QuantUniversity, LLC

www.quantuniversity.com

Machine Learning applications for Credit Risk

Presented By:

Sri Krishnamurthy, CFA, CAP

www.QuantUniversity.com

sri@quantuniversity.com

Credit Risk and scoring

Credit risk in consumer credit

Credit-scoring models and techniques assess the risk in lending to customers.

Typical decisions:

- Grant credit/not to new applicants
- Increasing/Decreasing spending limits
- Increasing/Decreasing lending rates
- What new products can be given to existing applicants ?

Working with mixed-data

- Gower similarity coefficient is a composite measure
- It takes quantitative (such as rating scale), binary (such as present/absent) and nominal (such as worker/teacher/clerk) variables.
- Gower distance is used for calculating distances when we have mixed types of variables (continuous and categorical)
- A linear combination using user-specified weights (most simple an average) is calculated to create the final distance matrix.
- The metrics used for each data type are described below:
  - Quantitative: range-normalized Manhattan distance
  - Ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
  - Nominal: variables of k categories are first converted into k binary columns and then the Dice coefficient is used (https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient)

Support in R

- daisy
- Pam
- agnes

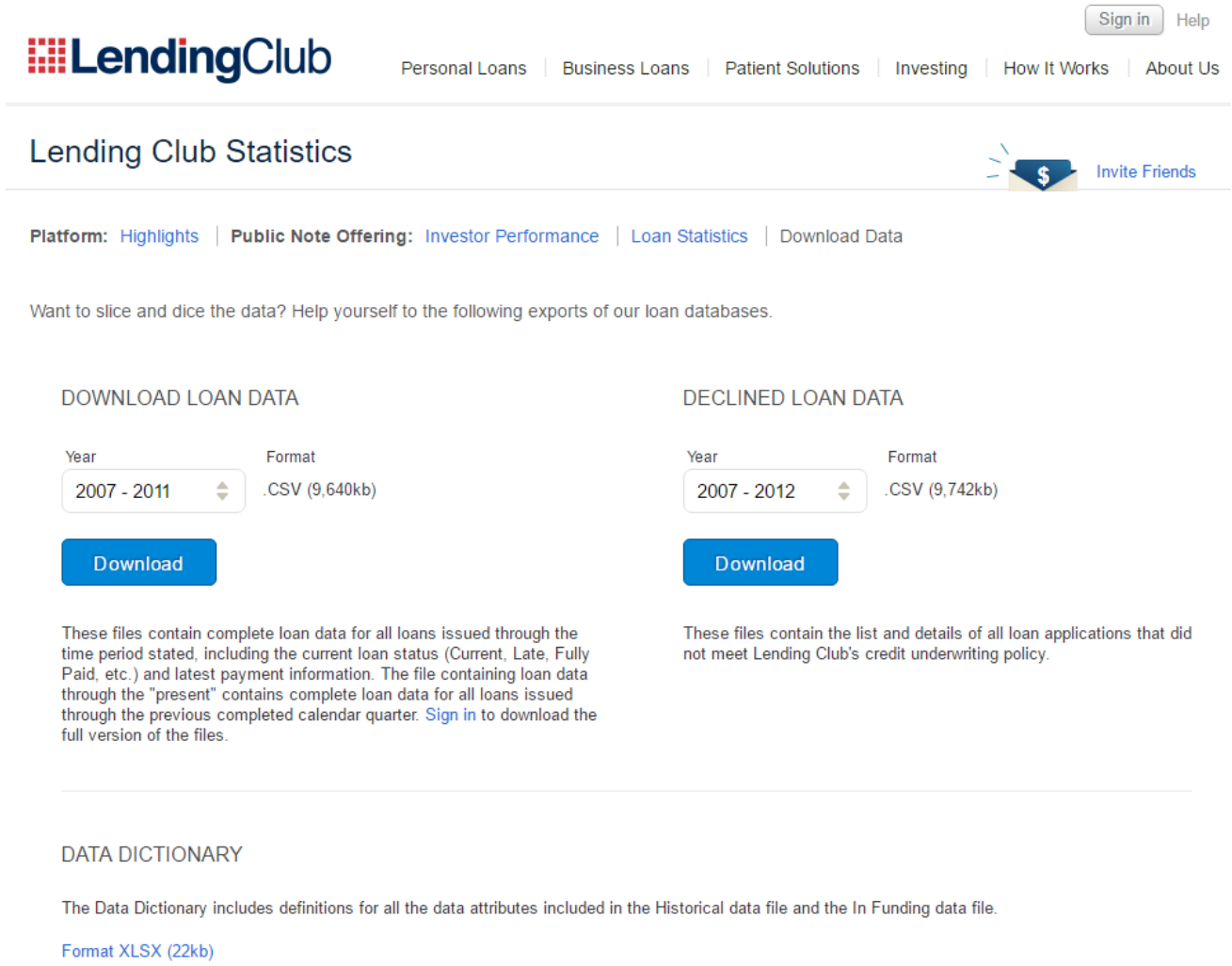
https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/pam.html

