Loan Purchase Prediction

1. Objective

The goal is to build a classification model that predicts whether a bank customer is likely to purchase a consumer loan. This will enable more targeted marketing campaigns, increasing conversion rates and reducing costs.

2. Dataset Description

The dataset contains information about customers' demographics, financial products, and service usage. The features include age, income, spending, account ownership, and a target variable indicating loan purchase.

Target variable: Has_Consumer_Loan (0: No, 1: Yes)

3. Feature Metadata

Customer_ID: Unique identifier for each customer.
Age: Age of the customer in years.
Years_Experience: Number of years the customer has been working.
Annual_Income: Customer's annual income (in $000).
ZIP_Code: Home ZIP code.
Family_size: Number of family members.
Avg_Spending: Average credit card spending per month (in $000).
Education_Level: 1 = Undergrad, 2 = Graduate, 3 = Advanced/Professional.
Mortgage: Value of mortgage loan (in $000).
Has_Securities_Account: 0 = No, 1 = Yes.
Has_CD_Account: 0 = No, 1 = Yes.
Uses_Online_Banking: 0 = No, 1 = Yes.
Has_CreditCard: 0 = No, 1 = Yes.
Has_Consumer_Loan: 0 = No, 1 = Yes (Target Variable).

4. Assignment Tasks

The assignment involves the following key steps:

1. 1. Import and inspect the dataset
2. 2. Visualize and explore data trends
3. 3. Clean and preprocess the dataset
4. 4. Apply feature engineering and transformation
5. 5. Build classification models: Logistic Regression, Decision Tree, KNN, Random Forest
6. 6. Tune model hyperparameters using cross-validation

7.   7. Evaluate models and choose the best one based on performance metrics

5. Performance Evaluation and Justification

The following table summarizes the evaluation metrics for the four models tested:

Model	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	0.857	0.686	0.762	0.968
Decision Tree	0.951	0.933	0.942	0.964
K-Nearest Neighbors	0.632	0.343	0.444	0.859
Random Forest	0.979	0.905	0.941	0.999

The model performance was evaluated based on four key classification metrics:
- Precision: measures the proportion of positive identifications that were actually correct.
- Recall: measures the proportion of actual positives that were correctly identified.
- F1 Score: the harmonic mean of precision and recall, providing a balance between the two.
- ROC AUC: reflects the model's ability to distinguish between classes.

Here's a breakdown and justification for each model's performance:

1. **Logistic Regression** performed decently across all metrics, especially ROC AUC (0.968). However, its recall (0.686) indicates it missed out on identifying a significant portion of actual loan takers. It is a reliable model but lacks in capturing enough positive cases for this use-case.

2. **Decision Tree** showed very strong results, with a high recall (0.933) and excellent F1 Score (0.942), making it a solid model. Its simplicity and interpretability make it suitable, although it might risk overfitting on unseen data.

3. **K-Nearest Neighbors (KNN)** underperformed significantly with the lowest recall (0.343) and F1 Score (0.444). This suggests that KNN is not an ideal model for this dataset, possibly due to high dimensionality or poor scaling impact.

4. **Random Forest** outperformed all other models. With a precision of 0.979 and ROC AUC of 0.999, it is not only excellent at correctly identifying loan takers but also maintains a low false-positive rate. Its balanced recall (0.905) and F1 Score (0.941) prove it is robust and generalizes well, making it the ideal choice for this problem.

✅**Conclusion:** Random Forest is the best-performing model for predicting consumer loan purchases. It achieves near-perfect classification results across all metrics, making it highly effective for optimizing bank marketing campaigns with minimal risk and maximum conversion.

6. Classification Report and Visual Comparison

Below is the classification performance of the best model (Random Forest Classifier):

Class	Precision	Recall	F1 Score
0	0.989	0.998	0.993
1	0.979	0.905	0.941
accuracy	0.988	0.988	0.988
macro avg	0.984	0.951	0.967
weighted avg	0.988	0.988	0.988

The following chart compares all classifiers visually based on their Precision, Recall, F1 Score, and ROC AUC scores:

Precision

Model	Value
Logistic Regression	0.86
Decision Tree	0.95
K-Nearest Neighbors	0.63
Random Forest	0.98

Recall

Model	Value
Logistic Regression	0.69
Decision Tree	0.93
K-Nearest Neighbors	0.34
Random Forest	0.90

F1 Score

Model	Value
Logistic Regression	0.76
Decision Tree	0.94
K-Nearest Neighbors	0.44
Random Forest	0.94

ROC AUC

Model	Value
Logistic Regression	0.97
Decision Tree	0.96
K-Nearest Neighbors	0.86
Random Forest	1.00