15/06/2021

Peer Lending Project
Stage 6 - Evaluation (deployment)
Group C

Background

This project is an analytical project for "GreatYields" company. The aim for us, as an analytical team, was to create a model which will yield the company higher returns than their annualized return of approximately 2% before entering the peer-to-peer loans market. We got a database of peer-to-peer loans from the "Soft Lending" platform. Throughout the project, we also had leading questions assigned to us by Walter, the company's CIO, that we will answer in the following paper.

The Data

The original data we received included 434,407 instances (different loans) and 150 features (columns). We got two snapshots on the same loans, one from 2018 and one from 2019.

Before starting to build our model, we needed to clean the data. Firstly, we compared the two snapshots and realized it would help us identify leakage - features that change throughout the loan lifespan and, therefore, can be problematic for our model. After that, we kept only the 2019 snapshot. Secondly, we removed features due to some reasons of irrelevance (based on business understanding), unavailable on new data, too many or too few unique values, correlation, and too many missing values.

Next, we cleaned the data while filtering out rows (loans). We kept only Defaulted and fully paid loans (matured loans). Around 80% of the loans are for 36 months. We decided to keep only loans for 36 months since they all matured, and the population's behavior can vary between the two loan terms.

We also dealt with categorical features, each with the method that suits it and fits them into a machine learning model. Finally, we performed a feature selection process to keep the knowledge-bearing variables only.

Model Selection

To better understand our data and investigate how the realized returns are distributed across both loan statuses (fully paid and default). We performed a thorough exploration of the data. Since the current return that "GreatYields" has is 2%, we decided to use this return as our baseline for comparison. Throughout the exploratory analysis, we noticed that 99% of fully paid loans have a realized return of more than 2% and an average return of 6%. In contrast, only 5% of the defaulted loans yield more than 2% and an average return of -12% (Appendix 1.1+1.2). Following this discovery, we decided to use a classification model - Gradient boosting, which classifies loans as fully paid or default. Thus we can choose to invest in loans that will be predicted as least likely to default.

Realized Return Calculation

Calculating the realized return is complicated due to two main factors. First, we need to consider the return on loans paid ahead of time. Second, we should treat defaulted loans that only part of them has been paid properly.

While calculating the realized return, we made two main assumptions:

- All payments are equal, each of p/m, are made in each of the m periods
- Each payment is immediately reinvested at a rate of 2%, the company's current average yearly yield.

The formula:

$$\frac{12}{T}\cdot\frac{1}{f}\left\{\left[\frac{p}{m}\cdot\left(\frac{1-(1+i)^{m}}{1-(1+i)}\right)\right]\cdot(1+i)^{T-m}-f\right\}$$

f - the full amount invested in a loan
p - full amount recovered from the loan (payments + recoveries)
t - nominal duration of the loan in months
m - actual duration of the loan – from first to last payments
i - interest reinvested rate

We will sum up the geometric series of loan repayments, subtract the original loan, and divide it by the original loan amount. To get the expected annual return, we will multiply by 12/t. Using this formula and under these assumptions, we can answer what the expected return will be and how the return is distributed for each grade (Appendix 2).

Performance Compared To Baseline

Each loan in the data we received from the "Soft Lending" platform is assigned with a grade (A-G). Where A represents the "safest" loans with an average return of 3.2% and standard deviation of 4.5%, and G represents the "riskiest loans" with an average return of -0.2% and standard deviation of 15.3% (i.e., a higher chance for extremely high and low return).

Next, we wanted to answer Walter's 2nd and 3rd questions, which were related to the amount of information our data contains and, according to that, the ability to outperform simply using a baseline model (Appendix 3.1) and yield better returns. We compared our model's performance to two baselines.

Firstly we needed to produce a higher average return than the current 2% that the company generates - a thing that we accomplished for sure since investing in all loans yields an average return of 2.8%. However, we should mention that the company's current investment standard deviation(risk) is unknown, and therefore we can't thoroughly compare the returns. Secondly, we compared our model to simply choosing loans by grade. The grade and the interest rate assigned for each loan are correlated. We researched how grade and interest rates are determined - Lending Club's interest rates consider credit risk and market conditions.

The final interest rate grade for each loan is the result of the following equation:

“Lending Club’s Base Rate (5.05%)+ Adjustment for Risk & Volatility [1]”

The Adjustment for Risk & Volatility is designed to cover expected losses and provide higher risk-adjusted returns for each loan grade increment from A to G.We assume that the loans graded A have the lowest probability to default, and this probability increases throughout the grades. We compare the average realized return and its standard deviation for every fraction of loans invested in whereas the fraction increases, we add loans that are more likely to default. For the grade model, we lower the 'threshold' by accepting loans with a lower grade, and for our model, we lower the threshold of the scores assigned for a loan to default. Our model achieves a higher average return and lowers standard deviation for every fraction than the grades model, however not by much (Appendix 3.2) .

