課題１デーブルデータ

高松 周平

問１

- ひとまず適当にSGDClassifierで分類をを行ってみました
- scikit-learnのサイト[1]で紹介されているフローチャートに従って手法を選びました。
- 結果は

CV of train : 0.3646056767403385
CV of test : 0.364603234942299

となり、線形な手法であるからか精度は低いように感じました。

問 2

有用な特徴量だけ捉える

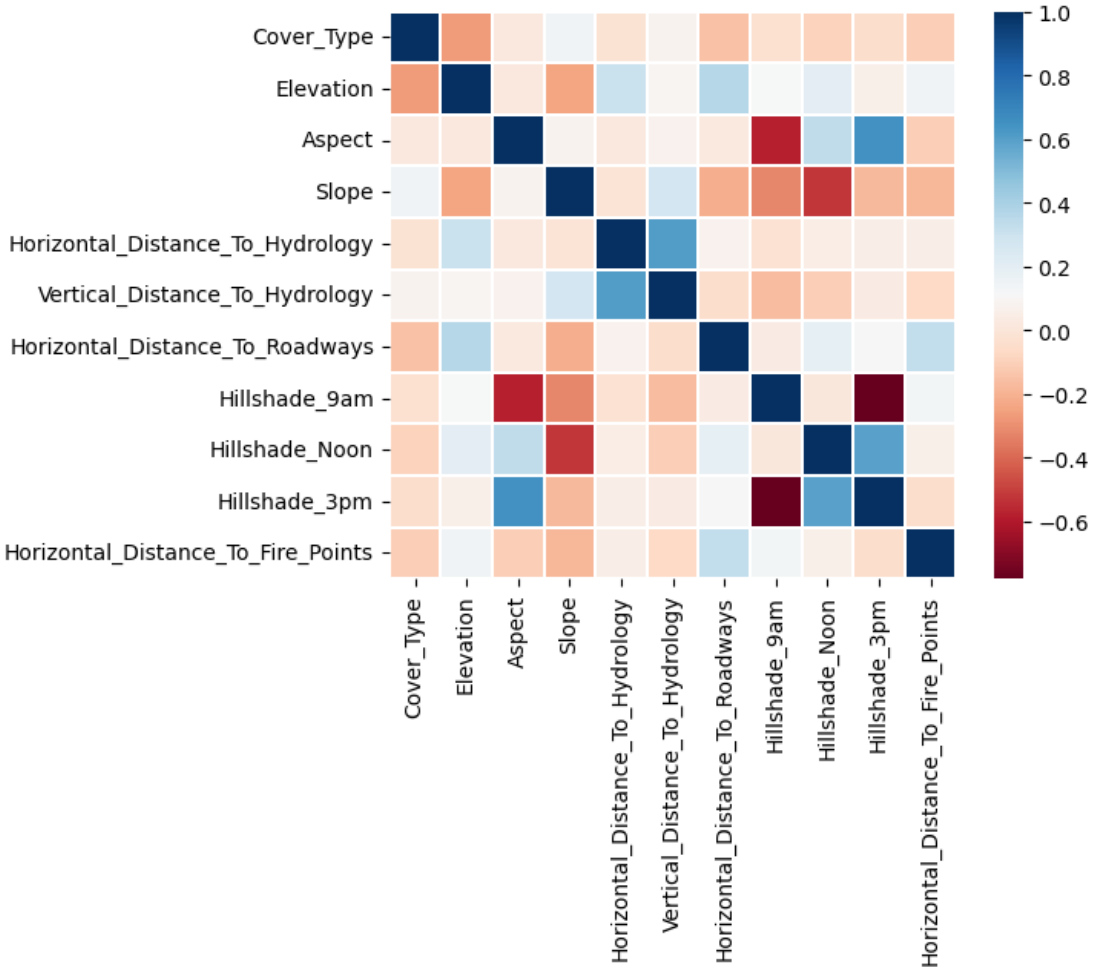
```
In [9]: covtype_df.describe().T
```

Out[9]:

