MATH-571
Project Proposal
Chetan Sharma
Narayan Dasari
Yash Agarwal

Project Proposal

Project Description and Research Goal

Credit risk has often been measured through consumer historical payment and income data. This study seeks to use loan consumer history, status, and ratio attributes to learn which attributes correlate with defaulting statuses. We plan to organize and analyze our data set using Principal Component Analysis and will construct models using Random Forest, Backwards step Logistic Regression and LDA. Eigenvalue decomposition and data reduction methods will be used to extract features. Our reduced dimension elements will be further analyzed using summated scales to find correlations between factor loan status. We will be looking to see what patterns this analysis can explain.

Questions Addressed

- How does Lending Club go about predicting whether a loan will default?
- Are there attributes amongst the consumer's history, status or ratings that influence this result?
- if there is a correlation between the consumer attributes and their likelihood of defaulting. ?
- Which analysis model will be best for analysis of dataset ?

Proposed Methodology

- Assumptions/limitations
    - Due to data availability, the time frame of this analysis is restricted to 2007-2016.
- Data Dictionary
    - After analyzing the initial dataset, variables that were missing a significant amount of the data were removed and the following 49 variables can be used to begin the research
- Data Collection
    - The initial dataset contained 49 variables with 407,770 observations. Our target variable is "loan_status" and this describes the current, delinquent, or paid statuses of the loan records. There are 8 different status of loan_status.
- Pre-Processing
    - Instances that included the loan_status of "Current", "Issued", or "In Grace" were removed as these represent loans that are not in a default status nor fully paid so would not apply to the goal of this analysis.
    - Our analysis focuses on loans that are either delinquent (Charged Off, Default, Late (16-30 days), Late (31-120 days)) or fully paid.
    - In order to split the data into a fully paid or default status, we will transform all the delinquent statuses into one group called "Delinquent"
- Data Analysis

- Principle Component Analysis : To choose the principal components, by running PCA analysis we will to account for 70% of the variation in the data, we require 10 components for both train and test set
- K mean clustering : K means clustering is a supervised machine learning technique. It aims to partition observations into k clusters with the nearest mean. This results into partitioning of the data space into Voronoi cells. Given a set of observations (x1, x2, x3, … , xn) where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k (≤ n) sets S = {S1, S2, …, Sk} so as to minimize the within-cluster sum of squares.
- Linear Discriminant Analysis : Linear discriminant analysis (LDA) is a statistical method used to find a linear combination of variables that separates two or more classes of objects. The goal is, utilizing computed means of each classes' variables and a common covariance matrix between all classes, to come up with a linear combination of variables that projects the distribution of the classes separately into one subspace. For our analysis, we will be using this method on Loan_Status to see if we can accurately separate the two groups "Delinquencies" and "Fully Paid".

Metrics to Measure Analysis

- We will be using analysis models like Principle component analysis, linear discriminant analysis, Canonical correlation etc . These include:
  - Covariance
  - Mean
  - Variance
  - Standard Deviation
  - Vector quantization , cluster analysis

Project Outline

Literature review and related work

- https://canvas.harvard.edu/courses/12656/files/2822174/download?verifier=cwyLD199GhxwqW1TKTESsPVfaaNJWX0lqZBDfSns&wrap=1
- https://is.cuni.cz/webapps/zzp/download/120269679
- http://cs229.stanford.edu/proj2015/199_report.pdf

Data sources and reference data

- https://www.kaggle.com/adityasheth/analysis-and-modelling-of-lending-club-loan-data/data

- https://www.lendacademy.com/policy-code-2-loans-lending-club/
- http://budgeting.thenest.com/open-trades-credit-report-23674.html
- The dataset was found on Kaggle.com and describes Lending Club's borrowing data for 887,379 distinct loans from the years 2007 to 2016 with 79 total variables

