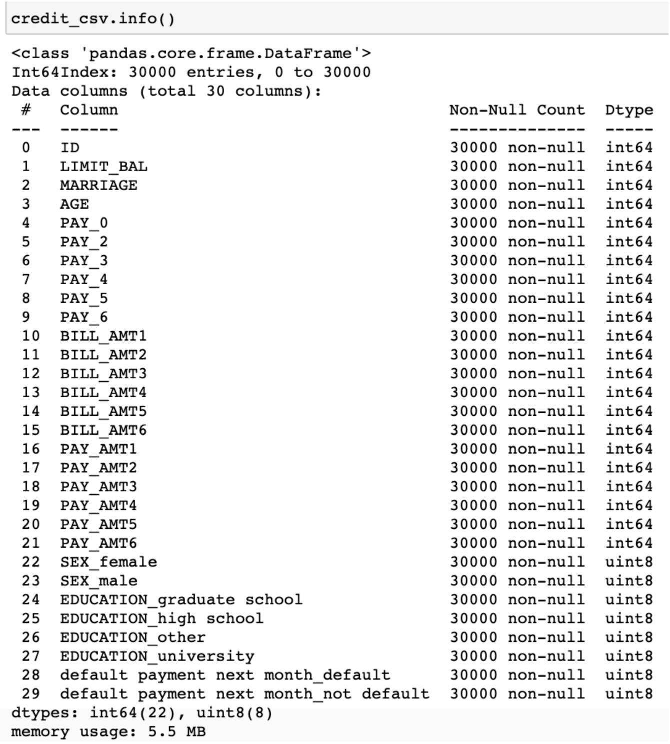
Course2 - Task3

Question

An increase in customer default rates is bad for Credit One since its business is approving customers for loans in the first place. This is likely to result in the loss of Credit One's business customers. You need to build a model that can better predict what credit limit a customer should be assigned.

There can be two ways to answer this question either to predict the loan defaulters or predict the credit limit anyone can get. We can see performance of the both below.

Predicting loan defaulters (Default Payment next month default)

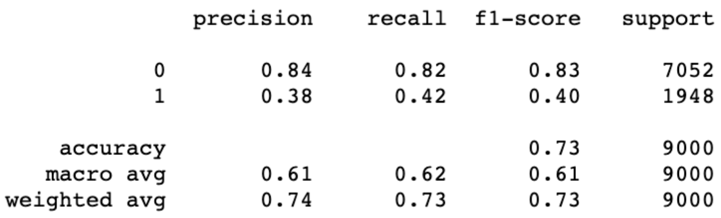


```
credit_csv.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 30000 entries, 0 to 30000
Data columns (total 30 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   ID                                  30000 non-null  int64
 1   LIMIT_BAL                           30000 non-null  int64
 2   MARRIAGE                            30000 non-null  int64
 3   AGE                                 30000 non-null  int64
 4   PAY_0                               30000 non-null  int64
 5   PAY_2                               30000 non-null  int64
 6   PAY_3                               30000 non-null  int64
 7   PAY_4                               30000 non-null  int64
 8   PAY_5                               30000 non-null  int64
 9   PAY_6                               30000 non-null  int64
 10  BILL_AMT1                           30000 non-null  int64
 11  BILL_AMT2                           30000 non-null  int64
 12  BILL_AMT3                           30000 non-null  int64
 13  BILL_AMT4                           30000 non-null  int64
 14  BILL_AMT5                           30000 non-null  int64
 15  BILL_AMT6                           30000 non-null  int64
 16  PAY_AMT1                            30000 non-null  int64
 17  PAY_AMT2                            30000 non-null  int64
 18  PAY_AMT3                            30000 non-null  int64
 19  PAY_AMT4                            30000 non-null  int64
 20  PAY_AMT5                            30000 non-null  int64
 21  PAY_AMT6                            30000 non-null  int64
 22  SEX_female                          30000 non-null  uint8
 23  SEX_male                            30000 non-null  uint8
 24  EDUCATION_graduate school           30000 non-null  uint8
 25  EDUCATION_high school               30000 non-null  uint8
 26  EDUCATION_other                     30000 non-null  uint8
 27  EDUCATION_university                30000 non-null  uint8
 28  default payment next month_default  30000 non-null  uint8
 29  default payment next month_not default  30000 non-null  uint8
dtypes: int64(22), uint8(8)
memory usage: 5.5 MB
```

Here Column 29 is the dependent variable and column 1 to 28 are the input variables.
The best data set is post discretization of AGE and LIMIT_BAL columns.

