Problem:

Credit one is a business that evaluates customers and gives approval for the loans and sets credit limits for the customers of other businesses. There is an increase in customer default rates. A better model should be built that can predict the credit limit for the customer.

Cleaning and Pre-processing:

The given data had some duplicated header rows which needed to be deleted. The column name “default payment next month” has been changed to “DEFAULT STATUS” to match the naming conventions of other columns

Encoding Data:

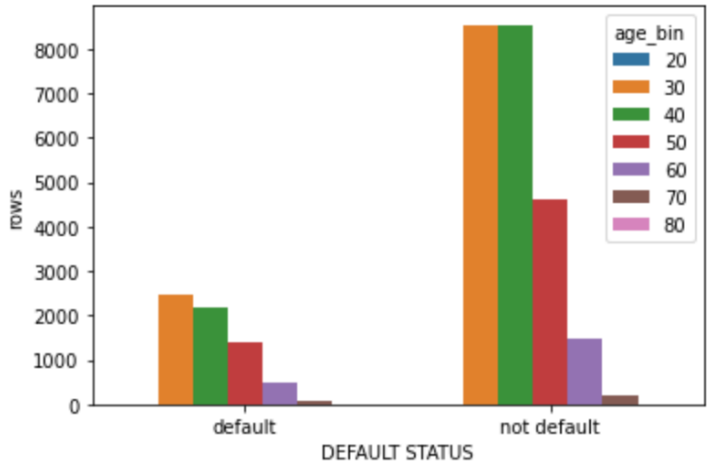
The data in three columns were of object data type. Column mapping function has been used to create columns which have numeric values for each distinct value in those columns.

```
# (default =1  ; non-default = 2)
df['DEFAULT']=df['DEFAULT STATUS'].map({'default': 1, 'not default' : 2})
#(Gender: male =1; female=2)
df['GENDER']= df['SEX'].map({'male':1, 'female': 2})
# Education 'graduate school': 1,'university': 2,'high school': 3, 'other' : 4
df['ED_LEVEL']= df['EDUCATION'].map({'graduate school': 1,'university': 2,'high school': 3, 'other' : 4})
```

Exploratory Data Analysis:

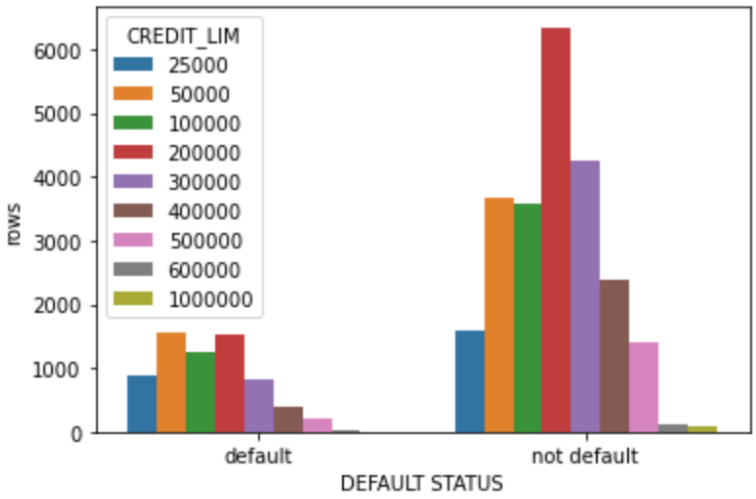
Two additional columns are created to hold discrete values for age and credit limit. These columns were created for EDA and to use them in classifier algorithms.

age_bin EDA:



The EDA above shows that there is not much relationship between age and default status. The age group 20-30 seems to have more customers. Both the default and non default status is high for that age group.

CREDIT LIMIT EDA:



The EDA above shows that there is not much relationship between credit limit and default status. There are more customers in the 100-200K and in the 25-50K credit limit. Both the default and non default status is high for the customers in these credit limit groups.

Predicting Credit Limit

Predicting Credit Limit (as a binned variable) using classifier algoirthms:

Credit Limit was kept as a dependent variable and all the other attributes were kept as independent variables. Three classifier algorithms are used.

Algorithm	Accuracy
Decision Tree Classifier	43%
Random Forest Classifier	39%
Gradient Boosting Classifier	48%

The Gradient Boosting Classifier provided better accuracy of 48% but it is still not a significant percentage to be able to predict the Credit Limit based on other available data.

Predicting Limit Bal (continuous data) using Regression algoirthms:

Limit Bal was kept as a dependent variable and all the other attributes were kept as independent variables. Three regression algorithms are used.

Random Forest Regressor 0.4669700418150324
Linear Regression 0.35153665570726883
Support Vector Regression -0.05035048153050248