Variables

- Input Variables

	LoanStatNew	Description
1	addr_state	The state provided by the borrower in the loan application
2	issue_d	The month which the loan was funded
3	policy_code	publicly available policy_code=1new products not publicly available policy_code=2
4	purpose	A category provided by the borrower for the loan request.
5	int_rate	Interest Rate on the loan
6	grade	LC assigned loan grade
7	annual_inc	The self-reported annual income provided by the borrower during registration.
8	annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
9	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
10	collections_12_mths_ex_me	Number of collections in 12 months excluding medical collections
11	delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
12	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.
13	dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
14	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.
15	funded_amnt	The total amount committed to that loan at that point in time.
16	home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
17	inq_last_6mths	The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
18	installment	The monthly payment owed by the borrower if the loan originates.
19	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.
21	mths_since_last_delinq	The number of months since the borrower's last delinquency.
22	mths_since_last_major_der	Months since most recent 90-day or worse rating
23	mths_since_last_record	The number of months since the last public record.
24	open_acc	The number of open credit lines in the borrower's credit file.
25	out_prncp	Remaining outstanding principal for total amount funded
26	pub_rec	Number of derogatory public records
27	revol_bal	Total credit revolving balance
28	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
29	total_acc	The total number of credit lines currently in the borrower's credit file
30	total_pymnt	Payments received to date for total amount funded
31	total_rec_late_fee	Late fees received to date
32	open_acc_6m	Number of open trades in last 6 months
33	open_il_6m	Number of currently active installment trades
34	open_il_12m	Number of installment accounts opened in past 12 months
35	open_il_24m	Number of installment accounts opened in past 24 months
36	mths_since_rcnt_il	Months since most recent installment accounts opened
37	total_bal_il	Total current balance of all installment accounts
38	il_util	Ratio of total current balance to high credit/credit limit on all install acct
39	open_rv_12m	Number of revolving trades opened in past 12 months
40	open_rv_24m	Number of revolving trades opened in past 24 months
41	max_bal_bc	Maximum current balance owed on all revolving accounts
42	all_util	Balance to credit limit on all trades
43	total_rev_hi_lim	Total revolving high credit/credit limit
44	inq_fi	Number of personal finance inquiries
45	total_cu_tl	Number of finance trades
46	inq_last_12m	Number of credit inquiries in past 12 months
47	acc_now_delinq	The number of accounts on which the borrower is now delinquent.
48	tot_coll_amt	Total collection amounts ever owed
49	tot_cur_bal	Total current balance of all accounts

- Output Variables

LoanStatNew	Description
loan_status	Current status of the loan

- Removed Variables

LoanStatNew	Description
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
mths_since_last_major_der	Months since most recent 90-day or worse rating
mths_since_last_record	The number of months since the last public record.
open_acc_6m	Number of open trades in last 6 months
open_il_6m	Number of currently active installment trades
open_il_12m	Number of installment accounts opened in past 12 months
open_il_24m	Number of installment accounts opened in past 24 months
mths_since_rcnt_il	Months since most recent installment accounts opened
total_bal_il	Total current balance of all installment accounts
il_util	Ratio of total current balance to high credit/credit limit on all install acct
open_rv_12m	Number of revolving trades opened in past 12 months
open_rv_24m	Number of revolving trades opened in past 24 months
max_bal_bc	Maximum current balance owed on all revolving accounts
all_util	Balance to credit limit on all trades
inq_fi	Number of personal finance inquiries
total_cu_tl	Number of finance trades
inq_last_12m	Number of credit inquiries in past 12 months

- Transformed Loan Status variables

loan_status records removed
Current
Issued
Grace Period

loan_satus OldValue	New Value
Current	Record Removed
Issued	Record Removed
Grace Period	Record Removed
charged Off	Deliquent
Default	Deliquent
Late (16-30 days)	Deliquent
Late (31-120 days)	Deliquent
Fully Paid	Fully Paid

- Transformed Employee length variable

emp_length Old Value	New Value
n/a	0
< 1 year	0.5
1 year	1
2 years	2
3 years	3
4 years	4
5 years	5
6 years	6
7 years	7
8 years	8
9 years	9
10+ years	10

Analysis Model

- Principle component analysis
- Linear discriminant analysis
- Canonical correlation
- K means clustering
- Naive Bayes
- Correspondence Analysis

Tools

- Software
  - Tableau
  - RStudio
- R Libraries
  - corrplot
  - psyc
  - car
  - CCA
  - MASS
  - QuantPsyc
  - leaps
- Project Management and Source Control
  - GitHub