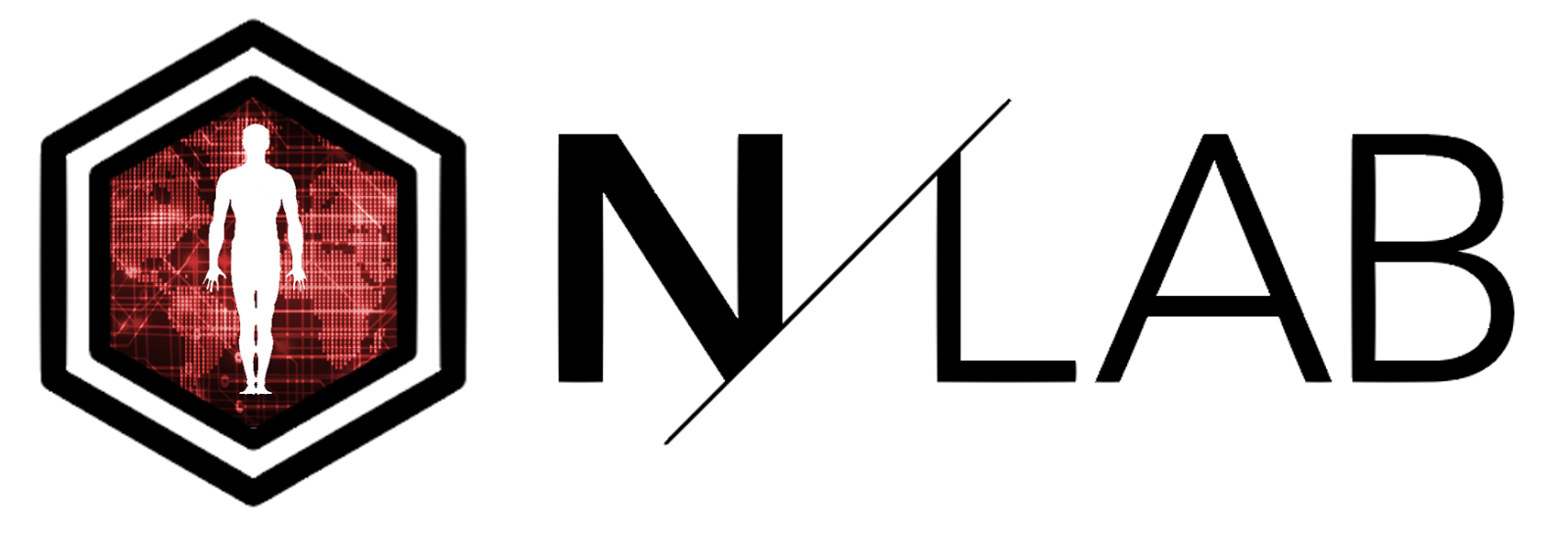
*Machine Learning and Predictive Analytics:* **Coursework**

****

# **~~Coursework: Continuous Risk Evaluation of Loans~~**

**~~DUE DATE: Friday 22th May, 23:59:59 Beijing time. Electronic submission via Moodle.~~**

# Overview

LendingClub is a peer to peer lending company based in the United States, which provides the "bridge" between investors and borrowers. Specifically, LendingClub enables borrowers to create [unsecured personal loans](https://en.wikipedia.org/wiki/Unsecured_debt) last for 36 months or 60 months. Investors can search and browse the loan listings on LendingClub website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, loan purpose, etc., and make money from interest. Borrowers are contracted to **pay back** principal and interest **monthly**, but they can also make payment larger than the monthly due amount and terminate the loan period **in advance**. However, not all monthly payments will be made in time and loans will be charged off if they are several months after due (see <https://help.lendingclub.com/hc/en-us/articles/216127747> for detail). Therefore, it’s important for LendingClub to evaluate the risk of a loan before it is issued and keep tracking the risk during the process of installment.

~~Now, LendingClub is considering mining the lending and payment data of loans issued between 2012-2013 (these loans have been completed before 2020, either full paid or charged off) to help with continuous risk evaluation of loans. Specifically, you~~ **~~task~~** ~~is:~~

**~~1)~~** ~~Developing data mining models (implemented in Python) to predict:~~

~~a. whether a loan will be a “~~**~~fully-paid~~**~~” or “~~**~~charged-off~~**~~” based on loan and borrower information before it’s issued,~~

b. **if** a loan has been issued, whether the borrower will pay back money next month during the process of monthly payment based on loan/borrower information and payment history.

**This section evaluates:**

→ Your ability to undertake cross-sectional and time-series predictive analytics in Python

~~→ Your ability to preprocess the data in case there are missing values and categorical features~~

~~→ Your ability to identify and create relevant features within a given business domain~~  
 2

**2)** Provide some insights into what differentiates a loan to be risky.

**This section evaluations:**

→ Your ability to interpret what is driving the predictions

→ Your ability to relate analytical outcomes to the business scenario

**3)** Provide an executive summary as to what activities can be taken to minimize risk of a loan based on your findings.

