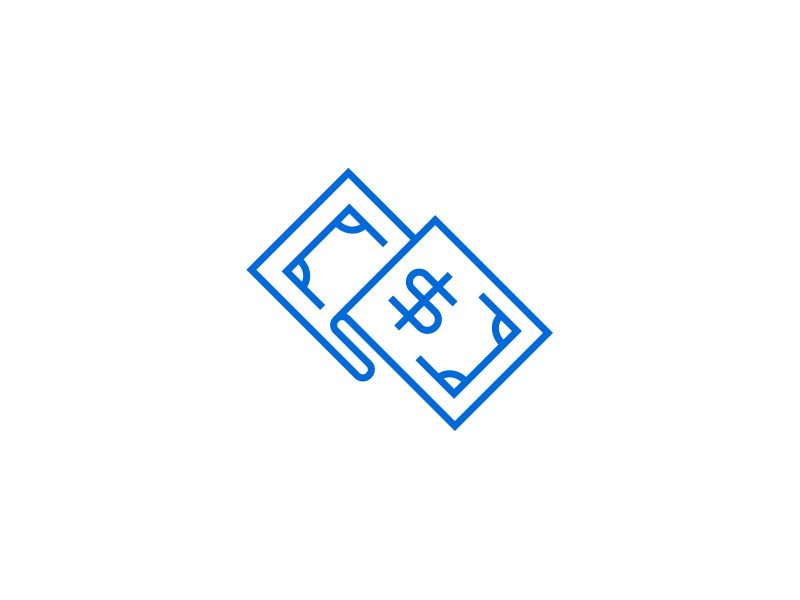
# **Credit Default Modeling**



With the progress of technology and implementation of Data Science in banking, it changes the face of the fintech industry. Most of the banking, financial sectors and social lending platforms are actively investing in lending. But financial institutions might face huge capital loss if they approved the loan without having any prior assessment of default risk. Financial institutions always need a more accurate predictive system for various purposes. Predicting loan defaulters is a crucial task for the banking industry. Banks have an immensely large amount of data like customer’s data, transaction behavior, etc. Data Science is a promising area to process the data and extract the hidden patterns using machine learning techniques. You are going to use statistical measures to preprocess the data and build an effective model that will predict the loan defaulter accurately.

# **OVERVIEW**

## PROBLEM STATEMENT

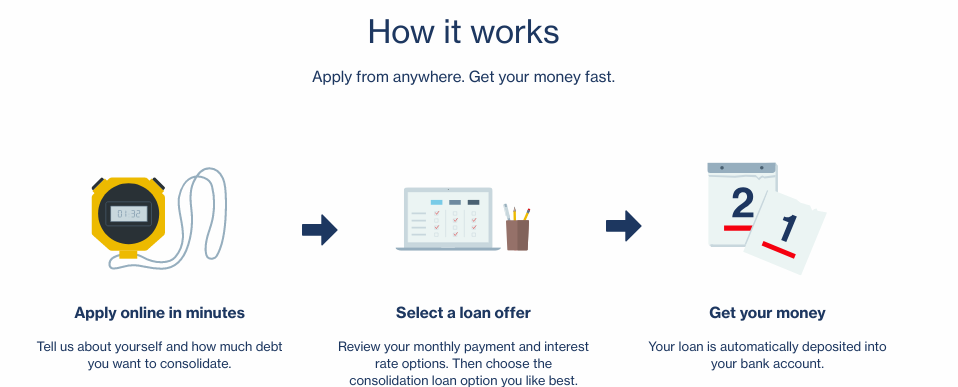
Lending Club (LC) is the world’s largest online marketplace connecting borrowers and investors and the lending network to register with the Securities and Exchange Commission (SEC). It is transforming the banking system to make credit more affordable and investing more rewarding. Lending Club operates at a lower cost than traditional bank lending programs and passes the savings on to borrowers in the form of lower rates and to investors in the form of solid risk-adjusted returns.

Peer to peer lending is the practice of lending money to individuals or businesses through online services that match lenders with borrowers. It is one the world’s largest marketplace connecting borrowers and investors and facilitates the payment of loans up to $40,000, where consumers and small business owners lower the cost of their credit and enjoy a better experience than traditional bank lending, and investors earn attractive risk-adjusted returns.

