DATA SCIENCE PROJECT DOCUMENTATION

Table Of Contents

1. Introduction

Name Of Project: LendingClub Loan Defaulters Prediction

Group Members: Ahmed Rashid, Eric Ngugi, Joy Angel

Date: 8th August, 2025

2. Executive Summary

This project aims to predict loan applicants that are likely to default in order to improve decision-making when it comes to selecting loan applicants for loan disbursement. By analysis of past historical loan data, key observations and insights are drawn so as to create a predictive model. Key steps included in this project are data cleaning, Exploratory Data Analysis(EDA), data preprocessing, model selection, model evaluation and deployment. Upon completion of the model, an interactive app is made available. The user can input the details of a new loan applicant and the model can predict whether the applicant is likely to default immediately after input.

3. Company Background

Company name: LendingClub

Introduction

LendingClub is a US-based financial technology company. The headquarters of the company are located in San Francisco, California. They offer financial services, specializing in digital banking and lending. It was originally launched as a peer-to-peer lending platform, connecting borrowers directly with investors. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC) and to offer loan trading as a secondary market.

Business Understanding

LendingClub company specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based in the applicant's profile.

4. Problem Statement

LendingClub needs a better way to decide which loan applications that they should

approve. If they approve an applicant who fails to repay the loan(default), the company gets a financial loss. If they reject an applicant who would have paid the loan, they miss out on profit. By studying past loan data, the goal is to find patterns that show an applicant is likely to default. This can help the company make better decisions such as rejecting risky applicants, changing loan terms or charging higher interests in order to reduce losses.

5. Objectives

- To identify risky loan applicants in order to make better decisions.
- To build a predictive model for loan default.
- To evaluate the performance of the model.
- To design a deployment plan for integrating the predictive model into the company.

6. Data Description

Source: Kaggle LendingClub loan data (2007-2018)

Number of Records: 396,030 rows (each representing a loan application)

Number of features: 27 columns

Feature Types

Numerical : loan_amnt, int_rate, installment, annual_inc, dti, open_acc, pub_rec, revol_bal, revol_util, total_acc, mort_acc, pub_rec_bankruptcies

Categorical : term, grade, sub_grade, emp_length, home_ownership, verification_status, issue_d, loan_status, purpose, earliest_cr_line, intial_list_status, application_type, emp_title, address

Target variable: loan_status

Key details on each feature

Feature	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
term	The number of payments on the loan. Values are in months and can either be 36 or 30
int_rate	Interest rate on the loan

installment	The monthly payment owed by the borrower if the loan originates
grade	LC assigned loan grade
sub_grade	LC assigned loan sub grade
emp_title	The job title supplied by the borrower when applying for the loan
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are RENT, MORTGAGE, OWN, OTHER
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income is verified by LC, not verified or if the income source was verified
issue_d	The month which the loan was funded
loan_status	Current status of the loan
purpose	A category provided by the borrower for the loan request
title	The loan title provided by the borrower
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application
addr_state	The state provided by the borrower in the loan application
dti	A ratio 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 self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
open_acc	The number of open credit lines in the borrower's credit file

pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
initial_list_status	The initial listing status of the loan. Possible values are W and F
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
mort_acc	Number of mortgage accounts
pub_rec_bankruptcies	Number of public record bankruptcies

Missing Data:

- *emp_title*: ~6% missing
- *emp_length*: ~5% missing
- *title*: ~0.5% missing
- *revol_util:* ~1% missing
- *mort_acc*: ~10% missing
- *pub_rec_bankruptcies*: ~0.2% missing

Data Transformations

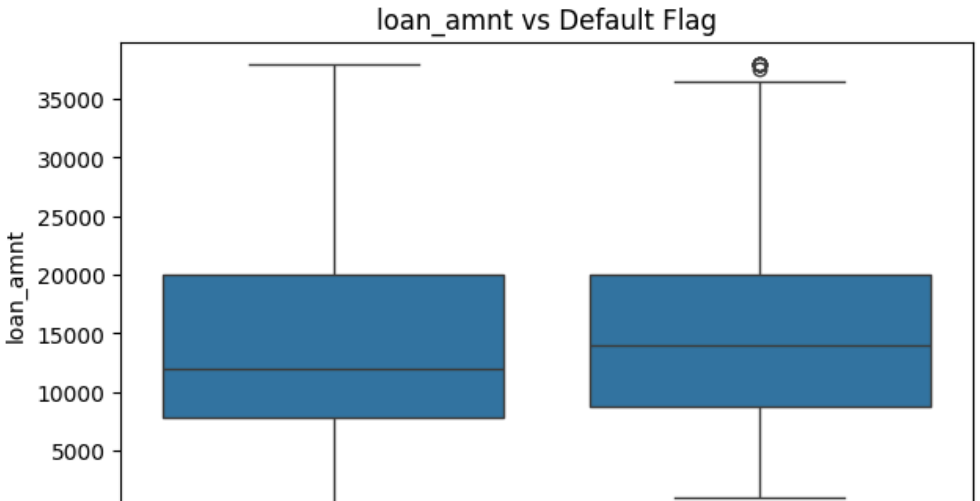
- Categorical variables are encoded using label encoding
- Numerical variables are normalized for modeling