Model analytic performances

Our key metric for evaluating the model results is "Recall" - the ratio of correct defaulted loans classified correctly out of all defaulted loans. Defaulted loans classified as fully-paid generate more losses (average realized return of -12%) than fully paid loans classified as default which prevents gain (average realized return of 6%). Therefore we want to maximize recall. We can also control recall by tuning the acceptable score threshold for a loan to default (Appendix 4.1). As we lower the threshold, we can prevent defaulted loans from being classified as fully paid, but we also lose fully paid loans in the process. The company's investment strategy should determine the threshold. Besides recall, an analytical metric, we measure our model's performance by the average realized return for every fraction of loans invested in (Appendix 4.2). We were hired to maximize profit, and therefore this is our primary goal (Appendix 4.3).

Investment Strategies

The 4th question Walter asked us was related to the average returns his company can get investing in peer-to-peer loans using our model. The answer to this question depends on the investment approach the company decides to choose and the budget they allocate to this investment. Following the decision, the threshold of the model is set accordingly:

- Choose the fraction of loans they want to invest in and take the X% of loans assigned by our model with the lowest probability to default (i.e., "best loans")
- Choose their desired risk and invest in the fraction of best loans respectively, as we mentioned above - investing in more loans increases the risk.

The fewer loans we invest in - the higher returns with lower standard deviation we get. This makes sense since as the fewer loans there are, there is a lower probability of default. Investing in more loans leads us to invest in more defaulted loans (average return of -12%), lowering our average return and increasing the risk.

[1] https://www.lendingclub.com/foliofn/rateDetail.action

Risk

As mentioned above, our data contains two snapshots of the loans from 2018 and 2019, and we don't have data on the loans throughout their lifespan. Therefore, we can't measure the risk entitled in such an investment by volatility as Walter initially asked in his 5th question. However, we can measure the risk of investment based on our model by calculating the standard deviation of the expected realized return of all the loans our model predicts as fully paid. This risk will vary by the fraction of loans (or absolute amount of money) the company will decide to invest in. As we tune our model to accept more loans by lowering the threshold for predicting a loan as fully paid, we invest in a higher fraction of the available loans, the amount of money invested rises and the absolute return increases. However, this is a tradeoff with the entitled risk since as we invest in more loans, the standard deviation of the expected realized return increases as well, i.e., the investment becomes riskier (Appendix 5.1). It is worth mentioning that while our model yields better results than selecting loans only based on their grade, the investment is still risky. The standard deviation (risk) is higher than the average expected return for almost every fraction of loans invested in. That means that there is a possibility of getting negative returns (Appendix 5.2).

Conclusion

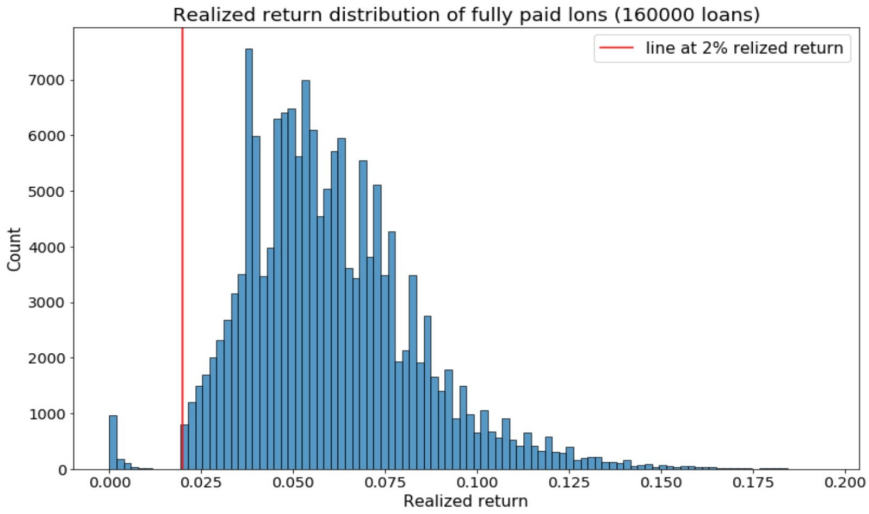
In conclusion, after performing analytical work on the data and using the model we created, we achieved better results than our two baselines. Hence, higher than the 2% realized return and slightly better than selecting loans only based on their grade. Still, as we noted, the investment in the peer-to-peer loans platform includes additional considerations such as risk, budget, and investment strategy that can affect whether the investment will be worthwhile or not.

