MCIT 591 – Final Project: Loan default predictor

Ian Park (ianspark@seas.upenn.edu) / Antonius Santoso (asantoso@seas.upenn.edu)

11th May 2019 - Update

Our final project is a ***program to predict loan default-rate based on machine learning algorithms***. The program is using Lending Club dataset from Kaggle (https://www.kaggle.com/wendykan/lending-club-loan-data) and the machine learning used is Logistic Regression.

The program has 2 parts:

1. Machine Learning program (back-end)

Machine Learning program is done in Python because it’s easier to perform data analyses and feature engineering as well as conduct machine learning operation in Python. This Python program will load the Lending Club data and use it to train and test ML model using Logistic Regression technique. The output of this program will be regression parameters and coefficients.

Milestones achieved: Performed data exploration, cleansing and analysis (Feature Engineering), Built ML model with Logistic regression, Refine ML model

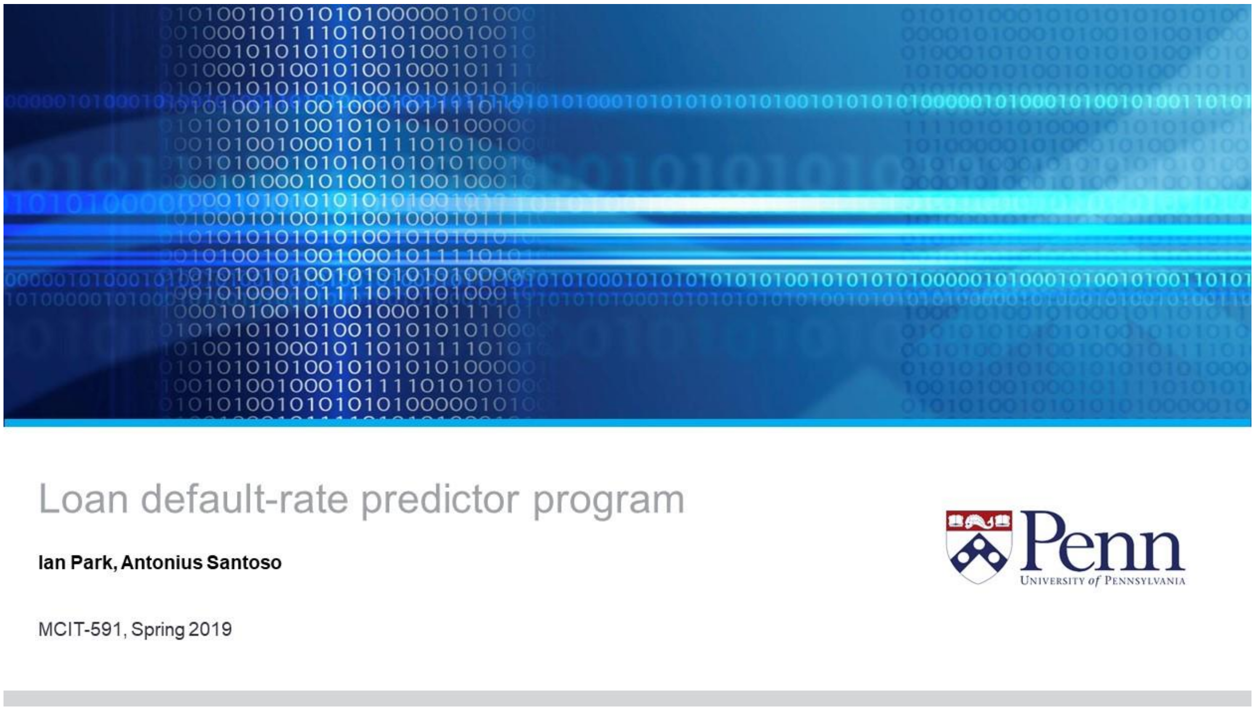
2. Loan Predictor program (front-end)

The loan predictor program is done in JAVA that will serve as main program for loan prediction. This program will receive Logistic regression parameters and coefficients from ML program and perform prediction of loan default based on user inputs.

