Project 4: Machine Learning

Presentation by Ethan Clark and Riley Sayre

Project Outline

The Problem: Insurance companies need help gauging risk when writing plans to avoid overspending.

Solution Proposal: Implement a machine learning model that can help predict if a member will become a high-cost claimant.

Results: Did the algorithm work?

The Problem

“High-cost claimants" are employees or covered dependents who cost their health plans $50,000 or more a year.

Solution Proposal

Create a machine learning algorithm that can help underwriters assess risk in order to more accurately predict occurrences of High Cost Claimants.

Sourcing the data

PatientID	age	gender	bmi	bloodpressure	diabetic	children	smoker	region	claim
1	39	male	23.2	91	Yes	0	No	southeast	1121.87
2	24	male	30.1	87	No	0	No	southeast	1131.51
8	19	male	41.1	100	No	0	No	northwest	1146.8
9	20	male	43	86	No	0	No	northwest	1149.4
10	30	male	53.1	97	No	0	No	northwest	1163.46
11	36	male	19.8	88	Yes	0	No	northwest	1241.57
12	37	male	20.3	90	Yes	0	No	northwest	1242.26
13	19	male	20.7	81	No	0	No	northwest	1242.82
17	35	male	34.1	90	No	0	No	southwest	1261.44
18	41	male	34.4	84	No	0	No	southwest	1261.86
19	49	male	35.4	97	Yes	0	No	southwest	1263.25
20	48	male	33.3	91	Yes	0	No	southeast	1391.53
21	45	male	23.2	85	Yes	0	No	southeast	1515.34
22	34	male	31.1	96	No	0	No	southwest	1526.31
23	18	male	35.5	100	Yes	0	No	southeast	1532.47
24	42	male	36.9	93	No	0	No	southeast	1534.3

Assigning dummy variables

	age	bmi	bloodpressure	children	claim	female	male	Diabetic? _No	Diabetic? _Yes	Smoker? _No	Smoker? _Yes	northeast	northwest	southeast
0	39	23.2	91	0	1121	0	1	0	1	1	0	0	0	1
1	24	30.1	87	0	1131	0	1	1	0	1	0	0	0	1
2	19	41.1	100	0	1146	0	1	1	0	1	0	0	1	0
3	20	43.0	86	0	1149	0	1	1	0	1	0	0	1	0
4	30	53.1	97	0	1163	0	1	1	0	1	0	0	1	0

Clustering data into bins

3 Cluster Bin

	Low Risk	Medium Risk	High Risk
Claims Under	9412.5	41350.8	inf
Claims Between	0 - 9412.5	9412.5 - 41350.8	41350.8 - inf
Claims Over	0	9412.5	41350.8
Cluster #	1	2	3

2 Cluster Bin

	Low Risk	High Risk
Claims Under	9412.5	inf
Claims Between	0 - 9412.5	9412.5 - inf
Claims Over	0	9412.5
Cluster #	1	2

Clustering Results

2 Cluster Bin

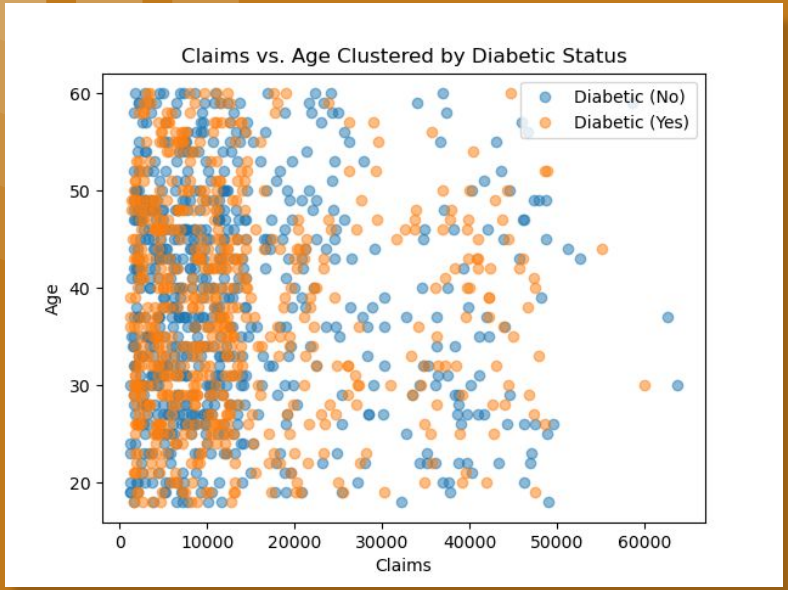
	precision	recall	f1-score	support
1	0.77	0.90	0.83	173
2	0.87	0.69	0.77	160
3	0.00	0.00	0.00	0
accuracy			0.80	333
macro avg	0.55	0.53	0.53	333
weighted avg	0.82	0.80	0.80	333