Therefore "GreatYields" will have to consider other considerations, especially the risk, along with the average return we forecasted.

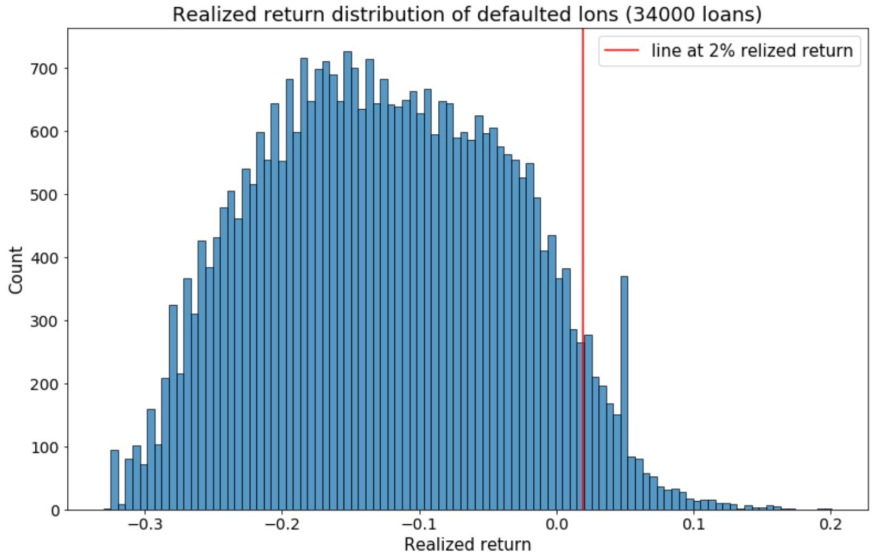
Appendices

Appendix 1 -Model selection

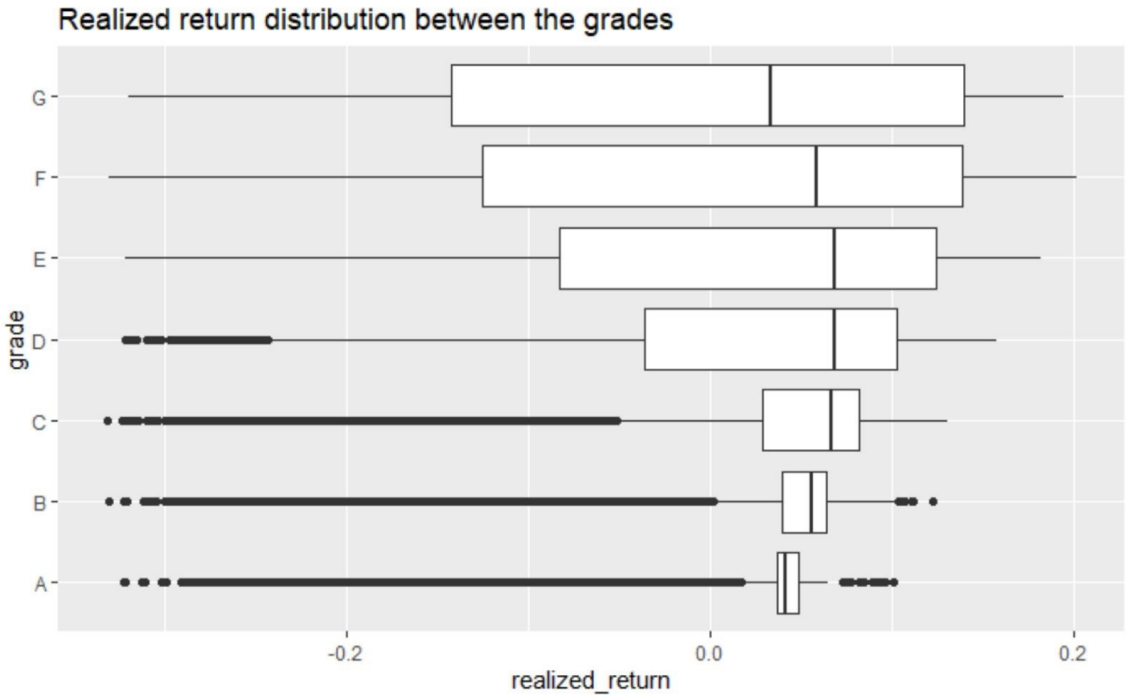
Appendix 1.1 - Fully paid average realized return distributions



