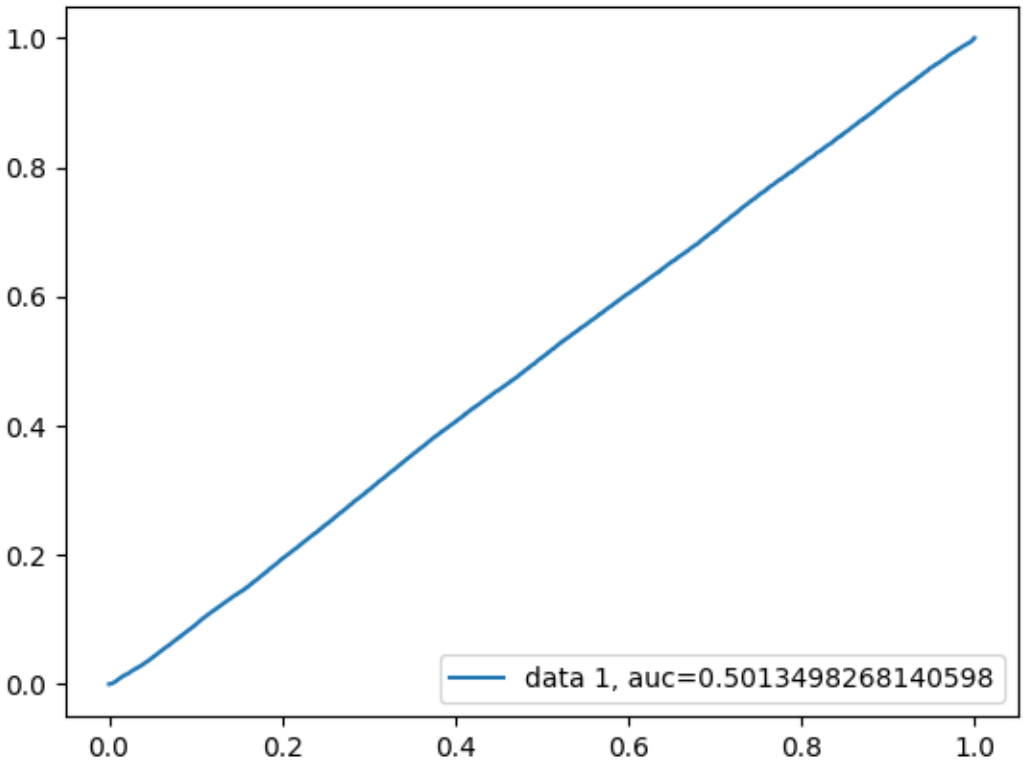
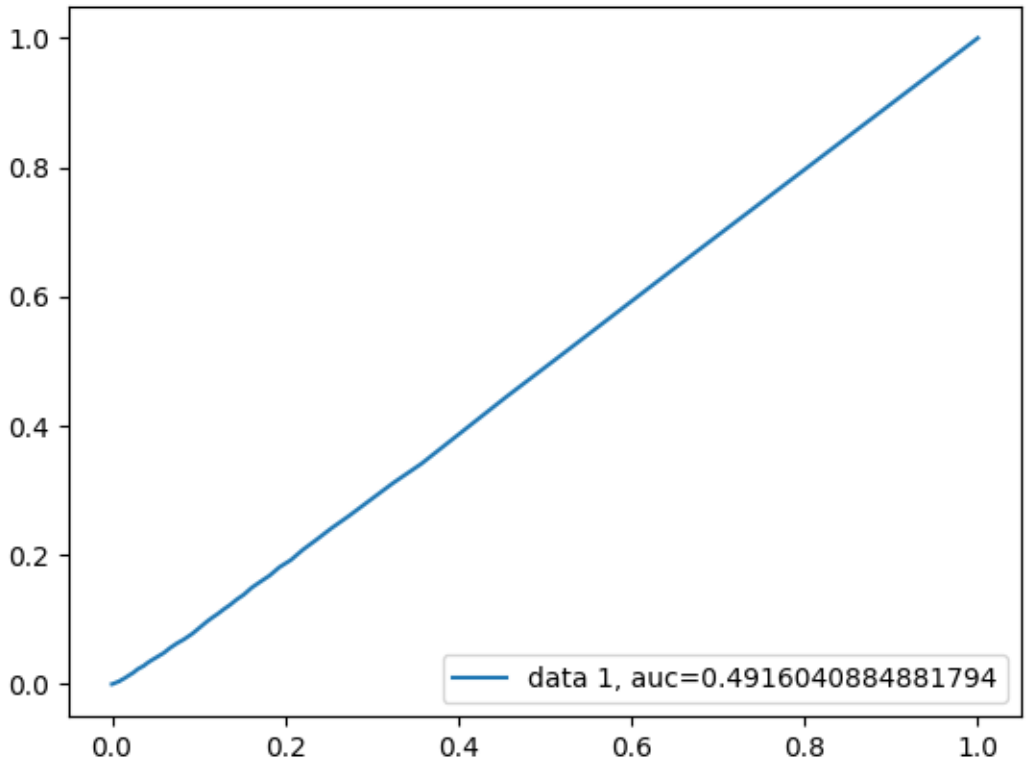
Since the dataset is too big,
the dataset has to be splited and the rows of the new dataset that we are using for the classification is 499999
From different classification method, the result is showed below:
logistic regression:
[[477326 55]
[22601 17]]