	count	mean	std	min	25%	50%	75%	max
Cover_Type	581012.0	2.051471	1.396504	1.0	1.0	2.0	2.0	7.0
Elevation	581012.0	2959.365301	279.984734	1859.0	2809.0	2996.0	3163.0	3858.0
Aspect	581012.0	155.656807	111.913721	0.0	58.0	127.0	260.0	360.0
Slope	581012.0	14.103704	7.488242	0.0	9.0	13.0	18.0	66.0
Horizontal_Distance_To_Hydrology	581012.0	269.428217	212.549356	0.0	108.0	218.0	384.0	1397.0
Vertical_Distance_To_Hydrology	581012.0	46.418855	58.295232	-173.0	7.0	30.0	69.0	601.0
Horizontal_Distance_To_Roadways	581012.0	2350.146611	1559.254870	0.0	1106.0	1997.0	3328.0	7117.0
Hillshade_9am	581012.0	212.146049	26.769889	0.0	198.0	218.0	231.0	254.0
Hillshade_Noon	581012.0	223.318716	19.768697	0.0	213.0	226.0	237.0	254.0
Hillshade_3pm	581012.0	142.528263	38.274529	0.0	119.0	143.0	168.0	254.0
Horizontal_Distance_To_Fire_Points	581012.0	1980.291226	1324.195210	0.0	1024.0	1710.0	2550.0	7173.0
Wilderness_Area_0	581012.0	0.448865	0.497379	0.0	0.0	0.0	1.0	1.0
Wilderness_Area_1	581012.0	0.051434	0.220882	0.0	0.0	0.0	0.0	1.0
Wilderness_Area_2	581012.0	0.436074	0.495897	0.0	0.0	0.0	1.0	1.0
Wilderness_Area_3	581012.0	0.063627	0.244087	0.0	0.0	0.0	0.0	1.0
Soil_Type_0	581012.0	0.005217	0.072039	0.0	0.0	0.0	0.0	1.0
Soil_Type_1	581012.0	0.012952	0.113066	0.0	0.0	0.0	0.0	1.0
Soil_Type_2	581012.0	0.008301	0.090731	0.0	0.0	0.0	0.0	1.0
Soil_Type_3	581012.0	0.021335	0.144499	0.0	0.0	0.0	0.0	1.0
Soil_Type_4	581012.0	0.002749	0.052356	0.0	0.0	0.0	0.0	1.0
Soil_Type_5	581012.0	0.011316	0.105775	0.0	0.0	0.0	0.0	1.0
Soil_Type_6	581012.0	0.000181	0.013442	0.0	0.0	0.0	0.0	1.0

- データの中身の分析を行いました。
- Wilderness_Area_0以降のデータはtrue or falseの二値であり、One-Hot Vectorであるとわかりました。

