Homework: Give me some credit data set (Report)

Data Preparation

for this process, I've drop some outlier data which is a data that was over than 99.99 percentage of RevolvingUtilizationOfUnsecuredLines, DebtRatio, MonthlyIncome,NumberRealEstateLoansOrLines, NumberOfDependents and also drop the age "0" in age column and I also replace Nan in MonthlyIncome and NumberOfDependents with median of these data.

Training

1. Gradient Boosting Classifier with Grid Search

   From Grid Search CV, the best parameter is max_dept=5 and n_estimators = 100

   a. Confusion matrix

Prediction Values \ Actual Values	Positive	Negative
Positive	41510	455
Negative	2433	573

   b. Classification Report

	precision	recall	f1-score	support
0	0.94	0.99	0.97	41965
1	0.56	0.19	0.28	3006
accuracy			0.94	44971
macro avg	0.75	0.59	0.63	44971
weighted avg	0.92	0.94	0.92	44971

2. MLP Classifier

a. Confusion matrix

Prediction Values \ Actual Values	Positive	Negative
Positive	41539	426
Negative	2448	558

b. Classification Report

	precision	recall	f1-score	support
0	0.94	0.99	0.97	41965
1	0.57	0.19	0.28	3006
accuracy			0.94	44971
macro avg	0.76	0.59	0.62	44971
weighted avg	0.92	0.94	0.92	44971

3. K-NN Classifier with Grid Search

From Grid Search CV, the best parameter is n_neighbors=5 and weights='uniform'

a. Confusion matrix

Prediction Values \ Actual Values	Positive	Negative
Positive	41847	118
Negative	2948	58

b. Classification Report

	precision	recall	f1-score	support
0	0.93	1.00	0.96	41965
1	0.33	0.02	0.04	3006
accuracy			0.93	44971
macro avg	0.63	0.51	0.50	44971
weighted avg	0.89	0.93	0.90	44971

Conclusion

from these three models, we can conclude that Gradient Boosting Classifier with Grid Search has the most score which is 0.63 and K-NN Classifier with Grid Search is the worst which give 0.50 of f1-score.