Maximise investment return in P2P lending

What is P2P lending?

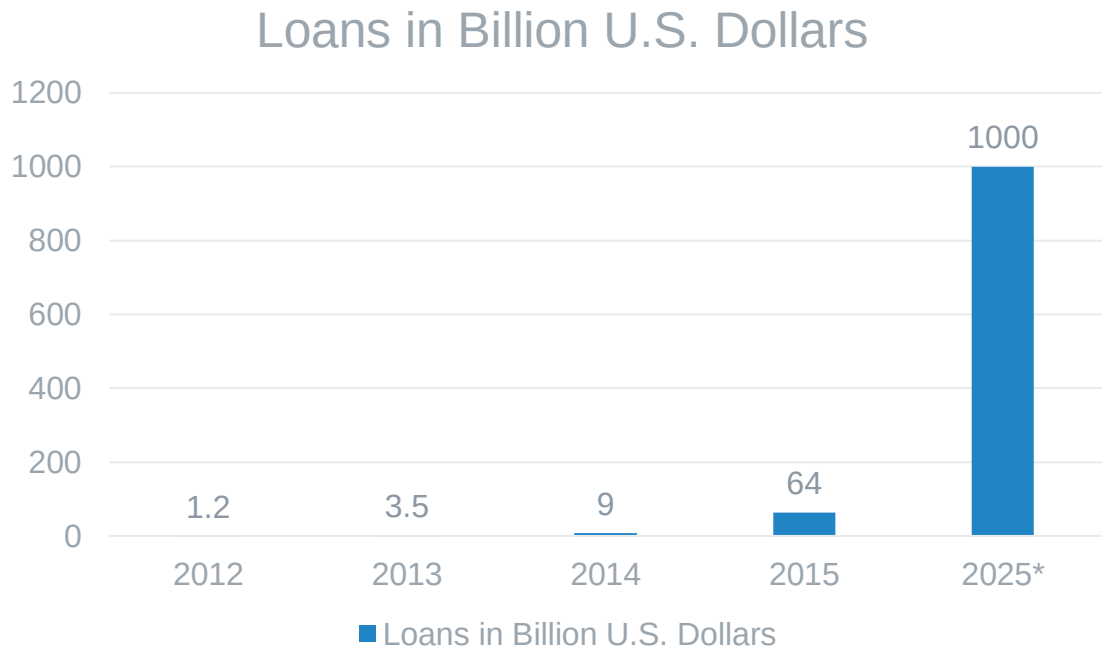
P2P lending gets rid of extra cost paid to traditional Financial Intermediary

- Individuals do not have to use an official financial institution (ie Bank) as an intermediary, effectively remove the middleman from the process. Think about the current DeFi
- P2P lending provides an alternative source of financing for borrowers as well as investment for investors

- Higher returns to the investors
- Opportunity of diversification among different loans and risk levels
- Investor can pick on his own in which loans to invest

- **High credit risk as these are usually unsecured loans and most borrowers who apply p2p lending as they are not able to borrow from banks**

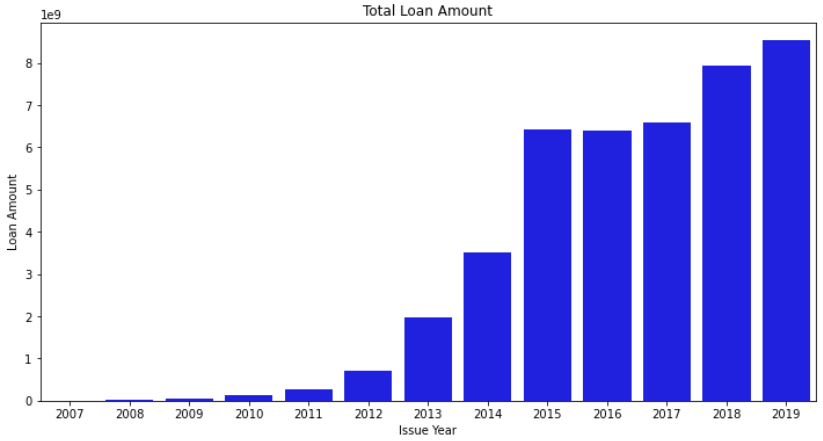
Value of global peer to peer lending from 2012 to 2025
(in billion U.S. dollars)