問 2



- それ以外のデータでの相関係数のヒートマップをプロットしました。
- 目的の変数とひときわ強い相関のあるパラメータはパット見ありませんでした

問２

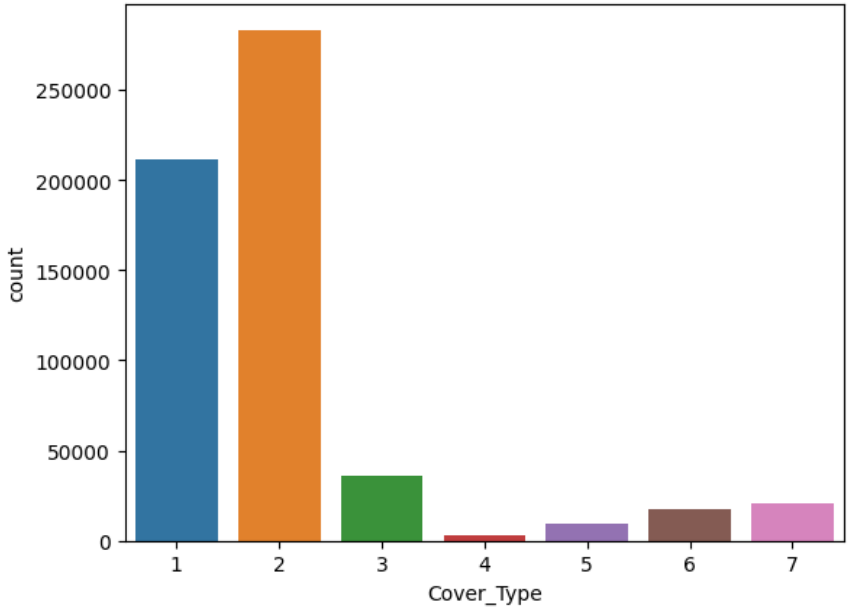
- 相関係数とソートすることで相関係数の強いパラメータを確認しました。

>　　>　　>　　>

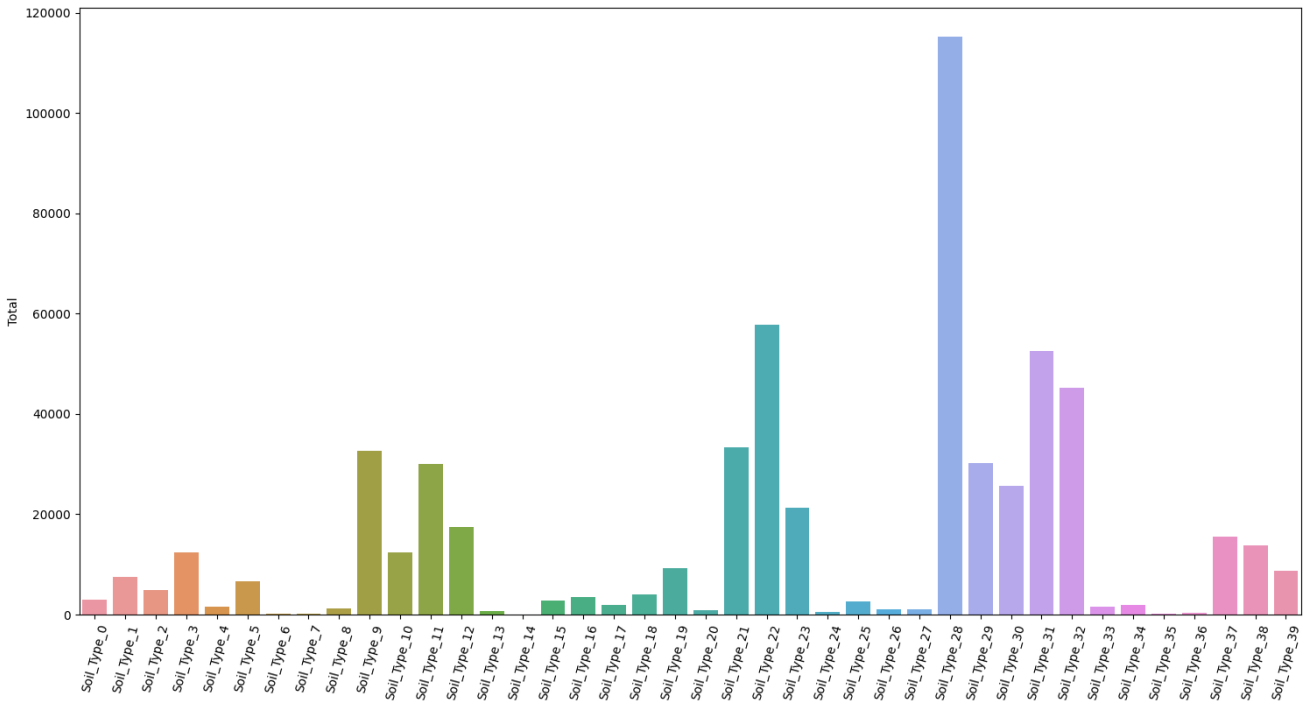
[21]:
```
covtype_df.corr().unstack().sort_values().drop_duplicates()
```

