Jaeha Huh

4/28/21

Intro to Machine Learning

Final Project Report: Heart Attack Analysis & Prediction

1. Introduction

- This is a data set for Heart Attack and each line represents patient information.
- Information such as the health, gender, and age of the subject is displayed, and the output variable is dependent variable, which means 1 is the expression of heart attack, and 0 is not expressed.
- Based on this information from individuals, this is a data set for diagnosis of the presence or absence of heart attacks.

Figure.1) Data ('heart.csv')

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

< Column information >

- age: Age of the patient
- sex: Gender of the patient
- cp: Chest Pain type chest pain type
  - Value 1: typical angina
  - Value 2: atypical angina
  - Value 3: non-anginal pain
  - Value 4: asymptomatic
- trtbps: resting blood pressure (in mm Hg)
- chol: cholestoral in mg/dl fetched via BMI sensor

- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- rest_ecg : resting electrocardiographic results
- thalach : maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak: previous peak
- slp: slope
- caa: number of major vessels (0-3)
  - Value 0: normal
  - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thall: Thal rate
- output: 0 = less chance of heart attack 1 = more chance of heart attack

Figure.2) General information such as observations, column types, and missing rates in the data

	Types	Counts	Distincts	Nulls	Missing_ratio	Uniques
age	int64	303	41	0	0.0	[63, 37, 41, 56, 57, 44, 52, 54, 48, 49, 64, 5...
caa	int64	303	5	0	0.0	[0, 2, 1, 3, 4]
chol	int64	303	152	0	0.0	[233, 250, 204, 236, 354, 192, 294, 263, 199, ...
cp	int64	303	4	0	0.0	[3, 2, 1, 0]
exng	int64	303	2	0	0.0	[0, 1]
fbs	int64	303	2	0	0.0	[1, 0]
oldpeak	float64	303	40	0	0.0	[2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, ...
output	int64	303	2	0	0.0	[1, 0]
restecg	int64	303	3	0	0.0	[0, 1, 2]
sex	int64	303	2	0	0.0	[1, 0]
slp	int64	303	3	0	0.0	[0, 2, 1]
thalachh	int64	303	91	0	0.0	[150, 187, 172, 178, 163, 148, 153, 173, 162, ...
thall	int64	303	4	0	0.0	[1, 2, 3, 0]
trtbps	int64	303	49	0	0.0	[145, 130, 120, 140, 172, 150, 110, 135, 160, ...

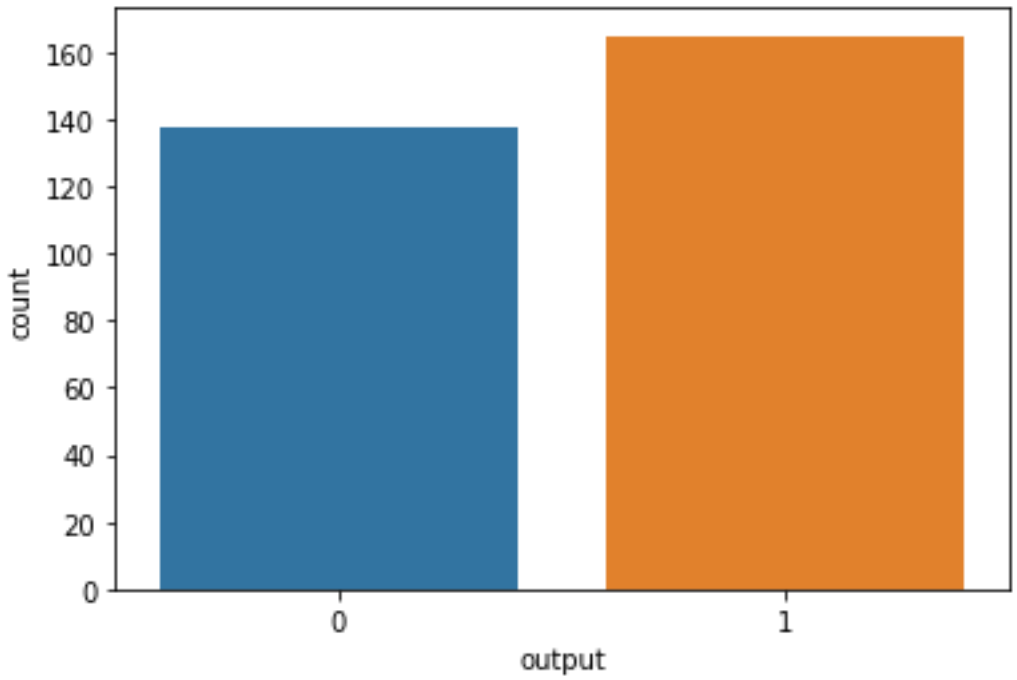
- It has a total of 14 variables and all variables have 303 observations without missing values.
- Among these variables age, chol, oldpeak, thalachh, trtbps are continuous variables and others are categorical variables.
- Types: Data Type of Column
  - Counts: Observations
  - Distincts: Number of eigenvalues (remove duplicates)

