

Lending Club

Case study

thucanhle149@gmail.com

Problem Statement

The aim is to identify patterns which indicate if a person is likely to default

Project Outlines

1. Exploratory Data Analysis (EDA)
2. Data Preprocessing
3. Suggestions and Recommendations

Part I

Data Cleaning

1. Examining the dataset

Shape

The dataset has 39,717 rows and
111 columns

Data types

There are 3 data types: Float,
Integer, and Object

Primary columns

- Column 'id' : unique id for each loan
- Column 'member_id': unqiue id for each person

2. Data Cleaning

2.1 Missing values (part 1)

- Dropping all null columns

2.2 Data Filtering (part 1)

- Deduplicate the dataset
- Dropping the unimportant columns

2.4 Invalid values

- Removing symbols from numeric columns
- Trimming leading spaces from categorical columns
- Check for column with wrong data types
- Check for columns with only one unique value

2. Data Cleaning

2.3 Missing values (part 2)

- Filling in missing values in columns

2.5 Data Filtering (part 2)

- Dropping current loan

2.6 Standardise values

- Removing outliers

Part II

Exploratory Data Analysis (EDA)

1. Univariate & Segmented Univariate Analysis

The background features a dark purple gradient with three large, semi-transparent overlapping circles. One circle is light blue at the top and magenta at the bottom. Another is magenta at the top and light blue at the bottom. A third, smaller circle is located in the bottom right corner, partially cut off by the frame.

Here are some key insights

01

14,5% defaulted in
the lending dataset

Details

02

*The percentage of
default loan is higher
in shorter term (36
months)*

03

The defaulted loans
tend to have *higher*
interest rates than
non-defaulted loans

04

Mortgage and Rent is
the home ownership
type that has the
highest number of
defaulted loan

05

Borrowers that have
*no information of
employment* has the
*highest rate of
defaulted loans*

06

The defaulted loans
has *less annual*
income than non-
defaulted loan.

2. Bivariate Analysis

- Comparing loan amount, funded amount and funded amount by investors
- Create cross table for pairs of columns



Here are some key insights

01

Loan amount \geq Funded
amount \geq Funded amount
by investors
-> *aligned to business sense*

02

For 36 months term,
borrowers have *Rent* is the
most popular while for 60
months term, *Mortgage* is
the most popular

03

Borrowers have *Mortgage* has better grade than *Rent*;

Borrowers have *Rent* type of home has a lot of loan in lower grade like B,C,D,E.

3. Derived metrics

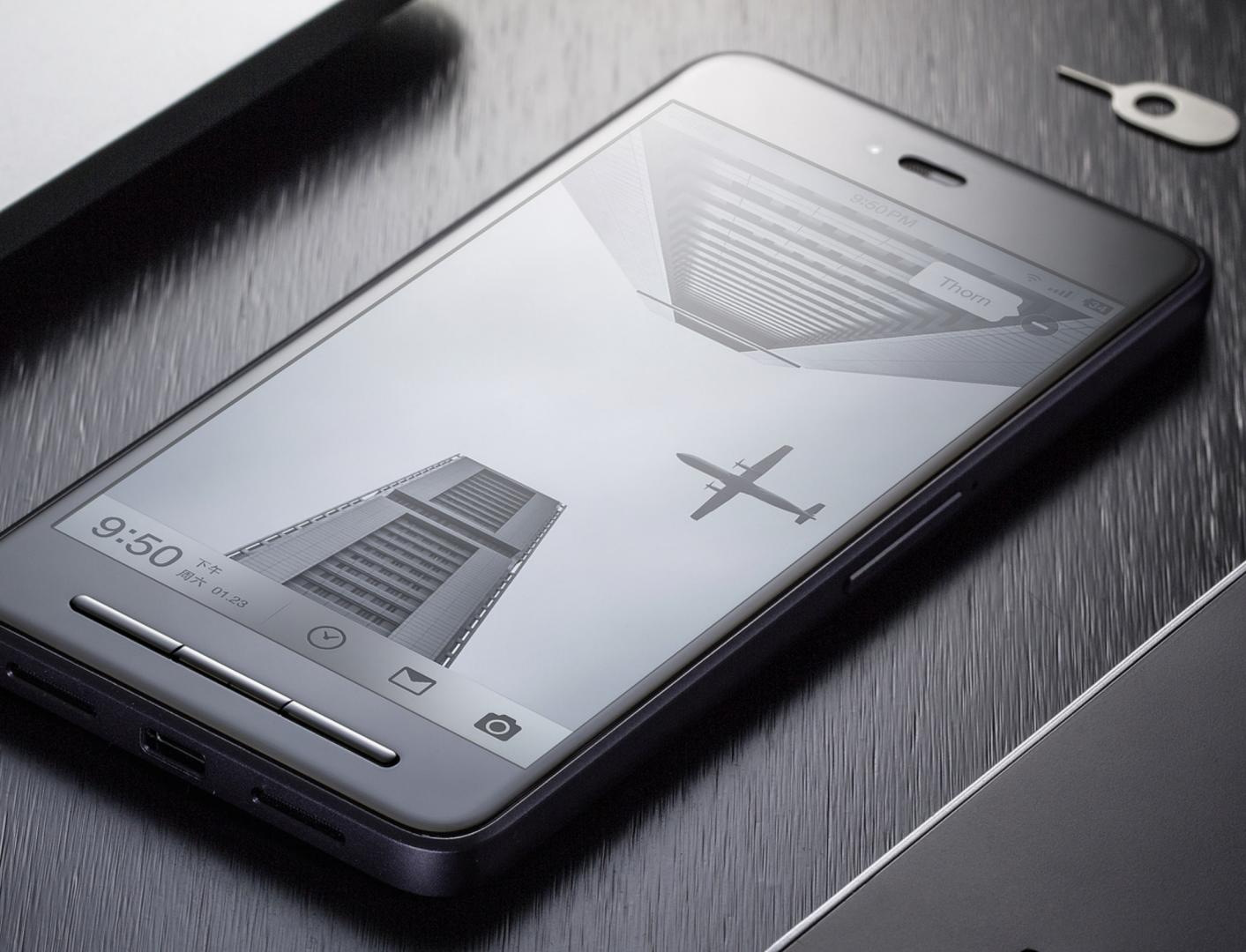
- Business driven: actual tenor
- Type driven: binning annual income
- Data driven: boolean column for loan status

Part III

Suggestions & Recommendations



Data Quality



01

Adding business-related column:

- year of the loan
- end date of the loan
- type of interest rate
- demographic of borrower

02

Missing values handling for better data quality:

- all null columns
- one unique value column

Further Data Preprocessing (optional)

Methods preparing data for ML models

- Standard Scaler (for numeric columns)
- One-hot encoding (for categorical columns)
- Dimensionality Reduction (PCA)
- ...

The end