Training Data Influence: Credit Risk Assessment

Winston Thov, Fiona Jin

Preliminaries:

Task: A logistic regression model will be used for the classification task of determining if the customer will pay their next payment on their loan or not based on demographics and payment behavior.

Dataset: https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients .

Features: credit limit, sex, marital status, age, payment history
Labels: If they defaulted on their payment or not

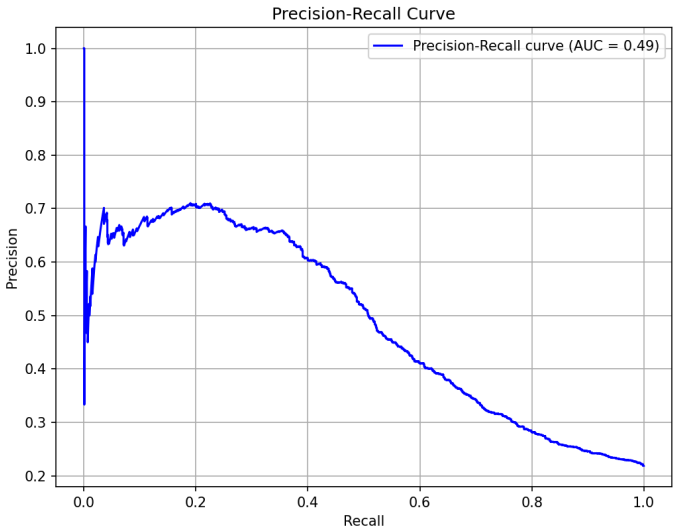
Splitting Dataset: 80/20 Split with random sampling, since it allows for a large amount of random data for both sets so that performance is correctly judged along with having a large variance in training data for the model to perform well. Furthermore as compared to time based splits it is more simple and easier to implement, so as to less over complicate the process.

Performance Report: The precision and recall both have are maintained to be high for when customers default on their loan, meaning a high f1 score of 0.89, but for when they do pay their loan back there is a high false negative rate shown by the low recall of 0.24 along with an increase in false positive rate shown by the decrease in precision to 0.70, overall give a very low f1 score. We will primarily be using the f1 score for this task.

Classification Report:

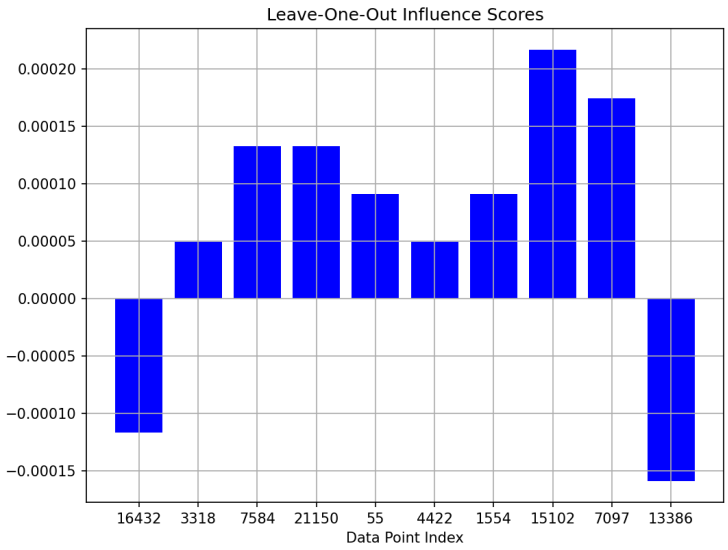
	precision	recall	f1-score	support
0	0.82	0.97	0.89	4687
1	0.70	0.24	0.36	1313
accuracy			0.81	6000
macro avg	0.76	0.61	0.62	6000
weighted avg	0.79	0.81	0.77	6000