* Estimated from Statista.com

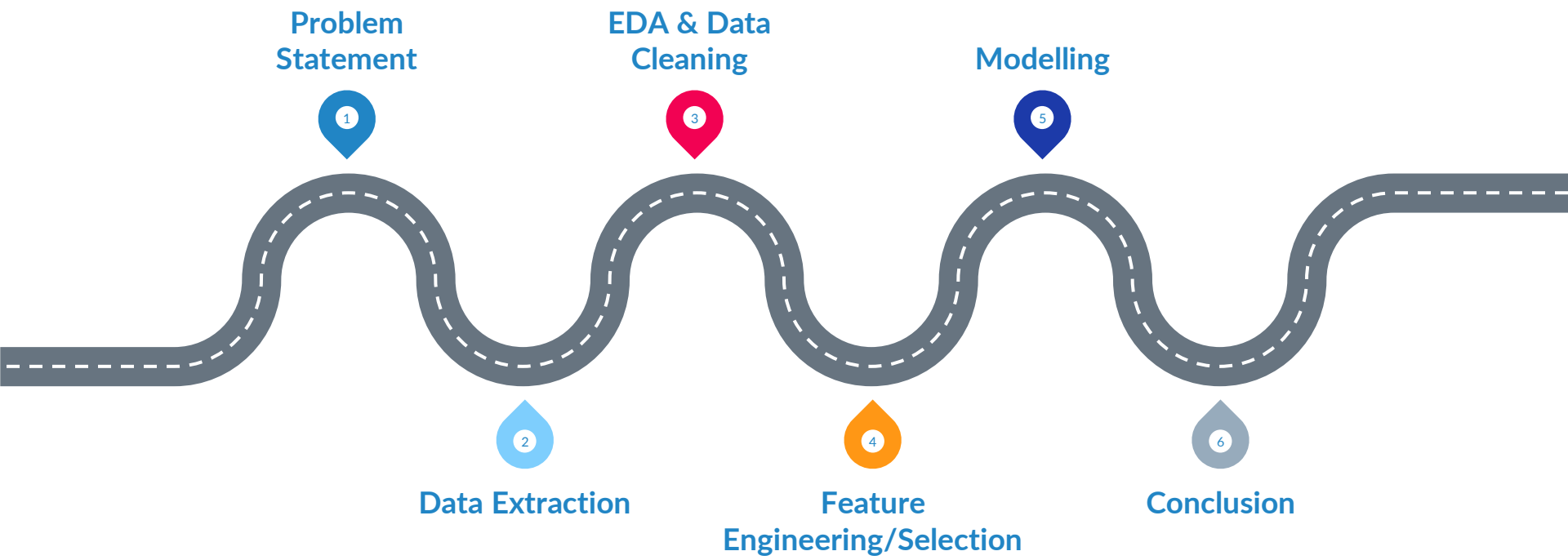
History of Lending Club, the world largest P2P lending plaform

- LendingClub is a peer-to-peer lending company, headquartered in San Francisco, California
- LendingClub was initially launched on Facebook as one of Facebook's first applications but after receiving series A funding round in 2007, it was developed into a full-scale P2P company.
- After a series of subsequent funding and partnership which help to steer the growth, the company IPO in December 2014.



- Borrowers would fill out an application detailing their credit history, loan details, employment status and other self-reported information by which Lending Club would assign a loan grade reflecting the quality of a loan.
- LendingClub publishes all its loans' information on its website for investors to determine which loans to invest