US Customers interested in a loan complete a simple application at LendingClub.com. LC uses online data and technology to quickly assess risk, determine a credit rating and assign appropriate interest rates and investors can review the loan applications, along with risk analysis provided by LendingClub, to determine how much of the loan request they are willing to fund. Qualified applicants receive offers in just minutes and can evaluate loan options with no impact to their credit score. Investors ranging from individuals to institutions select loans in which to invest and can earn monthly returns.

The entire process is online, using technology to lower the cost of credit and pass the savings back in the form of lower rates for borrowers and solid returns for investors.

Based on the setting of this platform, investors can make wrong decisions in some situations (the borrower received the loan even if they don’t have the ability to return the money). In this case, we want to have a better model that can predict the probability of full repayment..



Details of how it works can be found [here](https://www.lendingclub.com/public/how-peer-lending-works.action).

## PROJECT OVERVIEW

### OBJECTIVE

In this project, you are expected to play with the data provided by LC, conduct a set of exploratory analysis to identify the risky loan applicants at the time of loan application so that such loans can be reduced thereby cutting down the amount of credit loss, and try to apply various machine learning techniques to predict borrower’s default rate.

### MISSION

* Build ML model that predicts if a customer will pay back the laon to find the optimal model-based investment strategy and significant factors leveraging lending strategy. The company can utilise this knowledge for its portfolio and risk assessment. And thus minimise the risk of losing money while lending to customers.
* To analyse the relationship between default risk and these contributory variables, and the possible causes play a significant role in loan defaults. In other words, to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.
* Build a classifier that can accurately categorize a loan by grade.

## EXPECTED OUTCOME

## 1. A solid consolidated project documents and scripts of your ML models engineering.

## 2. Be able to understand and practice data science workflow.

## 3. Confidence in using data science toolbox and Jupyter notebook report

## 4. Be able to pinpoint abnormalities in the data.

## 5. Able to produce a well-organized report which contains relevant insights.

6. Mastering the machine learning skill in building model prediction for this business case and focus heavily on how to apply and evaluate the data science methods, and to suggest business strategies in a real-world business setting after you analyze the given data.

## 7. A summary of your process" - as in thought and work process, the logic / critical thinking behind it, how you solved the problem, tools used and reasoning.

SUBMISSION OF CAPSTONE PROJECT

You will need to deliver three components:

1. A clean version of your code (to Github) with a README giving instructions about how to run it.
2. A document describing your technical solution and evaluation results. You need to submit it in PDF format to Github along with the code, in the root directory.
3. An oral presentation describing your technical solution and evaluation results. You are required to use slide presentation (i.e. PowerPoint, Google Slide, Notebook...etc) and also have to submit your presentation to Github.



For the code and document, which must be submitted to Github, here are the instructions:

1. Put the files for this assignment in a dedicated folder in the group repository you created.
2. Tag your commit with “final version of ML track”.
3. Double-check that Lymeng is added as a collaborator for this repository.
4. Send a link to the tagged commit by email to Lymeng (**lymeng@mekongbigdata.com**) with Sokhna (**sokhna@mekongbigdata.com**) in cc.

## DATA

These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

**Declined Loan Data**

These files contain the list and details of all loan applications that did not meet Lending Club's credit underwriting policy.

In the real world, you may not be given a data dictionary. For this exercise, you are given the data dictionary to help you better understand the meaning of the data.

**Data dictionary**: [LCDataDictionary.csv](https://github.com/dosei1/Lending-Club-Loan-Data/blob/master/LCDataDictionary.csv)

They contain information on almost all the loans issued by LC. The only loans missing from these files are the few loans where LC was not authorized to release publicly the details of the transactions. The information available for each loan consists of all the details of the loans at the time of their issuance as well as more information relative to the latest status of loan such as how much principal has been paid so far, how much interest, if the loan was fully paid or defaulted, or if the borrower is late on payments etc.

* LOAN DATA: These files contain complete loan data for all loans issued through the time period stated, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter. You need to register an LC account and sign in to download the full version of the files.
* DECLINED LOAN DATA :These files contain the list and details of all loan applications that did not meet Lending Club's credit underwriting policy.

There are 115 data attributes in total, however, not all of them can contribute to data analysis or can be used as features for machine learning techniques. It is your responsibility to identify which features to use, or even better, create your own features based on the given attributes.

