Assignment Cover Sheet

Please fill out and insert at the top of your README or Google Doc as preferred.

Student Name and Number as per student card: Akash Sachin Nikam | 20054691

Programme: M.Sc. in Data Analytics

Lecturer Name: Alexander Victor

Module/Subject Title: Programming for Data Analysis Project

Assignment Title: Credit Default Prediction

By submitting this assignment, I am confirming that:

- **This assignment is all my own work;**
- **Any sources used have been referenced;**
- **I have followed the Generative AI instructions/ scale set out in the Assignment Brief;**
- **I have read the College rules regarding academic integrity in the QAH Part B Section 3, and the Generative AI Guidelines, and understand that penalties will be applied accordingly if work is found not to be my/our own.**
- **I understand that all work submitted may be code-matched report to show any similarities with other work.**

Note: Technical support is available to students between **0830- 2000 hrs (Mon-Fri), 0930-1630 (Sat) only**. There is no technical support after 2000 hrs. It is your responsibility to ensure that you allow time to troubleshoot any technical difficulties by uploading early on the due date.

Credit Default Prediction: Project Report

Section I: Introduction

The aim of this project is to conduct an extensive data-driven analysis of credit card default behavior based on a real-life dataset in a financial setting. Credit default prediction is a very vital process in financial institutions since it has a direct impact on risk management, lending policy, and profitability. The early detection of high-risk customers can enable banks and lenders to implement some interventions to minimize possible losses.

For the purpose of this analysis, the "Default of Credit Card Clients" dataset was chosen. It is part of the UCI Machine Learning Repository and contains demographic and financial data on 30,000 clients of a bank in Taiwan. It includes variables like age, sex, credit limit, billing amounts, repayment history, and whether or not the client defaulted on payment the next month.

The main aim of this project is to utilize programming and statistical methodology to:

- Investigate trends and patterns within the dataset using data visualisation
- Develop forecasting models to determine if a client will default
- Compare the performance of various classification algorithms

The project was implemented in Python using Jupyter Notebook and some of the most popular libraries, including pandas, matplotlib, seaborn, and scikit-learn. The project result can be applied to real-world credit scoring and financial risk management.

Section II: Data Description

Data utilized for this project is the "**Default of Credit Card Clients**" dataset, which is available via the UCI Machine Learning Repository. It consists of financial and demographic data of 30,000 individual credit card customers that had been gathered by a bank in Taiwan. The data are in tabular form with 25 columns (features), including input variables and a single binary target variable.

The target variable is default payment next month, which indicates whether the client defaulted (1) or not (0). Approximately 22% of the clients defaulted, so the dataset is slightly imbalanced. The other features are:

- Demographic attributes: like AGE, SEX (1 = male, 2 = female), MARRIAGE (1 = married, 2 = unmarried), and EDUCATION level
- Credit limit: LIMIT_BAL, credit given to the client
- Repayment status: PAY_0 to PAY_6, indicating the client's repayment status in the last six months
- Bill statement amounts: BILL_AMT1 to BILL_AMT6
- Historical payments: PAY_AMT1 to PAY_AMT6

The variables are numeric, and the dataset is therefore very suitable for statistical and machine learning analysis. The presence of both temporal and static data offers a fertile field for investigating determinants of credit default.

Section III: Method of Data Analysis

Analysis of the data was performed in Python with a Jupyter Notebook interface, and key libraries included pandas, matplotlib, seaborn, and scikit-learn. The analysis was performed in a systematic workflow of data cleaning, exploratory data analysis (EDA), model construction, and testing.

Data Preprocessing

Data was originally loaded from a CSV file. There was a header problem that was resolved by reloading the data with the appropriate header row. Basic checks revealed no missing values. The column names were tidied, and the target variable default payment next month was renamed to default for simplicity. All features were numeric and ready to be explored. The ID column was dropped as it was not predictive.

Figure 1: Distribution of Default vs Non-default Clients

Exploratory Data Analysis (EDA)

EDA involved developing descriptive statistics and plotting distributions in the form of histograms, boxplots, and count plots. Univariate analysis was conducted on significant features such as LIMIT_BAL (credit limit) and AGE. Categorical variables such as SEX, EDUCATION, and MARRIAGE were plotted against the default variable to observe group differences. A correlation heatmap was generated to determine relationships between variables, observing strong positive correlations between PAY_0 to PAY_6 and default.

Managing Class Balance

As the data was imbalanced (just 22% defaulters), techniques were used to increase sensitivity. One of them was to set class_weight='balanced' in classifiers to penalize misclassification of the minority class.

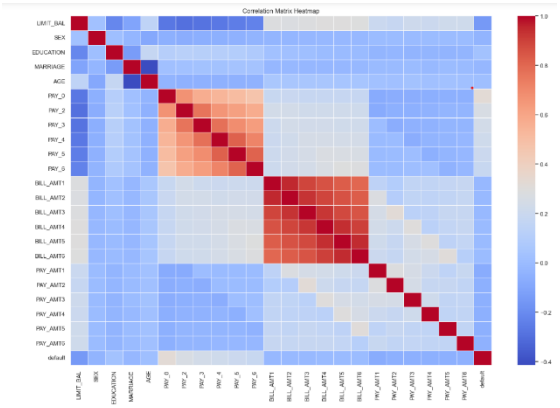


Figure 2: Correlation Matrix of All Features

Model Building

Three classification models from scikit-learn were used:

- Logistic Regression (plain and class-weighted)
- Decision Tree Classifier
- Random Forest Classifier

Models were trained effectively on 80% of the data and the remaining 20% data was used for testing. The accuracy, precision, recall, and F1-score were used to assess the models. Confusion matrices and classification reports were also generated for all the models. Finally, feature importance was extracted from the Random Forest model to identify the variables with the highest influence.