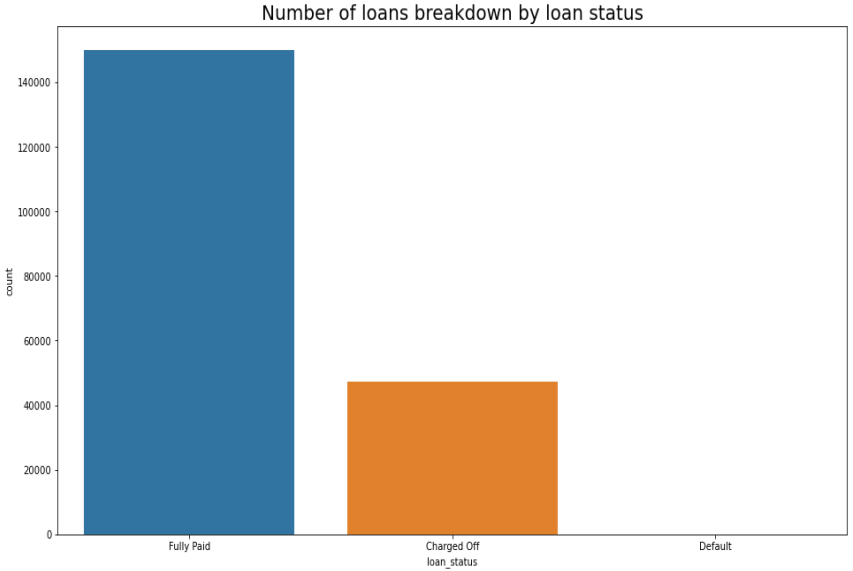
Roadmap


Problem Statement
1
EDA & Data Cleaning
3
Modelling
5
2
Data Extraction
4
Feature Engineering/Selection
6
Conclusion

Problem Statement:

Help P2P lenders to decide on the which loans to invest in their portfolio to obtain the best risk-reward (Sharpe Ratio) return.

Lending Club dataset is downloaded from Kaggle

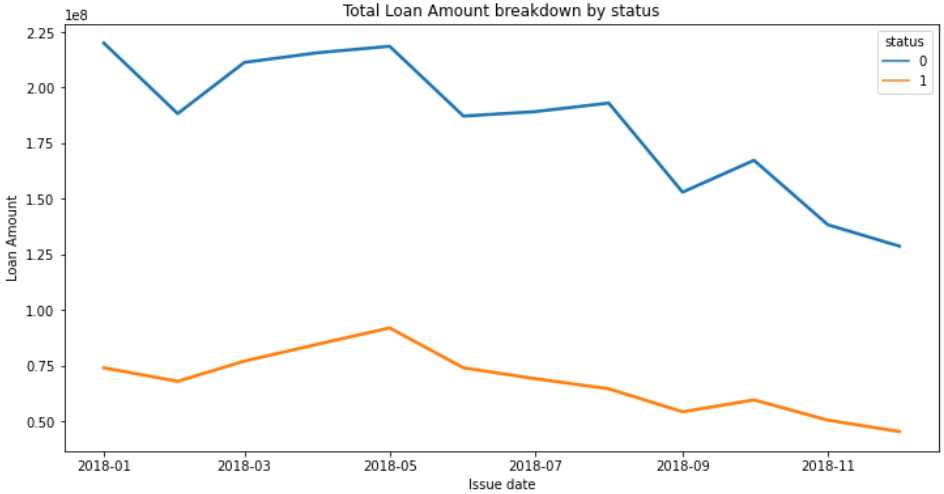


Note: Imbalance dataset which needed to be considered when training the model

- Though dataset is downloaded from 2007 to 2020, we will be using a subset of it (only year 2018 loans) due to the large size of the dataset and loans after 2018 might still be under current loan.
- Removed loans which are still current loan as we will not know if such loans will be defaulted.
- Charged off loans(write off) loans will be considered as default loans.