	Low Risk	High Risk
Claims Under	9412.5	inf
Claims Between	0 - 9412.5	9412.5 - inf
Claims Over	0	9412.5
Cluster #	1	2

3 Cluster Bin

	precision	recall	f1-score	support
1	0.77	0.90	0.83	173
2	0.75	0.66	0.70	143
3	0.50	0.12	0.19	17
accuracy			0.76	333
macro avg	0.67	0.56	0.58	333
weighted avg	0.75	0.76	0.74	333

	Low Risk	Medium Risk	High Risk
Claims Under	9412.5	41350.8	inf
Claims Between	0 - 9412.5	9412.5 - 41350.8	41350.8 - inf
Claims Over	0	9412.5	41350.8
Cluster #	1	2	3



Using Diabetes as a factor of HCC

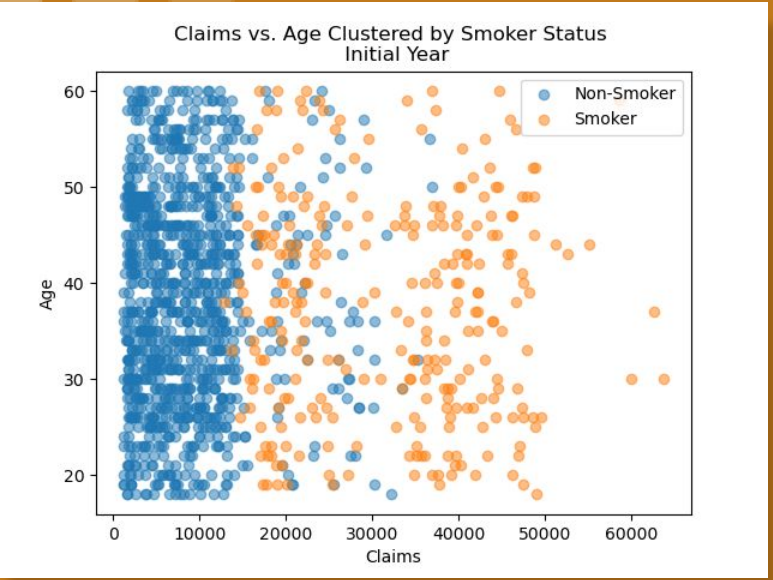
Our Model's results in numbers

	precision	recall	f1-score	support
0	0.48	0.73	0.58	165
1	0.45	0.22	0.30	168
accuracy			0.47	333
macro avg	0.46	0.47	0.44	333
weighted avg	0.46	0.47	0.44	333

Average claim for non-diabetic vs. diabetic

Diabetic?_Yes	
0	13406.117986
1	13235.990581

An accuracy of 47% when constrained by HIPAA and the lack of underwriting information is still an astounding rate and may help anyone with proper medical insurance administration via suggestions for testing.



Using Smoking as a factor to predict HCC

Our Model's results in numbers

	precision	recall	f1-score	support
0	0.92	0.96	0.94	255
1	0.87	0.74	0.80	78
accuracy			0.91	333
macro avg	0.90	0.85	0.87	333
weighted avg	0.91	0.91	0.91	333

Average claim for non-smoker vs. smoker

Smoker?_Yes	
0	8475.379962
1	32049.729927

This model is very accurate with a balanced accuracy of 91% with very few features involved.

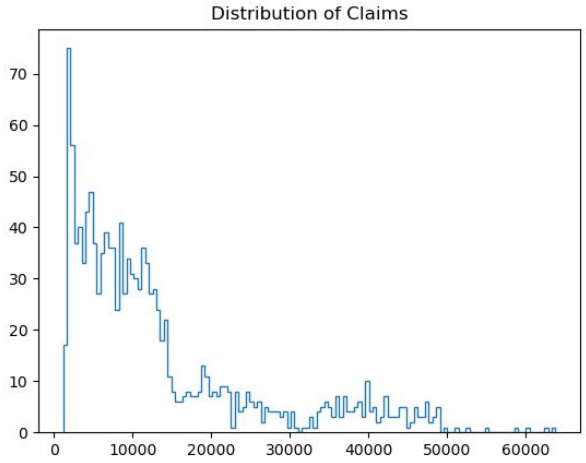
Bonus Application

The Problem: Write an insurance renewal for these claimants!

