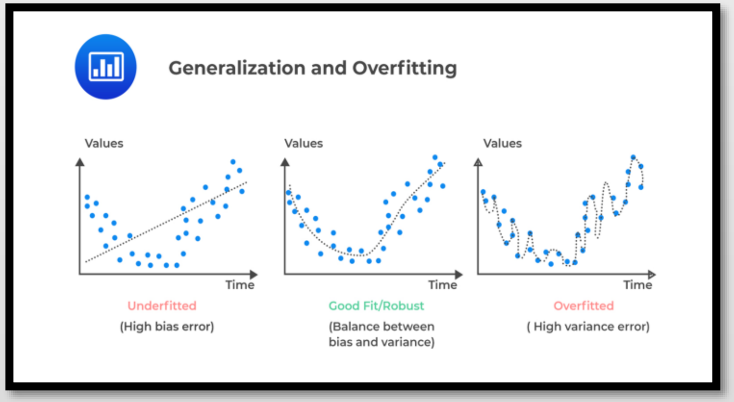
https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Model Optimization – Random Forest

•Random Forests is an ensemble learning method. Instead of one complex decision tree, it samples the data and constructs a multitude of simple decision trees.

•Unlike normal decision trees, Random Forests is robust against overfitting. SMOTE runs the risk of introducing noisy instances and overfitting problems.






```
rf_model = RandomForestClassifier(n_estimators=500, random_state=1567)

rf_model = rf_model.fit(X_train_scaled, y_train_sm)
```

Model Optimization: SMOTE + Random Forest Results

- Through SMOTE + Random Forests, the model classification report had low precision and high recall for the population with heart disease, which is preferable.

- Low precision: the model is incorrectly predicting heart disease for people who don't have it; many false positives.

- High recall: for the population that does have heart disease, the model is correctly identifying a majority of them; few false negatives.

- False positives will lead to further testing and safety precautions. A high number of false negatives may result in heart disease being untreated which might have serious consequences.

SMOTE

Accuracy Score : 0.7500656668626249

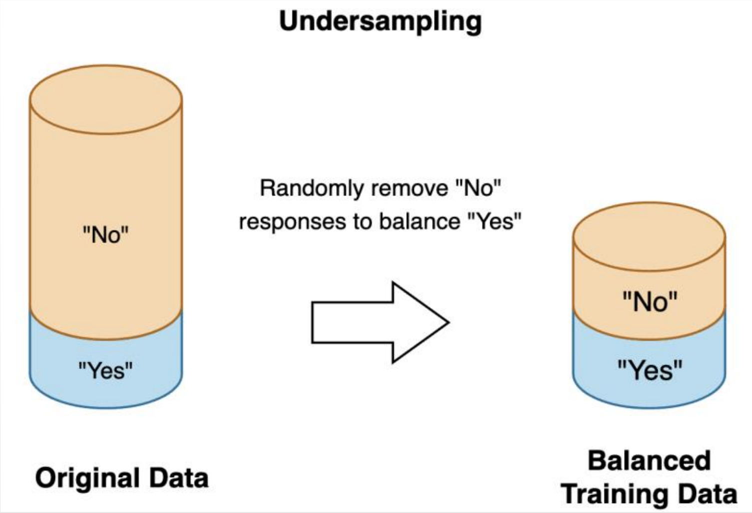
Classification Report

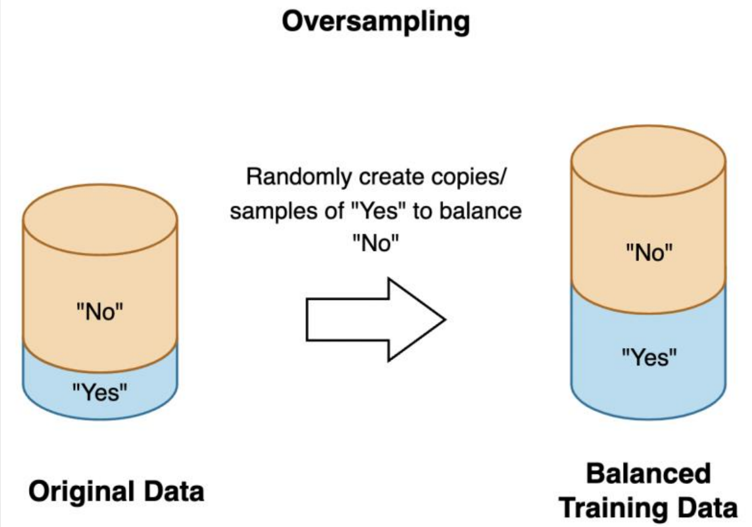
	precision	recall	f1-score	support
0.0	0.96	0.76	0.85	73152
1.0	0.20	0.64	0.30	6797
accuracy			0.75	79949
macro avg	0.58	0.70	0.58	79949
weighted avg	0.89	0.75	0.80	79949