Accuracy with Naive Data Set for Decision Tree Classifier (No Discretization)

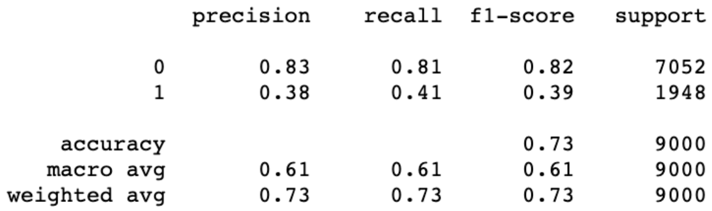


```
              precision    recall  f1-score   support

           0       0.84      0.82      0.83      7052
           1       0.38      0.42      0.40      1948

    accuracy                           0.73      9000
   macro avg       0.61      0.62      0.61      9000
weighted avg       0.74      0.73      0.73      9000
```

Accuracy Post Discretization for Decision Tree Classifier - AGE Column



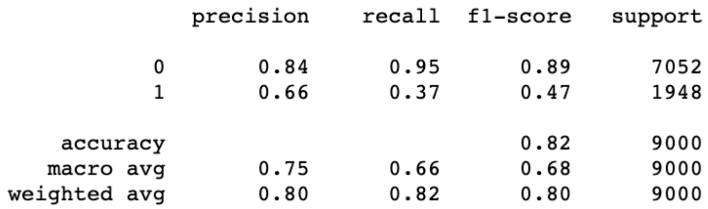
```
              precision    recall  f1-score   support

           0       0.83      0.81      0.82      7052
           1       0.38      0.41      0.39      1948

    accuracy                           0.73      9000
   macro avg       0.61      0.61      0.61      9000
weighted avg       0.73      0.73      0.73      9000
```

Accuracy Post Discretization for Decision Tree Classifier - AGE + LIMIT BAL Column

This accuracy is with the best data set so far identified.

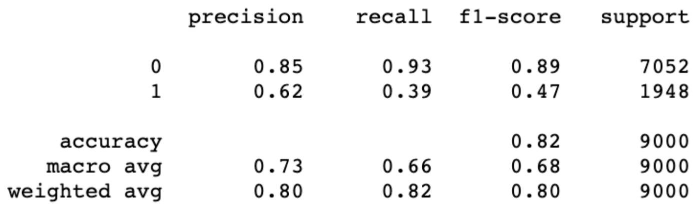


```
              precision    recall  f1-score   support

           0       0.84      0.95      0.89      7052
           1       0.66      0.37      0.47      1948

    accuracy                           0.82      9000
   macro avg       0.75      0.66      0.68      9000
weighted avg       0.80      0.82      0.80      9000
```

Accuracy with Best Data Set for Random Forest Classifier

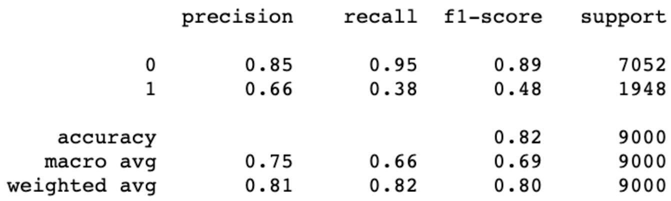


```
              precision    recall  f1-score   support

           0       0.85      0.93      0.89      7052
           1       0.62      0.39      0.47      1948

    accuracy                           0.82      9000
   macro avg       0.73      0.66      0.68      9000
weighted avg       0.80      0.82      0.80      9000
```

Accuracy with Best Data Set for Gradient Boosting Classifier



```
              precision    recall  f1-score   support

           0       0.85      0.95      0.89      7052
           1       0.66      0.38      0.48      1948

    accuracy                           0.82      9000
   macro avg       0.75      0.66      0.69      9000
weighted avg       0.81      0.82      0.80      9000
```

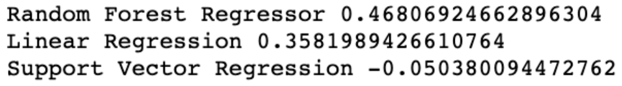
Inference

The accuracy improved from 0.74 precision to 0.80 precision post discretizing AGE & LIMIT_BAL columns. The accuracy didn’t improve much when trying Random Forest or Gradient Boosting Classifier.

Predicting Credit Balance (LIMIT_BAL)

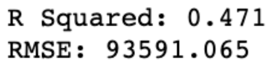
Refer to figure 1,Column 1 is the dependent variable and column 2 to 29 are the input variables.
The best data model for this data set is Random Forest Regressor.

Cross Validation Score for models



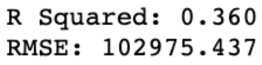
```
Random Forest Regressor 0.46806924662896304
Linear Regression 0.3581989426610764
Support Vector Regression -0.050380094472762
```

Random Forest Regressor



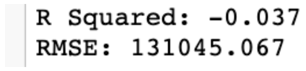
```
R Squared: 0.471
RMSE: 93591.065
```

Linear Regressor



```
R Squared: 0.360
RMSE: 102975.437
```

Support Vector Regressor



```
R Squared: -0.037
RMSE: 131045.067
```

Inference

Random Forest Regressor is the one with the lowest RMSE value. This one is gotten with the NAIVE data set with no discretization.
The Discretization of AGE, LIMIT_BAL, BILL_AMT & PAY_AMT columns is only making the performance of the model poor.

CONCLUSION

Since we have constructed a better model, it's best for CREDIT ONE to use the models created for predicting defaulters and amount of credit they can be approved before hand to improve the accuracy of the credit score and thereby helping clients.