## Loan Default Rate Prediction Model Based on XGBoost Algorithm

**1.Project Introduction-FX**

Background: Importance of loan default rate prediction

Objective: Establishing a loan default rate prediction model based on the XGBoost algorithm

**2. Data Collection and Preprocessing-Kaimi**

Data Source: Description of dataset and its features

Data Preprocessing: Data cleaning, handling missing values, feature selection, etc.

1. **Model Establishment-Cui**

Introduction to XGBoost Algorithm: Advantages and features of XGBoost

Dataset Splitting: Methodology for dividing dataset into training, validation, and testing sets

Model Training: Building and parameter tuning of XGBoost model

Model Evaluation

**4. Results Analysis and Conclusion-Max**

Model Results Presentation: Displaying model performance on the testing set

Results Analysis: Analyzing model predictions, discussing strengths, weaknesses, and areas for improvement

**2. Data Collection and Preprocessing**

**Slide 1**

Data Source:

* Lending Club Loan Dataset (<https://www.scaler.com/topics/data-science/loan-default-prediction/>)
* 20000 records, 15 columns
* Target: bad\_loan
* Features:

A screenshot of a computer program

Description automatically generated

## Dataset Feature Descriptions

The description for the features present in this dataset is -

* **id** - Unique ID of the loan application.
* **grade -** Loan grade.
* **annual\_inc -** The annual income provided by the borrower during registration.
* **short\_emp - 1** when the borrower is employed for 1 year or less.
* **emp\_length\_num -** Employment length in years. It ranges between 0 and 10, where 0 means less than one year and 10 means ten or more years.
* **home\_ownership -** Type of home ownership.
* **dti -** It is the Debt-To-Income Ratio that is calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s monthly income.
* **purpose -** A category for the loan request.
* **term -** The number of payments on the loan.
* **last\_delinq\_none -** 1 when the borrower had at least one event of delinquency.
* **last\_major\_derog\_none -** 1 when the borrower had at least 90 days of a bad rating.
* **revol\_util -** It is the revolving line utilization rate or the amount of credit the borrower uses relative to all available revolving credit.
* **total\_rec\_late\_fee -** Late fees received to date.
* **od\_ratio -** Overdraft ratio.
* **bad\_loan** - 1 when a loan was not paid.

Refrain from imputing in order to prevent data leakage. So, we prefer dropping the instances with missing values.

However, is missing values in home\_ownership related to the target, bad\_loan?

#if yes, then we should not just drop all the rows with missing home\_ownership values

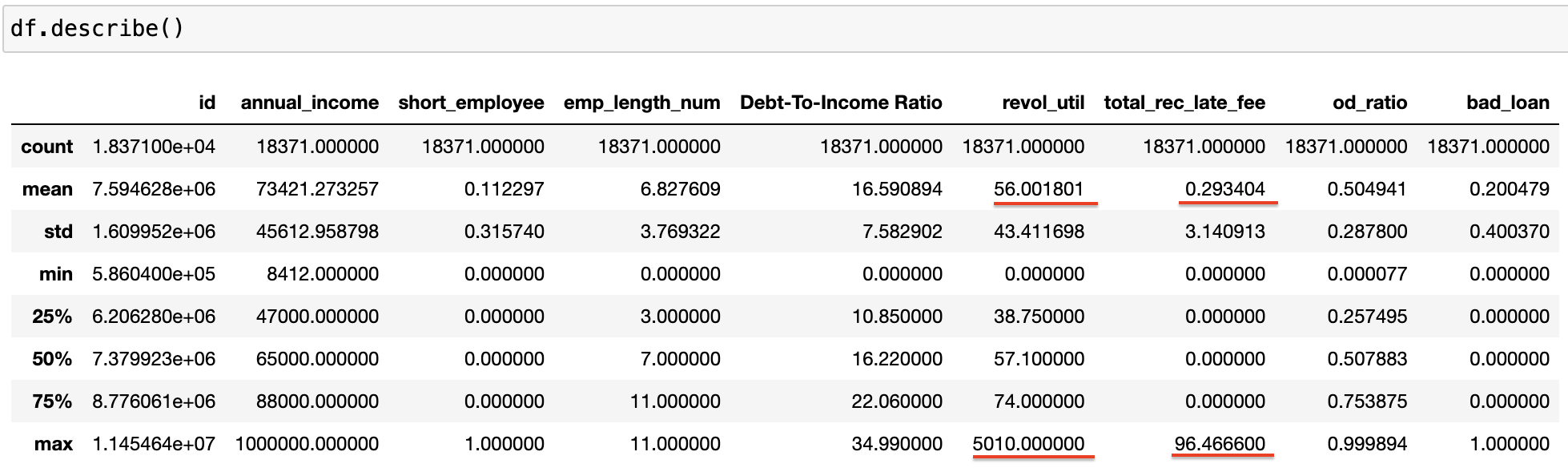
#if no, then we can drop those rows

#use chi-square test to test

Next, we do the same thing for Debt-To-Income Ratio

**Slide 2**

Outlier:



A graph with blue dots

Description automatically generated

A graph of a graph with blue dots

Description automatically generated with medium confidence

**Slide 3**

Distribution of Target and Categorical Feature Variables

A graph of a loan

Description automatically generated with medium confidence

A graph of a bar graph

Description automatically generated A graph of a home ownership distribution

Description automatically generated

A graph of a number of different colored bars

Description automatically generated with medium confidence

A graph of a long bar

Description automatically generated with medium confidence

A graph with blue and orange bars

Description automatically generated

**Slide 4**

Distribution Continuous Feature Variables

A graph of a comparison of a number of income

Description automatically generated with medium confidence

A graph of a number of different colored squares

Description automatically generated A blue and orange boxes with black lines

Description automatically generated A diagram of a credit utility rate

Description automatically generated A graph of a line graph

Description automatically generated with medium confidence A diagram of a comparison between two different colored squares

Description automatically generated