REPORT OF THE CASE STUDY

1. Business Context & Objective

The company is a consumer lender serving retail customers, many of whom have limited or thin credit histories. The key business risks are:

- Approving **high-risk customers** who later default
- Rejecting **good customers** and losing business

Using two datasets (current applicants and their previous loan history), the goal of this analysis was to:

- Understand customer and loan characteristics
- Identify drivers of default
- Quantify how previous loan behaviour (especially refusals) relates to current default
- Compare the company's default rate to an industry benchmark (10%)
- Provide recommendations to reduce credit risk

---

2. Data Used

Dataset 1 – Previous Loans (`credit_risk_previous_loans`)

- 1.67M records, 37 original columns (reduced to 26 after cleaning)
- Key variables:
    - `SK_ID_CURR` – customer ID
    - `NAME_CONTRACT_STATUS` – Approved / Refused / Canceled / Unused offer

- AMT_APPLICATION, AMT_CREDIT, AMT_ANNUITY, AMT_GOODS_PRICE
- DAYS_DECISION – how many days before current application the decision was made

Dataset 2 – Current Applicants (credit_risk_applicants)

- 307k records, 122 original columns (reduced to 73 after cleaning)
- Key variables:
    - TARGET – 1 = default / payment difficulty, 0 = repaid
    - Demographics: CODE_GENDER, DAYS_BIRTH, NAME_EDUCATION_TYPE
    - Financials: AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY
    - Employment: DAYS_EMPLOYED
    - Risk scores: EXT_SOURCE_2, EXT_SOURCE_3

The datasets were finally **merged on SK_ID_CURR** to analyse how previous loan behaviour links to current default.

CaseStudy1_1765134299 (1)

---

3. Data Preparation & Cleaning

3.1 Previous Loans Dataset

1. **Missing values**
    - Calculated missing % per column.
    - Dropped columns with **>40% missing**, including: AMT_DOWN_PAYMENT, RATE_DOWN_PAYMENT, RATE_INTEREST_PRIMARY, RATE_INTEREST_PRIVILEGED, NAME_TYPE_SUITE, multiple DAYS_* and

insurance flags.

2. **Imputation**
    - Numerical columns → **median**
    - Categorical columns → **mode**
    - After this, missing % for all remaining columns = 0%.
3. **Outliers**
    - Used boxplots for `AMT_CREDIT`, `AMT_ANNUITY`, `AMT_APPLICATION`, `AMT_GOODS_PRICE`.
    - Removed outliers using IQR rule (1.5×IQR beyond Q1/Q3).
4. **Feature engineering**
    - `DAYS_DECISION_ABS` = |`DAYS_DECISION`|; `YEARS_DECISION` = `DAYS_DECISION_ABS` / 365.25
    - `CREDIT_ANNUITY_RATIO` = `AMT_CREDIT` / `AMT_ANNUITY`
    - `APPLICATION_CREDIT_RATIO` = `AMT_APPLICATION` / `AMT_CREDIT`
    - Flags:
        - `IS_REFUSED` = 1 if `NAME_CONTRACT_STATUS` = “Refused”
        - `IS_APPROVED` = 1 if `NAME_CONTRACT_STATUS` = “Approved”

3.2 Applicants Dataset

1. **Missing values & drops**
    - Computed % missing for all 122 columns.
    - Dropped highly incomplete housing/real-estate columns (e.g. `APARTMENTS_AVG`, `COMMONAREA_*`, `YEARS_BUILD_*`, `TOTALAREA_MODE`, `EXT_SOURCE_1`, etc.)

where missing >40%, as they didn't add robust business insight.

2. **Imputation**
   - Numerical → **median**
   - Categorical → **mode**
   - Achieved 0% missing in retained columns.
3. **Fixing incorrect values**
   - `DAYS_BIRTH` (negative days) → `Age_years` = `−DAYS_BIRTH` / 365.25
   - Replaced `DAYS_EMPLOYED = 365243` (placeholder) with NaN, then created `YEARS_EMPLOYED` = `−DAYS_EMPLOYED` / 365.25
4. **Outlier treatment**
   - Used boxplots and IQR rule for `AMT_INCOME_TOTAL`, `AMT_CREDIT`, `AMT_ANNUITY` and removed extreme outliers.
5. **Feature engineering**
   - `CREDIT_INCOME_RATIO` = `AMT_CREDIT` / `AMT_INCOME_TOTAL`
   - Binary flags based on `NAME_CONTRACT_TYPE` (as implemented in the notebook):
     - `has_prev_refusal`
     - `has_prev_approval`

---

4. Exploratory Data Analysis (EDA)

4.1 Previous Loans

Univariate

- **Credit, Annuity, Application Amounts**
    - Histograms show right-skewed distributions: many small loans and a tail of large amounts.
    - After outlier removal, boxplots show more compact ranges and fewer extreme values.
- **Contract status & type**
    - Countplots show **“Approved”** as the largest group, followed by **“Canceled”** and **“Refused”**.
    - “Consumer loans” dominate, followed by “Cash loans”; revolving loans are fewer. CaseStudy1_1765134299 (1)

Bivariate

- **Credit vs Contract Status**
    - Boxplots indicate **refused applications tend to request higher credit** than approved ones.
- **Annuity vs Contract Status**
    - Refused and canceled applications often come with **higher annuity** obligations.
- **Contract Type vs Status**
    - Stacked countplots show approval and refusal rates **vary by loan product**; some product types are more likely to be refused.

