Imperial College London

Big Data in Finance: Final Assignment

10 % Annual ROI in the Peer to Peer Lending Market in 2018?

–

A machine learning-driven investment strategy

Author: Cyrill Studer

April 30, 2019

Abstract

In nowadays low-interest rate environment, alternative investments gain in importance for investors in order to uphold profit expectations. A promising multibillion-dollar market which experienced average annual growth rates of close to 50% is the peer to peer lending market. The goal of this paper is to evaluate data-driven and machine learning based investment strategies in the peer-to-peer lending market. The modeling of LendingClub data for Q3 and Q4 2017, one of the largest peer to peer lending platform in the US, shows that up to 3 % premium returns compared random loan selection and 7-10% absolute annual returns on investment for investment volumes between 50-300 million are visible.

Table of Contents

1 Introduction

The peer to peer lending market is a mostly uncollateralised debt market between borrowers and lenders where traditional intermediaries such as banks or other financial institutions are cut-off. Especially in developing markets, in the US and in the UK, the market for peer to peer lending is rapidly growing. With its high growth rate and the current low-interest rate environment, the peer to peer lending market becomes increasingly interesting for investors with regard to returns and diversification. The purpose of this paper is to make an explorative attempt to develop a data-driven and machine learning based investment strategy for institutional investors or ultra high net worth individuals based on LendingClub's Q3 and Q4 2017 data, one of the largest peer to peer lending platforms.

LendingClub was founded in 2006 and as of the beginning of 2019, the company has issued loans worth over $ 44 billions. The company requires each potential borrower to provide over 50 different data points in order to approve a potential borrower and classify them into 7 different risk categories and many more subcategories. Accordingly to the assigned category, the loans will be discreetly priced with an annual interest rate between 6.46%-29.00%. Lenders can choose whether to invest in individual loans or in categories of loans. LendingClub takes 1% in service fee of each borrower payment. (LendingClub, 2019)

The paper is being organized the following. First of all, the current state of data-driven investment strategies in the peer to peer lending market will be evaluated by a literature review. Next, the methodology of this paper's investment methodology will be outlined and a closer look at the data will be paid. Lastly, the resulting models will be described before a discussion will summarise the results and highlight points regarding implementation and limitations.

2 Literature Review

Already in 2012, a Forbes journalist pointed in one of his articles to the attractiveness of the peer to peer lending market with regard to realising close to two digits annual return (Barth, 2012). Following that, also in academia researchers increasingly paid attention to the topic.

On Kaggle, a data scientist community, people mainly use peer to peer lending data sets, popularly the data provided by the LendingClub, to model and increase the accuracy of default forecasting by using classification algorithm or to explore patterns via explanatory data analysis and visualisation. The development of investment strategies is rarer (Kaggle, 2019). Three

recent examples of developing investment strategies for peer to peer lending are the studies of Feiss et al. (2016, Cohen et al. (2018) and Guo et al. (2016).

Feiss et al. (2016), the work of five Stanford students, used the LendingClub data from 2007-2011 to develop a default probability-based investment strategy. By means of the linear discriminate analysis method, the student modeled the no-default probability of loans and developed an investment strategy accordingly. With this approach, the students managed to exceed the return of a random selection of loans within different credit rating buckets by 2-3%. Additionally, based on the gained findings from their investment strategy, Feiss et al. (2016) introduced a continuous pricing model that would allow LendingClub to price its loans more accurately.

Guo et al. (2016) developed a data-driven, real-time investment decision assessment tool based on default likelihood distances to past loans and portfolio optimization techniques. The researcher relied on data from the peer to peer lending platforms Prosper and LendingClub and achieved superior returns for small sized investors. Cohen et al. (2018), made use of the LendingClub data 2012-2014 to develop a similar trading strategy based on machine learning and portfolio optimization techniques which achieved 3% premium returns compared to a random loan selection.

This paper will use very recent data and range of different machine learning techniques to see whether these return numbers are still visible to achieve in nowadays peer to peer lending environment. Moreover, the aim is to develop strategies relying on default probability modeling for a largely diversified loan portfolio in order to minimalise volatility and unsystematic risk. Therefore, the strategies will be particularly suitable for institutional investors or ultra high net worth individual who want to diversify their portfolio while not compromising for returns.