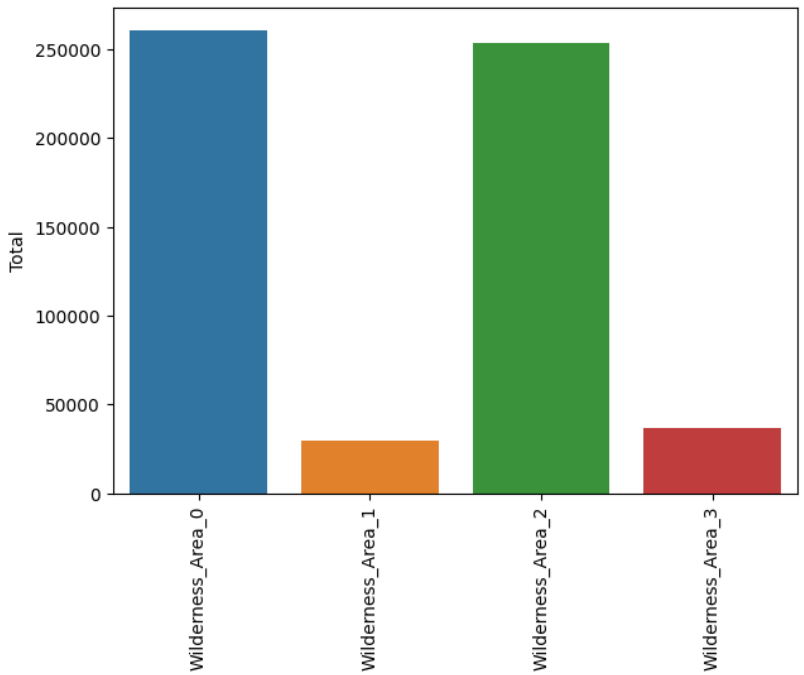
Out[21]:
```
Wilderness_Area_0                  Wilderness_Area_2                  -0.793593
Hillshade_3pm                      Hillshade_9am                      -0.780296
Wilderness_Area_3                  Elevation                          -0.619374
Hillshade_9am                      Aspect                             -0.579273
Slope                              Hillshade_Noon                     -0.526911
                                                                         ...
Soil_Type_28                       Wilderness_Area_0                   0.550549
Hillshade_3pm                      Hillshade_Noon                      0.594274
Horizontal_Distance_To_Hydrology   Vertical_Distance_To_Hydrology      0.606236
Hillshade_3pm                      Aspect                              0.646944
Cover_Type                         Cover_Type                          1.000000
Length: 1486, dtype: float64
```

