Objective:

To determine the correct predictive model to predict the default on credit card payment, using the data from a bank in Taiwan.

Data:

The data is available at https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset. This data is collection of customer data in a Taiwan bank. The data is donated by the Department of Information Management, Chung Hua University, Taiwan.

The data is checked for integrity, missing values and outliers and found that there is no missing value. I did not find any data integrity issues as well. Due to the nature of the financial data, it's not possible to determine the outliers. Almost every column can have a wide range of values.

Important Findings:

Detailed exploratory data analysis findings are in the final consolidated report. Following are the important findings.

- In the dataset given, 28% defaulted on the payment. That's much larger than 2.54% delinquency rate on credit card debt for all commercial banks in the U.S.
- Interestingly less educated people were more prompt in paying credit card bills!
- Using pearson correlation coefficient and heatmap, it's proven that the payment amount for previous month is correlated with repayment status for current month.

Machine Learning Models:

"One hot encoding" is applied to categorical features like EDUCATION. Then, the following machine learning models are used to predict the default on credit card payment.

1. Random Forest Classifier
2. K-Nearest Neighbors Classifier
3. Logistic Regression
4. Linear Regression
5. Decision Tree Classifier
6. Bagging Classifier
7. Gradient Boosting Classifier
8. SVM (Support Vector Machines)
9. XGBoost Classifier
10. Neural Networks

In all the models, I have tuned the parameters to obtain the best parameters to construct the model. I have also plotted the important features identified by each model, if the model supports the identification of features.

Linear regression is not suitable for classification problem such as this (default or no-default). Among the other models, SVM and Neural Networks performed poorly in terms of processing time.

Model	**Training Accuracy**	**Test Accuracy**	**AUC Score**
Random Forest	0.999	0.815	0.6555
K-Nearest Neighbors	0.816	0.7678	0.7678
Logistic Regression	0.810	0.808	0.6042
Decision Tree	0.824	0.819	0.6493
Bagging	0.980	0.806	0.641
Gradient Boosting	0.829	0.821	0.6587
SVM	0.994	0.780	0.5054
XGBoost	0.825	0.822	0.6553
Neural Networks	0.821	0.819	0.6421

*Highlighted models are top 3 models in terms of accuracy and AUC score.

Among the remaining seven models, Gradient Boosting Classifier, XGBoost Classifier and Random Forest Classifier are top 3 models. These are able to predict which customer would default on the payment next month, with better accuracy.

Among these three models, I recommend Random Forest Classifier because the accuracy and AUC score are almost at par with the other two models. It also takes many features into account while predicting the default. Gradient Boosting and XGBoost place very high weightage on just one feature (PAY_0) which is not a good idea because we can't let only one feature skew the prediction outcome.