Course 2 Task 3: Build and Evaluate Models
Credit One Report

Problem: 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.

Questions to Investigate:

1. How do you ensure that customers can/will pay their loans?
2. Can we approve customers with high certainty?

Based on a thorough EDA (previous submission), seven features were selected to predict the dependent variable (DV, credit): 'default', 'AGE', 'MARRIAGE', 'EDUCATION_graduate school', 'PAY_2', 'BILL_AMT5', 'PAY_AMT6'. These features are most correlated with the DV. Such selection will avoid the overfitting issue resulting from using all the possible variables as features.

The features and DV are used to build a regression model as the type of the DV is continuous. Three models were built so that the most appropriate one would be identified. Random Forest Regressor is selected as it has the highest accuracy score (see below) and it is used to make predictions what credit limit a customer should be assigned.

```
Random Forest Regressor 0.3605401264433903
Linear Regression 0.30485340464174787
Support Vector Regression -0.05103950390038747
```

Results shows the trained model performance was pretty good (R^2 = 0.89, close to 1.0**).** In contrast the test model performance was less satisfactory (R^2 = 0.34, farther from 1.0).

```
model.score(X_train,y_train)
0.8855209113732851
```

```
model.score(X_test,y_test)
 0.3441626132817739
```

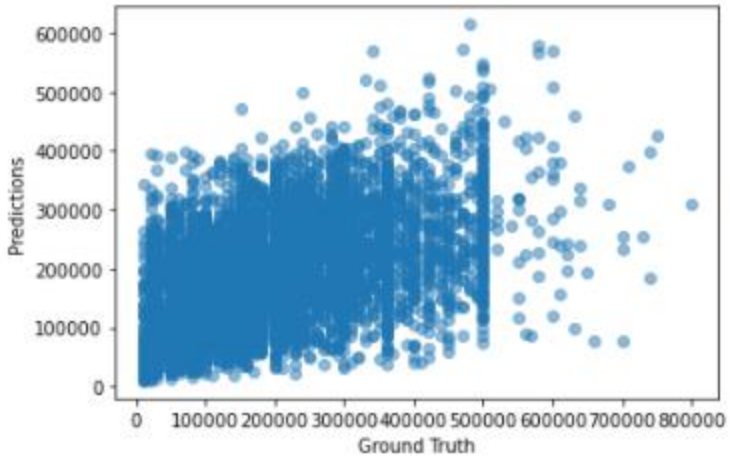
However, the selected features are the best that could be used for the analysis and the trained model performance was good. Predictions were proceeded to be made and results were evaluated. We have an accuracy score of approximately 34%.

```
r2_score(y_test,predictions)
0.3441626132817739
```

R^2 score tells us how well our model is fitted to the data by comparing it to the average line of the dependent variable. If the score is closer to 1, then it indicates that our model performs well versus if the score is farther from 1, then it indicates that our model does not perform so well. Parameter tuning was then applied to achieve higher accuracy of the model. Unfortunately, 0.34 is the highest accuracy that could be generated.

Finally, results were plotted to check a comparison between the known values in the test set (ground truth) and the predictions made by the model. It seems the association is good. Since higher numbers are better, correlation was run to obtain the coefficients. 0.59 is a good number and the model is retained.

```
plt.scatter(y_test, predictions, alpha = 0.5)
plt.xlabel('Ground Truth')
plt.ylabel('Predictions')
plt.show();
```




```
scipy.stats.pearsonr (predictions, y_test)
```

```
(0.5918249459815752, 0.0)
```

In order to ensure that customer can or will pay their loans, an appropriate credit limit should be assigned. With the current data, we are constrained to only use such attributes as default, age, marriage, education_graduate school, payment history, bill amount, and payment amount to make prediction. The 34% accuracy indicates that we cannot approve customers with high certainty.

From the customer perspective and informed by my analysis, step 3: data collection (BADIR Framework) can be further strengthened. Once better data are collected, the previously stated three models can be applied to build models and results will be evaluated to see if the accuracy is improved.

Specifically, I recommend the data of the following variables are collected.

- Number of times unemployed in past few years
- Current employment status (part-time, full-time, unemployed)
- The number of credit cards held
- Whether a customer has saving/checking bank accounts
- FICO score

Building and evaluating models is an iterative process. Using a better dataset, an EDA can be performed again to decide on new set of features that might make better predictions of the credit limit.