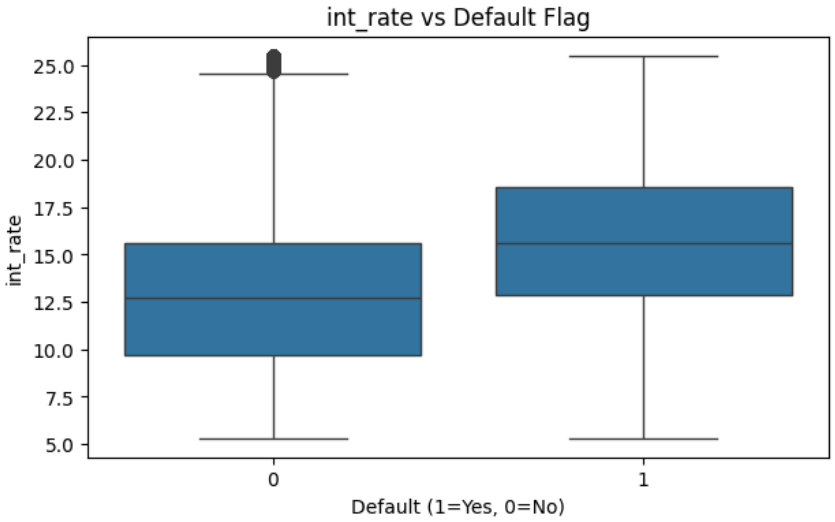
Initial Observations

It is observed that 19% of the loans in the dataset default.

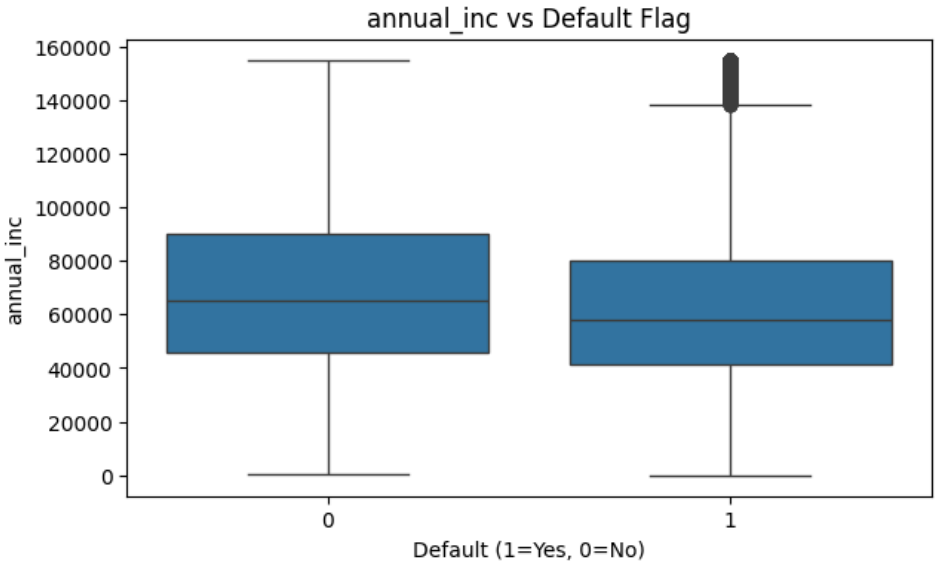
7. Exploratory Data Analysis

Here we can see that the greater the loan amount the greater the probability of defaulting.





Here we can see that the greater the interest rate the greater the probability of defaulting occurring.



Here we can see that the less the annual income, the greater the probability of defaulting.

8. Data Preprocessing

Encoding: loan_amnt was encoded using binary classification and renamed as default_flag. The other categorical features were encoded using get_dummies(pandas)

Scaling: loan_amnt, installment, annual_inc, dti, int_rate were the features scaled using StandardScaler and were used for deployment. The original model was trained using all the features using StandardScaler.

Splitting: Train-test split: 70/30

9. Modeling & Performance Metrics

Models used:

- Logistic Regression
- XG Boost
- Random Forest

Performance Metrics Table

MODEL	CLASS	PRECISION	RECALL	F1 SCORE	ACCURACY
Logistic Regression	0	0.81	0.99	0.89	80.83%
	1	0.54	0.06	0.11	
XGBoost	0	0.82	0.98	0.89	80.80%
	1	0.52	0.10	0.16	
Random Forest	0	0.81	0.99	0.89	80.82%
	1	0.54	0.06	0.11	

10. Deployment

Tools used:

- Python libraries - streamlit, matplotlib, pandas, scikit-learn

11. Recommendations

- The company should ensure that the model is retrained and maintained so as to improve performance over time.
- Methods of data collection should be improved so as to ease maintenance of the model and data completeness
- The company ought to benefit from the business model in order to promote future growth.