## MILESTONES

This part serves as a suggested project planning. Some parts can be later adjusted through discussions with mentor

**Step 1 : Project Overview and Setup**

In this step you will get an overview of the project that you have to complete by the end of the program. You are going to set up and upload your data into your own work environment on a local machine.

**Step 2 : Study Data and Data Preparation**

The data provided in this project is relatively formatted, however, you may need to deal with some null values or transform some categorical variables. You are also encouraged to seek external data sources related to lending club for P2P loan data.

In this step you will focus on investigating and understanding the data by using visualization tools or libraries packages in Python. You also need to prepare and clean your data for the next module.

**Deliverables:**

* Look at the data variables and try to understand the meaning behind it.
* Investigate and clean the data.
* Study each variable in comparison with target variables.

**Step 3 : Exploratory Data Analysis**

In this step you will focus on exploring the data by using statistical analysis in combination with visualization coded by libraries packages in Python . Based on the preprocessed data, you are expected to conduct a series of exploratory data analysis to get a taste of how the data looks like or how it is distributed using histogram or other visualization or statistical methods to testify your data normality and satisfy machine learning assumption . In this part you are required to perform various data visualisations to show what you have found and you also have to visualize your findings and to identify anomalies with python.

**Deliverable:**

* Create correlation between all relevant variables
* Provide a statistical detail of data.
* Create visualization from the data with python.
* Arrange a coherent Jupyter-Notebook report.

**Step 4 : Feature engineering**

Since there are over 100 data fields provided, you might not use them all, not even suggested to use all of them. It is your job to perform feature selection, feature engineering, and discover the features that are indicative of someone paying or defaulting on their loan.

**Deliverables:**

* Use feature engineering where necessary.
* Use dimensional reduction where necessary.
* Clean and process data for machine learning.

**Step 5 : Predictive Model for Default Rate**

Here comes the most exciting part! In this step you will focus on making a predictive model based on supervised learning. Based on the previous data exploration and feature engineering, you can predict whether someone will default based on some classification algorithms or predict someone’s probability of loan default using various classification models so as to avoid loans that are predicted to default.

**Deliverables:**

* Determine at least **3** ML algorithms and optimal number of clusters to apply. Possible algorithms include but are not limited to: **Logistic Regression, SVM, Naive Bayes, Random Forest, Ada Boost**, etc.
* Arrange a coherent presentation report of the results, a well-organized Jupyter-Notebook report

**Step 6 : Model Comparison and Evaluation and Final Report**

Your goal is to compare and identify the best model for predicting customer loan default and justify why. Here you are expected to come up with a set of evaluation criteria to compare different model’s performance and discuss their confusion metrics and evaluate your model with different metrics and improve your model. It is important to try to explain why the models give such results and what are their pros & cons under certain constraints.

**Deliverables:**

* Compare ML algorithms by using accuracy metrics (precision, recall …)
* Arrange a coherent presentation report of the results, a well-organized Jupyter-Notebook report

**Step 7 : Segmentation and Final Report**

In this step you will focus on grouping the dataset into different segments based on preferences and other feature engineering. You also need to provide a recommendation for best actions that can be taken to improve revenue from various segments, cost reduction initiatives for the segments found, marketing campaign or new product opportunities..etc for each segment based on your findings.

As this step, please write a final report in Jupyter notebook to include all your codes, graphs and writeup.

**Deliverable:**

* Using the ML algorithms to find loan segmentation.
* Give a meaningful representation for each cluster.
* Provide recommendations

REFERENCE

Here are some works done previously on this topic, feel free to take a look and grab things you think could be helpful to your work. You are also encouraged to research more on related work online. However, you MUST come up with your own ideas and mark all your references at the end of your final report. Here are some references for you to check:

* <https://www.lendingclub.com/info/statistics.action>
* <http://www.lendingmemo.com/lending-club-prosper-default-rates/>
* <https://www.lendingclub.com/public/how-peer-lending-works.action>
* <https://www.lendingclub.com/info/download-data.action>
* [Peer to Peer (P2P) Lending Problems and Potential Solutions](https://pdf.sciencedirectassets.com/280203/1-s2.0-S1877050919X00174/1-s2.0-S1877050919318265/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEGMaCXVzLWVhc3QtMSJIMEYCIQCs1489hTTp8C6Ez%2FWDYv2icSHifnWoqlngSDxbG2pzmAIhAOYhLHf9j0g11AajNB%2Fski7ArCb5km%2BlLJIO%2Bba7c85MKoMECMz%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEQBBoMMDU5MDAzNTQ2ODY1IgzeizGVTR4osU6GgiIq1wMgHNzfirkC%2B2eR8GMUEUjORc9W84nD%2FmeUd18sfqVnGTxIQOHw2zpbkclk0NJ%2B2pnSn%2FcExGlSWtWQ23ZZsAE%2Bbl6aTolyt8Y8sEyB%2FL4MnP8QdpiC93D01caro3%2BwexmPrS2HeHa2a6THZnRDf%2F%2FC5rFENholKbZVeJqH1zQyuXoMoiOHz%2ByxT9K8EcoieWBfmypIHrMWCZezSVqIPPnXU2sGuN7MjuUW6D8oxtWvRaKE%2FDyH8FvDF7B8bR3WucU%2B98uwMoVF50HrT0eJ4%2BzyleIRoPLTYeRoWpRqMElnzVcMWAy%2BixIqGItNdj6k6JOYjuuQpp%2BU7vrFt4%2FQLrFhiFyNckC9PnyM54NqmWwm02sXpItWsbLYBtBZ%2BB151nz7uAQ%2FUDx0MrCgPp2g4%2Bkbfk5fimROfZ1yHxhkAZJsFoVsHRaEwxNl97nXkMGAYvsJWIeYGJwf5OPnj1JI37A8n2WhUugFUhamW4ugfOdGrzpEmH2rmvqaI1O0i%2FoamnGUYH7qXA%2FyrTDPMY4nbaA5D%2BFH03Zmvs1w6BdHVOV5X0NXmp72wS2xvS%2BVYNx0T3Y5LNM1rpTZXx4kJB%2Bf7NOlW6BvdpdQGV6gWcm74MFMKsdrj7CuC48wptLUigY6pAEMjWEnsai%2F2XyviQ8CXB1QrsjFP70lDrHIGFD%2FbODIozHUnyxUAJkz9meaFQlM7hRlfiC6nEHA118Zj%2FrssPnAhZO8Vhq6NH5AWECRrguO3dqvrK2gRFdnWsGLEnyPPkgNXmVrdAEfIBE2iVXvp1Z8qS0rx8y%2F6%2BJdCpAFOrdzvhQcPRXk1iumsu4FcKn%2FbKh%2Fl8FAcg6p3p%2B69oER4oaP3Wf8RA%3D%3D&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20210930T042311Z&X-Amz-SignedHeaders=host&X-Amz-Expires=300&X-Amz-Credential=ASIAQ3PHCVTY6VIXNASZ%2F20210930%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Signature=f6650174a7f110a959e5b879c268c1095999e268aa378c1b8efe2621c4ea6351&hash=8151e2b736672de1610f2d0e0705400ad641cac6450368588d3ad1506638dd4c&host=68042c943591013ac2b2430a89b270f6af2c76d8dfd086a07176afe7c76c2c61&pii=S1877050919318265&tid=spdf-f3e07f76-b7ec-46c2-b46e-8c0379497969&sid=eb95fb07409b6144df783ef61ba185263867gxrqb&type=client)
* Referred to methodologies from stanford project: http://cs229.stanford.edu/proj2015/199\_report.pdf in which they tried to predict the loan status from financial information. They used TF-IDF for text processing.
* [Lending Club Personal Loan](https://www.lendingclub.com/loans/personal-loans)
* [FDIC Equal Housing Lender](https://www.fdic.gov/regulations/laws/rules/2000-6000.html)
* [Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending](https://www.researchgate.net/publication/320216578_Loan_default_prediction_by_combining_soft_information_extracted_from_descriptive_text_in_online_peer-to-peer_lending)
* [Towards Repayment Prediction in Peer-to-Peer](https://www.mdpi.com/2227-7390/7/11/1041/pdf)
* [Credit Risk Analysis in Peer to Peer](https://digitalcommons.bard.edu/cgi/viewcontent.cgi?article=1299&context=senproj_s2019)
* [Credit rating of online lending borrowers using recovery rates](https://ideas.repec.org/a/eee/reveco/v68y2020icp204-216.html)
* [Risk Prediction of Peer-to-Peer Lending Market by a LSTM Model with Macroeconomic Factor](https://dl.acm.org/doi/abs/10.1145/3374135.3385287)
* [Lending Club Review for New Investors](https://www.lendacademy.com/lending-club-review/)