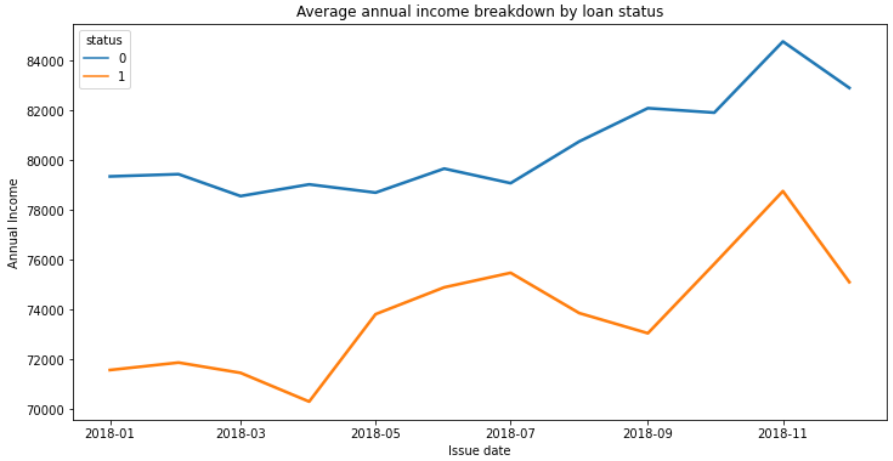
2

EDA: Loan amount decreases as the faster rate for non-defaulted loan



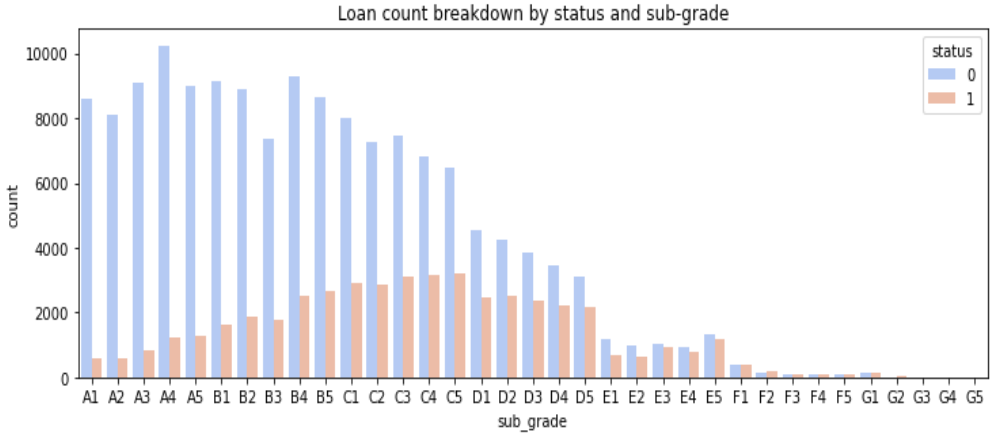
- Loan amount decreasing MoM with non defaulted loan decreasing at a faster rate which means % of defaulted loan in the portfolio is increasing with time.
- The cause of the overall loan decreased is due to a scandal in 2016 which lost some of the investors' confidence

EDA: Average annual income is lower for defaulted loan



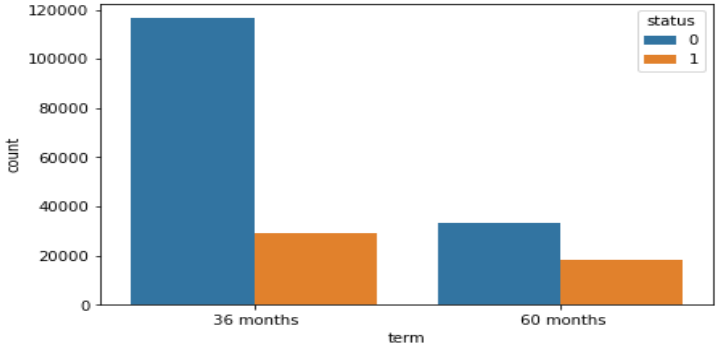
- While average annual income is lower for defaulted loan, annual income in general increase MoM
- This means that timing of the loan issuance might affect the accuracy the model.

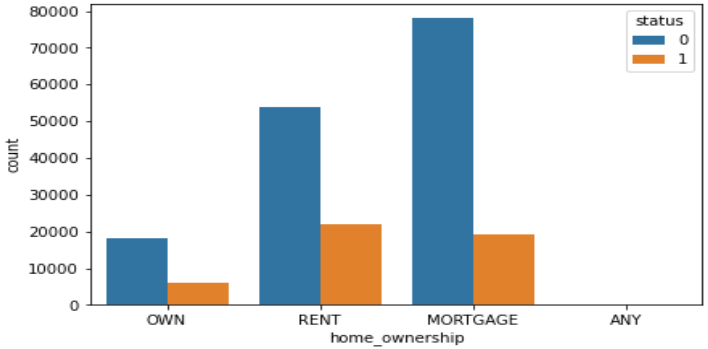
EDA: Probability of loan defaulting is higher for lower loan grade