3 Exploratory Data Analysis

The first step to define a promising investment strategy is to identify useful pattern in the data. The LendingClub data Q3 2017 will be used as the training/validation data set while the Q4 2017 data will be used as the out-of-sample test data. The cleaned training data set has 60,072 observations with information whether the customer defaulted or not defaulted and 48 additional customer/loan specific features. The data set is imbalanced with approximately 85% non-default observation. However, as the purpose of this study is to correctly predict the

majority class, the training data will not be over- or undersampled. Table 1 illustrates a summary statistic table subdivided by the seven assigned loan grades.

Table 1: Summary Statistics Training Data

	Ø Interest Rate (%)	Ø Default Rate (%)	Ø Loan Amount ($)	# Observations
Grade 0	7.02	4.45	13,772.84	9,958
Grade 1	10.55	9.75	13,694.26	17,316
Grade 2	14.36	17.07	15,482.53	21,065
Grade 3	19.07	26.35	15,991.89	7,791
Grade 4	25.12	33.85	16,910.96	3,001
Grade 5	30.26	40.55	18,795.11	1,041
Grade 6	30.89	48.55	21,340.87	900
Total	13.72	15.83	14,980.57	61,072

There are several insightful observations from the summary statistics in Table 1. First of all, it appears that approximately 80% of all loans fall within the top three graded loan buckets. Secondly, the average loan amount seems to increase in grade, which potentially leads paired with the fewer observations and the higher default rate to a higher variance in return of investment when investing in these loan buckets. Lastly, there seems to be, as expected, a direct relationship between the interest rates and the default rates. Therefore, as a next step, it makes sense to investigate this relationship further. Illustration 1 shows the average default rate (light grey bar) and the average default rate (light black bar) combined with the distribution of interest rates in form of a boxplot per grade on the left-hand side. On the right-hand side, one can find the development of the average interest rate (light black) and the average default rate (light grey) across seven loan buckets.

Illustration 1: Relationship Default Rates and Interest Rates

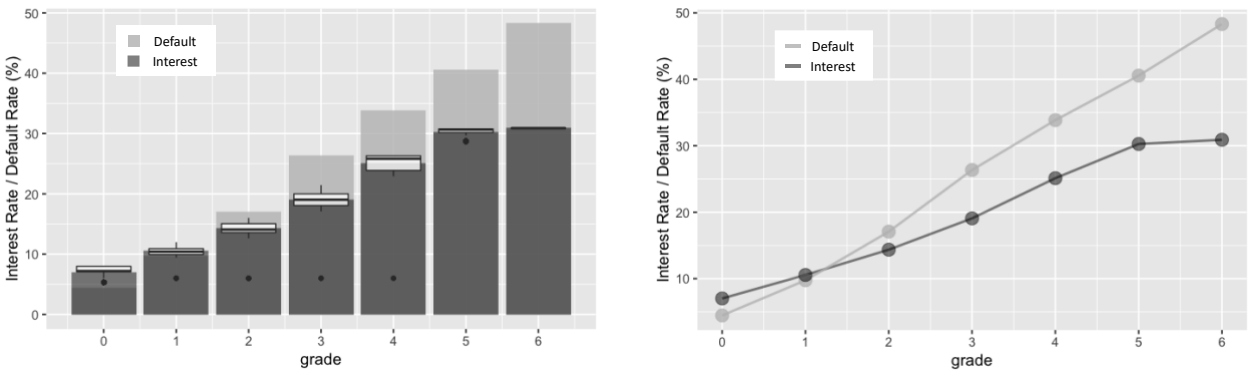
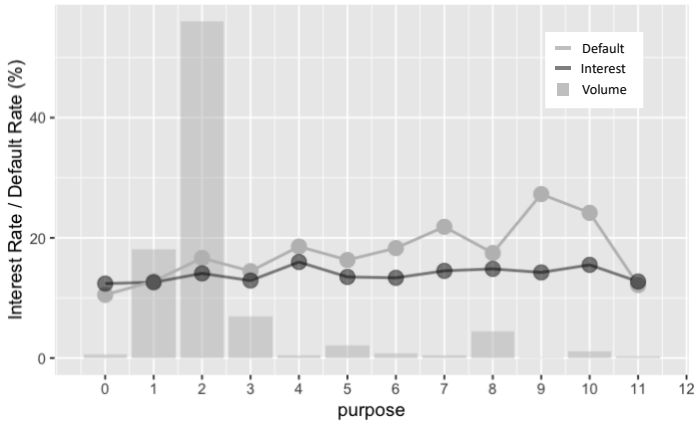


