Loan Default Risk Analysis

Exploratory Data Analysis — Banking & Financial Services

307,511	**122**	**8.07%**	**67 cols**
Loan Applications	Features Analyzed	Default Rate	With Missing Data

Academic Presentation | EDA Case Study | Danush Sinnan, Helen Tu, Xiyue Dai

Agenda

Overview of Analysis Sections

01 **Business Background & Goals**
Risk types, decision scenarios

02 **Data Overview & Quality**
Missing values, data structure

03 **Class Imbalance**
TARGET variable distribution

04 **Categorical Variable Analysis**
Gender, education, income type, etc.

05 **Numerical Variable Analysis**
Income, credit amount, age, etc.

06 **Previous Application Behavior**
Refusal history vs. default rate

07 **Correlation & Key Drivers**
EXT_SOURCE scores, age effects

08 **Conclusions & Recommendations**
Risk management strategies

Business Background & Objectives

Why EDA matters in loan risk assessment

Two Core Business Risks

False Rejection Risk

Rejecting a creditworthy applicant
-> Company loses business and interest revenue

False Approval Risk

Approving a client who later defaults
-> Company suffers direct financial loss

Analysis Objectives

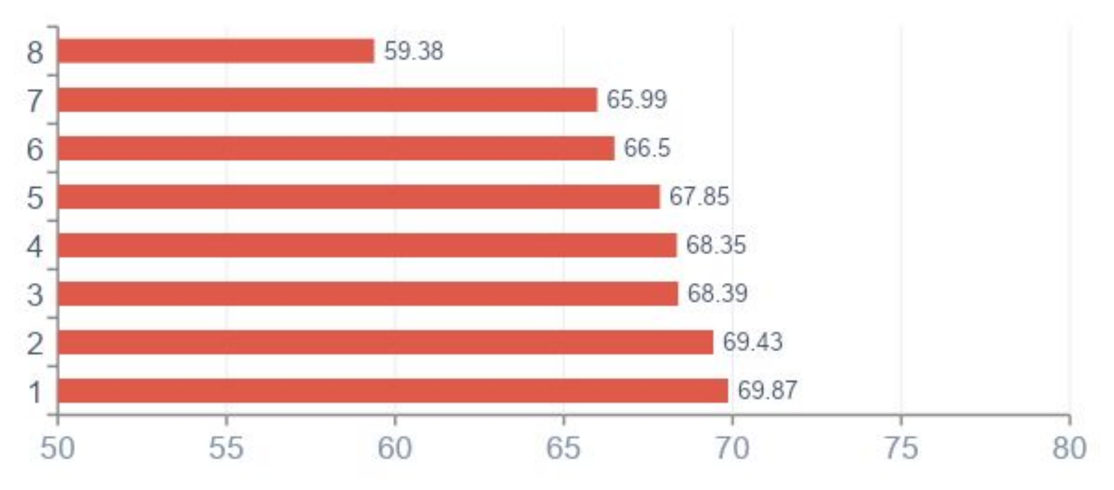
1. Identify client attributes strongly linked to default
2. Understand how loan features influence default risk
3. Provide data-driven basis for risk-based pricing
4. Select key features for downstream predictive modeling

Data Overview & Quality

Missing value analysis and handling strategy

307,511	122	67 cols	69.87%
Total Applications	Raw Features	Have Missing Data	Highest Missing Rate

