Determinants and Predictions of Risks of Diseases in Middle Ages—Statistical Models versus Neural Network Machine Learning Models

Lakshmi K. Raut, Visiting Fellow, University of Chicago[1]

2021-02-19

The objective

- Identification of factors and prediction of the risks of diseases are important for public policies and disease diagnosis in healthcare.
- The biomedical literature suggests that much of an individual's later life health outcomes is programmed at early stages of life. The programming is strongly modulated by the epigentic inputs created by the environment in mother's womb at prenatal stage and also by the environment at early postnatal ages.
- The most important epigenetic factor is stress of any kind – psychological, financial, social and chemical. Other significant factors are medical care, diets, smoking, substance use, and exercising. These modulating factors are important throughout life, with stronger effects imparted in early stages of life.
- I use the popular statistical model – **mulinomial logit model** – and the neural network model – **multi-layer deep learning model** – for identification of important factors and prediction of the risks of various diseases for individuals at their middle ages. I use the Health and Retirement Studies (HRS) data to that end.
- I compare their predictive powers using **confusion matrix indicator** and discuss various criteria used by these two methods to identify the important factors of the risks. I discuss pros and cons of these two types of models from prediction and inference viewpoint.

Models

- Let y denote a random variable denoting an individual's health status in middle age. The set of health statuses is {1,2,3,4,5}, where 1 = normal health, 2 = Cardiovascular disease, 3 = Cancer, 4 = Other single disease, 5 = comorbid diseases.
- X is a vector of individual characteristics such as demographic information, education, medical measurements, known as **regressors**.
- Data on individuals, $(y_i, X_i), 1 = 1, 2, \ldots, n$.
- **Goal**:
- To find a model $y = f(X)$ that best explains data in the sample
- To predict the health status y of an individual not in the sample given his/her measurements X – i.e., **out of sample** prediction.
- To have inference about which factors in X are most important.
- How to judge which model is best or how to choose between models.
- This problem is known as **classification** problem. y could be a probability mass function. There is an equivalent regression problem.

Models … continued

- Two approaches to get a model:

1. **Statistical Models** – *Multinomial Logistic Regression* model in particular.
2. **Machine Learning** or **Algorithms** such as **Multi Layer Neural Network** model also known as **Deep Learning** model or *Deep Neural Network Model*, **Tree based methods** (such as CART, Random Forest etc), **Support Vector Machines**.

- I will only briefly explain the Multinomial Logistic Regression model and the Multi Layer Neural Network model.

Multinomial Logistic Regression

- Assume that data (y, X) is generated from a known probability distribution with the conditional probability distribution $f(y|X, \beta)$ belonging to a **known** family of distributions with parameters β such as the exponential family.
- Multinomial Logistic Regression model comes down to the following probability model.

$$log\frac{Prob\,(Y = k|X)}{Prob\,(Y = 1|X)} = X'\beta, \;\; k = 2, \ldots, 5$$

Logit Models of Childhood factors

	cHLTH		College+		Init.HLTH	
	(1)	(2)	(1)	(2)	(1)	(2)
Intercept	0.220***	0.162	-2.205***	-3.981***	-1.028***	-1.132***
	(0.053)	(0.091)	(0.087)	(0.153)	(0.063)	(0.098)
White	0.282***	0.221***	0.227**	-0.076	0.226***	0.193**
	(0.053)	(0.066)	(0.077)	(0.089)	(0.057)	(0.067)
Female	-0.022	-0.012	-0.537***	-0.571***	-0.218***	-0.177***
	(0.044)	(0.053)	(0.057)	(0.063)	(0.044)	(0.050)
Childhood SES	0.841***		1.328***		0.222***	
	(0.054)		(0.058)		(0.051)	
Father's Education		0.038***		0.109***		0.021*
		(0.009)		(0.011)		(0.009)
Mother's Education		0.029**		0.139***		-0.000
		(0.010)		(0.013)		(0.009)
Father's Job		0.447***		0.714***		0.042
		(0.096)		(0.084)		(0.078)
Childhood Health			0.342***	0.407***	0.206***	0.217***
			(0.064)	(0.077)	(0.048)	(0.057)
College					0.149*	0.137*
					(0.059)	(0.064)
R^2	0.026	0.018	0.085	0.136	0.010	0.008
Num. obs.	9601	7457	9601	7457	9601	7457

