**🏦 Problem Statement – In Detail**

**1. What is the problem?**

The bank has tried to expand its market share by **issuing credit 💳and cash cards to many people**, including **unqualified applicants** (i.e., people who may not have the financial ability to repay).  
 At the same time:

* Many customers used the cards **excessively**, going beyond their repayment capability.
* This led to a **huge pile-up of unpaid credit**.

**2. Why is this a problem?**

* **High Default Risk:** Many cardholders are defaulting on payments.
* **Financial Loss:** The bank faces **revenue losses** from unpaid debt.
* **Trust Issues:** A financial crisis can damage **consumer trust** in the banking system.
* **Need for Prevention:** The bank must **predict future defaults** to manage this risk better.

**🎯 Goal of the Project**

The goal is **not just to classify** customers into risky or not risky, but to:

🔍 **Estimate the probability that a customer will default on their payment in the next month.**

This enables:

* **Risk control**
* **Better credit strategies**
* **Informed decision-making** on issuing cards or setting credit limits

**How will we solve the problem?**

We will use **data science and machine learning** techniques on a real-world credit card dataset. The dataset contains:

* Customer features like **credit limit, age, education, marital status**
* **Payment history** over 6 months
* **Bill statements and past payments**
* The **target variable**: Whether the customer defaulted the next month (1 = Yes, 0 = No)

We’ll analyze this data and build predictive models to estimate the **likelihood of default**.

**About the DATA**

This dataset is designed to predict whether a credit card customer will default on their next payment based on their demographic and financial history. Let's break down the description and explain each part clearly.

🎯 Objective (Problem Statement)

You are solving a binary classification problem, where the target variable is default payment (Y):

* Y = 1 → The person defaulted (failed to pay)
* Y = 0 → The person did not default

The goal is to use the other 23 variables (X1 to X23) to predict Y, i.e., whether a customer is likely to default on their next payment.

**The dataset covers:**

* Demographics (Gender, Education, Marital Status, Age)
* Financial profile (Total credit limit)
* Payment behavior (how late payments were, how much was billed, and how much was paid) — all over a 6-month historical period.

**Purpose of Using This Data**

**By analyzing these variables, the model will learn patterns of default, like:**

* People who regularly delay payments
* Those whose payments are lower than their bills
* Customers with low education or high outstanding bills
* Demographic groups that may default more often

**What We’re Trying to Do**

**We want to build a classification model (e.g., logistic regression, decision tree) that can:**

* Take in a new customer’s demographic and payment history
* Predict the probability that they will or will not default on their next credit card payment

**Project Lifecycle / Workflow (In Detail)**

**1. Problem Understanding**

* Understand business impact and the need for probability-based prediction.
* Define project objective: Predict the likelihood of defaulting next month.

**2. Data Collection & Understanding**

* Rows: 30000, Columns: 25
* Study the 23 independent variables (X1 to X23) and the target variable Y.
* Know what each column means:
  + Credit limit
  + Demographics (Gender, Education, Marital Status, Age)
  + Payment behavior (e.g., past due, payments made, bill amount)

**3. Data Cleaning**

* **Remove or fix invalid entries**:
  + Gender values in [1, 2]
  + Education values are not in [1, 2, 3, 4]