- Nulls: Number of missing values
- Missing_ratio: Missing Value Ratio
- Uniques: eigenvalues

2. Analysis

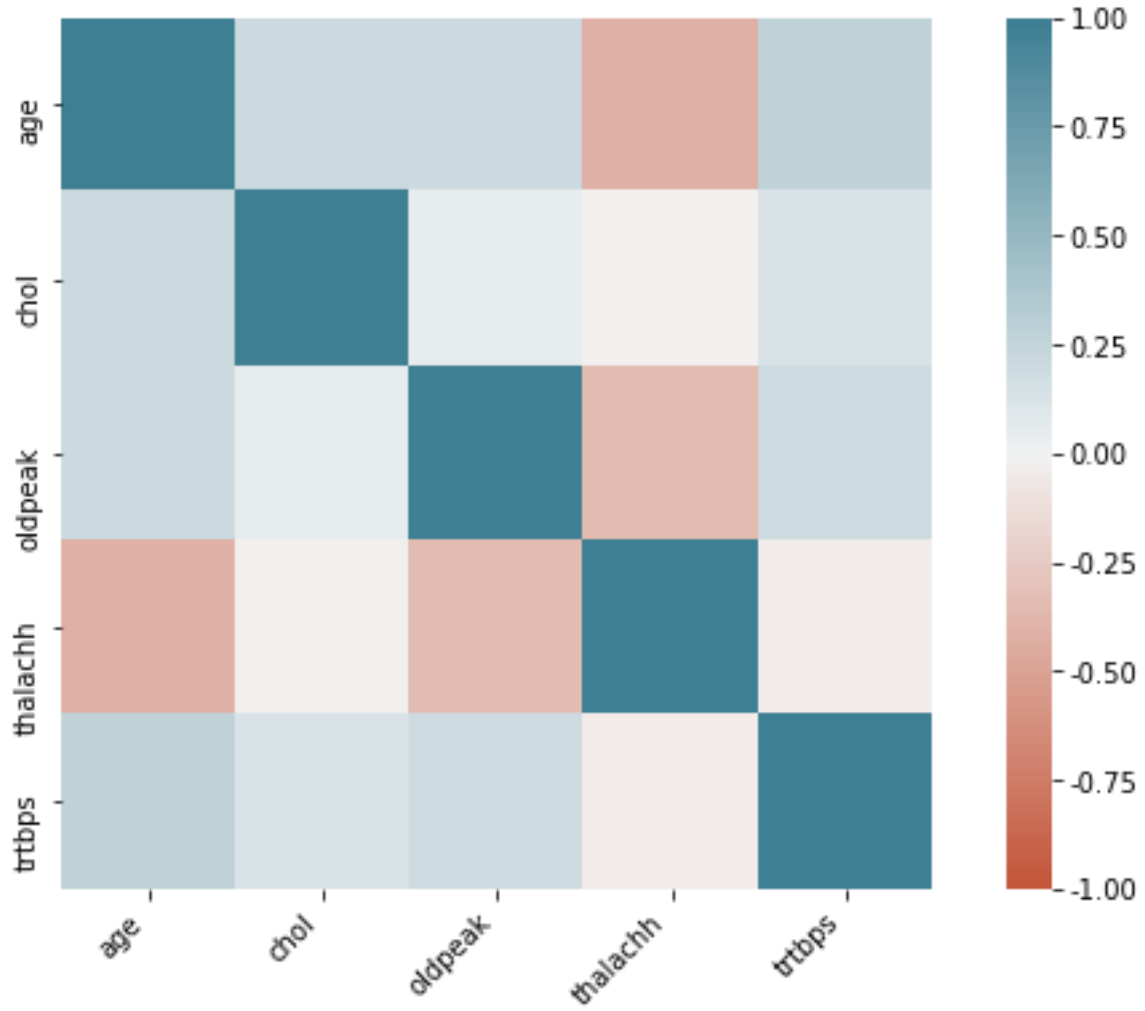
- To identify relationships between variables
- To determine which variables affect the heart attack, we determine the distribution of continuous variables according to output

