# Lending Club Loan Data Analysis
# Project Write-Up

## Introduction

This project focuses on predicting whether a loan will be fully paid or will default using historical data from Lending Club between 2007 and 2015. In the finance industry, predicting loan default is very important because wrong lending decisions can cause large financial losses. The dataset used in this project is challenging because it is highly imbalanced and contains many different financial and customer-related features.

## Problem Statement

For companies like Lending Club, correctly predicting whether a loan will default is very important. In this project, historical loan data from 2007 to 2015 is used to build a deep learning model that predicts the chance of default for future loans. The dataset is highly imbalanced and has many features, which makes this problem more complex.

## Objective

To create a model that predicts whether a loan applicant will repay the loan or default using historical loan data.

## Domain

**Finance**

## Dataset Description

The dataset contains information about borrowers and their loans. Some important features are:

credit.policy – Whether the borrower meets credit criteria

purpose – Purpose of the loan

int.rate – Interest rate

installment – Monthly installment

log.annual.inc – Log of annual income

dti – Debt-to-income ratio

fico – Credit score

days.with.cr.line – Length of credit history

revol.bal – Revolving balance

revol.util – Credit utilization rate

inq.last.6mths – Recent credit inquiries

delinq.2yrs – Past delinquencies

pub.rec – Public record issues

not.fully.paid – Target column (0 = fully paid, 1 = default)

# Methodology

## 1. Data Loading

The dataset was loaded using Python and Pandas. The structure, number of rows, columns, and data types were examined.

## 2. Data Cleaning

The dataset was checked for missing values. If any were found, they were handled using suitable methods like mean, median, or mode.

## 3. Feature Transformation

The categorical column “purpose” was converted into numerical form using one-hot encoding so that the model could understand it.

## 4. Exploratory Data Analysis (EDA)

EDA was performed to understand the data better:

- Distribution of defaulters and non-defaulters
- Relationship between features and target variable
- Correlation analysis using heatmaps

## 5. Handling Class Imbalance

The target column “not.fully.paid” was highly imbalanced. Techniques like sampling or class weighting were used to make sure the model learned both classes properly.

## 6. Feature Engineering

Highly correlated features were identified and removed to reduce redundancy and improve model performance.

## 7. Feature Scaling

Numerical features were scaled so that all values are in a similar range for better neural network training.

## Model Building

A deep learning model was built using Keras with TensorFlow backend:

- Input layer based on number of features
- Hidden layers with ReLU activation
- Dropout layers to avoid overfitting
- Output layer with Sigmoid activation for binary classification
- The model was trained using binary cross-entropy loss and Adam optimizer.
- Model Evaluation
- The model was evaluated using:
- Accuracy
- Confusion Matrix
- ROC-AUC Score
- Training and validation loss curves
- These metrics helped measure how well the model predicts loan default.

## Results :

The model was able to learn useful patterns from historical data and predict loan defaults with good performance. The ROC-AUC score showed that the model could separate defaulters and non-defaulters effectively.

## Conclusion

In this project:

- Data was cleaned and prepared
- Categorical features were encoded
- Important features were selected
- A deep learning model was built and trained
- Model performance was evaluated

This project shows how deep learning can be used in the finance domain to reduce risk and improve lending decisions.