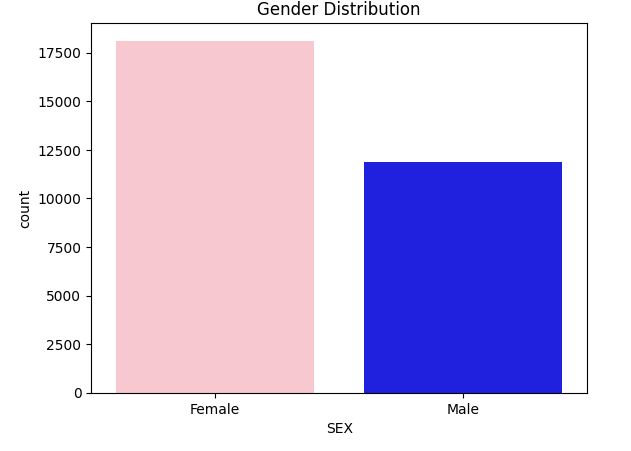
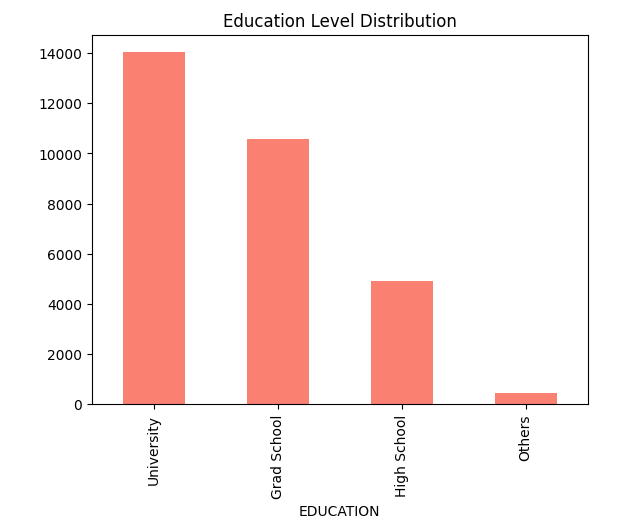
•In the Education column, we will replace 0, 5, and 6 with category 4 (Others), which is the most appropriate fallback

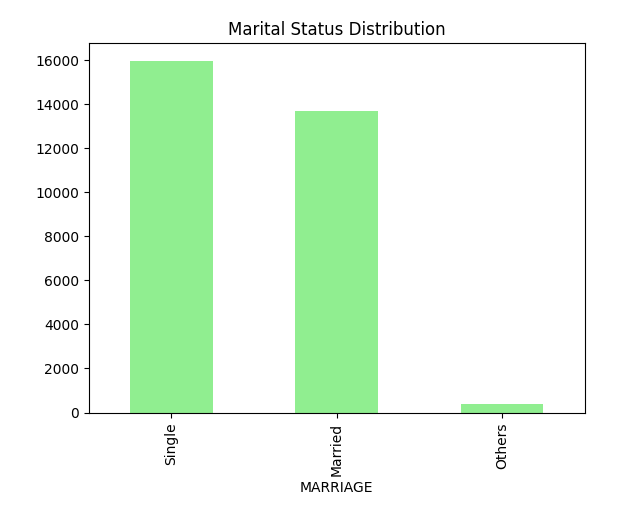
* + Marriage values are not in [1,2,3]

•In the Marriage column, we will replace the invalid 0 with 3 (Others), which is the safest fallback

* **Handle missing values** (if any)  
  There are no null or missing values in the dataset. All records are complete, which ensures data quality and reduces the need for imputation or cleaning
* **Check for duplicates, outliers**There are no duplicate ID values in the dataset.

**4. Exploratory Data Analysis (EDA)**

* Visualize:
* 
* 

****

**LIMIT\_BAL**

* **Heavily skewed** : Credit limits often vary drastically — from 10,000 to $10,00,000+.
* **Income-based differences** : High-income users may legitimately have very large credit limits, which show up as outliers.
* **Business logic applies** : What looks like an outlier statistically might be normal from a business or financial viewpoint.

**AGE** Normal distribution is not guaranteed: Ages are typically **right-skewed in financial datasets** (e.g., more young people with fewer older individuals).

Not necessarily an error: An 18-year-old and a 79-year-old are both valid ages.

**5. Statistical Analysis**

* **Descriptive Statistics**
* Average Limit balance is 167484.32266666667
* Average Age is 35.4855
* Minimum Limit balance is 10000
* Minimum Age is 21
* Maximum Limit balance is 1000000
* Maximum Age is 79
* Variance of Limit Balance is 16834455682.155386
* Variance of Age is 16834455682.155386
* Standard Deviation of Limit Balance is 129747.66156719506
* Standard Deviation of Age is 129747.66156719506
* Mode of Limit Balance is 0 50000
* Mode of Age is 0 29
* Range of Limit Balance is 990000
* Range of Age is 990000
* IQR of Limit Balance is 190000.0
* IQR of Age is 13.0
* Skew of Limit Balance is 0.992866960519544
* Skew of Age is 0.992866960519544
* kurtosis of Limit Balance is 0.536262896398668
* kurtosis of Age is 0.536262896398668
* **Final Summary and Key Inference:**
* The customer base primarily consists of working-age individuals around 35 years old, with most between 28 to 41 years, showing a slightly younger, right-skewed age distribution. This suggests the bank is targeting economically active individuals with potentially stable income.
* Credit limits show high variability, ranging from 10,000to10,00,000, with an average limit of around $1.67 lakhs. The distribution is positively skewed, indicating that while most customers have moderate limits, a few have very high credit lines, possibly reflecting a mix of middle-income and high-net-worth individuals.
* **Key Insight**: The bank's credit card portfolio serves a young, financially active population, with wide credit exposure, highlighting the need for risk-based segmentation and credit scoring models to manage high-limit accounts prudently.