Illustration 1 shows that the spread of interest rates within each grade category is small. Moreover, the average default rate across the different loan buckets follows an approximately linear and continuous trend. The combination of categorical pricing and the apparently continuous and linear default probability of loans raises the question of whether all loans are priced efficiently. In case this hypothesis proves to be true, there would be different risk levels which are equally priced. Consequently, there would be potential for arbitrary returns and the foundation of a promising investment strategy.

As outlined in Illustration 2, another major observation is that there seem to be significant differences in average default rates (light grey) and the corresponding interest rate (light black) among the differently grouped purposes of a loan.

Illustration 2: Comparison of Loans by Purpose



However, the distribution of the loan volume among the different purposes seems to be strongly concentrated on purpose 1, 2 and 3. For these purposes the gap between average default rate and average interest rate is marginal.

4 Methodology

Based on the gained insights from the exploratory data analysis, the next step will be to design a methodology in order to develop a profit maximalising investment strategy.

First of all, a preselection of considered loans for the investment strategy will be made. Even though the 36 months duration loans and the 60 months duration loans seem to be approximately equally well priced with regard to duration, average default rate and average interest rate, this investment strategy will only focus on 36 months loans. The 36 months loans make up approximately 2/3 of all loans which are equal to 43,209 observations. The reason for

focusing only on shorter-period term loans is the reduced exposure to changes in macroeconomic conditions and to uncertainty. Due to the unequal distribution among the investment purposes, this feature will not be considered in order to predefine the potential loan investment pool.

As a second step, based on the insights from the average interest rate and average default rate relationship and the potential arbitrary returns, different models shall be developed in order to predict the probability of a loan not defaulting. The applied modeling technique will be reduced models which perform well for default forecasting purposes. The four probability-prediction-suitable machine learning techniques linear discriminate analysis (LDA), logistical regression, gradient boosting and random forest were chosen to be tested for this purpose. Each technique was applied with the full set of 48 numerical predicting features and with a lasso-reduced subset of eight predicting features. Furthermore, all predicted probabilities were the result of a ten-fold cross-validation. The selection of a subset of predictors was done by using the first 7,500 data point of the training data set while the other 35,709 data points were used for the cross-validation probability prediction.

As a third step, a function was developed in order to choose the profit-maximising machine learning prediction technique with the corresponding threshold of above which non-default prediction rate to invest into a loan. Firstly, the function calculates the investment volume, the profit, the annual return on investment and the accuracy for each cutoff threshold between 1and 100 %. Thereby, the investment volume is defined as the sum of the loan amounts with a negative default prediction. The accuracy is defined as the number of not defaulting loans out of the number of invested loans. The profit term underlies certain assumptions. The profit is the result of the realised revenue through the collected interest rates minus the loss caused by defaulted loans. To simplify the calculation, for the revenue side, it is assumed that the interest rates are paid annually for loans which do not default. For the loss side, it is assumed that defaulting loans have to be completely charged off and do not pay any interest. This is a generalisation and oversimplification which has to be made due to the lack of accurate data and missing data regarding default amount and default point in time. However, comparing the average profits and return of investment of the model with the by the LendingClub published numbers shows that this simplification affects the accuracy of the model only marginally in terms of the absolute values and doesn’t affect the results in terms of generated return premium compared to the random selection of loans (LendingClub, 2019b). The annual return on investment is then calculated based on the before made definition of profit and investment amount:

$$Annual\ Return\ on\ Investment = \sqrt[3]{1 + \frac{Profit}{Investment\ Amount}}$$

As a last step, the function loops through the realised profits of the different cutoff thresholds and returns the profit maximising cutoff threshold with the corresponding profit, investment amount and annual return on investment. Based on this information, the best the performing machine learning prediction technique with the corresponding threshold will be chosen.