Appendix 1.2 - Default average realized return distributions



Appendix 2 -Realized Return distribution between the grades



Appendix 3 - Performance Compared To Baseline

Appendix 3.1 - grade baseline

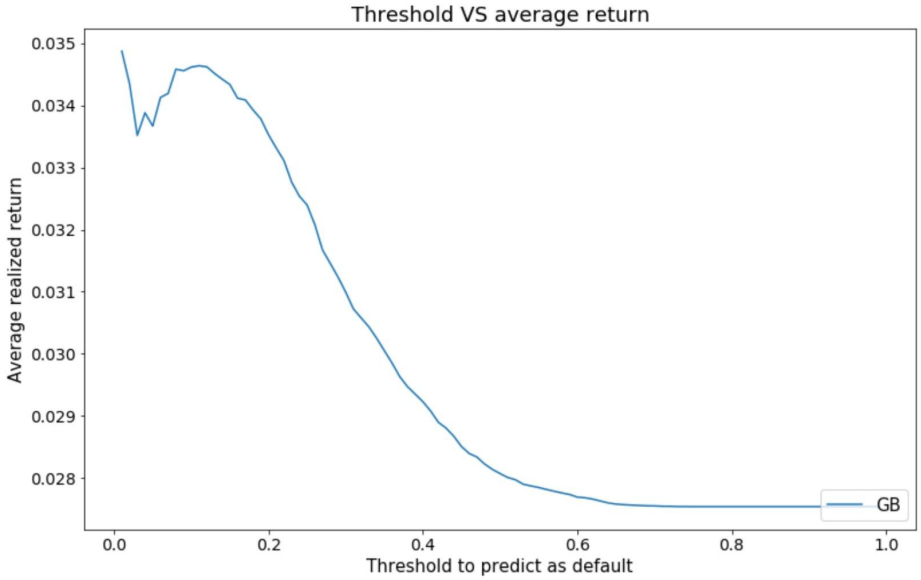
grade/ Realized Return	A	B	C	D	E	F	G
mean	0.0323	0.0305	0.0257	0.0194	0.0162	0.0046	-0.002
median	0.0409	0.0558	0.0658	0.0679	0.0666	0.0514	0.04
std	0.0451	0.0708	0.0925	0.113	0.1296	0.144	0.1536
min	-0.3227	-0.3303	-0.3235	-0.3212	-0.321	-0.3199	-0.3197
max	0.0957	0.1108	0.1294	0.1575	0.1813	0.201	0.1942
count	41611	67470	54659	22288	6352	1434	333
loans_frac	0.2143	0.3475	0.2815	0.1148	0.0327	0.0074	0.0017

Appendix 3.2 - Our model VS grade model

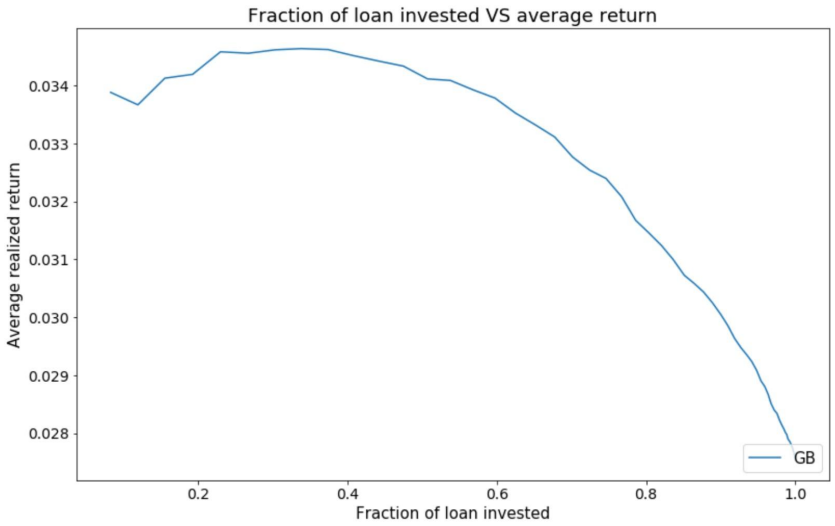
	Loan fraction	Grades mode Mean return	GB model Mean return	Grades model SD return	GB model SD return
A	0.2143	0.03230	0.03540	0.04510	0.04291
A-B	0.5618	0.03080	0.03395	0.06290	0.06093
A-C	0.8433	0.02880	0.03087	0.07470	0.07355
A-D	0.9581	0.02760	0.02892	0.08040	0.07996
A-E	0.9908	0.02720	0.02798	0.08260	0.08121
A-F	0.9982	0.02710	0.02765	0.08330	0.08235
A-G	1	0.02700	0.02700	0.08350	0.08350