Top 8 Columns by Missing Data Rate



Handling Strategy

> 60% missing

Drop the column entirely

30-60% missing

Impute with median / mode

OWN_CAR_AGE

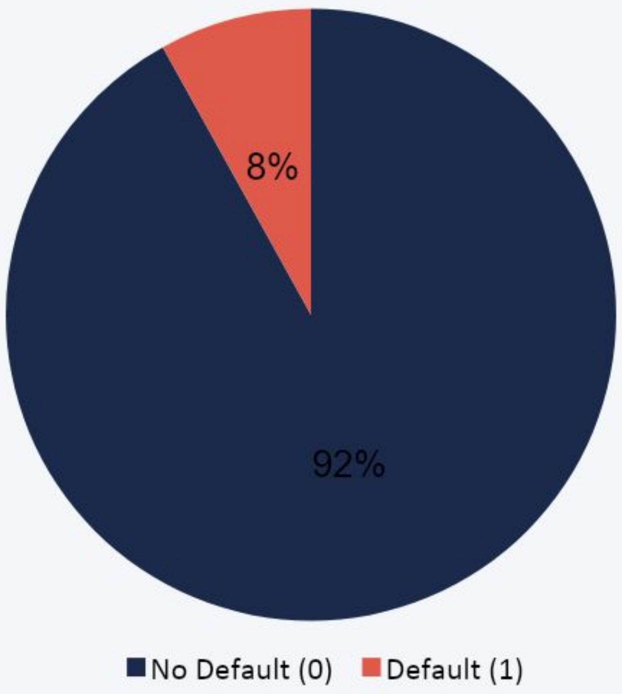
Fill 0 for clients without a car

DAYS_EMPLOYED

Replace 365,243 outlier with NaN

Class Imbalance in Target Variable

Distribution of TARGET: 0 = repaid on time, 1 = defaulted



8.07%

Default Rate

Severe Imbalance

Non-defaulters outnumber defaulters 11.4x

Modeling Impact

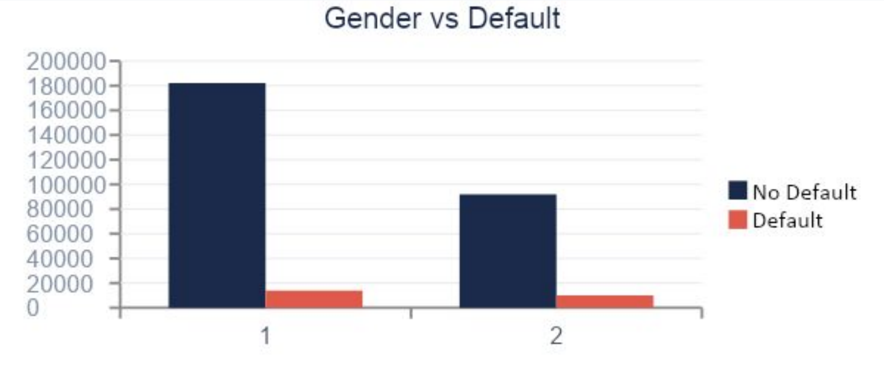
A naive model predicting all 0s gets 92% accuracy but is useless

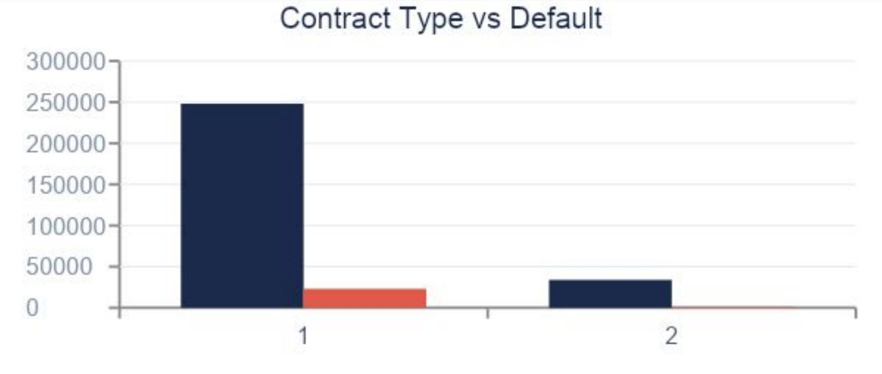
Solution

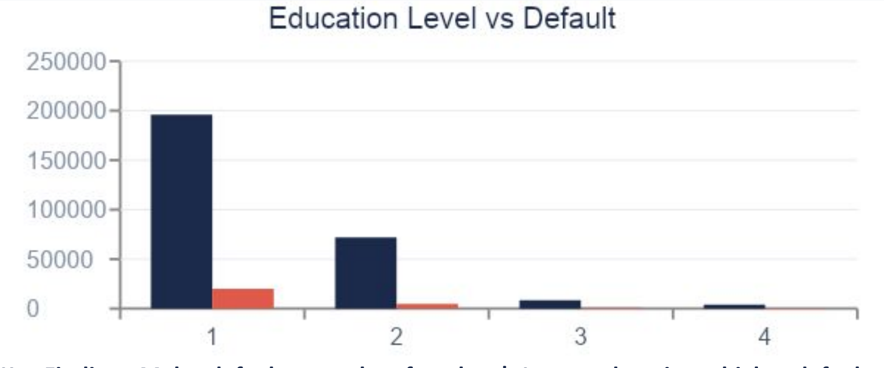
Apply SMOTE oversampling or adjust class_weight during modeling