Figure.3) distribution of dependent variable output



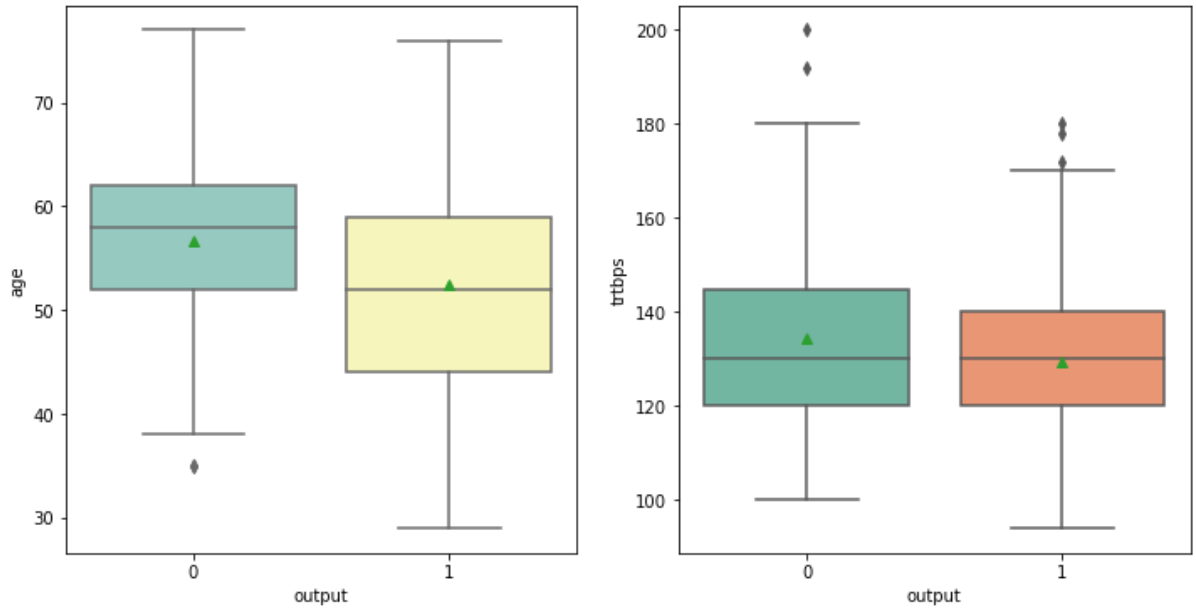
- The ratio of 1 is a little higher.