Appendix 4 -Model analytic performances

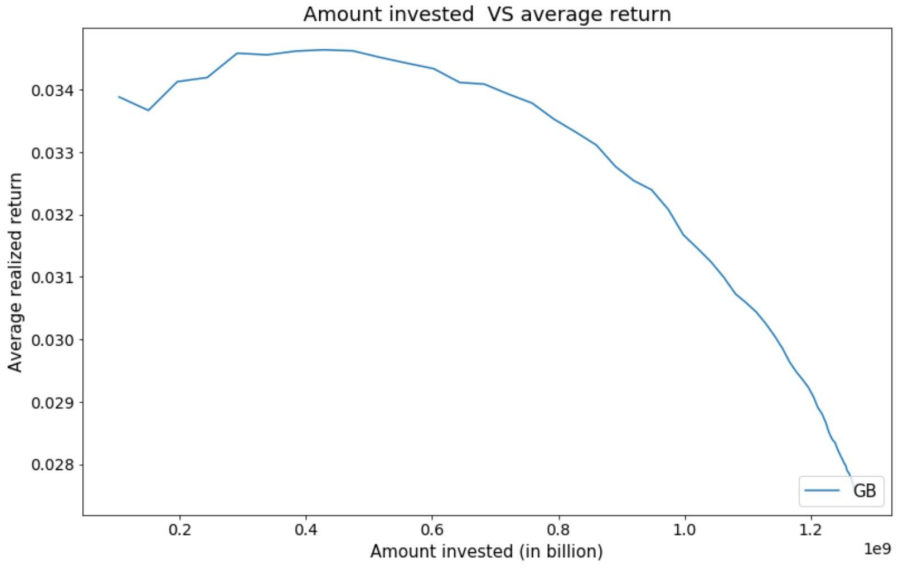
Appendix 4.1 - threshold to predict loans as defaulted VS average realized return



Appendix 4.2 - Average realized return as a function of the fraction of loans invested in

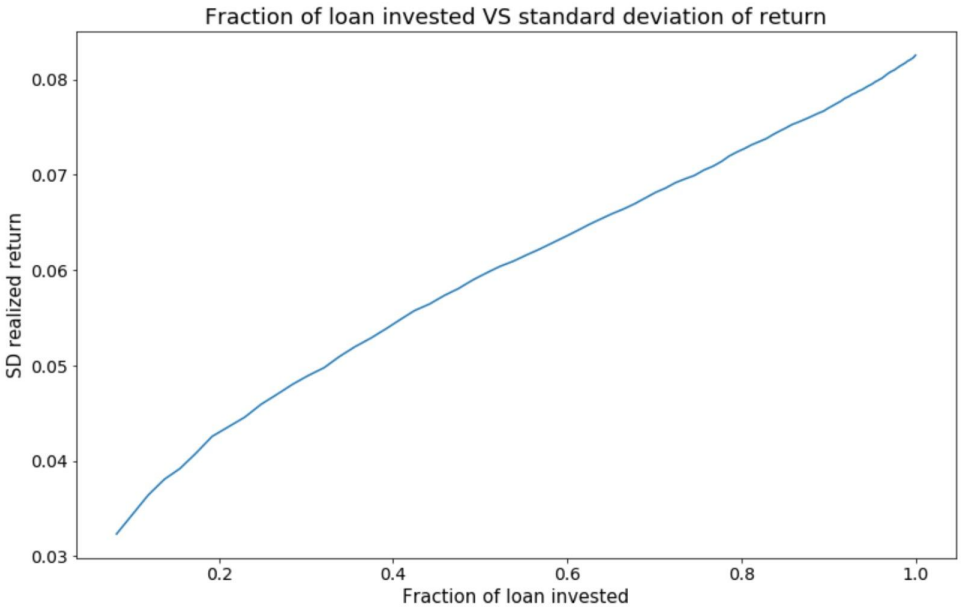


Appendix 4.3 - Average realized return as a function of the amount invested in



Appendix 5 - Risk

Appendix 5.1 - SD of realized return as a function of loan fraction invested in



Appendix 5.2- The tradeoff between investing in a higher loan fraction and the risk entitled in such an investment