Obviously, there is a trade-off between the investment amount, the profit and the return of investment. The return of investment is expected to increase with the threshold while the investment amount is expected to decrease with an increasing threshold. Consequently, it is expected that the profit will increase with the threshold as long as the effect of choosing more profitable loans dominates the effect of losing volume before it will start to decrease when the latter effect takes over.

5 Models

As a next step, a short reasoning about the suitability of each of the four in the previous chapter mentioned machine learning techniques for the particular task at hand will be followed by the depiction of results of the above-described methodology. In order to be able to benchmark the obtained results, an investment in a random selection of loans on the training data will on average lead to an ROI of 6.65% with an investment amount of $ 261m and a profit of $ 56.39m.

The logistic regression is a linear model specifically designed for probability prediction and classification tasks. The model with the full set of predicting features led to a profit of slightly above $ 56m and an annual ROI of 6.86%. The model with the reduced set of independent variables led to a profit of slightly above $ 54.86 m and an annual ROI of 6.55%. It is important to see that the model with 45 independent variables is likely to suffer from multicollinearity and a lack of normally distributed variable which could affect the stability of the results.[1]

The linear discriminate analysis is similar to the logistic regression a linear, multivariate function that allocates observations to different groups or assigns a probability of belonging. The linear discriminate model with all 45 independent features led to an annual ROI of 7.12 % and $ 57.21m profit while the model with 8 predictors led to 6.55% annual ROI and $ 54.86m profit.

[1] However, scaling the independent variables did not change the obtained results.

Gradient boosting step-wise builds a decision tree to minimalise a defined loss function in order to correctly classify/assign probabilities to different classes. It can be particularly suitable for unbalanced data sets. The gradient boosting algorithm with the full set of predictors led to an annual ROI of 7.33% and $ 58.7m in profit while the model with the restricted set of predictors achieved an annual ROI of 6.75% and a $ 56.56m profit.

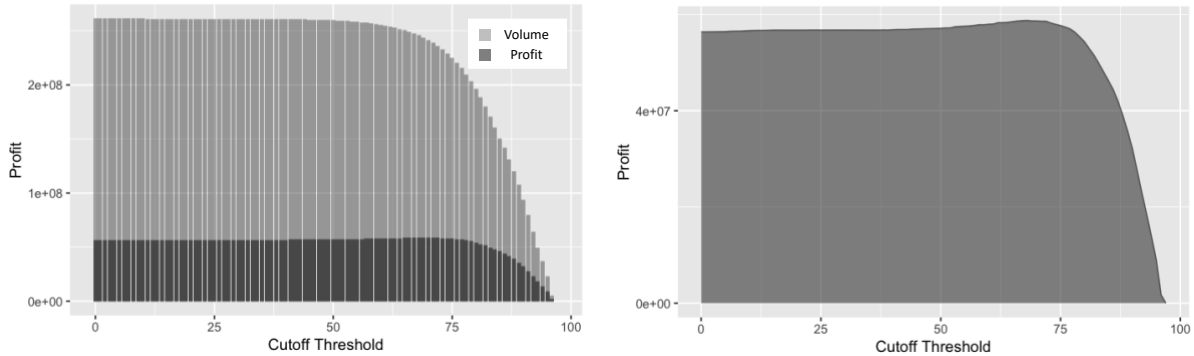
Random forest is a machine learning technique suitable for probability prediction that reduces the overfitting tendency of decision trees by building a multitude of trees with different subsets of predictors. With a maximal tree depth of four, the random forest led to an annual ROI of 6.89% and $56.46m profit with the full set of predictors and an annual ROI of 6.52% and a profit of $ 54.87 m with the reduced set of predictors. Table 2 below summarises the results:

Table 2: Result Training Data

	Annual ROI full model (%)	Profit full model ($m)	Profit max. cutoff value full model (%)	Annual ROI reduced model (%)	Profit reduced model ($m)	Profit max. cutoff value reduced model (%)
Logistic Regression	6.86	56.00	65	6.55	54.86	76
LDA	7.12	57.21	70	6.55	54.86	59
Gradient Boosting	7.33	58.77	68	6.75	56.56	67
Random Forest	6.89	56.46	73	6.52	54.87	76

The results show that gradient boosting with the full set of predictors and a cutoff threshold of 68% no-default probability seems to be the most promising technique in order to maximalise absolute profits when investing into the peer to peer loans of LendingClub's platform. As a next step, the goal is to investigate further how the relationship between investment amount, profit and annual return on investment works for the gradient boosting algorithm across different thresholds in order to develop a set of investment strategies.

Illustration 3: Choice of Ideal Cutoff Threshold for Overall Profit Maximalisation



The gradient boosting technique with a cutoff threshold of 68% leads to a marginal boost in absolute profits of 4.22% compared to randomly investing (threshold=0) into the full market ($ 56.39m versus $ 58.77m) for the training data. The same strategy lowered the required investment capital by 5.75% ($ 261m to $246m). Overall, the annual return on investment could be increased by 10.06% (6.66% versus 7.33%). This investment strategy of using gradient boosting with a cutoff threshold of 68% no-default probability is particularly interesting for investors who want to invest large volumes and want to be well diversified. This strategy will be tested under the naming strategy 1 on the test data. Next, the focus will lie on achieving a high annual return on investment without neglecting the investment volume. Illustration 4 highlights the relationship between annual return on investment and investment volume.

Illustration 4: Tradeoff of the Annual ROI and the Investment Volume with regard to the Cutoff Threshold

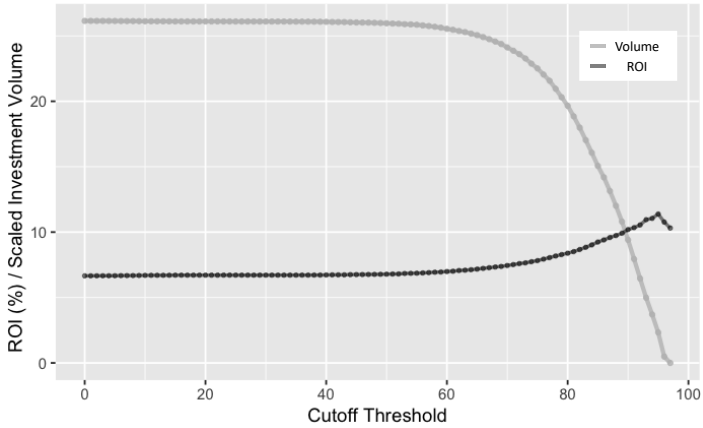


Illustration 4 shows that a threshold cutoff around 90 leads to an annual return on investment of approximately 10% at a still sizable investment volume. With regard to concrete numbers on the training data set, the gradient boosting technique with a cutoff threshold of 90 leads to a decrease in absolute profit compared to random investing of 42.95% ($ 56.39 m versus $ 32.17 m) but also to a decrease of the required capital of 63.98% ($ 261 m to $ 94 m). Overall, the selection of more profitable loans leads to an increase in the annual return on investment of 53% (6.66% versus 10.19%). Gradient boosting with a threshold of 90% will be tested as strategy 2 on the test data set.

Additionally, there should also be a strategy which aims to maximalise the annual return of investment without paying any attention towards investment volume restrictions. In the training data set this is the case at the cutoff threshold of 95% non-default probability. With this threshold, the required capital decrease by 91.06% ($ 261 m to $ 23.33 m) and the profit decreases by 84.02% ($ 56.39 m versus $ 9.01 m) while the annual return of investment

increases by 70.87% (11.38% versus 6.66%). Gradient boosting with a threshold cutoff of 95% will be particularly suitable for investors with a relatively small investment volume who aim for the highest ROI while assigning less relevance to very high diversification. This investment strategy will be referred to as strategy 3.

Lastly, there should also be made an attempt to specify the cutoff threshold for each grade to see whether this leads to superior alpha generation.