Milestones achieved: Developed JAVA classes (DefaultPredictor, FileReader, Loan, LogRegCalculator), Built JUnit test cases, Built user input interaction along with explanation and 2 different examples of default rate

MCIT 591 – Final Project: Loan default predictor

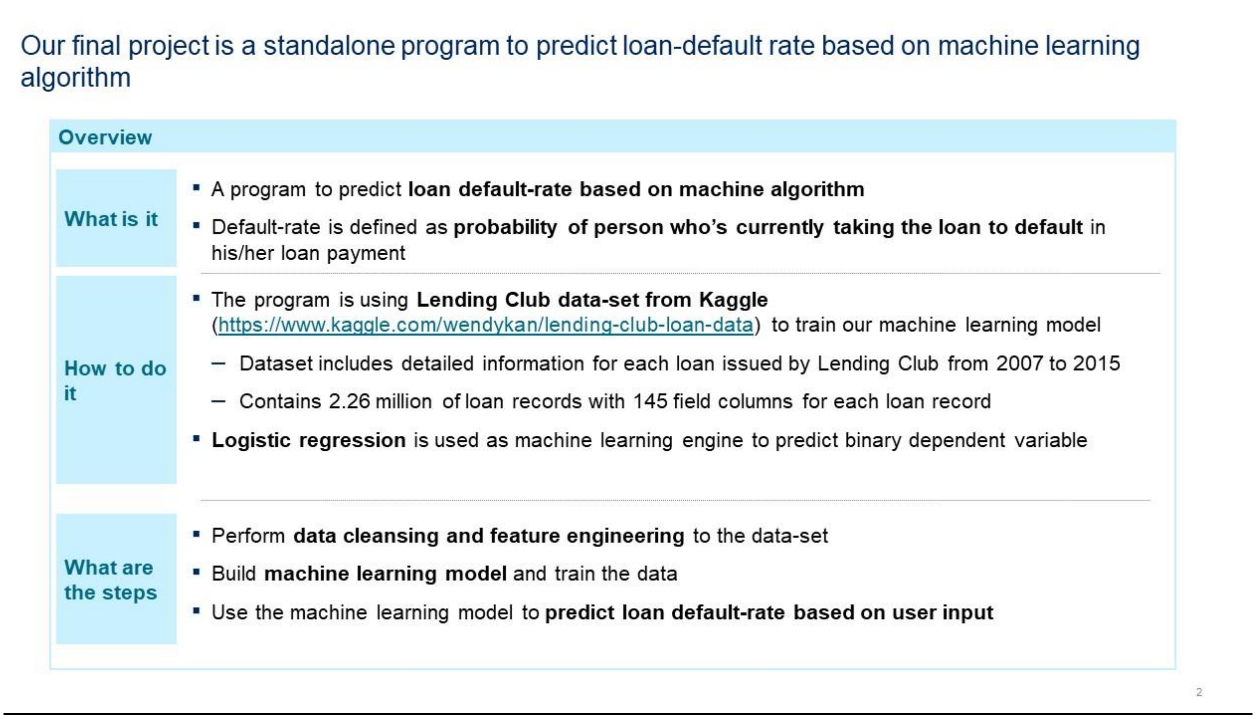
Ian Park (ianspark@seas.upenn.edu) / Antonius Santoso (asantoso@seas.upenn.edu)



Loan default-rate predictor program

Ian Park, Antonius Santoso

MCIT-591, Spring 2019



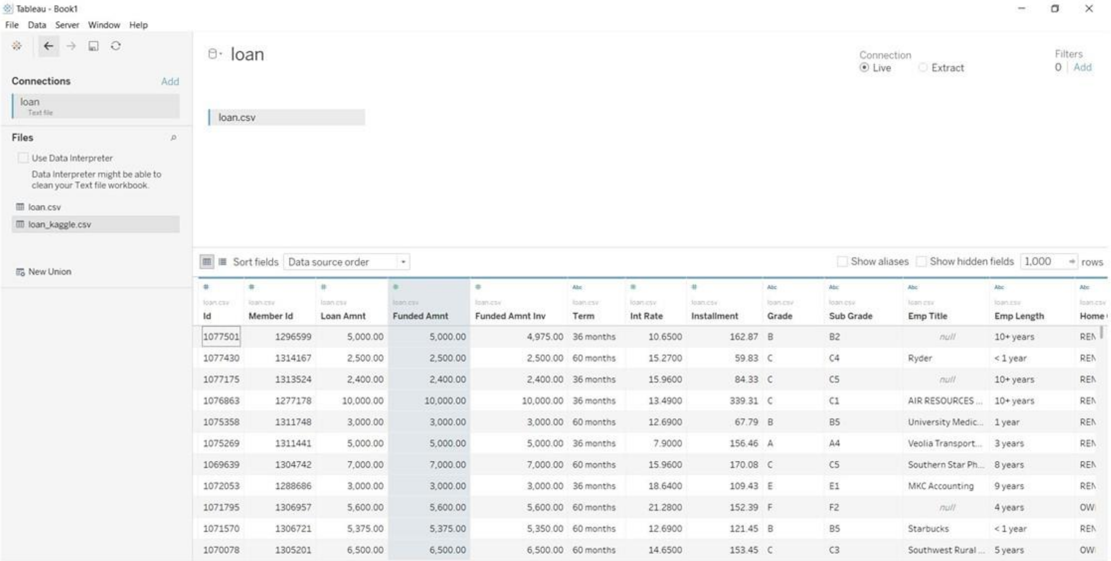
Our final project is a standalone program to predict loan-default rate based on machine learning algorithm

