Sentiment Analysis of Amazon

1. The columns in dataset are: Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount', 'Class'], dtype='object')
2. We have 31 columns in the train dataset in which V1 to V28 columns are the columns that are generated after PDA. Class is the target column.
3. The shape of training dataset (227845, 31)
4. The shape of testing dataset (56962, 30)
5. There are no null values in the training and testing datasets
6. We observe that there is no other brand in the dataset other than amazon hence we can drop brand column
7. From the heat map correlation matrix no variable that are positively correlated. Also Class is somewhat negatively corrlated to the V12, V14 and V17 and Time is correalted to V3.
8. Within the 2 days we observe that maximum fraud had occurred at 11th to 12th hr. of the 1st day more than 35 and at 2nd to 3rd hr of the second day more than 25
9. We observe that the maximum amount of fraud is done for 1 euro to 500 euros
10. We observe that the maximum frauds have occurred in peak time and low during night and amount in range of 2500 euros.
11. Based on the box plot we get the interquartile range after standardizing data as 0.28 and finding upper bound we get value as 1.5 * 0.28 + 0.046 = 0.47. Hence, we drop values greater than 0.75
12. We can Drop the Time column as we converted it the time interval of 48 hrs
13. The shape of after data processing training dataset (211504, 31)
14. The shape of after data processing testing dataset (52877, 30)
15. The no fraudulent class is 2.1 Lakhs and fraudulent class is just 337. The dataset is highly imbalance
16. Reports for the models:

accuracy	precision	f1	recall	
GaussianNB	0.93	0.970	0.92	0.88
LogisticRegression	0.95	0.980	0.95	0.93
SVC	1.00	0.870	0.86	0.86
RandomForestClassifier	1.00	0.950	0.88	0.82
XGBClassifier	1.00	0.970	0.90	0.85
ANN	1.00	0.999	0.58	0.55

17. In Annamolly detection the non-fraudulent transactions (blue) generally cluster towards higher anomaly scores, reflecting a higher likelihood of belonging to the normal distribution.
18. The fraudulent transactions (red) are scattered across the entire range of anomaly scores. Some fraudulent transactions might have high anomaly scores, significantly deviating from the normal behavior. However, others might have anomaly scores that overlap with the non-fraudulent transactions, making them more challenging to detect purely based on this anomaly score.