Figure.3) correlation visualization by continuous variables



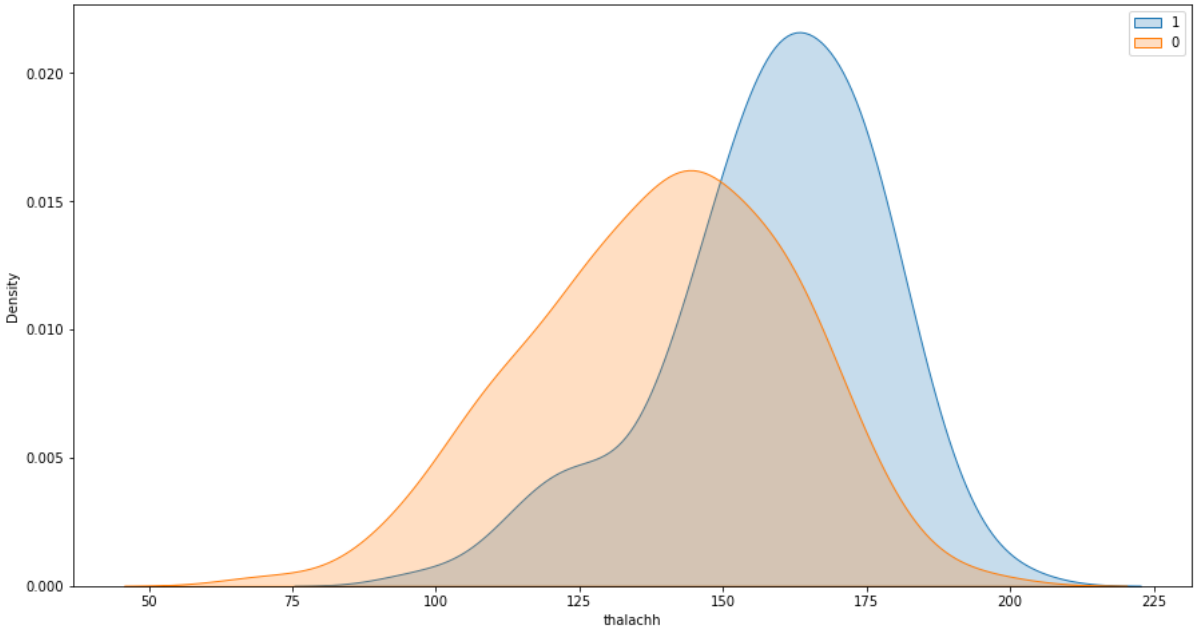
- Red indicates strong negative correlation between variables and blue indicates strong positive correlation between variables.
- A noticeable correlation is that age and thalachh are strongly negative. In addition, age and trtbps have a slight positive correlation.

Figure.4) box plot of output and age (figure 4.1); bot plot of output and trtbps (figure 4.2)



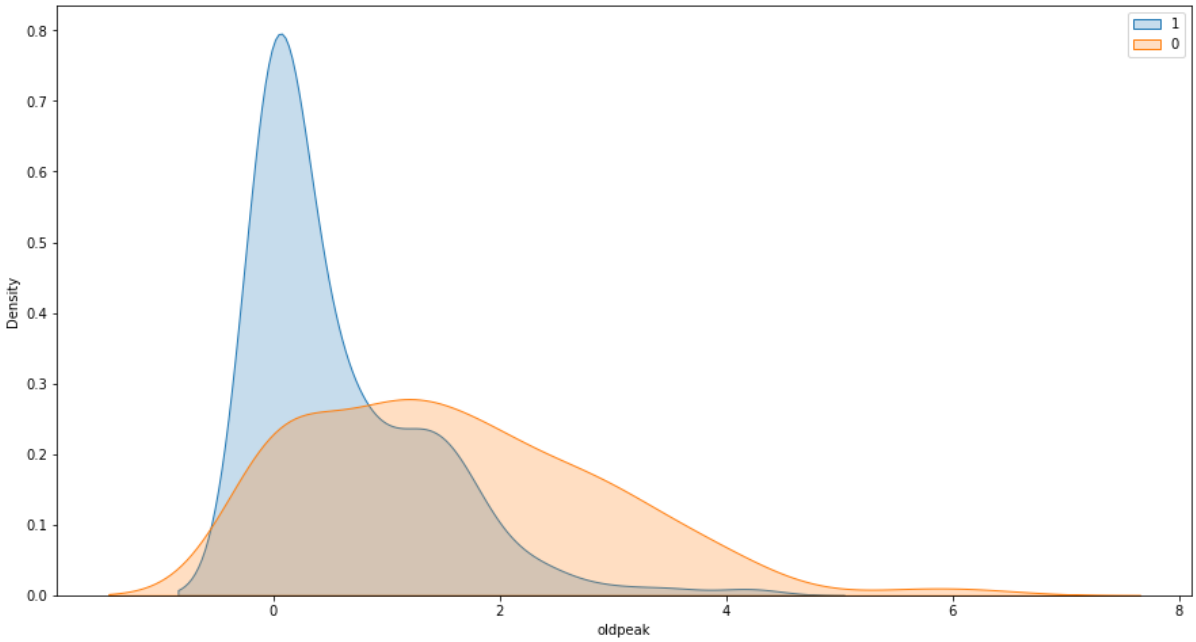
- If output is 1, it is distributed where the average age is lower than 0.
- If output is 1, trtbps tends to be lower than 0.