問 2

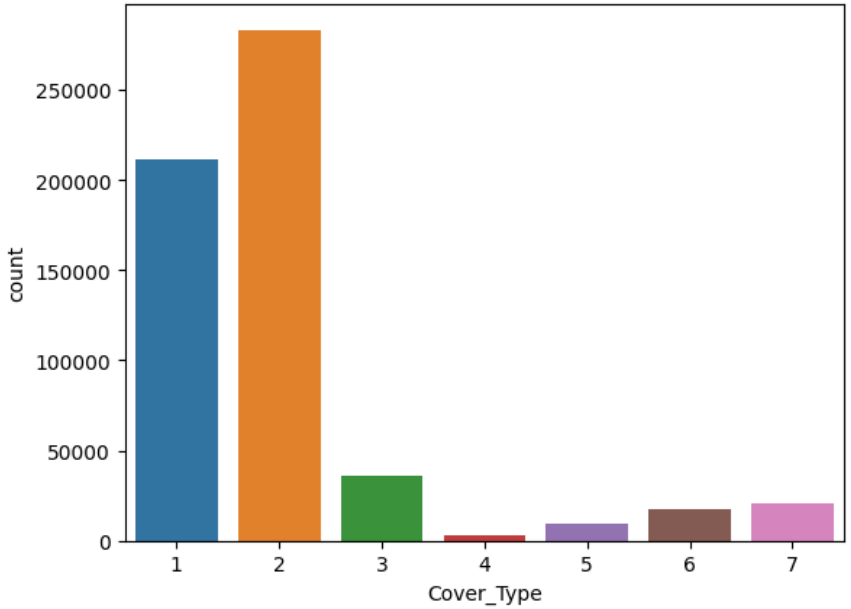


- One-Hot Vectorの三種類のパラメータをプロットしてみました。
- ほぼfalseであるパラメータもあることが確認できました。

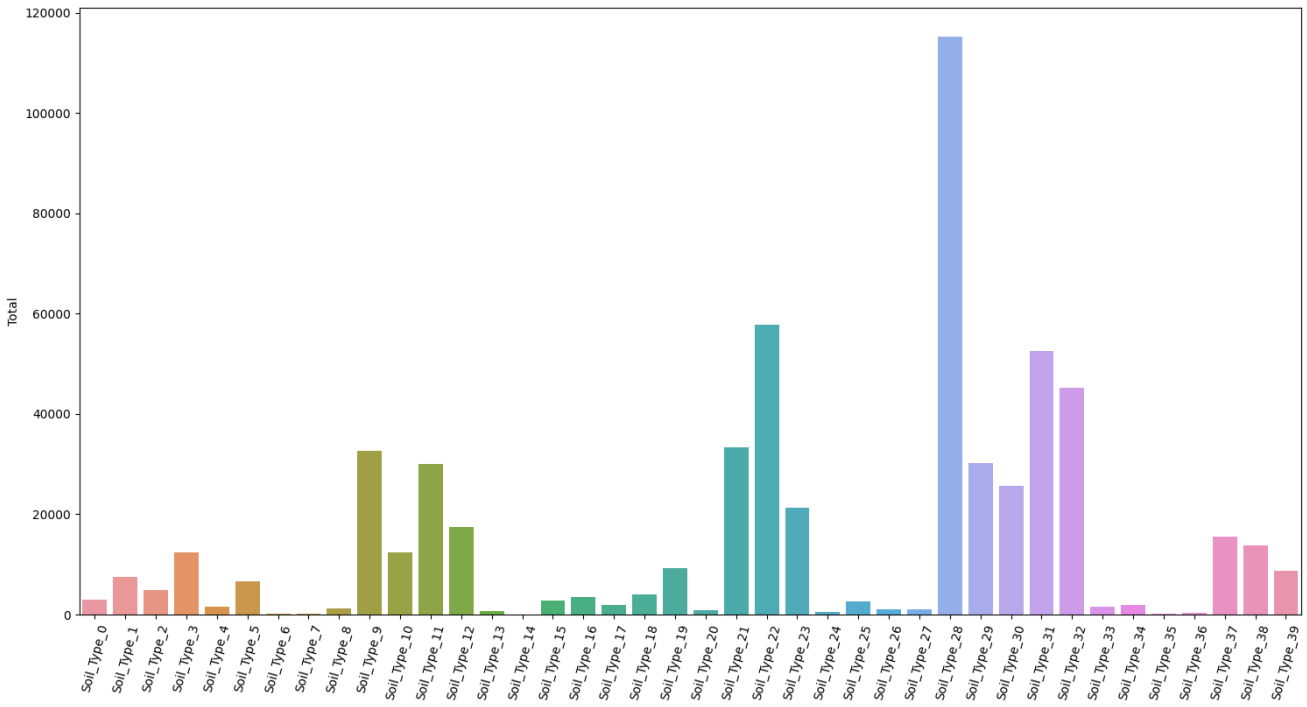


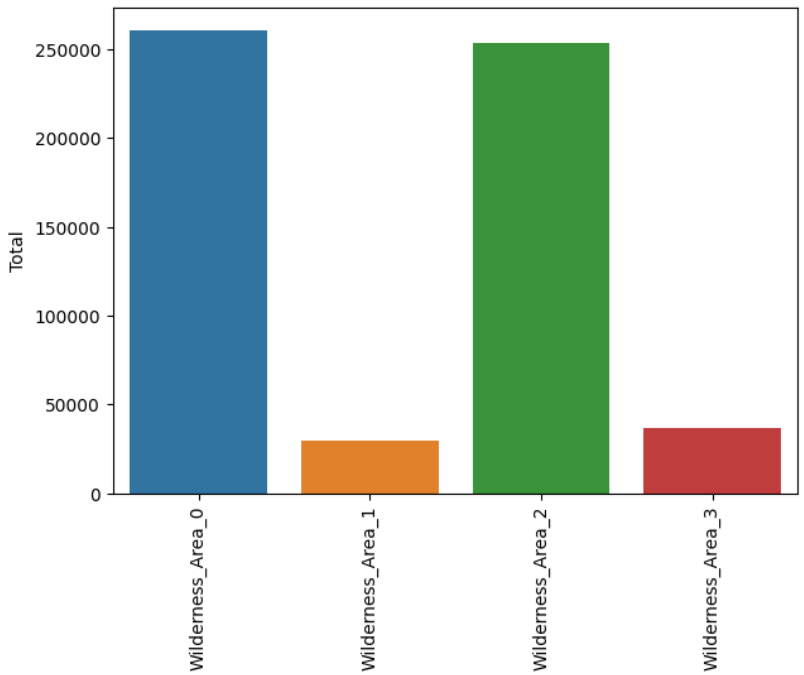


問 2



- One-Hot Vectorの三種類のパラメータをプロットしてみました。
- ほぼfalseであるパラメータもあることが確認できました。





問 2

- 規格化を行いました

[33]:
```
scaler = StandardScaler()
standard_covtype_df = covtype_df.copy(deep=True)
standard_covtype_df[['Elevation', 'Aspect', 'Slope',
       'Horizontal_Distance_To_Hydrology', 'Vertical_Distance_To_Hydrology',
       'Horizontal_Distance_To_Roadways', 'Hillshade_9am', 'Hillshade_Noon',
       'Hillshade_3pm', 'Horizontal_Distance_To_Fire_Points']] = scaler.fit_transform(standard_covtype_df[['Elevation', 'Aspect', 'Slope',
       'Horizontal_Distance_To_Hydrology', 'Vertical_Distance_To_Hydrology',
       'Horizontal_Distance_To_Roadways', 'Hillshade_9am', 'Hillshade_Noon',
