# A close-up of a logo Description automatically generated

# Lending Club Loan Status Prediction

Introduction:

Lending Club is a financial services company headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission, and to offer loan trading on a secondary market.

Our Purpose is to provide a case study which gives us an idea on how Lending Club real world business problem is solved. In this case study we’ll create a predictive analysis i.e. Machine Learning Models to get in-depth and understanding of how Lending Club uses risk analytic strategies to determine if the borrower is worthy of receiving the loan all while reduce the risk of losing money

## Business Overview:

Lending Club faces two types of risks:

* *If the borrower is likely repaying the loan, then not approving the loan results in a loss of business to the company.*
* *If the borrower is likely NOT to repay the loan, i.e. they’re likely to default, then approving the loan will potentially lead to financial loss for the company.*

Data Provided contains information about applicant’s loan status and which is divided into two different datasets, one dataset for approval of the loan and second dataset is rejection of the loan. The data further also contains several features that determines the weight of the borrower’s credit worthiness.

Once those objectives are met Lending Club evaluates per borrower’s basis and provide one of two following decisions:

* Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
  + Fully paid borrowers have fully paid the loan (the principal and the interest rate)
  + Current: borrowers are in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
  + Charged-off: borrowers have not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan.
* Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

## Business Objectives:

Some lending companies tends to approve loans (with very high interest rate) to ‘risky’ borrowers who risk of a company’s financial loss The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default causing the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

To determine if Lending Club making correct analytic loan decisions to borrowers and reduce financial loss; we’ll need to adopt a predictive model analysis to project how accurate Lending Club makes their lending decision.

Lending Club makes these business lending decisions based on these driving factors:

* FICO (Credit Scores)
* Payment History
* DTI (Debt-To-Income)
* Risk Score
* Employment
* Income

Our projection goal is to predict the accuracy of borrowers that were offered and rejected by the Lending Club.

Determine any borrowers who were approved with the loan that had defaulted on the company’s loan.

Detect any potential false-positive and/or true-negative based on company’s loan decision (i.e. a borrower whom with good credit score, middle-class income and never misses a payment for over past 7 years gets denied of a loan)

## Dataset Source:

Data comes in two sources: first dataset shows list of approved borrowers including the borrower’s demographic, income, loan information, and financial details.

Second dataset shows list of borrowers who were rejected by the company includes the borrower’s application information such as loan request amount information.

Noted that the dataset provided below are historical data ranged from 2007 thru 2018.

Dataset with borrowers’ approval of the Loan:

id,

member\_id,

loan\_amnt,

funded\_amnt,

funded\_amnt\_inv,

term,

int\_rate,

installment,

grade,

sub\_grade,

emp\_title,

emp\_length,

home\_ownership,

annual\_inc,

verification\_status,

issue\_d,

loan\_status,

pymnt\_plan,

url,

desc,

purpose,

title,

zip\_code,

addr\_state,

dti,

delinq\_2yrs,

earliest\_cr\_line,

fico\_range\_low,

fico\_range\_high,

inq\_last\_6mths,

mths\_since\_last\_delinq,

mths\_since\_last\_record,

open\_acc,

pub\_rec,

revol\_bal,

revol\_util,

total\_acc,

initial\_list\_status,

out\_prncp,

out\_prncp\_inv,

total\_pymnt,

total\_pymnt\_inv,

total\_rec\_prncp,

total\_rec\_int,

total\_rec\_late\_fee,

recoveries,

collection\_recovery\_fee,

last\_pymnt\_d,

last\_pymnt\_amnt,

next\_pymnt\_d,

last\_credit\_pull\_d,

last\_fico\_range\_high,

last\_fico\_range\_low,

collections\_12\_mths\_ex\_med,

mths\_since\_last\_major\_derog,

policy\_code,

application\_type,

annual\_inc\_joint,

dti\_joint,

verification\_status\_joint,

acc\_now\_delinq,

tot\_coll\_amt,

tot\_cur\_bal,

open\_acc\_6m,

open\_act\_il,

open\_il\_12m,

open\_il\_24m,

mths\_since\_rcnt\_il,

total\_bal\_il,

il\_util,

open\_rv\_12m,

open\_rv\_24m,

max\_bal\_bc,

all\_util,

total\_rev\_hi\_lim,

inq\_fi,

total\_cu\_tl,

inq\_last\_12m,

acc\_open\_past\_24mths,

avg\_cur\_bal,

bc\_open\_to\_buy,

bc\_util,

chargeoff\_within\_12\_mths,

delinq\_amnt,

mo\_sin\_old\_il\_acct,

mo\_sin\_old\_rev\_tl\_op,

mo\_sin\_rcnt\_rev\_tl\_op,

mo\_sin\_rcnt\_tl,

mort\_acc,

mths\_since\_recent\_bc,

mths\_since\_recent\_bc\_dlq,

mths\_since\_recent\_inq,

mths\_since\_recent\_revol\_delinq

num\_accts\_ever\_120\_pd,

num\_actv\_bc\_tl,

num\_actv\_rev\_tl,

num\_bc\_sats,

num\_bc\_tl,

num\_il\_tl,

num\_op\_rev\_tl,

num\_rev\_accts,

num\_rev\_tl\_bal\_gt\_0,

num\_sats,

num\_tl\_120dpd\_2m,

num\_tl\_30dpd,

num\_tl\_90g\_dpd\_24m,

num\_tl\_op\_past\_12m,

pct\_tl\_nvr\_dlq,

percent\_bc\_gt\_75,

pub\_rec\_bankruptcies,

tax\_liens,

tot\_hi\_cred\_lim,

total\_bal\_ex\_mort,

total\_bc\_limit,

total\_il\_high\_credit\_limit,

revol\_bal\_joint,

sec\_app\_fico\_range\_low,

sec\_app\_fico\_range\_high,

sec\_app\_earliest\_cr\_line,

sec\_app\_inq\_last\_6mths,

sec\_app\_mort\_acc,

sec\_app\_open\_acc,

sec\_app\_revol\_util,

sec\_app\_open\_act\_il,

sec\_app\_num\_rev\_accts,

sec\_app\_chargeoff\_within\_12\_mths,

sec\_app\_collections\_12\_mths\_ex\_med,

sec\_app\_mths\_since\_last\_major\_derog,

hardship\_flag,

hardship\_type,

hardship\_reason,

hardship\_status,

deferral\_term,

hardship\_amount,

hardship\_start\_date,

hardship\_end\_date,

payment\_plan\_start\_date,

hardship\_length,

hardship\_dpd,

hardship\_loan\_status,

orig\_projected\_additional\_accrued\_interest,

hardship\_payoff\_balance\_amount,

hardship\_last\_payment\_amount,

disbursement\_method,

debt\_settlement\_flag,

debt\_settlement\_flag\_date,

settlement\_status,

settlement\_date,

settlement\_amount,

settlement\_percentage,

settlement\_term

Here’s Dataset column values for the loans that were rejected:

Amount Requested,

Application Date,

Loan Title,

Risk\_Score,

Debt-To-Income,

Ratio,

Zip Code,

State,

Employment

Length,

Policy Code

\*\* Two datasets shown to compare those whom Lending Club accepted and rejected the loan, will use these two data for factoring the company’s decisions (i.e. factors that impact the approval or denial of loans)

\*\*Because dataset is very large this requires the data to be scaled down to smaller multiple datasets via partitioning, set datatypes and reduce unneeded columns

## Data Analytic Tools

Following Requirements and tools needed to provide a predictive model to LendingTree lending decision.

Primary language, and platform required for our choice to write our analysis in:

A logo with a snake

Description automatically generated

Platform(s) Required use for creating in notebook formats (\*Note: using Google Collab to run PySpark module):

A logo with grey circles and black text

Description automatically generated

A yellow circle logo with black text

Description automatically generated

Data sources platform require for import and store data (\*Data Stored in Google Drive due to large size):

A logo for a company

Description automatically generated

Another component requires to handle this large dataset for, reviewing datasets (to extract needed data) modifying columns, partition data into chunks.

A logo with text and a star

Description automatically generated

Libraries require for data wrangling, cleaning, and creating model and running tests on the dataset:

A logo of a person

Description automatically generated with medium confidence

A blue and black text

Description automatically generated

A logo of a company

Description automatically generated

A blue letters with a circle and a blue and orange circle

Description automatically generated with medium confidence

A logo with a letter f

Description automatically generated

*(\*\*possibility…??)*

Required platform to use for our Visual Data Display:

A logo with blue and orange crosses

Description automatically generated