Categorical Variable Analysis

Comparing default rates across key demographic and loan categories



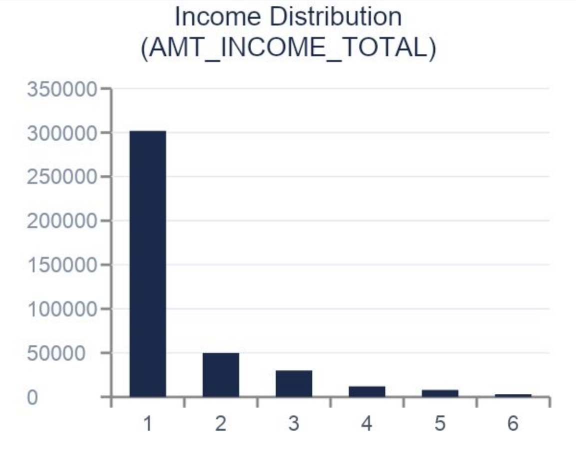


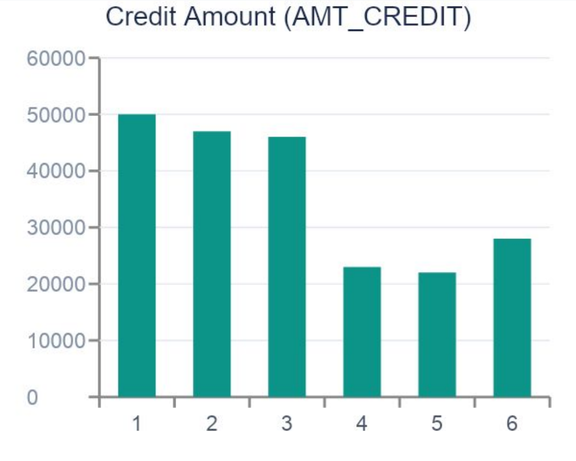


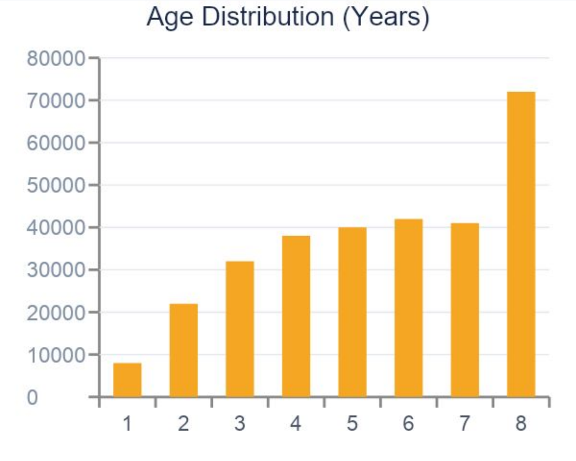
Key Finding: Males default more than females | Lower education = higher default rate | Unemployed clients show the highest default proportion