Overview	
What is it	▪ A program to predict **loan default-rate based on machine algorithm** ▪ Default-rate is defined as **probability of person who's currently taking the loan to default** in his/her loan payment
How to do it	▪ The program is using **Lending Club data-set from Kaggle** (https://www.kaggle.com/wendykan/lending-club-loan-data) to train our machine learning model – Dataset includes detailed information for each loan issued by Lending Club from 2007 to 2015 – Contains 2.26 million of loan records with 145 field columns for each loan record ▪ **Logistic regression** is used as machine learning engine to predict binary dependent variable
What are the steps	▪ Perform **data cleansing and feature engineering** to the data-set ▪ Build **machine learning model** and train the data ▪ Use the machine learning model to **predict loan default-rate based on user input**

MCIT 591 – Final Project: Loan default predictor

Ian Park (ianspark@seas.upenn.edu) / Antonius Santoso (asantoso@seas.upenn.edu)

Lending Club Data set from Kaggle has rich features (e.g. loan term, interest rate, income, etc.) to train ML model to make prediction



loan
Text file

☐ Use Data Interpreter

Data Interpreter might be able to clean your Text file workbook.

loan.csv

loan_kaggle.csv

loan

Connection: ● Live ○ Extract

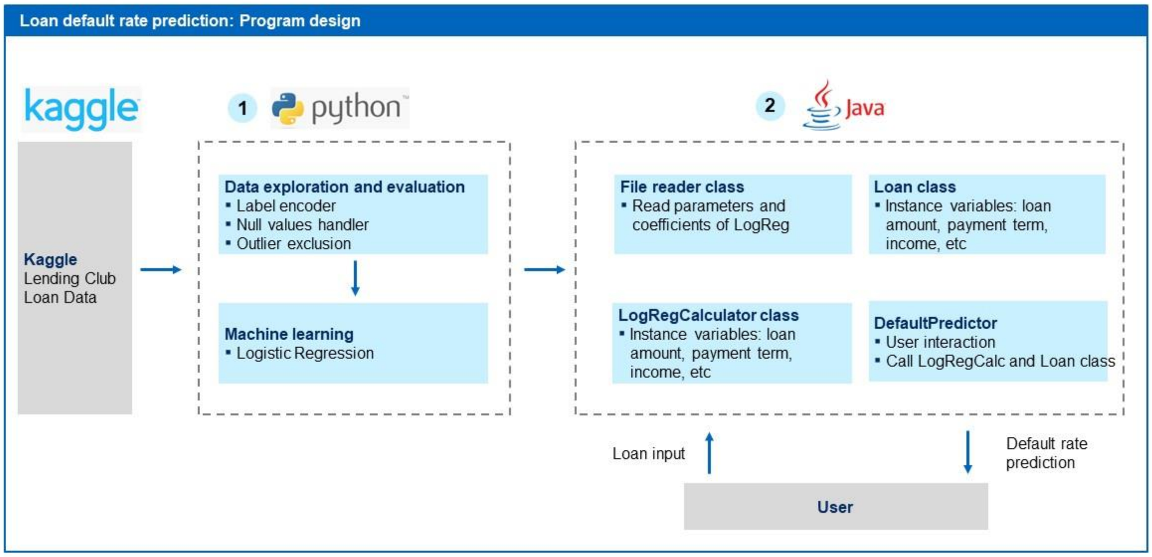
Filters: 0

Sort fields: Data source order

☐ Show aliases ☐ Show hidden fields 1,000 rows

Id	Member Id	Loan Amnt	Funded Amnt	Funded Amnt Inv	Term	Int Rate	Installment	Grade	Sub Grade	Emp Title	Emp Length	Home
1077501	1296599	5,000.00	5,000.00	4,975.00	36 months	10.6500	162.87	B	B2	null	10+ years	REN
1077430	1314167	2,500.00	2,500.00	2,500.00	60 months	15.2700	59.83	C	C4	Ryder	< 1 year	REN