***p < 0.001; **p < 0.01; *p < 0.05

Statistical models

Estimates from the Multinomial Logistic Regression model of disease risks

	2-Cardiovas	3-Cancer	4-other	5-Comorbid
Intercept	-0.214	-5.014 ***	-0.759 ***	0.359 *
	(0.192)	(0.675)	(0.203)	(0.182)
white	-0.593 ***	0.465	0.306 ***	0.011
	(0.077)	(0.309)	(0.092)	(0.080)
female	-0.209 **	1.138 ***	0.520 ***	0.544 ***
	(0.065)	(0.230)	(0.069)	(0.064)
childhood SES	-0.045	-0.668	-0.137	-0.084
	(0.103)	(0.367)	(0.109)	(0.106)
childhood Health	0.103	0.006	-0.290 ***	-0.359 ***
	(0.073)	(0.228)	(0.072)	(0.067)
college+	0.066	0.343	-0.101	-0.118
	(0.083)	(0.248)	(0.090)	(0.088)
bmiH	0.712 ***	0.178	0.239 ***	0.818 ***
	(0.067)	(0.198)	(0.066)	(0.066)
CES-D	0.262	0.301	0.909 ***	1.454 ***
	(0.156)	(0.471)	(0.149)	(0.134)
cognitive scores	-0.001	0.033	0.003	-0.018 **
	(0.007)	(0.023)	(0.007)	(0.007)
smoking	0.044	0.321	0.201 **	0.330 ***
	(0.065)	(0.201)	(0.067)	(0.064)
exercising	-0.118	-0.343	-0.111	-0.585 ***
	(0.086)	(0.247)	(0.088)	(0.077)
AIC	22938.165	22938.165	22938.165	22938.165

*** p < 0.001; ** p < 0.01; * p < 0.05.

Deep Neural Network (Multi Layer Perceptron)

- Neural network is a highly parameterized universal function approximator of the form $\hat{y} = f(x; w)$, x is a set of inputs, and w is a vector of parameters– same form as a statistical model. The **problem** is to design a *neural network architecture* of the approximating function $y = f(x, w)$ and find a suitable *learning algorithm* to learn the parameter values w of the network using a training set of examples. This trained network can then be used to predict y for an individual given his characteristics x.
- The popularity and wide applicability of neural network lies in the fact that it designs the approximator in a hierarchy of functions, of the form:

$$\hat{y} = f(x; w) \equiv f^L_{w^L} \circ \ldots \circ f^1_{w_1}(x).$$

  Each function corresponds to a layer of artificial neurons.

Neural network (continued)

Consider a simple neural network architecture It has three layers – layer 0: input layer, layer 1: hidden layer, and layer 2: output layer. Last layer is denoted as L. Layer 0 has three input neurons. The second layer has 4 neurons. and the last layer has two neurons corresponding to the two output levels, in our case probability of two events, giving the hierarchical function specification of the form:

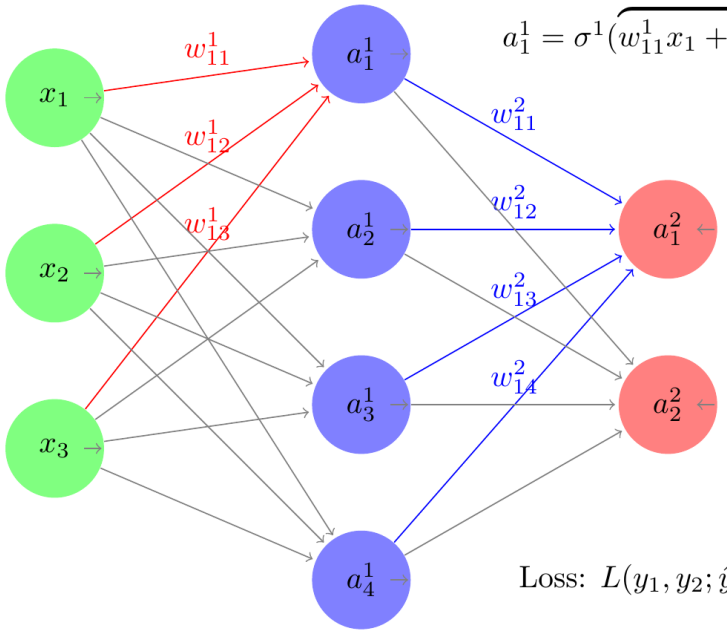
$$f(x; w) = \sigma^2\left(z^2\left(\sigma^1\left(z^1(x, w^1)\right), w^2\right)\right) \equiv f^2_{w^2} \circ f^1_{w^1}(x).$$

Function $z^i(a^i, w^i) = w^i \cdot a^{i-1}$ at each layer i is a linear aggregator. The function σ^i is a squashing function of the same dimension as z^i, known as *activation function*.