**This section evaluates:**→ Your ability to interpret existing analysis  
→ Reason within a given business scenario in order to undertake analytics

# Dataset

* ~~Each student is assigned with an individual dataset of loans issued at a specific month between 2012 to 2013. Each dataset includes two parts: 1) information about loans and borrowers; 2) payment history of the corresponding loans. You can download your assigned dataset from Moodle in the data folder with the filename~~ **~~[data\_studentid.rar]~~**~~.~~
* ~~Since the dataset is based on real world data errors may still exist. As in reality, as an analyst you are expected to check the validity of your analytics as you go. If something doesn't look right and there is a possibility it is caused by data errors, you need to check and potentially document/correct.~~
* ~~Other than cleaning, the data has not been altered and your knowledge of this dataset schema etc. remains valid.~~
* ~~Data dictionary can be found in LCDataDictionary.xlsx~~

# What you need to submit

* A report. **8 pages** maximum. See the report requirement section for the structure and to understand what is expected.
* A all code / notebooks / tableau files. At a minimum I expect:
  + A commented Juypter notebook containing the risk evaluation component
* A supporting documentation document which should include at a minimum:
  + Instructions on how to run the risk evaluation model
  + Any additional cleaning you have undertaken (in most cases I expect this to be a statement that no extra cleaning was required)
  + A table detailing which python method/tableau file/… that was used to create each figure in the report.

# ~~How you need to submit~~

* ~~Except for your coursework report, all other files must be zipped into one file~~
* ~~Submission is electronically via Moodle. There will be two links for submission. Please submit your report via the Turnitin link, and submit your zipped file via another link.~~
* **~~DUE DATE: 22th May, 23:59:59 Beijing time.~~**

# Report Requirements

~~Your report should contain the following.~~ **~~Other than the executive summary~~**~~, page counts are only indicative and you can change the lengths as you see fit. Remember 8 pages is the maximum, it is the quality of the report that is important not the length. Your report should contain both text and all graphs. Remember how (i.e. a statement/pointer to the code/tableau file) each figure was created should be included in a table outside of the report in the supporting document.~~

* Executive Summary **[1 page] 。0.75**
* Description and discussion on summary statistics of important variables **[1.5-2 page] 1.125-1.5**
* Technical report on the developed purchase prediction model. In all cases discuss things that were tried including the evaluation / reasoning as to why they were not finally included. You should at least includes a section on: **[4.5 pages] 3.375**
  + Features  
    **List and justify** which features you tried, which you kept/discarded and why
  + Prediction approach:   
    Describe and justify the selected machine learning approaches (set of steps including any preprocessing and the final predictive algorithm) you tried/used.
  + Evaluation
  + Summary
* An insight report **[2 pages]**  
  This should detail the difference between loans which are high-risky and those are low-risky and be split into these parts:
  + A summary aimed at risk minimization.   
    This should detail the insights in the context of the business and in non-technical language. These insights should be backed and linked to analytics detailed in the second part of the insight report.
  + Your profiling of loans which are high-risky and those are low-risky.
  + A technical report detailing how these insights were derived.

# Some notes/hints:

* Marks will be given for demonstrating understanding and critical thinking of the approaches attempted as well as for the final outcome. For instance, the report should not be of the form: "I did this, I did that" but more focused and critical. I.e. "The final solution was [something]. This was compared against [something else] to investigate [another thing] with [this thing] showing improved performance likely due to [a good reason]."
* Marks will be given for clear documentation within your code and my ability to easily run it. This reflects the real-world situation where the person who writes the code if typically not the one that ends up running it and/maintaining it in the long run. Note, however, that in code documentation should be concise.

# Marking Guidance

Marking will be by section. The provisional relative weights for these sections are included below, along with indications of what will be looked for.

**Executive summary [10%]**

→ Clarity and faithfulness as to the content being summarized

→ Appropriateness of the report for the intended audience

**Initial analysis of the data [5%]**

→ Demonstrated understanding and critical thinking in undertaking the task

→ Appropriateness/justification of approach

→ Correctness of implementation

→ Clarity of report and documentation

**Technical report on the developed purchase prediction models:** **Feature selection, Feature extraction, the prediction approach & evaluation [50%]**

→ Demonstrated understanding and critical thinking in undertaking the task

→ Appropriateness/justification of approach

→ Correctness of implementation

→ Clarity of report and documentation

**Insight report part 1 [10%]**

→ Clarity and faithfulness as to the content being summarized

→ Appropriateness of the report for the intended audience

**Insight report part 2 [25%]**

→ Demonstrated understanding and critical thinking in undertaking the task

→ Appropriateness/justification of approach

→ Correctness of implementation

→ Clarity of report and documentation