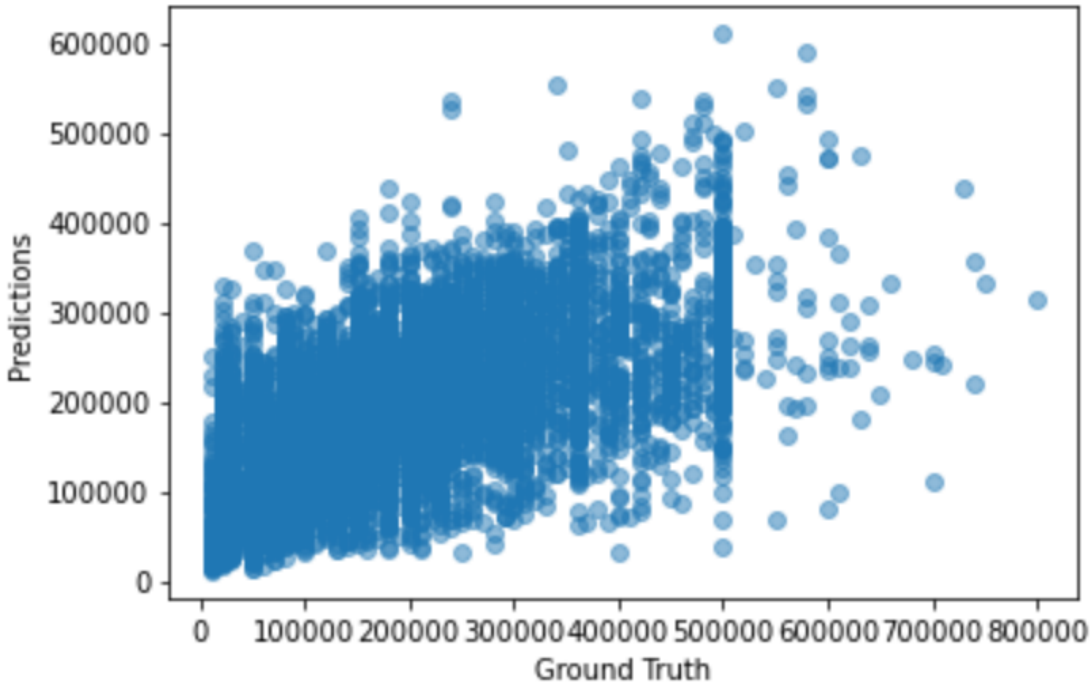
Since the regression scoring was given a value of r2. The value which is close to 1 is selected as the best model to further explore and plot. In our case, Random Forest Regressor is used.

For the Random Forest Regressor the following is the output of the model.

- R Squared : 0.466
- RMSE: 94018.193

The R Squared value of 46% indicated that the model can predict limit balance based on the other attributes with 46% accuracy.

The RMSE value of $94018.19 indicates that the deviation between the predicted limit balance to actual balance is $94000. It is a large amount for the credit limit.



Predicting Default Status:

Predicting Default Status using classifier algorithms:

Default was kept as a dependent variable and all the other attributes were kept as independent variables. Three classifier algorithms are used.

Algorithm	Accuracy
Decision Tree Classifier	82%
Random Forest Classifier	80%
Gradient Boosting Classifier	82%

The Decision tree Classifier provided better accuracy of 82% for predicting Default Status .

Predicting Default Status using Regression algorithms:

Random Forest Regressor 0.1823640937540296
Linear Regression 0.12062633892179626
Support Vector Regression -0.08516928276399234

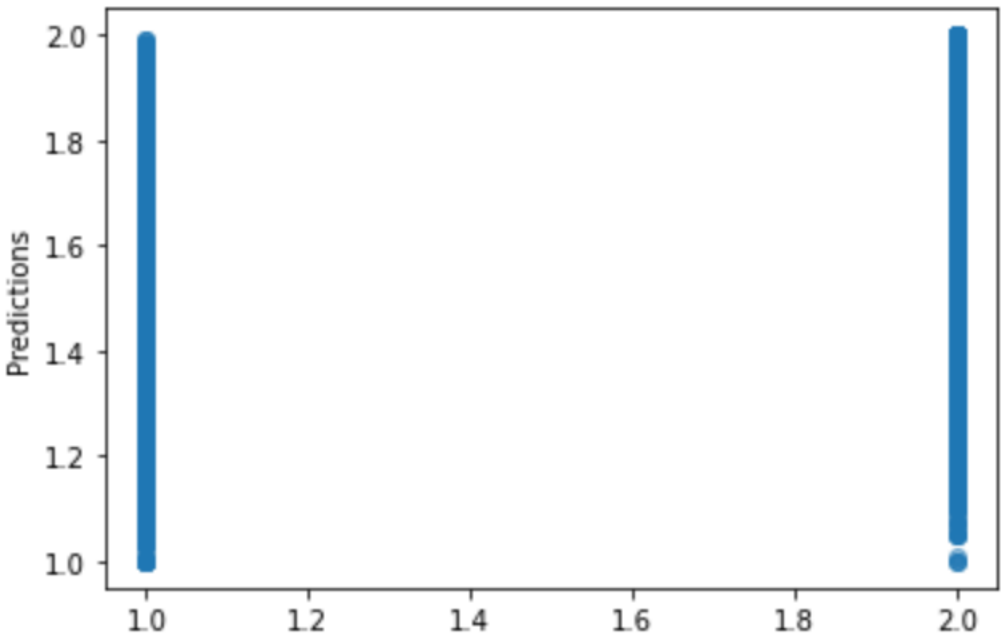
Random Forest Regressor, the following is the output for the regression algorithm.

R Squared : 0.171
RMSE: 0.375

The R Squared value of 0.171 indicates that the default status can be predicted with 17% accuracy.

The RMSE value of 0.375 indicates the deviation between the predicted default status to actual default status.

```
R Squared : 0.172
RMSE: 0.375
```



Conclusion:

Both the classifier algorithm and the regression algorithm predicted the limit balance with almost similar accuracy. Though it is not a statistically significant percentage, the prediction accuracy stood close to 45% across different algorithms.

Interesting point to note is that RMSE in Random Forest Regressor for the limit balance came to 94018.193. Also the plot has too many outliers for the high value. Eliminating the outlier datas with very high value may produce better prediction of limit balance.

The classifier algorithm is better at predicting default status with 82% accuracy, compared to the regression algorithms.