https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/CLARA

Getting to know the Lending Club Dataset

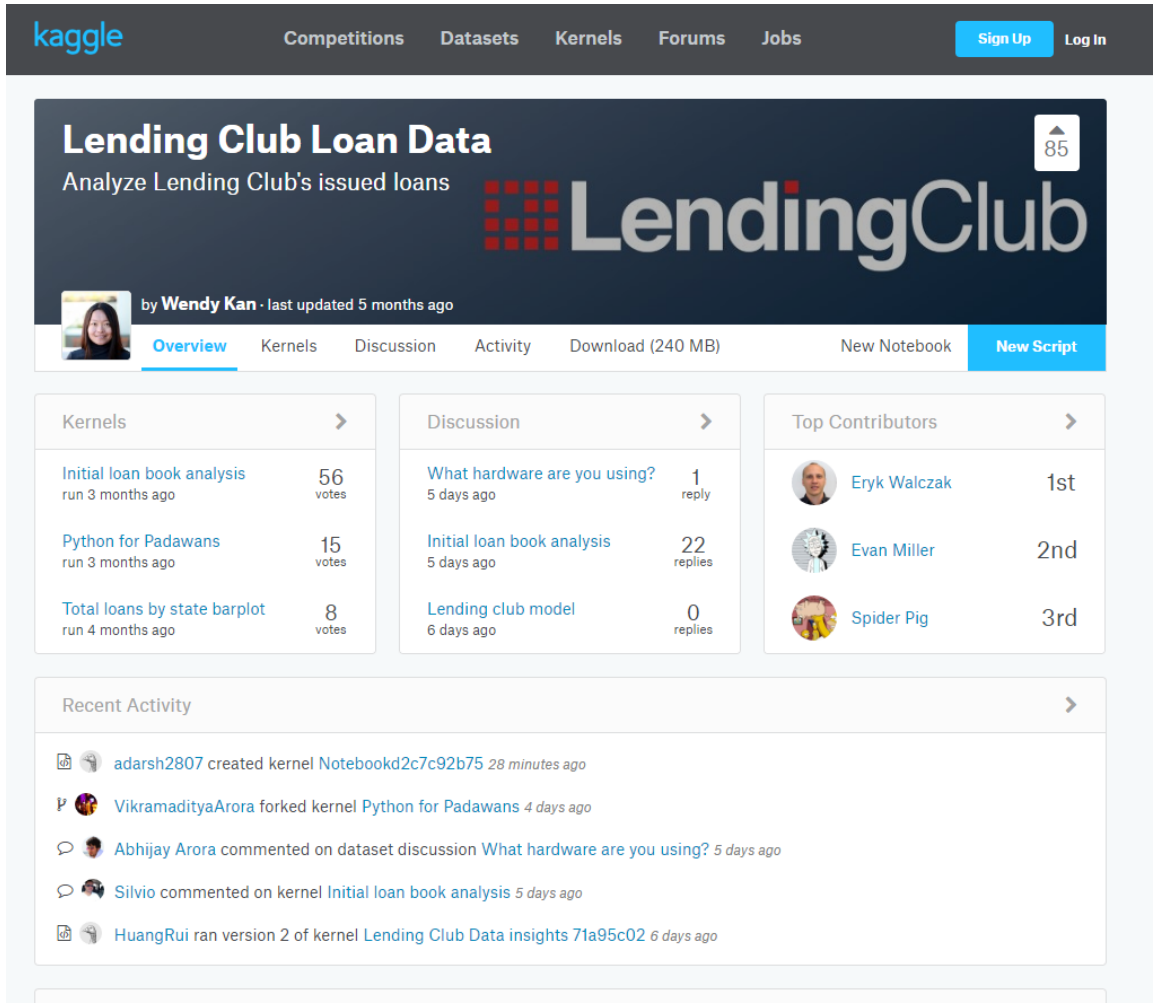
Lending club

The Data



https://www.lendingclub.com/info/download-data.action

The Data



https://www.kaggle.com/wendykan/lending-club-loan-data

Variable description

Variable name	Variable type	Descriptions
addr_state	factor	Name of states
annual_inc	numerical	The self-reported annual income provided by the borrower during registration.
delinq_2yrs	numerical	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
dti	numerical	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.
emp_length	factor	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.
funded_amnt	numerical	The total amount committed to that loan at that point in time.
funded_amnt_inv	numerical	The total amount committed by investors for that loan at that point in time.
grade	factor	LC assigned loan grade
home_ownership	factor	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
inq_last_6mths	numerical	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
installment	numerical	The monthly payment owed by the borrower if the loan originates.
int_rate	numerical	Interest Rate on the loan
issue_d	factor	The month which the loan was funded
loan_amnt	numerical	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.
loan_status	factor	Current status of the loan
purpose	factor	A category provided by the borrower for the loan request.
sub_grade	factor	LC assigned loan subgrade
term	factor	The number of payments on the loan. Values are in months and can be either 36 or 60.
verification_status	factor	Current verification status

Objective

- Calculate dissimilarity between observations.
- Select algorithm to group observations together
- Choose the best number of clusters
- Visualize clusters on reduced dimensions

Selecting number of clusters

- Partitioning around medoids (PAM) is used in this case.
- PAM is an iterative clustering procedure with the following steps:
  - Step 1: Choose k random entities to become the medoids.
  - Step 2: Assign every entity to its closest medoid (using the distance matrix we have calculated).
  - Step 3: For each cluster, identify the observation that would yield the lowest average distance if it were to be re-assigned as the medoid. If so, make this observation the new medoid.
  - Step 4: If at least one medoid has changes, return to step 2. Otherwise, end the algorithm.

Visualization with reduced dimension

- One way to visualize many variables in a lower dimensional space is with t-distributed stochastic neighborhood embedding (t-SNE)
- This method is a dimension reduction technique that tries to preserve local structure so as to make clusters visible in a 2D or 3D visualization.
- https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding