Project Report: Customer Default Risk Prediction

---

1. Introduction

The objective of this project is to predict **customer loan default risk** using a synthetic dataset of **10,000 records and 20 features**.

The dataset contains both **numerical** and **categorical** variables such as age, income, loan_amount, credit_score, education, employment, etc. The **target variable** is target_default_risk, where:

- 0 = Customer does not default
- 1 = Customer defaults

This project demonstrates the **end-to-end Data Science workflow**, including **EDA, preprocessing, feature engineering, model training, hyperparameter tuning, and evaluation**.

---

2. Exploratory Data Analysis (EDA)

- Checked dataset shape (10000, 20) and feature data types.
- Found both numeric and categorical features with missing values.
- **Numeric insights:**
    - income and loan_amount are skewed with outliers.
    - credit_score is mostly between 500–800.
- **Categorical insights:**
    - "Bachlors" typo in education fixed to "Bachelors".
    - Employment type mostly "salaried".
- **Correlation:**
    - Positive correlation between income and loan_amount.
    - Negative correlation between credit_score and default risk.
- **Target balance:** ~60% non-default, ~40% default → handled with **SMOTE oversampling**.

---

3. Data Preprocessing

- **Missing values:** filled numeric with **median**, categorical with **most frequent**.
- **Categorical encoding:**
  - **Ordinal encoding** for education.
  - **One-hot encoding** for nominal variables (employment, etc.).
- **Scaling:** Applied **StandardScaler** for numeric features.
- **Feature engineering:** extracted signup_year and created income_per_dependent.
- **Outliers:** capped extreme values at 1st/99th percentile.

---

4. Modeling & Results

We trained and tuned **5 machine learning models** and **1 deep learning model (ANN)**.

4.1 Logistic Regression (Tuned)

- **Best Params:** C=1, solver=lbfgs
- **Results:**
  - Accuracy: **93.4%**
  - Precision: 96.18%
  - Recall: 90.83%
  - F1 Score: 93.43%

4.2 Decision Tree (Tuned)

- **Best Params:** max_depth=5, min_samples_split=2
- **Results:**
  - Accuracy: **93.6%**
  - Precision: 94.11%
  - Recall: 93.46%
  - F1 Score: 93.79%

4.3 Support Vector Machine (SVM) (Tuned)

- **Best Params:** C=1, kernel=linear
- **Results:**
  - Accuracy: **94.2%**
  - Precision: 96.81%
  - Recall: 91.71%
  - F1 Score: 94.19%

4.4 Random Forest (Tuned)

- **Best Params:** n_estimators=200, max_depth=15, min_samples_split=2
- **Results:**
  - Accuracy: **94.3%**
  - Precision: 95.78%
  - Recall: 92.98%
  - F1 Score: 94.36%

4.5 XGBoost (Tuned)

- **Best Params:** learning_rate=0.05, max_depth=7, n_estimators=300
- **Results:**
  - Accuracy: **96.2%**
  - Precision: 96.75%
  - Recall: 95.80%
  - F1 Score: 96.28%

4.6 Artificial Neural Network (ANN)

- 3 hidden layers (Dense layers, ReLU activations, Sigmoid output).
- Optimizer: Adam, Loss: Binary Crossentropy.
- Trained with early stopping.
- **Test Accuracy: 95.4%**

---

5. Model Evaluation

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	93.4%	96.2%	90.8%	93.4%
Decision Tree	93.6%	94.1%	93.5%	93.8%
SVM	94.2%	96.8%	91.7%	94.2%
Random Forest	94.3%	95.8%	93.0%	94.4%
XGBoost	**96.2%**	96.8%	95.8%	96.3%
ANN	95.4%	95%+	95%+	95%+

Key Takeaways:

- All models performed well (>93% accuracy).
- **XGBoost achieved the highest accuracy (96.2%)**.
- **ANN performed competitively (95.4%)**, showing deep learning can be effective.
- Logistic Regression and Decision Tree are weaker baselines.
- Ensemble methods (Random Forest, XGBoost) dominate.

---

6. Conclusion

- The dataset required **cleaning, encoding, scaling, outlier treatment, and feature engineering**.
- EDA revealed skewness, outliers, and class imbalance, which were successfully addressed.
- After **hyperparameter tuning**, ensemble models performed best.

- **XGBoost emerged as the most accurate model (96.2%)**, followed by **ANN (95.4%)**.
- These models provide a reliable system for predicting customer default risk.