Model Optimization: Undersampling

- How does it work?
  - "Opposite" of oversampling
- Undersampling model
  - Results & classification report

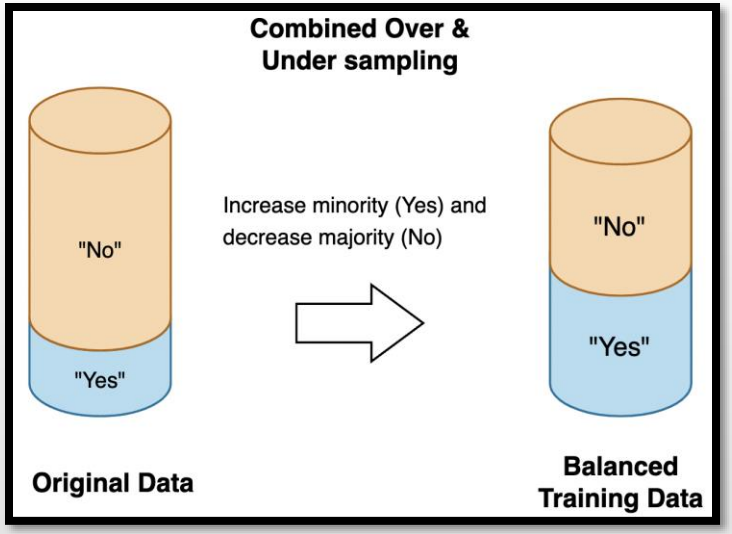
	precision	recall	f1-score	support
False	0.96	0.78	0.86	170404
True	0.21	0.62	0.32	15869
accuracy			0.77	186273
macro avg	0.58	0.70	0.59	186273
weighted avg	0.89	0.77	0.82	186273





Model Optimization: Combining Over & Undersampling

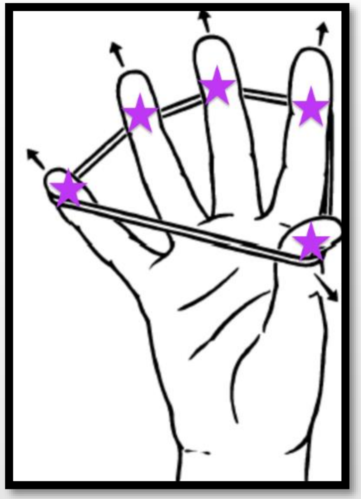
- Goal: “Balance” Bias Created By Either Method
    - Oversampling might add non-useful data
    - Undersampling might remove useful data
- Results & Classification Report
    - Why didn’t this help much?
    - Traded off precision for recall
- Which Model Is “Best”?
    - Model became very sensitive

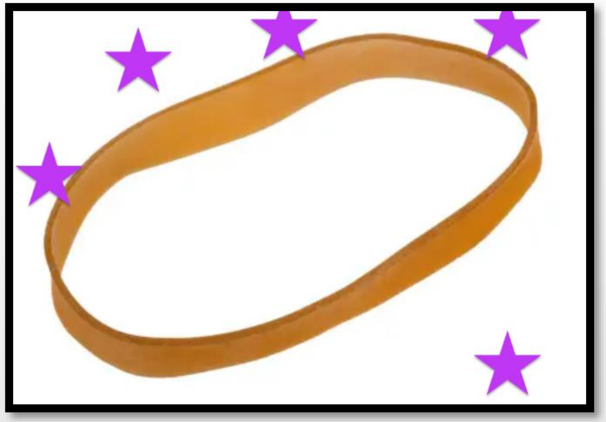


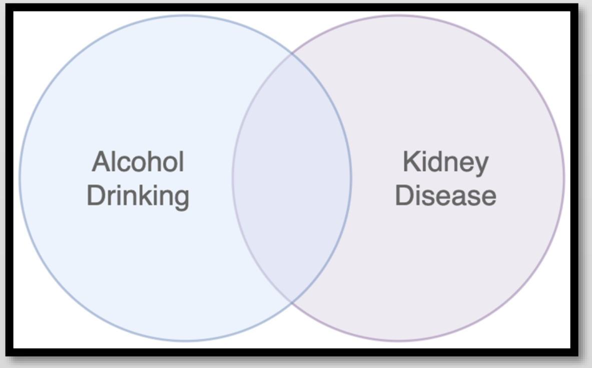
	precision	recall	f1-score	support
0	0.96	0.74	0.84	141962
1	0.19	0.67	0.30	13266
accuracy			0.73	155228
macro avg	0.58	0.70	0.57	155228
weighted avg	0.89	0.73	0.79	155228