Numerical Variable Distributions

Income, credit amount, and age distributions





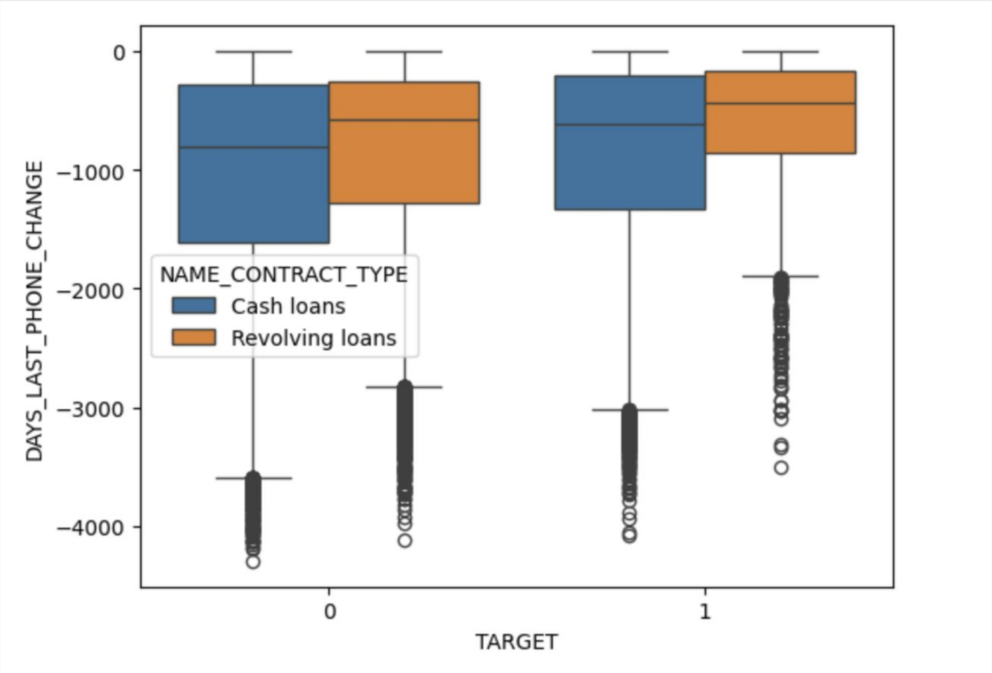


Data Quality Warning — DAYS_EMPLOYED Anomaly

DAYS_EMPLOYED contains a large spike of values equal to 365,243 — equivalent to 1,000 years. This is a system placeholder used to flag non-employed clients.
Fix: Replace all 365,243 values with NaN, then handle via imputation or create a binary is_employed feature.

Variable Analysis

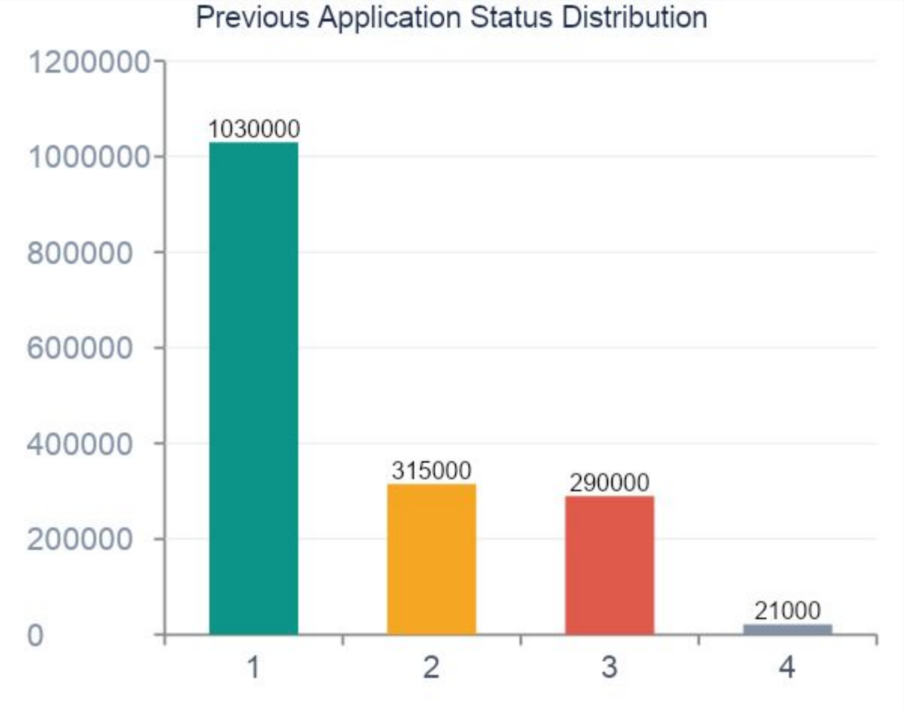
Comparing default rates v.s. recent phone change by contract types.