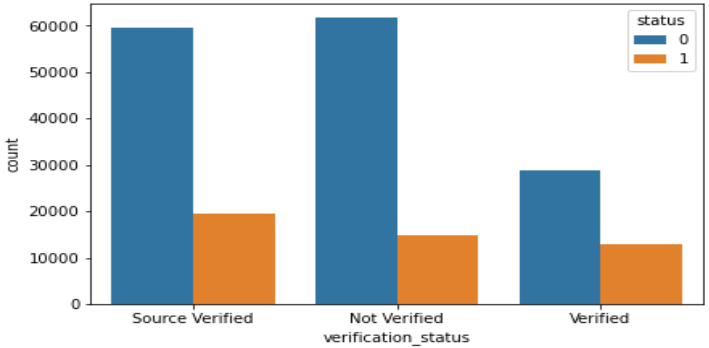


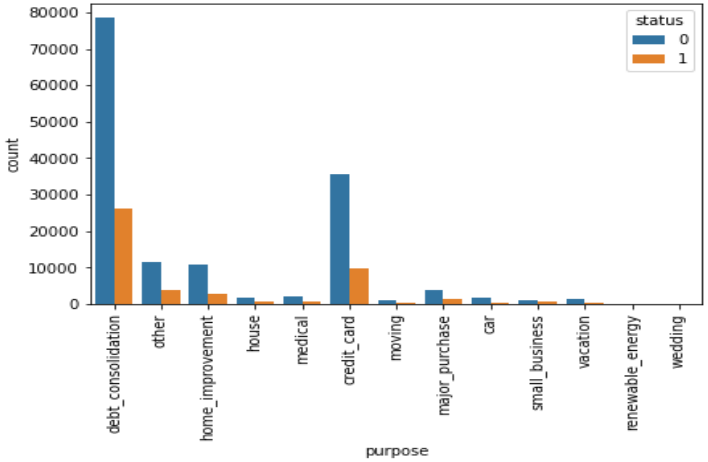
- Majority of the loan are within grade A,B and C. Loan grading is correlated with the loan status as % of defaulted loans is higher for lower graded loans.
- Loan grading will be a good predictor for loan default. However, need to take note if there are any changes to how the loan is graded by LendingClub as it will affect the modelling.

EDA: Loan count breakdown for different categorial features









Steps to clean features before training the model

Replacing null values

- Dropped features which have more than 70% null values.
- For features with less than 10% null values, delete the affected rows with null values.
- For the rest of the features with null values, we replaced these null values based on our best understanding of the data.

Prevent data leakages

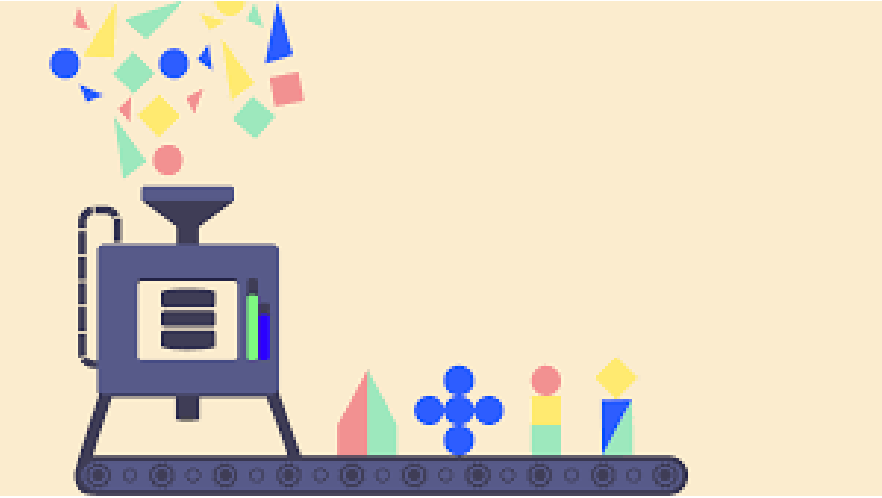
- Removed forward looking features such as next payment date, recoveries and collection status.
- Removed features with no value add to our model. Eg, loan and member ID.

Clean data types

- Checked that data types are correct. Ie, dates and numeric features.
- Encode ordinal features to numerical value.
- Performed one hot encoding for nominal features.