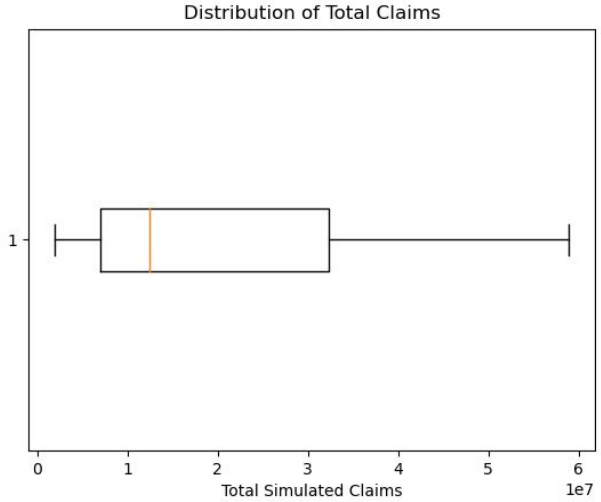
Solution Proposal: Implement a Monte Carlo Method Convoluted with a Mixed Gaussian Distribution for claims and simulate claimant lives one year forward

Results: Yes! We have a reasonable estimate for monthly premiums!

Code Time!



100 Simulations




```
# Retrieving expected claims
Estimated_Cost = np.quantile(Totals_df['Total'],0.9)
print(Estimated_Cost)
```

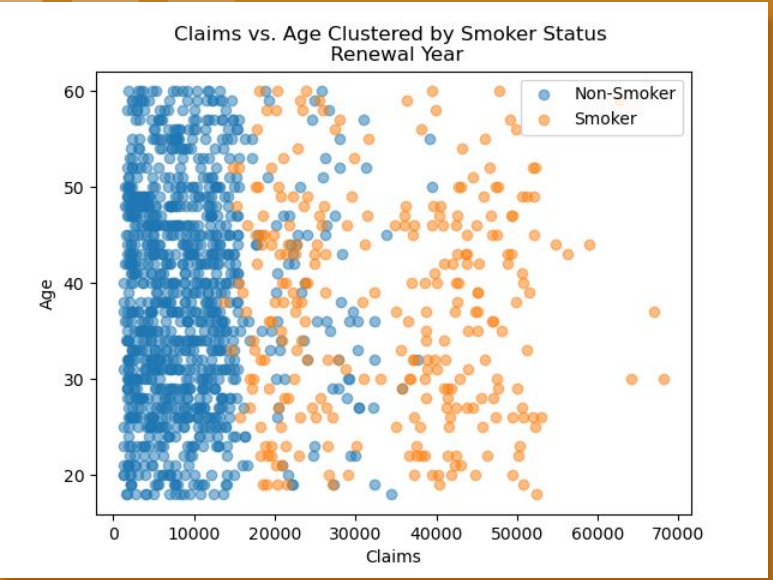
```
38928667.281243764
```

```
# Need to turn a profit, was
Budget = Estimated_Cost*1.15
print(Budget)
```

```
44767967.37343033
```

```
# Now to calculate premiums
Num_Subs = len(renewal_df['age'])
Months = 12
# Now Including the average medical cost growth trend
Monthly_Premium = Budget/Num_Subs/Months*1.07
Monthly_Premium
```

```
2996.854672770924
```



Using Smoking as a factor to predict HCC. (Pt. 2)

Our Model's results in numbers

	precision	recall	f1-score	support
0	0.93	0.95	0.94	255
1	0.83	0.77	0.80	78
accuracy			0.91	333
macro avg	0.88	0.86	0.87	333
weighted avg	0.91	0.91	0.91	333

Average claim for non-smoker vs. smoker

Smoker?_Yes	
0	9068.674858
1	34293.266423

This model is very accurate with a balanced accuracy of 91% with very few features involved. Great Bias example!

Questions?