1077175	1313524	2,400.00	2,400.00	2,400.00	36 months	15.9600	84.33	C	C5	null	10+ years	REN
1076863	1277178	10,000.00	10,000.00	10,000.00	36 months	13.4900	339.31	C	C1	AIR RESOURCES ...	10+ years	REN
1075358	1311748	3,000.00	3,000.00	3,000.00	60 months	12.6900	67.79	B	B5	University Medic...	1 year	REN
1075269	1311441	5,000.00	5,000.00	5,000.00	36 months	7.9000	156.46	A	A4	Veolia Transport...	3 years	REN
1069639	1304742	7,000.00	7,000.00	7,000.00	60 months	15.9600	170.08	C	C5	Southern Star Ph...	8 years	REN
1072053	1288686	3,000.00	3,000.00	3,000.00	36 months	18.6400	109.43	E	E1	MKC Accounting	9 years	REN
1071795	1306957	5,600.00	5,600.00	5,600.00	60 months	21.2800	152.39	F	F2	null	4 years	OW
1071570	1306721	5,375.00	5,375.00	5,350.00	60 months	12.6900	121.45	B	B5	Starbucks	< 1 year	REN
1070078	1305201	6,500.00	6,500.00	6,500.00	60 months	14.6500	153.45	C	C3	Southwest Rural ...	5 years	OW

Loan default-rate prediction program has 2 parts: ML program at back-end with Python and Loan predictor program at front-end with Java



Loan default rate prediction: Program design

Kaggle Lending Club Loan Data

Data exploration and evaluation
- Label encoder
- Null values handler
- Outlier exclusion

Machine learning
- Logistic Regression

File reader class
- Read parameters and coefficients of LogReg

Loan class
- Instance variables: loan amount, payment term, income, etc

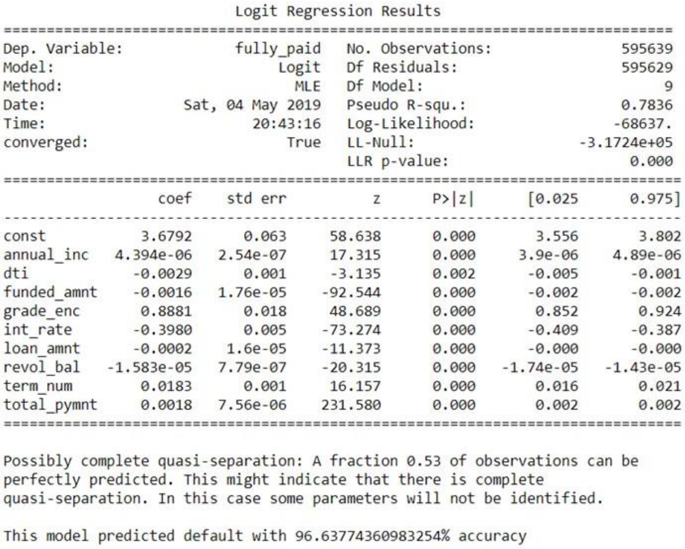
LogRegCalculator class
- Instance variables: loan amount, payment term, income, etc

DefaultPredictor
- User interaction
- Call LogRegCalc and Loan class

MCIT 591 – Final Project: Loan default predictor

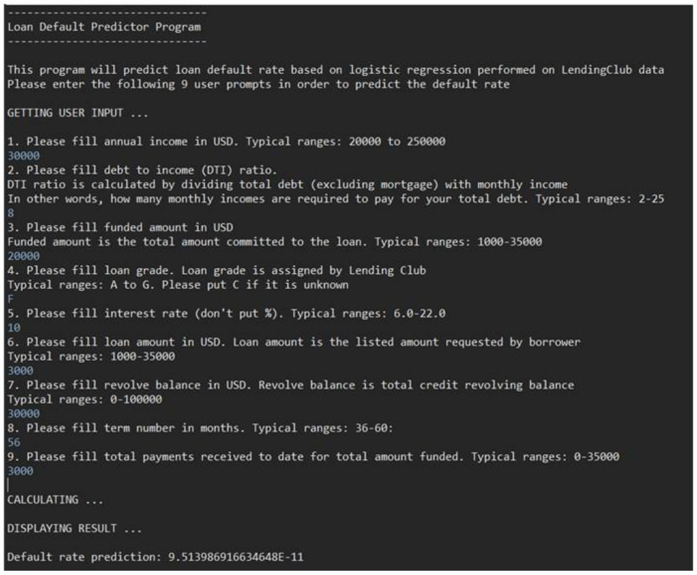
Ian Park (ianspark@seas.upenn.edu) / Antonius Santoso (asantoso@seas.upenn.edu)