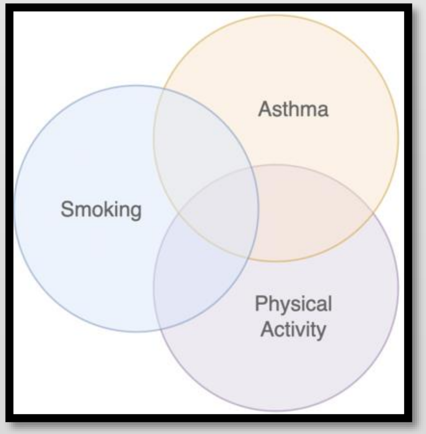
Model Optimization: Combining Over & Undersampling

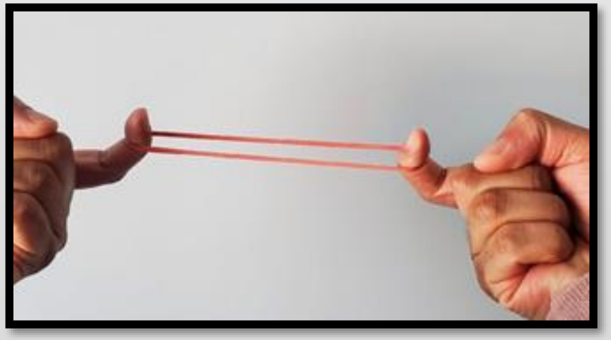
- Try different resampling methods
  - Rubber band analogy for weights
- Some feature columns capture overlapping information
  - Leads to unintentionally over/ under weighting variables