Figure.5) Distribution of output and thalachh



- For thalachh, the distribution varies significantly depending on the value of output
- If output is 1, thalachh is large on average around 170.
- If output is 0, it is densely distributed at values near 140
- The variable is likely to contribute greatly to the classification of output.

Figure.6) Distribution of output and oldpeak



- Oldpeak also shows a distribution that varies significantly depending on the value of output.
- If output is 1, it is distributed intensively about 0 on average and shows a small variance.
- If output is 0, it is mainly distributed at values near 1, but the overall variance is large, showing a wide distribution.
- The variable is also clearly classified according to output, which is expected to contribute greatly to the classification of output in the future.

3.Supervised analysis

< Modeling >

- Modeling steps use Logistic Regression, SVM, Neural Networks

< Process >

- Logistic Regression and SVM divide into two main parts.
- First, proceed 5-fold CV for train set. This will override the training set and the validation set internally at a ratio of 8:2.
- Cross Validation (CV) is necessary because the results predicted for the test set without Cross validation may be biased against the test set. Therefore, it should be determined that the Mean accuracy score obtained through CV is the actual performance of the model.
- After that, the final evaluation is carried out by predicting a separate test set..
- We also compare the above process by dividing the case of normalizing continuous variables and the case of non-normalizing raw data.
- However, in the case of Neural Networks (NN), val_accuracy is measured using the validation_split without using CV, and evaluation is carried out with the test set.

< Hyper-parameter Tuning >

- Logistic Regression and SVM replace important hyper-parameters and proceed k-fold again to explore near-optimal hyper-parameters.
- For Neural Networks, we control the number of layers and the number of neurons to explore suitable models and regulate epochs.
- (Epoch is when an entire dataset is passed forward and backward through the neural network only once)

< Best Model Selection >

- In both Logistic Regression and SVM, the best model is selected for the highest accuracy based on K-fold verification results.
- Neural Networks select the best model if validation accuracy is high.

Figure.7) split the train set and test set

```
Train set: (212, 13) , (212,)
Test set: (91, 13) , (91,)
```

◆ Logistic Regression

i. K-fold Cross Validation results if not normalize:

   A. Mean test set accuracy_score: 0.8304540420819491

ii. Test set evaluation if not normalized:

   A. Accuracy: 0.8791208791208791

iii. K-fold Cross Validation results if normalize:

   A. Mean test set accuracy_score: 0.8256921373200443

iv. Test set evaluation if normalized:

   A. Accuracy: 0.8681318681318682

Hyper-parameter Tuning:

- Tuned to non-normalized cases because it performed better without normalization.

C = 0.1:

⇨ Mean test set accuracy_score: 0.8212624584717609

C = 0.5:

⇨ Mean test set accuracy_score: 0.8305647840531561

C = 1:

⇨ Mean test set accuracy_score: 0.8304540420819491

C = 3:

⇨ Mean test set accuracy_score: 0.8210409745293467

C = 5:

⇨ Mean test set accuracy_score: 0.8211517165005537

According to Hyper-parameter Tuning, C = 0.5 model has the highest accuracy, so we normalize the C = 0.5 model.

- Accuracy: 0.8571428571428571

◆ SVM

i. K-fold Cross Validation results if not normalize:

   A. Mean test set accuracy_score: 0.5852713178294573

ii. Test set evaluation if not normalized:

A. Accuracy: 0.6703296703296703

iii. K-fold Cross Validation results if normalize:

A. Mean test set accuracy_score: 0.7972314507198228

iv. Test set evaluation if normalized:

A. Accuracy: 0.8681318681318682

Hyper-parameter Tuning:

- Tuned to normalized cases because it performed better than non-normalized cases.
- Tunes kernel with different types

linear kernel :

⇨ Mean test set accuracy_score: 0.8255813953488372

polynomial kernel:

⇨ Mean test set accuracy_score: 0.8066445182724251

Gaussian radial basis function(Gaussian rbf) kernel :

⇨ Mean test set accuracy_score: 0.7972314507198228

Sigmoid kernel:

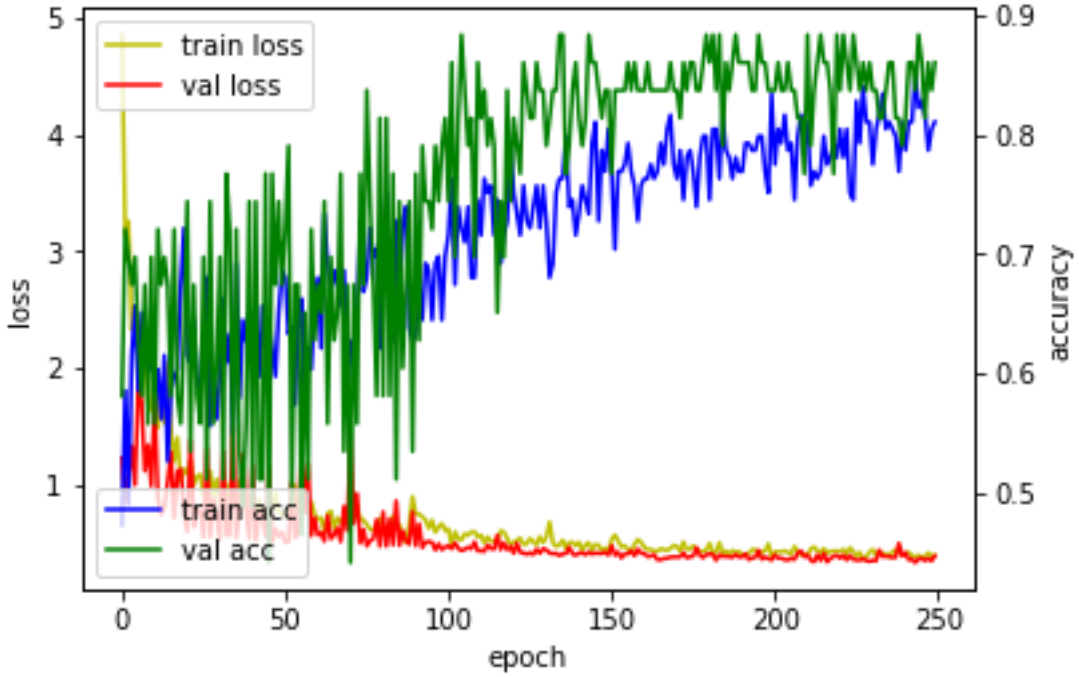
⇨ Mean test set accuracy_score: 0.7973421926910299

According to Hyper-parameter Tuning, linear kernel model has the highest accuracy, so we normalize the linear kernel.