Multivariate & Correlation

- Scatter of `AMT_CREDIT` vs `AMT_ANNUITY` coloured by contract type shows distinct repayment patterns for different loan categories.
- `APPLICATION_CREDIT_RATIO` vs `NAME_CONTRACT_STATUS` highlights mismatches between requested and granted credit.
- Correlation analysis:

- `IS_REFUSED` has its strongest positive correlations with:
    - `CREDIT_ANNUITY_RATIO` (~0.14)
    - `AMT_CREDIT` (~0.10)
    - `AMT_APPLICATION` (~0.08)
- Larger loans and higher credit-to-annuity ratios are associated with higher refusal likelihood.
  CaseStudy1_1765134299 (1)

4.2 Current Applicants

Univariate

- **Income (`AMT_INCOME_TOTAL`)**
    - Right-skewed: most customers have moderate income, with a small group at very high income levels.
    - Log-transformed distribution is more symmetric and suitable for statistical tests.
- **Credit Amount & Annuity**
    - Boxplots show a wide range of loan sizes and repayment burdens.
- **Education & Gender**
    - Most customers are “Secondary / secondary special” or “Higher education.”
    - Gender distribution is skewed towards females in this dataset.
      CaseStudy1_1765134299 (1)

Bivariate (with TARGET)

- **Income vs Default**

    - Boxplot of `AMT_INCOME_TOTAL` by `TARGET` suggests **defaulters tend to have somewhat lower incomes** than non-defaulters.
- **Credit Amount vs Default**
    - Boxplot indicates defaulters often have **slightly higher loan amounts** than non-defaulters.
- **Education vs Default**
    - Countplot shows default proportion differs across education levels – customers with lower education appear more likely to default.
- **Employment Length vs Default**
    - Histogram of `YEARS_EMPLOYED` by `TARGET` suggests **shorter employment history is associated with higher default risk**.
- **External Scores vs Default**
    - Bar charts of mean `EXT_SOURCE_2` and `EXT_SOURCE_3` by `TARGET` show:
        - **Non-defaulters have significantly higher external scores** than defaulters.

Multivariate & Correlation

- **Income × Credit × TARGET**
    - Log–log scatter shows high-credit + low-income combinations are more concentrated among defaulters.
- **Age × Employment × TARGET**
    - Younger, less stable (low `YEARS_EMPLOYED`) customers show higher default.
- **Correlation with TARGET**
    - Strongest negative correlations: `EXT_SOURCE_2`, `EXT_SOURCE_3` (higher score → lower default).

- Positive correlations: `DAYS_BIRTH`, `DAYS_EMPLOYED`, some region and phone-change variables.
- Top 10 correlated features with `TARGET` were plotted to highlight the most predictive variables.
 CaseStudy1_1765134299 (1)

---

5. Hypothesis Testing Results

All tests used the cleaned data prepared above.

5.1 Do defaulters have significantly lower income?

- **Test:** Two-sample t-test on log income (`LOG_INCOME`) between
  - Group 0: `TARGET = 0` (non-defaulters)
  - Group 1: `TARGET = 1` (defaulters)
- **Result:**
  - $t \approx 3.62$, $p \approx 0.0003$ (< 0.05)
- **Conclusion (business):**
  - There is a **statistically significant income difference** between defaulters and non-defaulters.
  - Non-defaulters have **slightly higher incomes**, supporting the view that **lower income customers are more likely to default**.
 CaseStudy1_1765134299 (1)

5.2 Is default rate different across genders?

- **Test:** Chi-square test of independence between `CODE_GENDER` and `TARGET`.

- **Result:**
  - p-value ≈ 1.65×10^{-212} (< 0.05)
- **Conclusion:**
  - **Gender and default are not independent.**
  - Default rates differ significantly between male and female customers, meaning gender is a relevant segmentation variable (though it should be used carefully for fairness/compliance reasons).
    CaseStudy1_1765134299 (1)

5.3 Are education level and default correlated?

- **Test:** ANOVA using encoded `NAME_EDUCATION_TYPE` (`EDU_CODE`) by `TARGET`.
- **Result:**
  - $F \approx 713.2$, $p \approx 6.28 \times 10^{-157}$ (< 0.05)
- **Conclusion:**
  - **Education level statistically affects default risk.**
  - Lower education levels are associated with higher default probability; higher education tends to be safer.
    CaseStudy1_1765134299 (1)

5.4 Do previous loan rejections predict higher current default probability?

(Using merged dataset `df_merged`.)