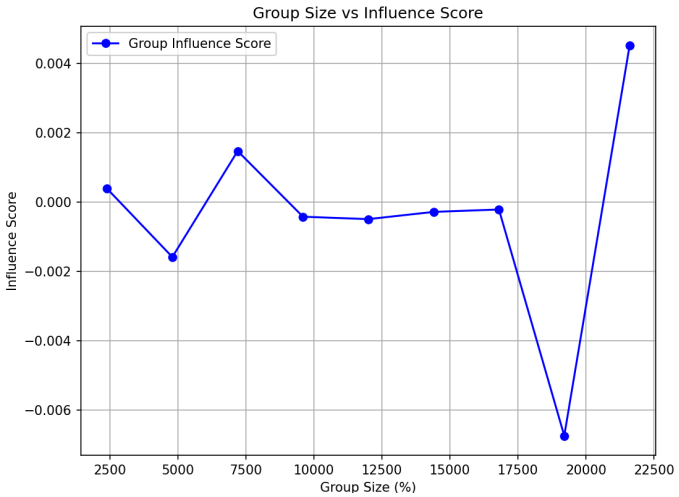
```
Accuracy: 0.8108
Confusion Matrix:
[[4551  136]
 [ 999  314]]
```



Brute Force LOO influence:

We can see high variance within the LOO influence scores within the data ranging from a very low -0.00015 at instance 13386 to +0.00020 at instance 15102 in this small sample which is a relatively large difference of 0.00035. To add to this these values are quite small due to the large nature of the dataset meaning each data instance has very little impact on the actual model alone



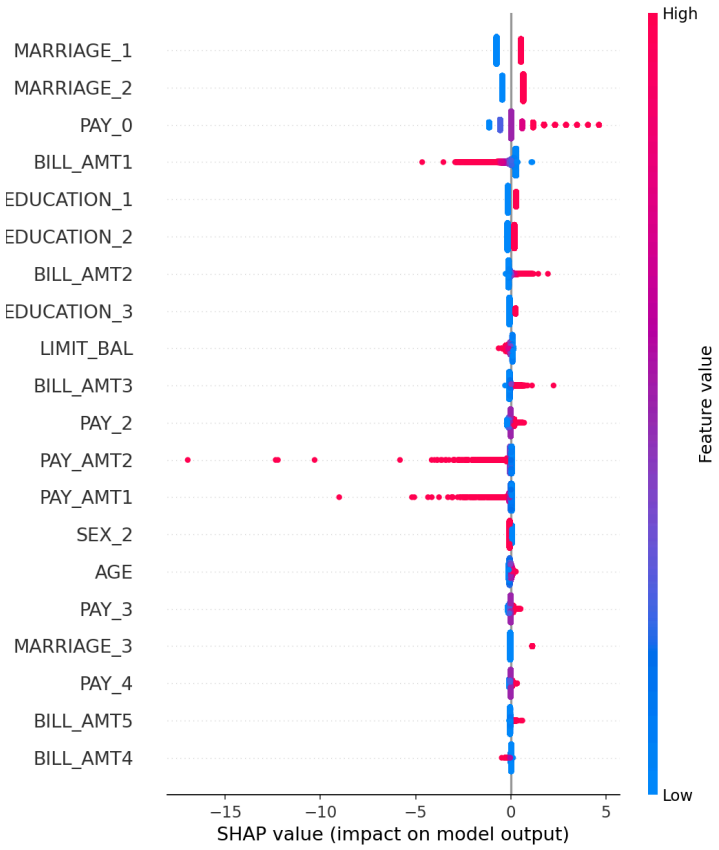


Group Based influence:

For calculation of group based influence we used groups that are 10% the size of the entire dataset, with 10 different groups. The scores are reported here:

	Group Size (%)	Influence Score
0	2400	0.000380
1	4800	-0.001594
2	7200	0.001464
3	9600	-0.000431
4	12000	-0.000500
5	14400	-0.000292
6	16800	-0.000222
7	19200	-0.006750
8	21600	0.004500

Shapley Values:



Feature is ranked by influence
Spread of Shapley values (Horizontal axis)
Dots represent data points.

For features such as marriage and education, the points are tightly packed, meaning the impact is consistent across most data instances, the opposite is true for attributes with more horizontal span such as payment amount.

Contribution Statement

Note: ChatGPT was used for boilerplate code

Winston Thov:
Report writing
Interpretation of model outputs
Error debugging

Fiona Jin:
Code Writing
Graph generation for metrics