**Descriptive Statistics on Categorical Data**

Gender Frequency:

Count Percentage

SEX

Female 18112 60.37

Male 11888 39.63

Education Frequency:

Count Percentage

EDUCATION

University 14030 46.77

Grad School 10585 35.28

High School 4917 16.39

Others 468 1.56

Marital Status Frequency:

Count Percentage

MARRIAGE

Single 15964 53.21

Married 13659 45.53

Others 377 1.26

Modes:

Gender: Female

Education: University

Marital Status: Single

Gender vs. Education:

EDUCATION Grad School High School Others University

SEX

Female 6231 2927 298 8656

Male 4354 1990 170 5374

Marital Status vs. Education:

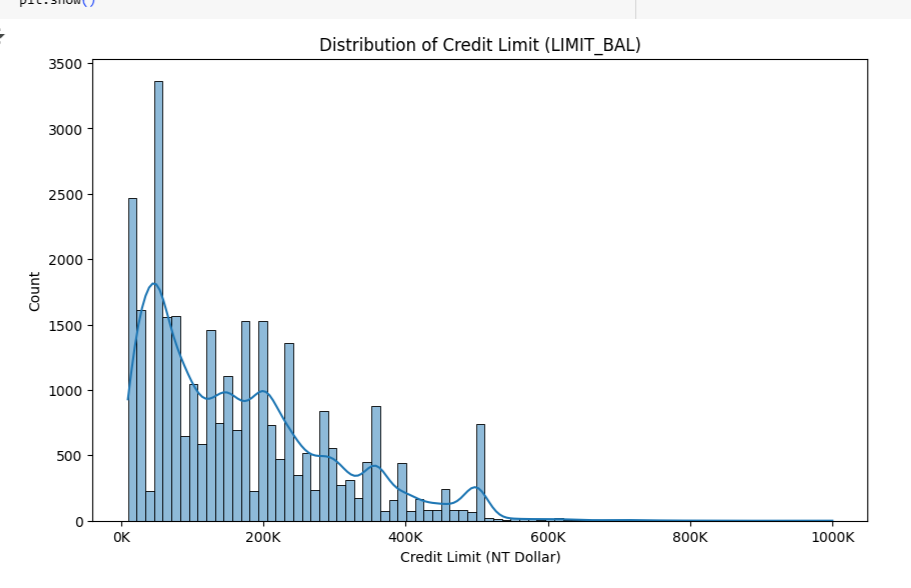
EDUCATION Grad School High School Others University

SEX

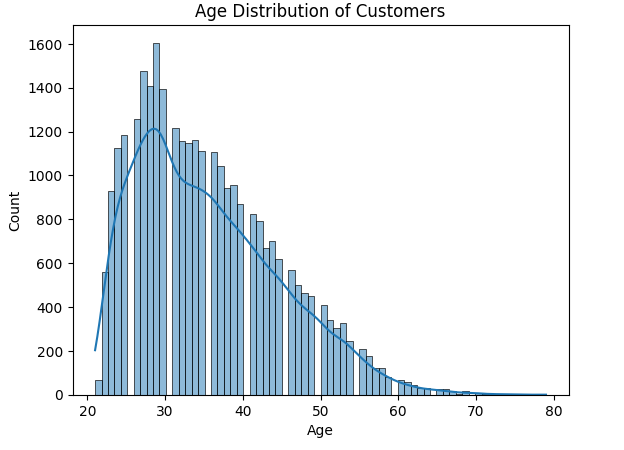
Married 4354 1990 170 5374

Single 6231 2927 298 8656

* **Probability distributions** (e.g., normal/exponential for payments)



* The credit limit (LIMIT\_BAL) distribution is **right-skewed**, indicating that **most** customers have **low to moderate credit limits**, while a small number of customers have very high limits. This is typical in credit card datasets due to risk management and customer segmentation practices.



