Project Report

---

1. Methodology, Approach, and Model Selection Rationale

The primary objective of this project was to build a model that predicts the likelihood of loan default, leveraging data cleaning, exploratory data analysis (EDA), and predictive modelling techniques. The methodology included several key stages:

- **Data Cleaning**: Addressed missing values through imputation, removed unnecessary columns, and handled categorical data via encoding methods. This ensured the data was prepared in a consistent, usable format.
- **EDA**: Conducted analysis to understand the target variable's distribution and relationships with other features, using visualisations to detect patterns, outliers, and important variables.
- **Feature Engineering**: Created additional features, such as income-to-loan ratio, to enhance model performance.
- **Model Selection**: Started with a baseline model to establish initial performance metrics, followed by a Random Forest model, selected due to its robustness in handling tabular data and feature importance assessment capabilities. SMOTE was applied to address class imbalance, ensuring accurate predictions across both loan default classes.

The choice of Random Forest was justified by its interpretability, feature importance insights, and ability to handle complex interactions among features without extensive preprocessing.

---

2. Advantages and Limitations of the Chosen Model

- **Advantages**:
    - **Interpretability**: Random Forests provide insights into feature importance, allowing the business to understand key drivers of loan defaults.
    - **Robustness**: The model performs well with complex datasets and can handle non-linear relationships.
    - **Imbalance Handling**: With SMOTE applied, the model manages imbalanced datasets effectively, which is crucial for accurate default prediction.
- **Limitations**:
    - **Computationally Intensive**: Random Forests require more memory and processing power, especially with large datasets.
    - **Lack of Real-time Adaptability**: Although effective in batch processing, Random Forests may be slower in real-time environments compared to simpler models.

---

3. Architecture of the Final Solution

The solution architecture integrates a predictive model hosted via FastAPI and a user interface through Streamlit:

- **Data Pipeline**: Data preprocessing is conducted in the Jupyter Notebook, where missing values are imputed, categorical features encoded, and SMOTE applied for class balancing.
- **Model Pipeline**: The trained model is saved and deployed through FastAPI, which handles prediction requests in real time.

- **Frontend Interface**: Streamlit is used as a frontend to collect inputs from the user and communicate with the FastAPI backend, providing an accessible and interactive experience for end-users.

---

4. Considerations on Deployment and Scalability

To ensure efficient deployment and scalability, the following considerations were made:

- **Containerisation**: Docker can be used to package the FastAPI application, making it easier to deploy on cloud services.
- **Load Balancing**: For high volumes, load balancers can distribute prediction requests across multiple instances, preventing delays.
- **Microservices Architecture**: FastAPI can be integrated as a microservice within the company's broader IT infrastructure, allowing it to interact with other services, such as customer data management or alerting systems.
- **Use in Business as Usual (BAU)**: The model can be embedded within loan processing workflows to score applications as they are submitted, providing real-time insights and aiding decision-making. Automating this prediction process allows risk assessment teams to focus on high-risk cases.

---

5. Estimated Impact and Return on Investment (ROI)

- **Impact**: By predicting loan defaults accurately, the business can take proactive measures to minimise risk exposure. This leads to a higher quality loan portfolio, reduces potential losses from defaults, and optimises customer screening.
- **ROI**: Implementing this predictive model in BAU operations can potentially reduce default rates by targeting high-risk loans. With early identification, the company may adjust terms, provide financial counselling, or deny applications to minimise loss. The model's cost-saving and profit-generating potential make it a valuable asset for the business.