Use of Principal Component Analysis (PCA) for dimensionality reduction

- Reduce the number of input variables without sacrificing much of the explainability
- Keep 90% of the explainability
- Reduced input features by 40%



Selected Light GBM Classifier as the best model

	Model_name	train_auc	test_auc	auc_variance	train_precision	test_precision	train_recall	test_recall
0	Logistic regression	0.729	0.731	0.004	0.377	0.377	0.663	0.665
1	CatBoostClassifier	0.735	0.731	0.005	0.373	0.369	0.693	0.687
2	Light GBM	0.735	0.729	0.008	0.371	0.368	0.694	0.689

Model Selected due to the following:

- Train and test AUC are very similiar and all 3 models generalised quite well (no overfitting) based on the auc variance between the train and test data.
- Selected Light GBM Classifier as it is the fastest and has the best recall score (TP/(TP+FN)).

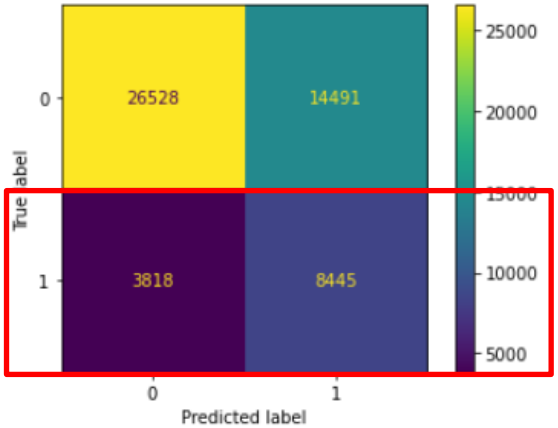
Selected Light GBM Classifier as the best model

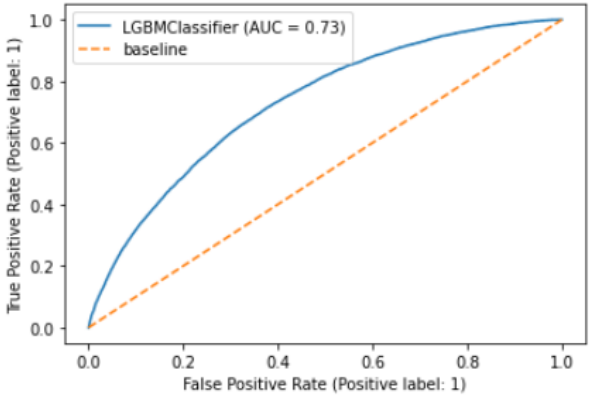
Classification Report

	precision	recall	f1-score	support
0	0.87	0.65	0.74	41019
1	0.37	0.69	0.48	12263
accuracy			0.66	53282
macro avg	0.62	0.67	0.61	53282
weighted avg	0.76	0.66	0.68	53282

Confusion Matrix

<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisp





- We want the best recall score as we aimed to reduce the number of false negative (ie actual defaulted loan but predict as non-default) which may lead investors to investing in defaulted loan

- Investing in defaulted loan will cost investors to lose their investment which will affect investors' portfolio's risk adjusted annual return (measured by sharpe ratio).

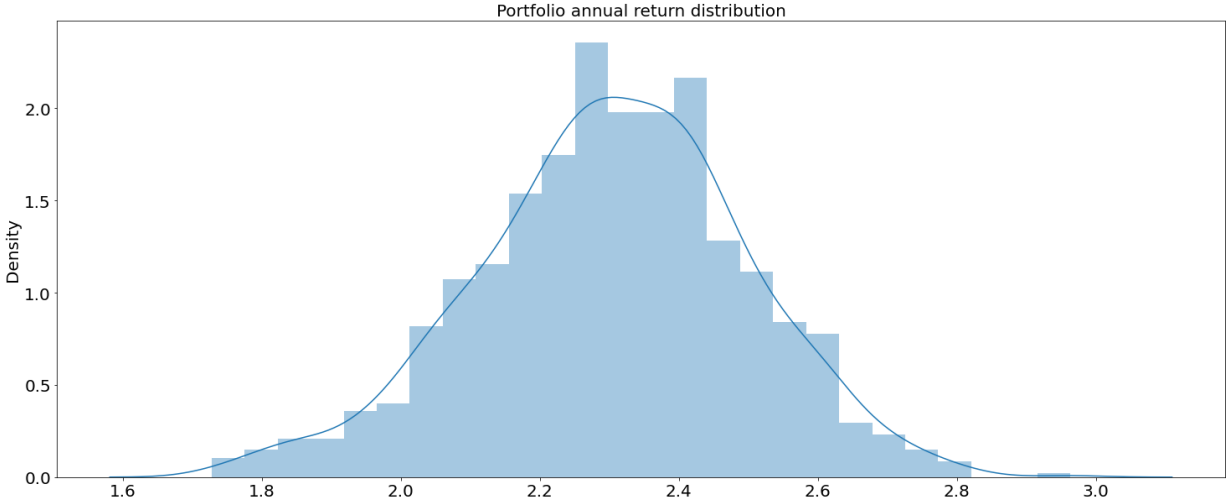
Problem Statement:

Help P2P lenders to decide on the which loans to invest in their portfolio to obtain the best risk-reward (Sharpe Ratio) return.

Steps for portfolio simulation

Number of loans to invest	Select loans to invest	Calculate returns
For the purpose of this presentation, we assume that an investor wants to invest in 10,000 loans	Randomly select 10,000 loans from the set of loans 1,000 times to simulate the portfolio return distribution	From the 1,000 times selection, calculate the sharpe ratio of the portfolio return based on the average annual return and standard deviation

Portfolio return for randomly selected loans w/o model

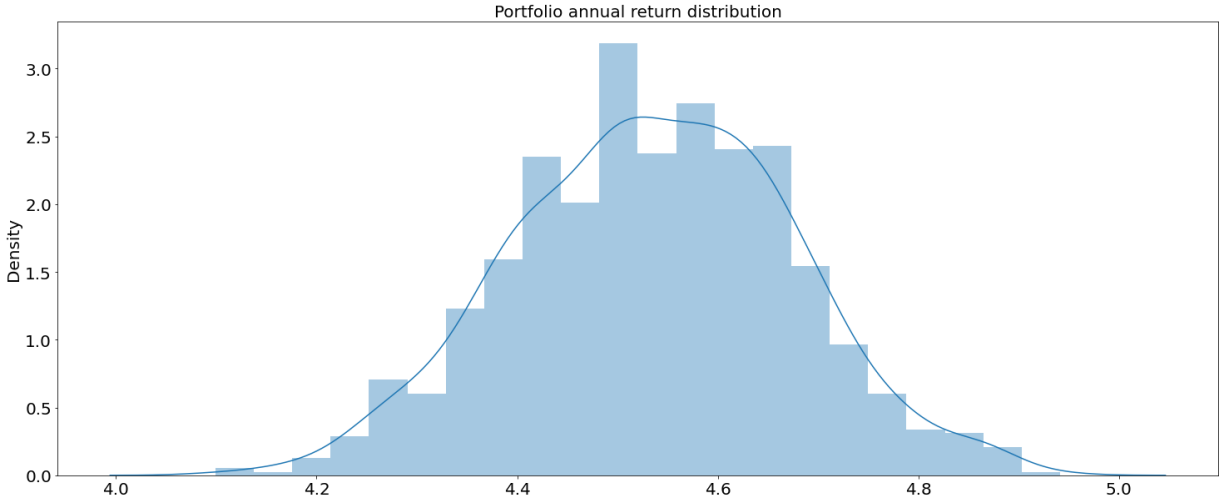


Average Annual Return: 2.3%
Standard Deviation: 0.19
Sharpe Ratio: 4.21

Assumptions:

1. Risk free rate is at 1.5% based on current US 10 treasury bond yield
2. 0 recovery rate for defaulted loan. Full loan amount is defaulted
3. No interest is earned from defaulted loan as we removed any forward-looking features from the data

Portfolio return for randomly selected loans with model



Average Annual Return: 4.5%
Standard Deviation: 0.14
Sharpe Ratio: 21.7

Note: Loan is randomly selected based on a population that was predicted to be non defaulted loans

5x

Increase in sharpe ratio when using model for loan selection

Future Work and Improvements

1. Use longer time period and more features to train the model better. We can also run more grid search to better tune the hyperparameters for each model.

2. Consider how time period will affect the features (such as annual income which will increases over time due to inflation) which are used to train the model.

3. Use of external features such as macroeconomic and market data which are know to be great predictor of bond default rate.

4. Consider time series split and check if the model generalise well with future loans. Afterall, the main purpose of this model is to have the ability to predict future loan default rate.

Thank You!

Q & A