- **Feature created:**
  - `PREV_REJ` = 1 if any previous application had `NAME_CONTRACT_STATUS = "Refused"`.

  - `DEFAULT` = `TARGET` (current default flag).
- **Test:** Chi-square test of independence on contingency table of `PREV_REJ` vs `DEFAULT`.
- **Result:**
  - p-value ≈ 0.0 (< 0.05).
- **Conclusion (business):**
  - **Yes. Previous loan rejections are strongly associated with higher current default probability.**
  - Customers who were refused earlier are **significantly more likely to default** on their current loans.
  - Previous refusal history should be a key input in risk scoring and approval rules. CaseStudy1_1765134299 (1)

5.5 Is the company's default rate higher than the industry benchmark?

- **Benchmark:** Industry default rate assumed at **10%**.
- **Method:** One-sample proportion z-test using:
  - Observed company default rate = mean of `TARGET` in `df_merged`
  - Benchmark proportion = 0.10
- **Result:**
  - p-value ≈ 0.089 (> 0.05)
- **Conclusion (business):**
  - We **cannot conclude** that the company's default rate is significantly different from the 10% industry benchmark.
  - Statistically, the company appears to be **roughly in line with the industry**, not clearly worse or better at this significance level.

CaseStudy1_1765134299 (1)

5.6 Additional Hypothesis Tests on Previous Loans

- **Credit Amount vs Status (Approved vs Refused)**
  - Two-sample t-test (log credit): t ≈ 169, p ≈ 0.0
  - **Refused applications request significantly different (and generally higher) credit amounts** than approved ones.
- **Contract Type vs Status**
  - Chi-square on `NAME_CONTRACT_TYPE × NAME_CONTRACT_STATUS`: p ≈ 0.0
  - **Approval/refusal strongly depends on loan product type**.
- **Goods Category vs Status**
  - Chi-square on `NAME_GOODS_CATEGORY × NAME_CONTRACT_STATUS`: p ≈ 0.0
  - Certain purchase categories are **more likely to be refused**, indicating segment-wise risk.
- **Contract Type vs Credit Amount (ANOVA)**
  - F ≈ 32566.8, p ≈ 0.0
  - Different contract types carry **very different mean loan sizes**, which must be factored into risk and pricing.
    CaseStudy1_1765134299 (1)

---

6. Key Drivers of Default (from your analysis)

Based on EDA and hypothesis tests:

1. **Income level**

- Lower income customers are more likely to default.

2. **Loan size and burden**

    - Higher `AMT_CREDIT`, `AMT_ANNUITY`, and high `CREDIT_INCOME_RATIO` are linked to worse outcomes.

3. **Previous loan history**

    - Prior **refusals** are a strong indicator of future default risk.

4. **External risk scores** (`EXT_SOURCE_2`, `EXT_SOURCE_3`)

    - Strong negative correlation with `TARGET`; low scores signal high risk.

5. **Education level**

    - Lower education categories carry higher default risk.

6. **Employment stability**

    - Shorter `YEARS_EMPLOYED` and younger customers with unstable careers show higher risk.

7. **Product and goods category**

    - Certain loan products and goods categories have systematically higher refusal and risk patterns.

---

7. Business Recommendations

All recommendations below follow directly from the analyses you implemented:

1. **Use previous refusal history as a key risk input**

    - Add **“any previous refusal”** as a strong risk flag; tighten approval criteria or reduce limits for such customers.

2. **Tighten policies for low-income, high-loan customers**

   - Cap CREDIT_INCOME_RATIO (e.g., maximum loan relative to income).
   - For customers with low income + high requested credit, require stronger documentation or collateral.

3. **Segment by education and employment stability**
   - Higher risk for low education / short employment → consider:
     - Lower initial credit limits
     - Step-up limits after good repayment behaviour

4. **Leverage external scores more aggressively**
   - Use EXT_SOURCE_2 and EXT_SOURCE_3 as primary inputs into risk scorecards.
   - For very low external scores, either decline or price with higher interest and additional checks.

5. **Product-wise and category-wise policy refinement**
   - Review loan products and goods categories with high refusal/default rates.
   - Consider stricter criteria or higher pricing for those specific segments.

6. **Benchmark monitoring**
   - Continue to monitor default rate vs the 10% industry benchmark.
   - Current rate is not significantly worse; with the above actions, the company should aim to be **better than** the benchmark in future periods.

---

8. Limitations & Next Steps

- Analysis is based on historical data only; no predictive model (like logistic regression) was built in this notebook.

- Some potentially useful real-estate variables were dropped due to very high missingness.
- Future work can include:
    - Building a formal **credit scoring model** using the key drivers identified
    - Stability tests over time (cohorts)
    - ROI / profitability analysis by segment, not just default rate

---