A neural network is called **deep neural network model** or **deep learning model** because it can have more than one hidden layer to get better approximation compared to the original neural network models with only one hidden layer.

Graph of a Deep Neural Network Model

0: Input layer 1: Hidden layer 2: Output layer



$$a_1^1 = \sigma^1(\overbrace{w_{11}^1 x_1 + w_{12}^1 x_2 + w_{13}^1 x_3 + b_1^1}^{z_1^1})$$

$$a_1^2 = \sigma^2(\overbrace{\sum_{j=1}^{4} w_{1j}^2 a_j^1 + b_1^2}^{z_1^2}) \equiv \hat{y}_1$$

$$a_2^2 = \sigma^2(\overbrace{\sum_{j=1}^{4} w_{2j}^2 a_j^1 + b_2^2}^{z_2^2}) \equiv \hat{y}_2$$

Loss: $L(y_1, y_2; \hat{y}_1, \hat{y}_2)$; $L = 2$ is the last layer

MLP architechture.

Predictive Accuracies

- First statistical model — Multinomial Logistic Regression Model
- Then Neural Network Model

Confusion matrix of the multinomial Logistic Regression Model (numbers)

Predicted/Actual	Normal	Cardiovas	Cancer	Other	Comorbid	Total
Normal	2017	215	0	4	516	2752
Cardiovas	1142	262	0	1	424	1829
Cancer	83	6	0	0	25	114
Other	1018	87	0	14	470	1589
Comorbid	1000	171	0	11	953	2135

Accuracy = 38.56 percent.

Confusion matrix of the multinomial Logistic Regression Model (in percent)

Predicted/Actual	Normal	Cardiovas	Cancer	Other	Comorbid
Normal	73.29	7.81	0	0.15	18.75
Cardiovas	62.44	14.32	0	0.05	23.18
Cancer	72.81	5.26	0	0.00	21.93
Other	64.07	5.48	0	0.88	29.58
Comorbid	46.84	8.01	0	0.52	44.64

Accuracy = 38.56 percent.

Confusion matrix of the deep learning model (numbers)

Confusion matrix of the deep learning model.

Predicted/Actual	normal	Cardiovas	Cancer	Other	Comorbid	Total
Normal	2018	281	7	160	286	2752
Cardiovas	476	1001	8	105	239	1829
Cancer	35	8	46	10	15	114
Other	443	165	7	724	250	1589
Comorbid	438	196	9	124	1368	2135

Accuracy = 61.25 percent.

Confusion matrix of the deep learning model. (in percent)

Confusion matrix of the deep learning model.

Predicted/Actual	normal	Cardiovas	Cancer	Other	Comorbid
Normal	73.33	10.21	0.25	5.81	10.39
Cardiovas	26.03	54.73	0.44	5.74	13.07
Cancer	30.70	7.02	40.35	8.77	13.16
Other	27.88	10.38	0.44	45.56	15.73
Comorbid	20.52	9.18	0.42	5.81	64.07

Accuracy = 61.25 percent.

Policy Exercise: Disease risks

Three types of white males:

- type 1: poor childhood cSES, cHLTH, no college degree, high BMI, (smoke, no exercise), average biomarkers.
- type 2: good childhood cSES, cHLTH, college degree, high BMI, (smoke, no exercise), average biomarkers.
- type 1: good childhood cSES, cHLTH, college degree, high BMI, (does not smoke, does exercise), average biomarkers.

Using Multinomial Logit Model

Group	Normal	Cardiovas	Cancer	Other	Comorbid
type 1	0.225	0.192	0.008	0.173	0.403
type 2	0.289	0.278	0.007	0.131	0.295
type 3	0.389	0.319	0.005	0.129	0.159

Using Neural Network Model TO-DO

Discussions

- Usefulness for Policy Analysis.
- Statistical models have well developed theories for inference about which inputs are important and other properties like how good the estimates are. But very *poor* performance at prediction. Some nonlinearity can be introduced to get better predictive performance, but it does not come close to the performance of the machine learning algorithms.
- Machine Learning models are good at prediction, but poor at identifying which inputs or regressors are important. Recently there are attempts to do so such as using Shapley Value criterion, but they are not useful in the multi outcome scenarios such as in the present case.

The End

Happy birthday, Professor Parikh.

Thank You.