- Accuracy: 0.8681318681318682

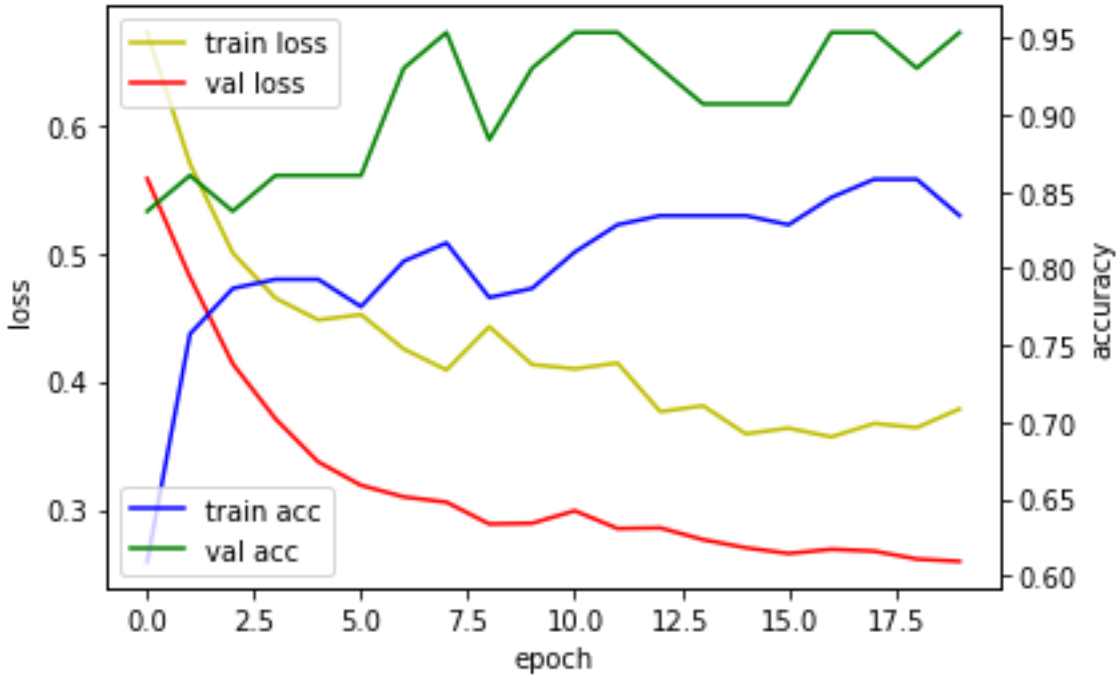
◆ Neural Network

Figure.8) Visualize loss and accuracy as learning progresses If not normalized



- This is the loss & accrual change table for the train & validation set as the epoch progresses.
- As epoch progresses, train and val loss converges gradually downward.
- At the same time, accuracy initially perturbates in a very large range and causes large errors, but after approximately 150 epochs, it typically converges to values above 0.8 indicating that the learning is progressing well.

Figure.9) Visualize loss and accuracy as learning progresses if normalized



- It is the same structure as the NN Model without normalization, but when normalized, it quickly reaches a figure of 0.8 or higher in just a small number of epochs.
- We can see that the model is very sensitive to normalization.
- In addition, the validation acid is easily above 0.9 whereas the training acid is around 0.85, resulting in slight overfitting of the validation set.

4.Table of Results

Comparision Table

Model 1	Hyper-paramter	Logistic Regression	
	C	5-fold CV	Evaluation
Normalization X	0.1	0.8212	-
	0.5	**0.8305**	0.8571
	1	0.8304	**0.8791**
	3	0.821	-
	5	0.8211	-
Normalization O	1	0.8256	0.8681

Model 2	Hyper-paramter	Support Vector Machine	
	kernel	5-fold CV	Evaluation
Normalization O	linear	**0.8255**	**0.8681**
	poly	0.8066	-
	rbf	0.7972	-
	sigmoid	0.7973	-
Normalization X	linear	0.5852	0.6703

Model 3	Hyper-paramter	Neural Network	
	Epoch	Validation Accuracy	Evaluation
Normalization O	20	**0.9535**	**0.8571**
Normalization X	250	0.8605	0.8131

5. Why

< Result >

- The results of each model can be seen briefly with the table.

- Logistic Regression yields the highest accuracy of CV results when C=0.5 when not normalized. Although the test set accuracy drops by 0.02 at evaluation, the data shows that cross-validation results with k-fold CV are more reliable than with single test set accuracy due to a small amount of data.

- Support Vector Machine changes the hyperparameter called kernel. The corresponding model exhibits a large deviation of accuracy depending on whether it is normalized or not. In case of non-normalization, the figure is lower than 0.5-0.6, but the performance during normalization is significantly improved to 0.8 levels. Therefore, hyper-parameter tuning was also performed in the normalized case, recording the highest CV results in linear among the various kernels.

- In the Neural Network, the model is constructed by loading three Dense layers and a Dropout layer to regulate overfitting. Basically, artificial neural networks are known to be sensitive as models that require regularization, and results show that regularization is also important. In the case of non-normalization, the validation accrual reaches 0.8 levels after about 200 epochs, but when normalization is performed, even the same model structure reaches high accuracy in 10 epochs in an instant. The number of epochs was set at an appropriate level because additional learning could lead to overfitting problems. As mentioned, the normalized model showed good learning performance, and the found problem was that the equilibrium of the validation set reached a level of about 0.9 or higher, but the train accuracy exceeded 0.85 levels, showing rather overfitting the validation set.