       'Hillshade_3pm', 'Horizontal_Distance_To_Fire_Points']])
standard_covtype_df
```

Out[33]:

	Cover_Type	Elevation	Aspect	Slope	Horizontal_Distance_To_Hydrology	Vertical_Distance_To_Hydrology	Horizontal_Distance_To_Roadways	Hi
0	5	-1.297805	-0.935157	-1.482820	-0.053767	-0.796273	-1.180146	
1	5	-1.319235	-0.890480	-1.616363	-0.270188	-0.899197	-1.257106	
2	2	-0.554907	-0.148836	-0.681563	-0.006719	0.318742	0.532212	
3	2	-0.622768	-0.005869	0.520322	-0.129044	1.227908	0.474492	
4	5	-1.301377	-0.988770	-1.616363	-0.547771	-0.813427	-1.256464	
...	...	...	...	...	...	...	...	
581007	3	-2.012130	-0.023740	0.787408	-0.867697	-0.504653	-1.437962	
581008	3	-2.029988	-0.032675	0.653865	-0.952383	-0.590424	-1.446299	
581009	3	-2.047847	0.029873	0.386780	-0.985317	-0.676194	-1.449506	
581010	3	-2.054990	0.128163	0.119694	-0.985317	-0.710502	-1.449506	
581011	3	-2.058562	0.083486	-0.147392	-0.985317	-0.727656	-1.464256	