1. Python: Machine Learning Engine

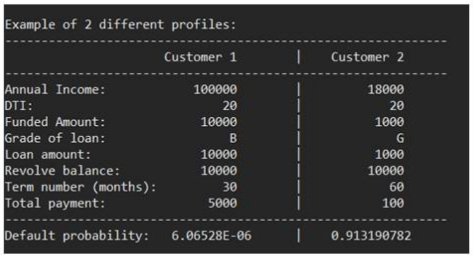


```
                           Logit Regression Results
==============================================================================
Dep. Variable:             fully_paid   No. Observations:               595639
Model:                          Logit   Df Residuals:                   595629
Method:                           MLE   Df Model:                            9
Date:                Sat, 04 May 2019   Pseudo R-squ.:                  0.7836
Time:                        20:43:16   Log-Likelihood:                -68637.
converged:                       True   LL-Null:                   -3.1724e+05
                                        LLR p-value:                     0.000
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const           3.6792      0.063     58.638      0.000       3.556       3.802
annual_inc   4.394e-06   2.54e-07     17.315      0.000     3.9e-06    4.89e-06
dti            -0.0029      0.001     -3.135      0.002      -0.005      -0.001
funded_amnt    -0.0016   1.76e-05    -92.544      0.000      -0.002      -0.002
grade_enc       0.8881      0.018     48.689      0.000       0.852       0.924
int_rate       -0.3980      0.005    -73.274      0.000      -0.409      -0.387
loan_amnt      -0.0002    1.6e-05    -11.373      0.000      -0.000      -0.000
revol_bal   -1.583e-05   7.79e-07    -20.315      0.000    -1.74e-05   -1.43e-05
term_num        0.0183      0.001     16.157      0.000       0.016       0.021
total_pymnt     0.0018   7.56e-06    231.580      0.000       0.002       0.002
===============================================================================

Possibly complete quasi-separation: A fraction 0.53 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.

This model predicted default with 96.63774360983254% accuracy
```

2. JAVA: Default predictor with user input



```
------------------------------
Loan Default Predictor Program
------------------------------

This program will predict loan default rate based on logistic regression performed on LendingClub data
Please enter the following 9 user prompts in order to predict the default rate

GETTING USER INPUT ...

1. Please fill annual income in USD. Typical ranges: 20000 to 250000
30000
2. Please fill debt to income (DTI) ratio.
DTI ratio is calculated by dividing total debt (excluding mortgage) with monthly income
In other words, how many monthly incomes are required to pay for your total debt. Typical ranges: 2-25
8
3. Please fill funded amount in USD
Funded amount is the total amount committed to the loan. Typical ranges: 1000-35000
20000
4. Please fill loan grade. Loan grade is assigned by Lending Club
Typical ranges: A to G. Please put C if it is unknown
F
5. Please fill interest rate (don't put %). Typical ranges: 6.0-22.0
10
6. Please fill loan amount in USD. Loan amount is the listed amount requested by borrower
Typical ranges: 1000-35000
3000
7. Please fill revolve balance in USD. Revolve balance is total credit revolving balance
Typical ranges: 0-100000
30000
8. Please fill term number in months. Typical ranges: 36-60:
56
9. Please fill total payments received to date for total amount funded. Typical ranges: 0-35000
3000
|
CALCULATING ...

DISPLAYING RESULT ...

Default rate prediction: 9.513986916634648E-11
```



```
Example of 2 different profiles:
-----------------------------------------------------------
                          Customer 1        |    Customer 2
-----------------------------------------------------------
Annual Income:              100000          |         18000
DTI:                            20          |            20
Funded Amount:               10000          |          1000
Grade of loan:                   B          |             G
Loan amount:                 10000          |          1000
Revolve balance:             10000          |         10000
Term number (months):           30          |            60
Total payment:                5000          |           100
-----------------------------------------------------------
Default probability:    6.06528E-06         |   0.913190782
```