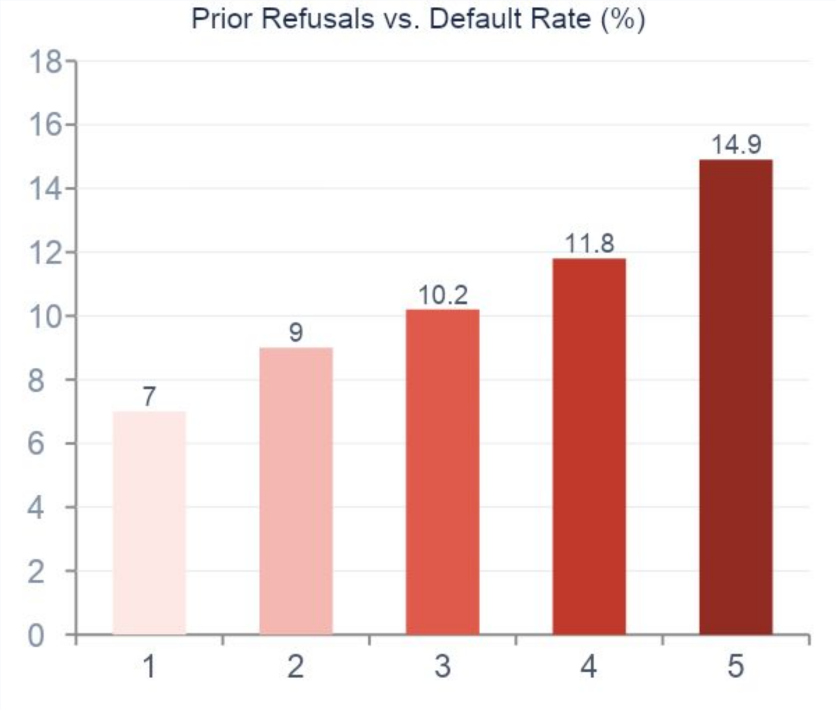


Key Finding: Recent phone changes is associated with higher default risk|Relationship is robust across contract types

Previous Application Behavior

How prior loan history predicts current default risk

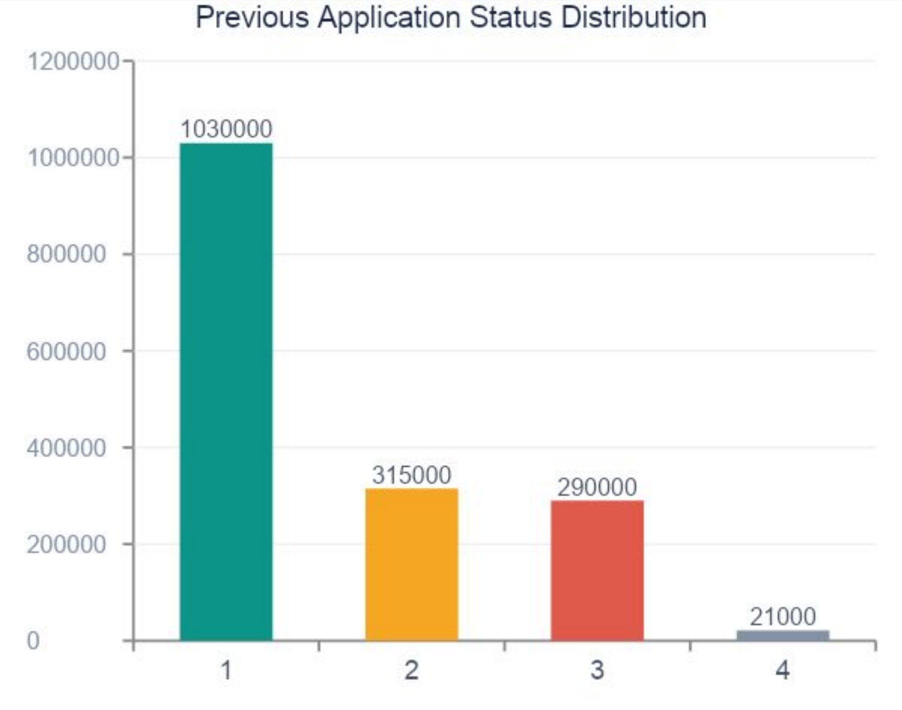


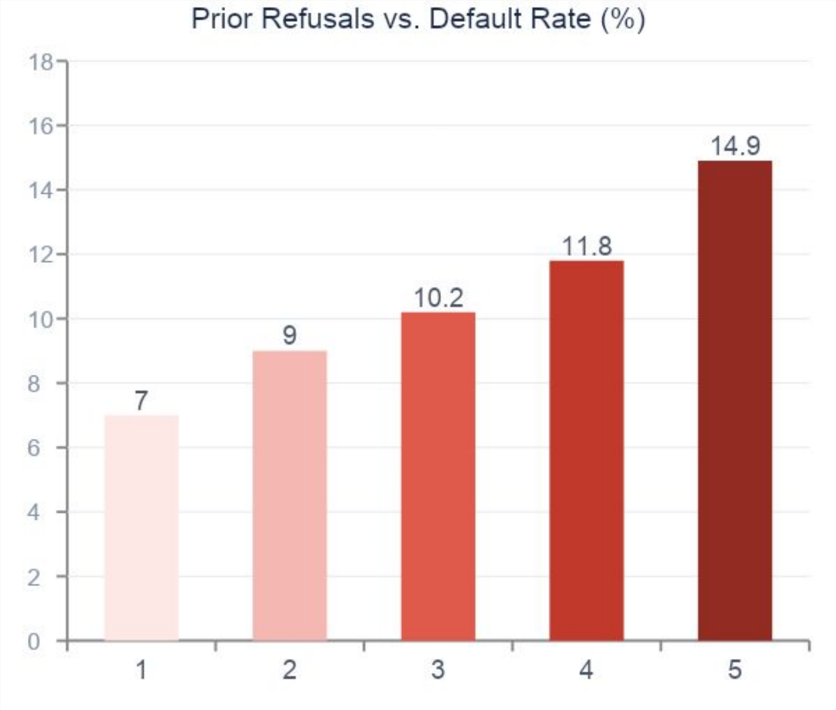


Key Finding: **Clients with 5+ prior refusals default at 14.9% — more than double the 7% rate of clients with no refusal history**

Previous Application Behavior

How prior loan history predicts current default risk

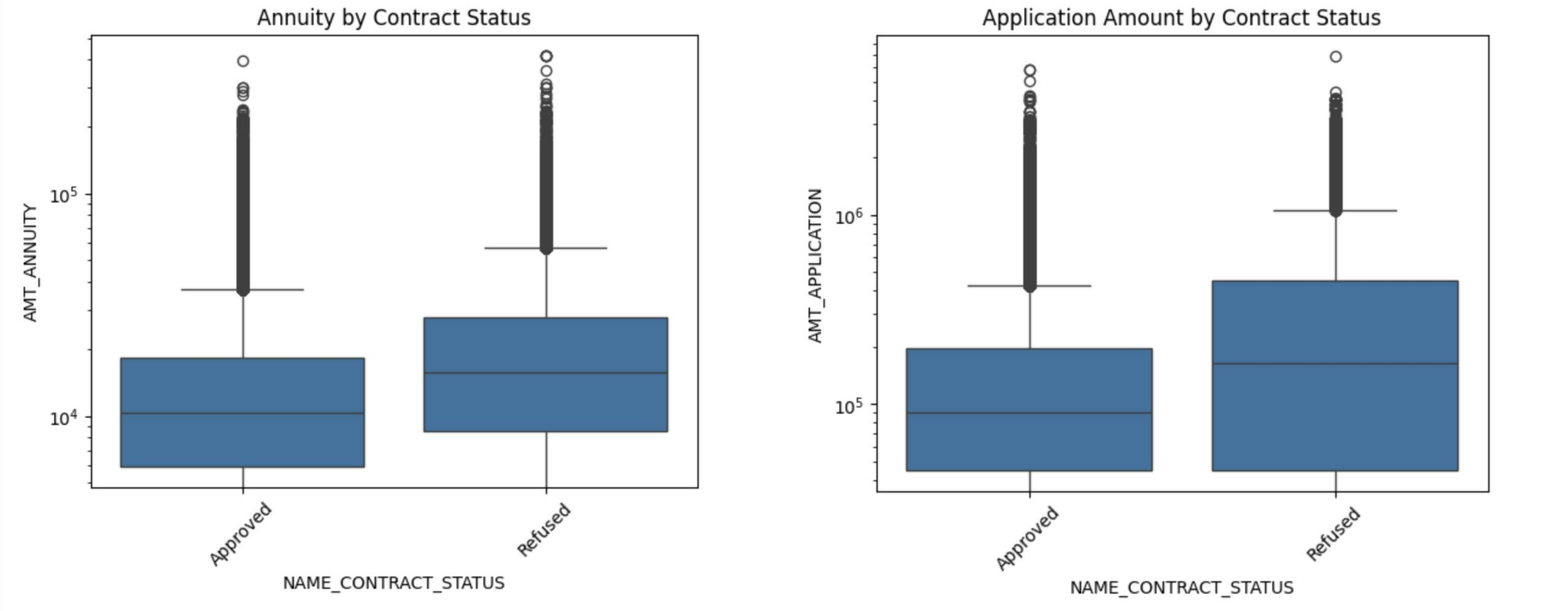




Key Finding: Clients with 5+ prior refusals default at 14.9% — more than double the 7% rate of clients with no refusal history

Previous Application Behavior

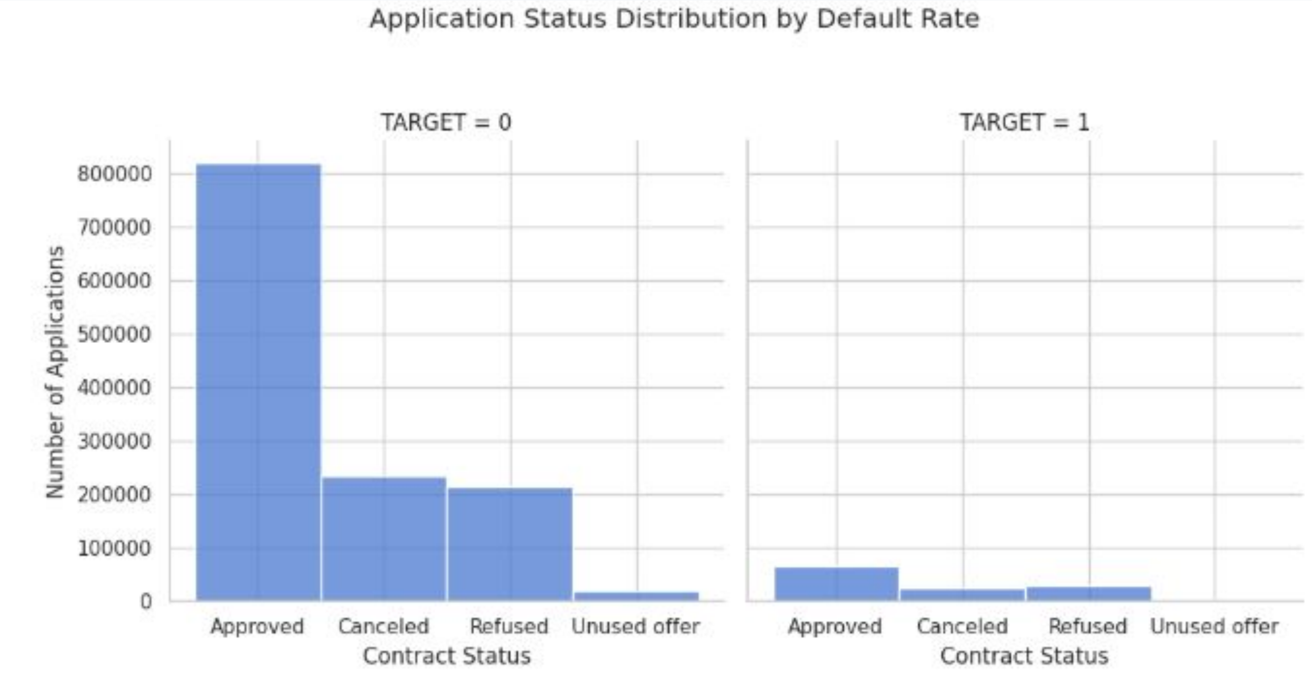
How specific features correlates with default risk(focusing on 'Approved' and 'Refused' in 'NAME_CONTRACT_STATUS)



Key Finding: Refused applications have a higher median annuity and a higher requested credit amount than approved ones

Merging Current and Previous Application Data

How feasible is it to merge the current and previous application datasets to find new insights in our analysis



