Silver Sterling Credit Card Default Risk

1.What would be your dependent and independent variables ?

- Dependent - 'SEX', 'EDUCATION','MARRIAGE', 'AGE', 'PAY_1', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT6', 'PAY_AMT1', 'PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6'
- Independent - default payment next month

2. How would you check for any outlier?

- Outliers can be detected as points on a boxplot. As outliers depends on the business problem, in this case extreme values represent the high-profile customers.
  It can be identified by the formula: (75^{th} percentile + K * IQR) where K =1.5 or 2.5 depending on the business

3. How would you check for multicollinearity amongst the variables?

- Check for multicollinearity can be done through VIF, in this case features with VIF more than 5 have been removed.

4. What statistical method would you use here and why?

- Methods like Logistic Regression, Decision Tree and Random Forest were used but Random Forest had the highest accuracy of 0.8196 and also solves the problem of overfitting.

5. Explain the statistical output in business terms

- Accuracy of 0.8196 suggests that 81.96 % times the defaulters and non defaulters were correctly identified. This could be useful in lending out credit cards to the right customers (8 out of 10) and lowering the risk of defaulting customers.

6. From the perspective of risk management, what is more important? The overall classification result credible and non-credible results or the result of predictive accuracy of the estimated probability of default?

- Assumption: Overall classification is distinguishing the customers as credible or non credible based on 1s' and 0s' where as predictive accuracy is the estimated probability of a customer defaulting (i.e a probability value between 0-1)
- In this case the Overall classification is what matters as we have to classify the defaulters and non defaulters and not to predict just the probability of them defaulting.

7. Are these variables good enough for a sound model or any some other attributes could have added value?

- Only 82% of the variance in data has been identified, other variables which could be useful are:
  1. Occupation of the customer
  2. Additional Loan taken
  3. Additional Credit Card user

* Codes for the analysis can be found in the python notebook