Project Proposal: Payment Propensity Prediction Model

Authored by:

1. SivaRamaKrishna Yarra
2. Sekhar Maruthi Nalluri
3. Sai Keerthi Baaghi
4. Bibi Reddy Bhuvaneshwari

---

1. Background, Motivation, and Significance

Financial institutions must accurately predict customer payment behavior to maintain **cash flow**, **minimize default risks**, and **develop effective collection strategies**. A **Payment Propensity Prediction Model** assists businesses in determining whether a debtor will meet payment deadlines, allowing them to prioritize outreach and improve recovery rates efficiently.

This project leverages **machine learning techniques** to provide actionable insights for debt collection companies while **reducing operational costs** and **enhancing customer service**. Through predictive analytics, businesses can identify **high-risk customers** and implement **proactive collection strategies**, ensuring optimized financial planning.

We will use a **structured dataset** obtained from the **UCI Machine Learning Repository**, which contains historical payment patterns, demographic information, and financial data.

---

2. Research Questions

1. **What variables significantly influence a customer's ability to complete their payment?**
2. **How accurately can machine learning models predict customer payment behavior?**
3. **Which customer segments exhibit high or low payment propensity, allowing for tailored collection strategies?**

These research questions are crucial in identifying predictors of payment behavior and **enhancing debt collection techniques** through data-driven decision-making.

---

3. Dataset

The dataset used in this study is sourced from the **UCI Machine Learning Repository** and comprises **30,000 records** with **25 variables** covering customer demographics, payment history, and credit information.

Key Variables:

- **ID:** Unique customer identifier
- **LIMIT_BAL:** Credit limit (in NT dollars)
- **SEX:** Gender (1 = Male, 2 = Female)
- **EDUCATION:** Level of education (1 = Graduate, 2 = University, 3 = High School, 4 = Others)
- **MARRIAGE:** Marital status (1 = Married, 2 = Single, 3 = Others)
- **AGE:** Customer's age (years)
- **PAY_0 - PAY_6:** Repayment status for the past **six months** (-1 = Paid on time, 1 = Delayed 1 month, etc.)
- **BILL_AMT1 - BILL_AMT6:** Amount of bill statements for the past **six months**
- **PAY_AMT1 - PAY_AMT6:** Amount of previous payments for the past **six months**
- **Target Variable:** `default.payment.next.month` (1 = Default, 0 = No Default)

This dataset offers a **strong foundation** for training a **classification model** to predict customer payment behavior.

---

4. Methodology

Our objective is to develop a **supervised machine learning classification model** for predicting customer payment behavior. The project follows these key steps:

Step 1: Data Preprocessing

- Handle **missing values** and **outliers**.
- Normalize numerical data to maintain structured datasets.
- Encode categorical variables appropriately.

Step 2: Exploratory Data Analysis (EDA)

- Identify patterns, distributions, and correlations among key features.
- Conduct **feature importance analysis** to determine variables with the highest impact on payment behavior.

Step 3: Model Selection

- Implement and compare different machine learning models:
    - **Logistic Regression**
    - **Decision Trees**
    - **Random Forest**
    - **Gradient Boosting Algorithms** (XGBoost, LightGBM, CatBoost)

Step 4: Evaluation Metrics

Each model will be evaluated based on:

- **Accuracy**
- **Precision, Recall, and F1-score**
- **ROC-AUC Score** (Receiver Operating Characteristic - Area Under Curve)

Step 5: Hyperparameter Tuning

- Apply **Grid Search** and **Random Search** techniques to optimize model parameters for enhanced performance.

These methods are chosen as they are **highly effective** in **classification tasks** and are widely used in **financial risk modeling**.

---

5. Expected Outcomes

- A **highly accurate predictive model** for **customer payment behavior**.
- Identification of key factors influencing **default risk**.
- Development of **data-driven debt collection strategies**.
- A foundation for further **financial analytics research**.

---

6. Conclusion

By leveraging **machine learning and predictive analytics**, this project will help financial institutions optimize their **collections strategy**, improve **cash flow management**, and

minimize default risks. The insights derived will be valuable in creating **customer-centric repayment plans**, ensuring financial stability, and enhancing overall **operational efficiency**.

Project Timeline

Phase	Task	Duration
Phase 1	Data Collection & Preprocessing	2 Weeks
Phase 2	Exploratory Data Analysis (EDA)	2 Weeks
Phase 3	Model Development & Testing	4 Weeks
Phase 4	Model Evaluation & Optimization	3 Weeks
Phase 5	Report & Final Presentation	2 Weeks

---

7. References

- UCI Machine Learning Repository: **Default of Credit Card Clients Dataset**
- Research papers on **financial risk prediction** and **credit scoring models**

Prepared By:

📌 **SivaRamaKrishna Yarra, Sekhar Maruthi Nalluri, Sai Keerthi Baaghi, Bibi Reddy Bhuvaneshwari** 📅 **Date:** 13-02-2025
📍 **Institution/Organization:** Depaul University