Section IV: Results

The findings of the analysis are laid out in two broad sections: model performance results and exploratory findings.

Exploratory Data Analysis Results:

The target variable default was imbalanced, with about 22% of the clients defaulted and 78% not defaulted. Histograms indicated that the majority of the clients had credit limits of less than 200,000 and ages ranging from 25 to 40 years. Boxplots indicated the existence of high-value outliers in bill amount and credit limit features.

Counts plots indicated default was just more common among male customers, university or high school graduates, and singles. These differences were not radical, though. The heatmap of correlations demonstrated that the most highly correlated with the default variable were the repayment history variables (PAY_0 to PAY_6).

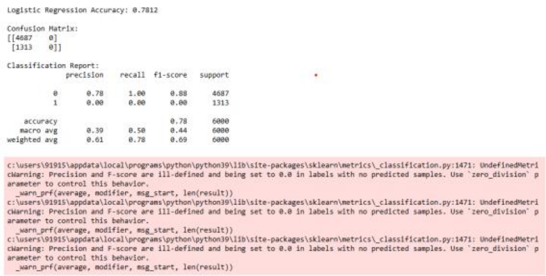
Model Performance Results:

Three models were built and examined:

- Logistic Regression (Unweighted) was very accurate (78%) but did not predict any defaults because of class imbalance.
- Weighted Logistic Regression made recall among defaulters 73%, yet reduced accuracy to 56%.
- Decision Tree provided a trade-off between interpretability and performance with 73% accuracy and 40% recall.
- The best among all models was Random Forest with accuracy 81.5%, precision 65%, and recall 34% for the default class.

A feature importance plot from the Random Forest model showed that the most predictive of default

were repayment status (PAY_0), credit limit, and recent payment and billing amounts.



```
Logistic Regression Accuracy: 0.7812

Confusion Matrix:
[[4687    0]
 [1313    0]]

Classification Report:
              precision    recall  f1-score   support

           0       0.78      1.00      0.88      4687
           1       0.00      0.00      0.00      1313

    accuracy                           0.78      6000
   macro avg       0.39      0.50      0.44      6000
weighted avg       0.61      0.78      0.69      6000

c:\users\91915\appdata\local\programs\python\python39\lib\site-packages\sklearn\metrics\_classification.py:1471: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
c:\users\91915\appdata\local\programs\python\python39\lib\site-packages\sklearn\metrics\_classification.py:1471: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
c:\users\91915\appdata\local\programs\python\python39\lib\site-packages\sklearn\metrics\_classification.py:1471: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Figure 3: Logistic Regression Confusion Matrix and Classification Report

These findings indicate that trends of financial behavior—particularly recent payment behavior—are the best predictors of credit risk in this sample.

Section V: Discussion

The study showed many significant observations and patterns consistent with real-life expectations for credit risk forecasting.

One of the most important observations was that long-term and, more recently, repayment behavior is very predictive of credit default. The PAY_0 feature (repayment status in the most recent month) was most highly correlated with default and also scored as the most important feature by the Random Forest model. This makes sense according to financial practice, where recent repayment behavior is often seen to signal short-term risk.

Lower credit limits (LIMIT_BAL) also increased the probability of default for customers, either because they had less room to borrow or because they were more financially constrained. Bill amount and payment amount in six months also played an important role in model performance, suggesting that trends of debt build-up or unstable payments can identify risky customers.

Demographic variables like AGE, SEX, and MARRIAGE were weaker predictors of default. While there were small differences (e.g., default rates slightly higher for single or male customers), each of these factors by itself was not a good predictor. This reflects that behavior data rather than demographic data is more informative in predicting default risk.

From a model performance perspective, Random Forest classifier achieved the best balance overall among precision, recall, and accuracy. It excelled especially at not making false positives, which in financial situations where misclassifying stable clients as defaulters hurts customer relationships.

Nonetheless, there were a few project limitations. First, although the Random Forest model worked extremely well, it had a modest recall on defaulters at just 34%, which meant that a high proportion of defaulters went uncaptured. Second, the data, although rich, is static and historic—it lacks any dynamic or economic context, i.e., unemployment rates or inflation. Adding more contextual or behavioral features (e.g., transactional data) would enhance predictive power still further.

Section VI: Conclusion

The objective of this project was to utilize programming and data analysis competencies to investigate and forecast credit card default behavior using a real-life dataset. By cleaning, exploring, visualizing, and modeling the data, we were capable of gaining insightful information on client behavior and financial risk drivers.

The research found that the most recent payment status, especially PAY_0, was the most predictive of default. Credit limits and billing/payment amounts were also important features, while demographic variables had little predictive value.

Of the models attempted, Random Forest worked best with 81.5% accuracy and a balance between precision and overall generalization. Logistic regression, while interpretable, was hindered by the unbalanced dataset. Decision trees provided an interpretable alternative with slightly worse performance.

While the models worked nicely, they can be improved even more by incorporating additional behavioral or economic features, and also by employing more advanced techniques such as hyperparameter tuning or ensemble boosting methods.

The outcomes of this project are useful for lending organizations that aim to enhance credit risk assessment and formulate more advanced, data-driven lending policies.

Section VII: References

Lichman, M. (2013). UCI Machine Learning Repository: Default of Credit Card Clients Dataset. [online] Available at: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients [Accessed 1 Apr. 2025].

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, pp.2825–2830. Available at: https://scikit-learn.org/stable/

Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), pp.90–95. Available at: https://matplotlib.org/

Waskom, M. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. Available at: https://seaborn.pydata.org/

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). [online] Available at: https://pandas.pydata.org/ Van Rossum, G., & Drake, F.L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.