* Why Is the age Column Right-Skewed in Credit Card Data?
* **More young to middle-aged users:** Most credit card users are between 20 and 50 years. This creates a concentration of values in the lower age range.
* **Fewer elderly users:** Fewer people above 60–70 actively apply for or use new credit cards. But some older users (e.g., 75, 80+) still exist and stretch the upper tail, causing right skew.
* **Regulatory and risk reasons:** Credit card issuers often target working-age individuals due to income stability, reducing the number of very old cardholders.

**Hypothesis tests**:

1. **Two-Sample t-test (LIMIT\_BAL vs Default Status):**

**Null (H0)**: Mean LIMIT\_BAL is the same for both defaulters and non-defaulters

**Alternative (H1)**: Mean LIMIT\_BAL differs between the two groups

* **T-statistic:** 28.95
* **P-value:** 0.0000
* **Conclusion:** Reject the null hypothesis — there is a **significant difference** in mean credit limits between defaulters and non-defaulters.

1. **Chi-square Test (Gender vs Default Status):**

**Null (H0)**: Proportions of defaulters are the same for males and females

**Alternative (H1)**: Proportions differ between males and females

* + - Chi-square Test Statistic: 47.70879689062111
    - P-value: 4.944678999412044e-12
* **Conclusion:** Reject the null hypothesis — gender and default are dependent.

1. **Pearson Correlation (Numeric vs Numeric):**

**Null Hypothesis (H0)**: No linear correlation between the two variables.

**Alternative Hypothesis (H1)**: A linear correlation exists (positive or negative).

* + - * **Correlation coefficient (r):** 0.1447
      * **P-value:** 0.0000
* **Conclusion:** Reject the null — a **significant but weak positive correlation** exists between the two numeric variables tested.

**6. Feature Engineering**

* **label encoding** to categorical features (e.g., Gender, Education)
* Scale features using **StandardScaler**

**7**. **Feature Selection**

* Selected all the Features from the dataset

**8. Model Building**

Use classification models:

* Logistic Regression
* Decision Tree Classifier

Train with:

* train\_test\_split
* Evaluate with:
  + Accuracy
  + Precision/Recall
  + AUC-ROC
  + Confusion Matrix

**9**. **Handling Imbalanced Data and Model Evaluation Strategies**

* To address the challenge of imbalanced data in the dataset, I implemented a variety of resampling techniques and robust classification strategies.
* Specifically, I used **SMOTE**, **SMOTEENN**, **SMOTETomek**, and **ADASYN** to oversample the minority class and generate more balanced training data.
* For modeling, I employed both **Logistic Regression** and **Decision Tree** classifiers, along with ensemble learning techniques such as **Bagging** and **Boosting** to improve predictive performance and model stability.
* Additionally, **GridSearchCV** was used to optimize hyperparameters for better generalization.
* The models were evaluated using key classification metrics, including the **classification report** (precision, recall, F1-score) and the **ROC AUC score**, ensuring a comprehensive assessment of model effectiveness, particularly on the minority class.

**10. Model Evaluation**

* Compare models using performance metrics
* Look at **feature importance** (which features drive default behavior?)
* Select the best model

**11. Conclusion**

* To effectively handle class imbalance and improve model sensitivity toward defaulters (minority class), I implemented a custom two-layer ensemble approach. The first layer involved training 12 Gradient Boosting models on different balanced subsets of the training data, each with varied hyperparameters to ensure diversity.
* Their outputs were then combined and fed into a Logistic Regression meta-classifier (second layer), which learned how to optimally aggregate the predictions.
* To fine-tune the classification decision boundary, I used a precision-recall curve to determine the best threshold based on F1-score, a metric that balances precision and recall.
* However, since the project's goal prioritizes catching more defaulters, I slightly lowered the threshold (from 0.58 to 0.40), increasing recall at the expense of some precision, effectively flagging more high-risk individuals.
* I further enhanced the model with a cost-sensitive rule-based override, ensuring defaulters with high-risk payment behavior (PAY\_1 ≥ 2) are not missed.
* This ensemble strategy effectively addresses the challenges of imbalanced data, maximizes recall on the minority class, and incorporates domain knowledge for better real-world impact.