Background on peer-to-peer lending

Peer-to-peer lending refers to the practice of lending money to individuals (or small businesses) via online services

Interest rates are usually set by an intermediary platform

The intermediary platform generates revenue by collecting a one-time fee on funded loans

(from borrowers) and by charging a loan servicing fee to investors.

LendingClub issues :

Amount : loans between $1,000 and $40,000

Duration : 36 or 60 months

Categories of Loan : grades A, B, C, D, E, F, and G (A 🡪 Safest)

Note: The safer the loan the lower the interest rate, and so investors have to balance risk and return when deciding which loans to invest in.

Current - being reimbursed in a timely manner.

Late - 16 and 120 days overdue.

Default - more than 121 days.

Charged-Off - LendingClub decided that the loan will not be paid off

Means 5 months after the term of each loan has ended, it is fully paid or charged-off

Imp Note: For this analysis we are considering expired loans not current on going loans

Example : For example, for an analysis carried out in April 2018, this implies looking at all 36-month loans issued on or before October 31, 2014 (36 months + 5 months) and all 60-month loans issued on or before October 31, 2012.

Terms:

resp. less - likely to default

resp. high – good returns

Therefore, to avoid unsuccessful investments, our goal is to estimate which loans are more likely to default and which will yield low returns

**we investigate several machine learning tools and show how one can use historical data to construct informed investment strategies**

Jasmin would seek to construct a portfolio:

1. with the highest possible return,
2. subject to constraints imposed by her risk tolerance and diversification requirements

(e.g., no more than 25% of loans with grades E or F)

In this case study, we investigate the extent to which using predictive models can increase portfolio performance.

1. estimating the predictive ability of models on out-of-sample data
2. estimate the return to expect from an investment

Note : Even with a seemingly good predictive model in hand, estimating the return of an investment requires additional analysis

**Objective:**

1. **Decision**

* How much to allocate
* Which Loan to invest

1. Best decision / worst decision

objective—to make as much money as possible

Split the data

* + first part to make her decisions
  + second to evaluate them

Note: her decisions will be which loans in the second part to invest in, and evaluating her decision will require looking at the actual outcome of those loans and seeing how much

money they returned

This looks easy but it is difficult to calculate exactly ‘‘how much money’’

Some loans are 36 months long, and some are 60 months long

loans default at different times— early or late

repaid early

1. jbk

Whether a loan will default.

Whether a loan will be paid back early.

If it defaults, how soon will this happen?

If it is paid back early, how soon will it be paid back?

1. Check all attributes and get imp

Some attributes are related to :

the borrowers’ - FICO score, employment status, annual income),

platform’s decisions - loan grade, interest rate

loan performance - status, total payment

categorical

numeric

Some attributes are constantly updated while some are set once and for all

Note : there is no variable in the data set with the return information of each loan. Instead, one needs to carefully calculate the return by using the data available

**actual return or annualized return**

The relevant variables for calculating the return are the loan status, the total payment, the funded amount, the fees generated, and the loan duration.

**Leakage:** leakage is a situation in which a model is built using data that will not be available at the time

the model will be used to make a prediction

Ex: here we are calculated total amount paid based on loan default, loan early repayment, and time of

aforementioned events

when investing in future loans, Jasmin would not have access to the total amount that would eventually be repaid on that loan

data has hanged timely so leakage

therefore should not be used to create an investment

strategy

**Cleaning**

**Remove loans that are in continuation**

**Remove leakage**

**When we take loans in continuation the set has more shorter loans than longer (over sampling of shorter loans) because those will only be the expired loan**

**So the analysis is little not correct , but we can not directly remove it so we keep it and also test extra with removing the short loan**

**Remove outliers**

box-and-whisker plots for continuous variables

histograms for discrete variables

we can not remove outliers that appear in target variables or in leakage variables

**Save in Pickle**

**Data Exploration**

**Observation :**

the lower the grade, the higher the probability of default, and the higher the average interest rate LendingClub charges

**Predictive models for default**