581012 rows × 55 columns

問２

- ロジスティック回帰を行いました。
- 結果は

CV of train : 0.6629045478895632
CV of test : 0.6633563677357728

- となり比較的高い予測精度となりました。

問２

- ロジスティック回帰を行いました。
- なおmaxのイテレーションは500回としましたが、頭打ちだったようです。
- 結果は

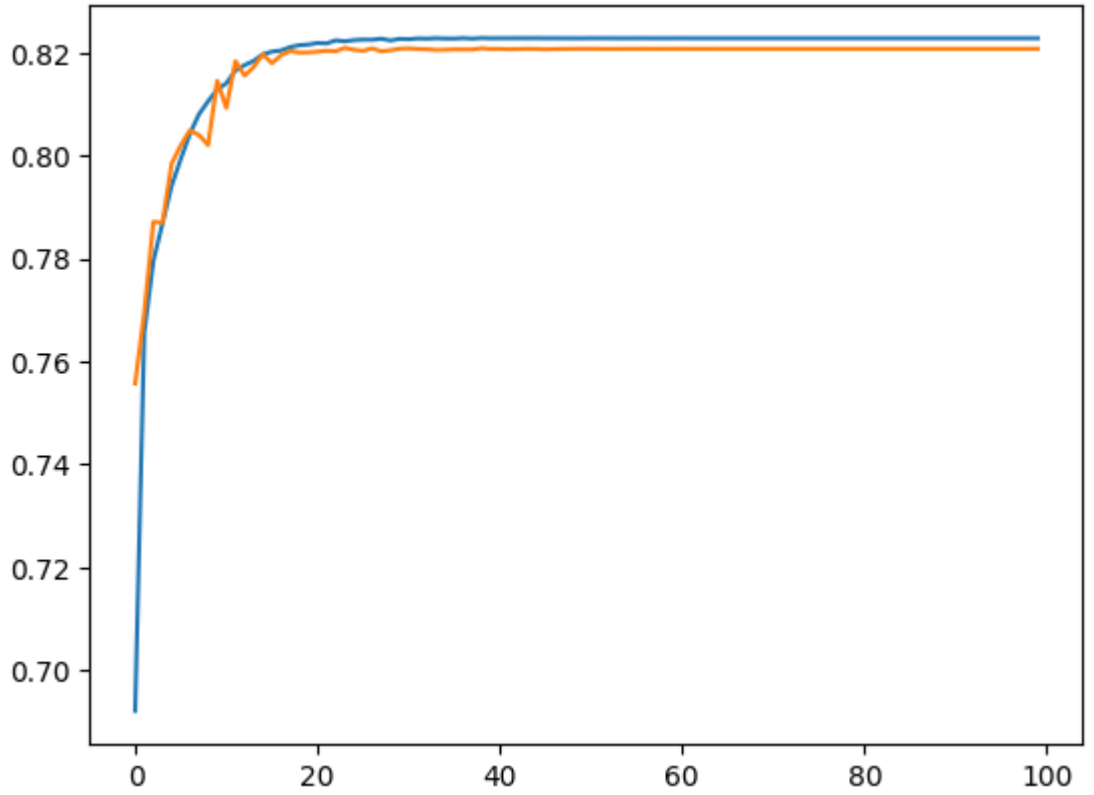
CV of train : 0.6629045478895632
CV of test : 0.6633563677357728

- となり比較的高い予測精度となりました。

問２

- シンプルなMLPを用いて予測を行いました。
- 結果は

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_5 (Dense)              (None, 54)                2970
_________________________________________________________________
dense_6 (Dense)              (None, 48)                2640
_________________________________________________________________
dense_7 (Dense)              (None, 32)                1568
_________________________________________________________________
dense_8 (Dense)              (None, 16)                528
_________________________________________________________________
dense_9 (Dense)              (None, 8)                 136
=================================================================
Total params: 7,842
Trainable params: 7,842
Non-trainable params: 0
_________________________________________________________________
```



CV of test ：0.8213729421787733

- となり割と高い予測精度となりました。

参考サイト

- 参考にしたサイトは多くありますが、すべてを写経したわけではなく自分なりに書き直しています。
- https://www.kaggle.com/code/mdriponmiah/eda-and-data-visualization-for-beginners
- https://www.kaggle.com/code/devanshiipatel/forest-cover-type-classification