**Credit Risk Prediction** **&** **Customer Segmentation**

*Prepared by: Truong Nguyen Thanh Tu*

*Date: April 27, 2025*

**1. Executive Summary**

This report analyzes 33,000 loans totaling $312 million with the primary goal of assessing and predicting default risk.

* **Overall default rate:** 21.82% (7,108 loans)
* **Average interest rate:** 11.01%
* **Credit distribution:** 65% of borrowers are in Grades A–B, indicating a generally high-quality pool.
* **Recommended model:** Extra Trees Classifier, achieving **Precision 0.96**, **Recall 0.93**, **F1-score 0.95** in detecting high-risk borrowers.

**2. Introduction**

As the consumer lending market grows, accurate risk-assessment tools are vital to minimize financial losses. This report will:

1. Explore and describe borrower characteristics.
2. Identify which features most strongly influence default.
3. Segment borrowers by risk profile.
4. Build and evaluate a predictive model.

**3. Data Collection**

Data were sourced from internal loan records covering 33,000 borrowers, including these fields:

* **Age** (20–144 years; mean 27.73)
* **Annual Income** (mean ≈ $66 075; extreme outliers up to $6 000 000)
* **Home Ownership** {RENT, MORTGAGE, OWN, OTHER}
* **Employment Length** (years; mean 4–5)
* **Loan Purpose** {EDUCATION, MEDICAL, VENTURE, PERSONAL, HOME IMPROVEMENT, …}
* **Credit Grade** A–G
* **Loan Amount**
* **Interest Rate**
* **Loan-to-Income Ratio**
* **Has Prior Default** {Y, N}
* **Credit History Length** (years)
* **Loan Default** {0 = no default, 1 = default}
* **Cluster** (segments 0–3 via clustering)

**4. Methodology**

1. **Exploratory Data Analysis (EDA)**
   * Descriptive statistics, histograms, box-plots to reveal distributions and outliers.
2. **Risk Analysis**
   * Calculated default rates by feature (e.g., Credit Grade, Loan Purpose, LTI).
3. **Customer Segmentation**
   * Applied K-Means to form four borrower clusters (0–3).
4. **Model Building**
   * Compared Random Forest vs. Extra Trees; selected Extra Trees for faster training and better generalization.
5. **Evaluation**
   * Assessed on held-out test set using Accuracy, Precision, Recall, and F1-score; prioritized reducing false negatives.

**5. Data Analysis**

**5.1 Demographics & Financials**

* **Age:** Mostly 20–40; one erroneous outlier at 144.
* **Annual Income:** Right-skewed due to high outliers; majority around $55 000.
* **Employment Length:** Most between 2–10 years; a few exceed 14 years.

**5.2 Credit & Home Ownership**

| **Credit Grade** | **Proportion (%)** | **Default Rate (%)** |
| --- | --- | --- |
| A | 33.08 | 9.96 |
| B | 32.08 | 16.28 |
| C | 19.82 | 20.73 |
| D–G | 15.02 | 59–98 |

* 65% are in Grades A–B.
* Grades F and G (small share) have extremely high default rates (70–98%).
* **Home Ownership:** 50.48% rent, 41.26% mortgage, 7.93% own, 0.33% other.

**5.3 Loan Purpose & LTI**

* Top purposes: EDUCATION (19.81%), MEDICAL (18.63%), VENTURE (17.55%), PERSONAL (16.95%).
* **Loan-to-Income (LTI) ratio:** mean 0.17; default risk rises sharply when LTI > 0.3.

**5.4 Prior Default & Credit History**

* Borrowers with prior defaults default again 37.81% of the time.
* Short credit histories (< 1 year) see default rates above 50% on average.

**6. Findings and Insights**

1. **Highest-risk factors:** Low Credit Grade (F–G), high LTI (> 0.3), prior default, and purposes MEDICAL or DEBT CONSOLIDATION.
2. **Risky segments:**
   * **Cluster 3:** default rate 43.23%, predicted default probability 71%.
   * **Cluster 1:** default rate 28.65% despite higher incomes—driven by larger loans and higher rates.
3. **Low-risk segment:**
   * **Cluster 0:** default rate 5.53%, probability 38%, characterized by low LTI and higher incomes.
4. **Moderate-risk segment:**
   * **Cluster 2:** default rate 20.32%, lower income and relatively high LTI.

**7. Recommendations**

* **Approve first:** Applicants with Grades A–B, LTI < 0.2, credit history > 2 years.
* **Adjust pricing:** Charge higher rates for MEDICAL and DEBT CONSOLIDATION loans.
* **Monitor closely:** Clusters 3 and 1—implement stricter review and ongoing follow-up.
* **Feature reduction:** Remove or consolidate low-impact Loan Purpose categories to simplify the model.

**8. Conclusion**

We’ve identified the key drivers of default risk and segmented borrowers by their risk profiles. The Extra Trees Classifier delivers strong predictive performance and is well-suited for integration into the credit-decision process.