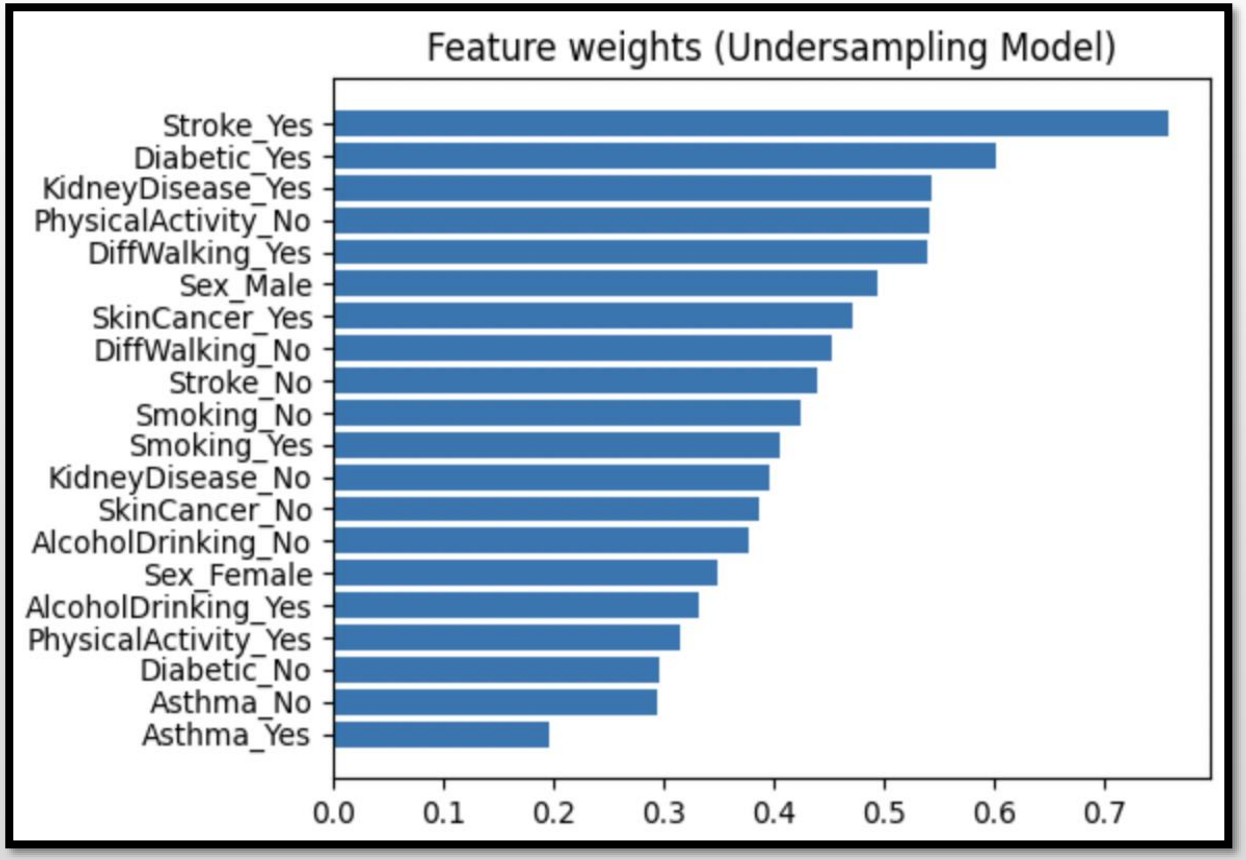


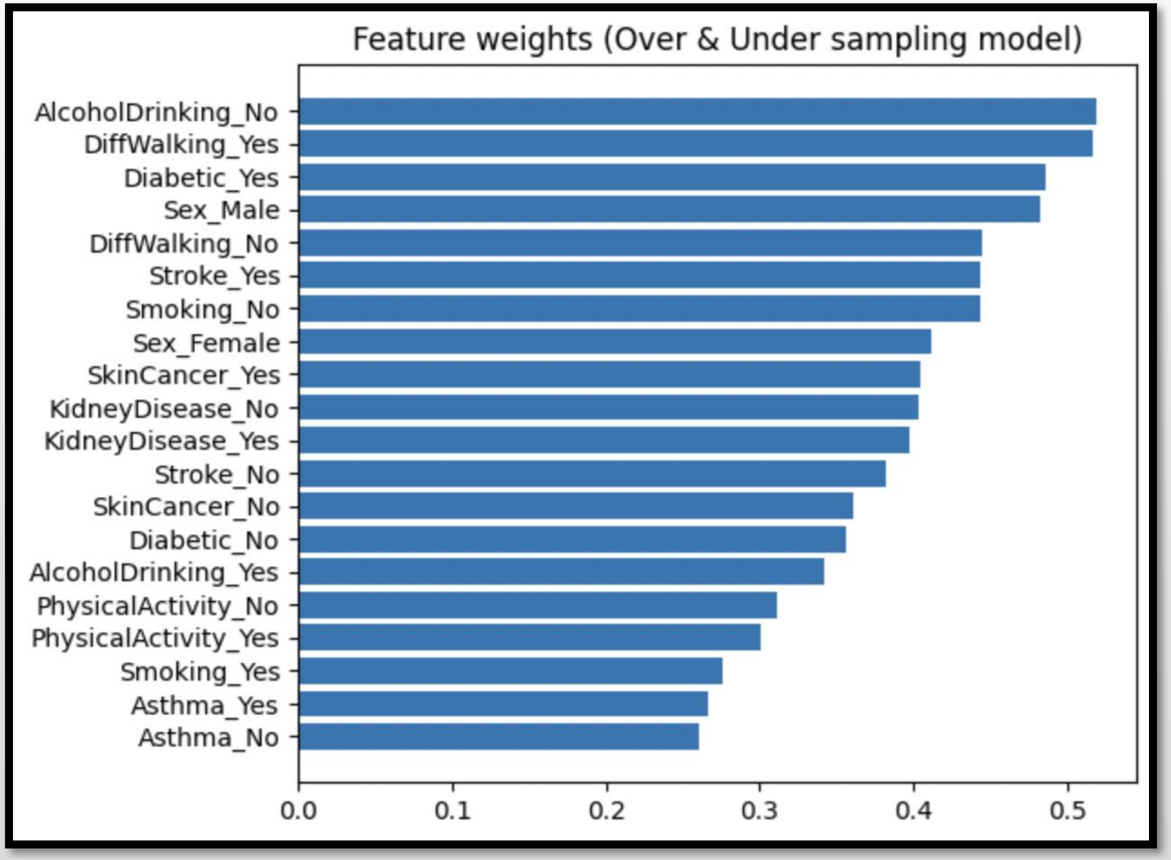


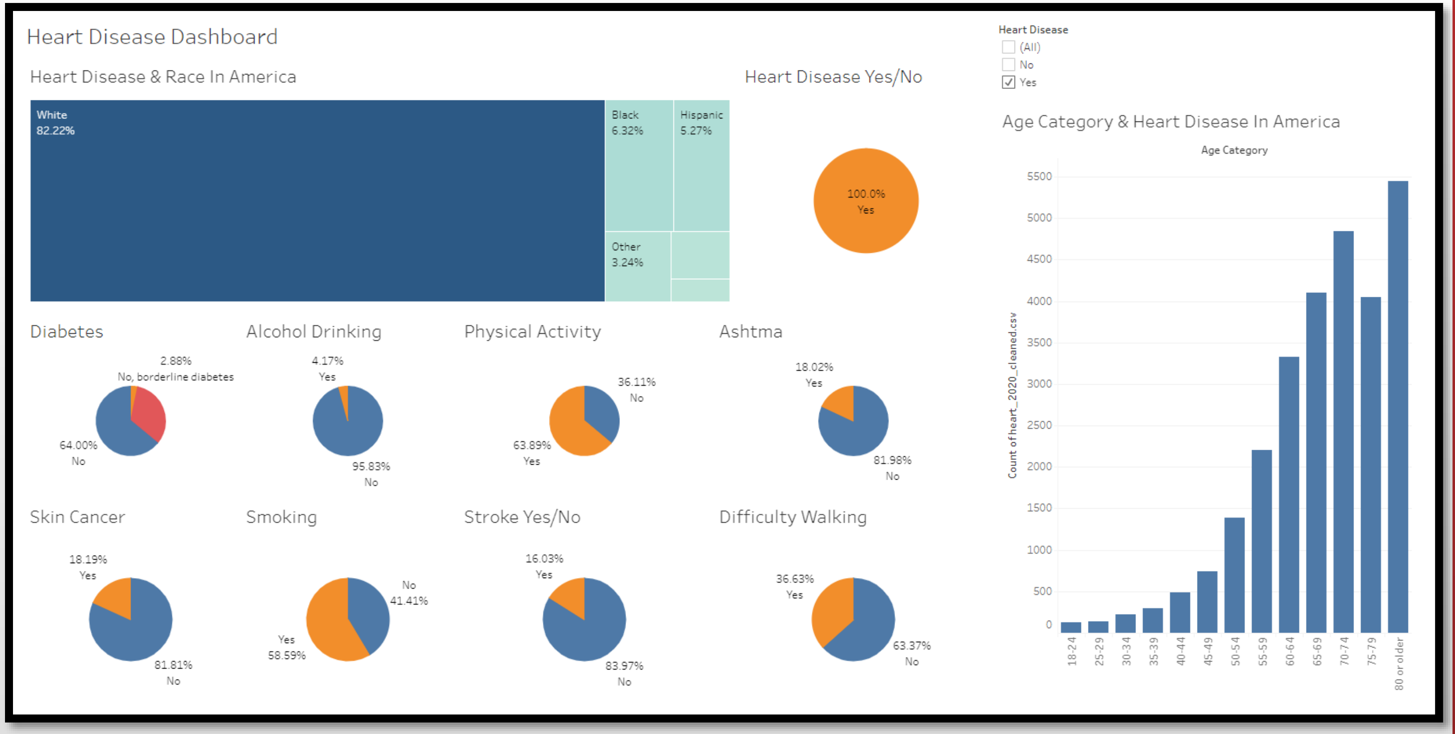




Further Model Improvements






Heart Disease Dashboard
Heart Disease & Race In America
White
82.22%
Black
6.32%
Hispanic
5.27%
Other
3.24%
Heart Disease Yes/No
100.0%
Yes
Heart Disease
(All)
No
Yes
Age Category & Heart Disease In America
Age Category
Count of heart_2020_cleaned.csv
Diabetes
2.88%
No, borderline diabetes
64.00%
No
Alcohol Drinking
4.17%
Yes
95.83%
No
Physical Activity
36.11%
No
63.89%
Yes
Ashtma
18.02%
Yes
81.98%
No
Skin Cancer
18.19%
Yes
81.81%
No
Smoking
No
41.41%
Yes
58.59%
Stroke Yes/No
16.03%
Yes
83.97%
No
Difficulty Walking
36.63%
Yes
63.37%
No
18-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80 or older

Thank You!