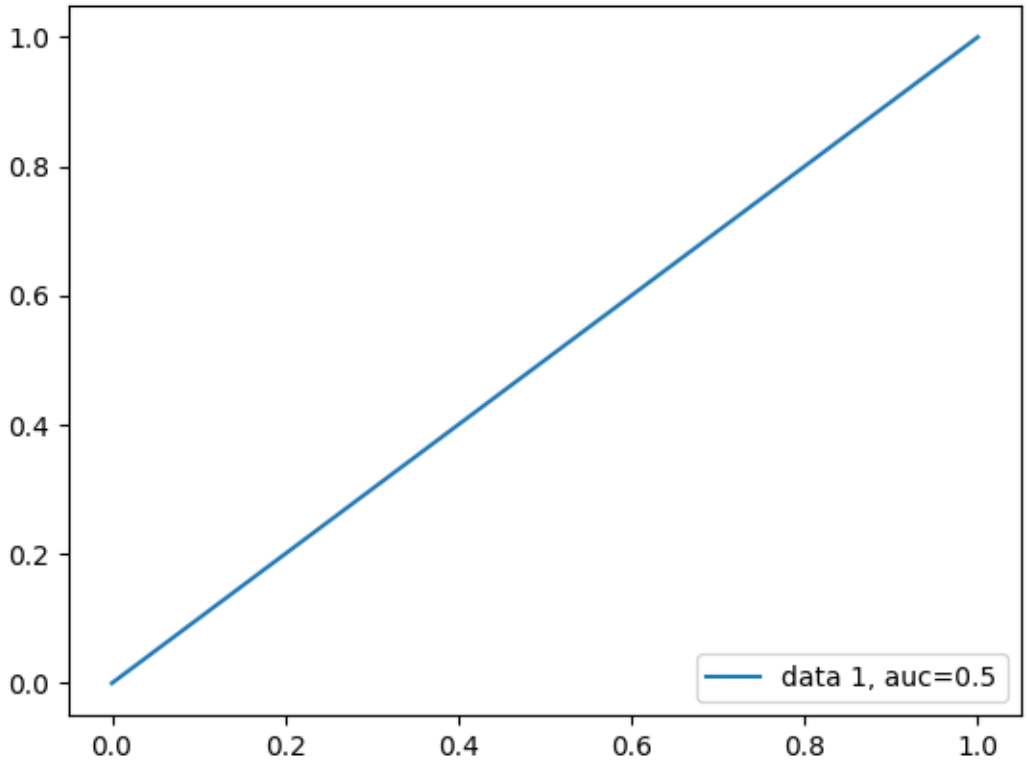


random forest:

```
[[468069   9312]
 [ 20719   1899]]
```



```
Neural Network:
[[477381      0]
 [ 22618      0]]
```



TPOT :

```
from sklearn.naive_bayes import BernoulliNB
clf = BernoulliNB(alpha=0.01, fit_prior=True)
clf.fit(X, Y)

y_pred = clf.predict(X_test)
print (y_pred)
```

we get BernoulliNB as the best pipeline

```
N    478328
Y     21671
Name: Y, dtype: int64
['N' 'N' 'N' ... 'N' 'N' 'N']
[[477381      0]
 [ 22387    231]]
477381
The matrix for Q12005 file

[['Q12005', 21671, 231, 499999, 231, 0]]
```

The BernoulliNB has the highest accuracy amount these classification model. However the true positive is pretty low which means it still really hard for the model identify the delinquent status correctly.

We only use a portion of the data, if we try to use all the data, the model might gives us a better result since classification problem requires lots of data to train the model