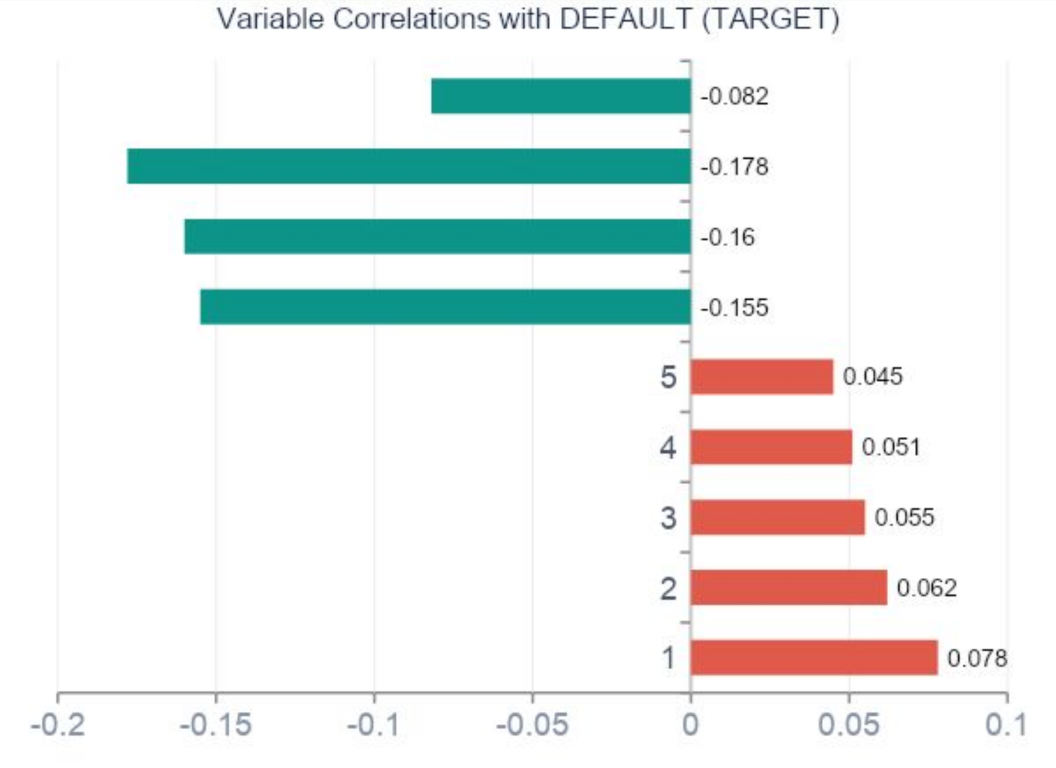
Merging Data Warning — SK_ID_CURR Anomaly

The current application dataset only has 307511 entries while the merged dataset has over 1000000 entries. This means that merging on SK_ID_CURR did not give us a dataset that matches one-to-one. Any IDs that were not in the current application dataset but were in the previous application dataset seem to be automatically given TARGET = 0 even if that loan applicant might have defaulted on their payments.

Key Finding: A merged dataset of current and previous application data on IDs of loans in our samples is unreliable for analysis

Correlation & Key Driver Variables

Which features are most predictive of default?



Strongest Protective Factor

EXT_SOURCE_3 -0.178

External credit score 3. Higher score = much lower default probability. Top single predictor.

Second Strongest Factor

EXT_SOURCE_2 -0.160

External credit score 2. Works best combined with EXT_SOURCE_3.

Age Effect

AGE_YEARS -0.082

Older clients default less. Clients under 30 need closer scrutiny.

Risk Indicator

DAYS_BIRTH +0.078

Stored as negative days; larger absolute value = younger client. Inverse of AGE_YEARS.

Conclusions & Recommendations

Key EDA findings and business action items

Severe Class Imbalance

Only 8% default — must address before modeling

Education Matters

Secondary-educated clients default more than graduates

Gender Risk Gap

Male clients carry a higher default rate than females

Age is Protective

Older applicants are more reliable; young clients need review

History Predicts Risk

More prior refusals = significantly higher default risk

Credit Scores are Key

EXT_SOURCE 1/2/3 are the strongest predictors overall

Business Recommendations

- Apply higher rates or lower credit limits for high-risk groups (young, low-education, multiple refusals)
- Prioritize EXT_SOURCE 1/2/3 as core features in any predictive model built on this data
- Use SMOTE or class_weight balancing before model training to avoid biased predictions

Thank You!

Questions & Discussion Welcome

Q & A

Data Source: Home Credit Default Risk Dataset
EDA Analysis by Jacob Frias Koehler