6 Results

After developing the investment strategies, the performance of the different strategies needs to be tested on previously untouched test data. Therefore, the gradient boosting algorithm was trained with Q3 2017 data to forecast default probabilities of Q4 2017. It is important to see that for the test data set all loans with a duration of 36 months were taken into account which was not the case in the training data set as some of the data was used for feature selection. Therefore, the absolute values investment amount and profit are not directly comparable to the training data results. However, the annual ROI is. The following results were achieved:

Table 3: Results Test Data

	Cutoff Treshold (%)	Investment Amount ($m)	Profit ($m)	Annual ROI (%)	Default Rate (%)
No Startegy	0	351.11	85.83	7.48	10.02
Strategy 1	68	339.59	65	7.86	9.01
Strategy 2	90	166.61	53.12	9.56	4.40
Strategy 3	95	62.22	21.66	10.36	2.42

The above results show that strategy 1 with its aim of absolute profit maximisation could not be confirmed for the out-of-sample test data. This is not highly surprising as the achieved premium was marginal in the training phase. Strategy 2 and strategy 3, on the other hand, could confirm their respective expected annual returns on investment. Also the investment amount for strategy 2 & strategy 3 could as expected with the larger amount of data be increased.

The idea of specifying a cutoff threshold specifically for each grade did not produce any reasonable results. A potential reason for the failure of this approach could be the too small size of the data set which does not allow to model the default prediction accurately enough for each grade.

7 Conclusion

In this last section, the achieved results will shortly be reflected and some considerations regarding the limitations and implementation of the resulting strategies will be raised.

The present paper showed that machine learning-driven investment strategies in the peer to peer lending market are able to achieve annual return rates of 7-10 % for investment volumes between $ 50-300m in nowadays market environment. Therefore, the peer to peer lending market is a valuable alternative in order to diversify institutional investor's or ultra high net worth individual's portfolio.

When it comes to the implementation of the developed investment strategies, it's important to see that in order to be well diversified and avoid exposure to high volatility by defaulting borrower's, investment volumes above $ 50m are advisable. Additionally, it is important to see that the entirety of the model only relies on very recent data. This is on the one hand side useful, due to the refection of similar market conditions, can however be fatal in case of a significant change in market conditions such as a radical change in interest rates or an economic crisis. Running different stress test simulations before implementing the proposed investment strategy is advisable.

Moreover, studies have already shown that hooking up to LendingClub's API and implement an automated loan buying strategy does work (Feiss et al., 2016). However, in case this procedure is or becomes highly popular, execution speed will become critical and would have to be investigated further. Also the transferability of the strategy to other platforms in order to achieve the required volume would need to be investigated further. Moreover, it is important to see that at least in the US, the realised returns on LendingClub's platform are subject to personal income taxes rather than investment taxes which could affect the attractiveness of the strategy.

With regard to technical limitations, the study can legitimately be accused of being distorted by relying on simplified assumptions as outlined in chapter 4. Although not one to one comparable, the resulting absolute numbers regarding realised annual ROIs are potentially slightly (0-1%) to high compared to the numbers published by the LendingClub. However, the resulting increase in alpha of 2-3% compared to the random selection are relative measure and therefore valid. Moreover, the numbers are in accordance with previous findings.

References

Barth, C. (2012). Looking for 10% yields? Go online for peer to peer lending. Retrieved April 23, 2019 from https://www.forbes.com/sites/chrisbarth/2012/06/06/ looking-for-10-yields-go-online-for-peer-to-peer-lending/#d8a4f0e3a8fa

Cohen MC, Guetta CD, Jiao K, Provost F (2018) Data-driven investment strategies for peer-to-peer lending: a case study for teaching data science. *Big Data* 6:3, 191–213, doi: 10.1089/big.2018.0092.

Feiss, A., Metha, V., Morris, S., Solitario, J. & Van de Graff, C. (2016). P2P loan selection. Retrieved April 23, 2019 from http://stanford.edu/class/msande448/2016/final/ group4.pdf

Guo, Y., Zhou, W., Luo, C., Liu, C., & Xiong, H. (2016). Instance-based credit risk assessment for investment decisions in P2P lending. *European Journal of Operational Research*, *249*(2), 417-426, doi: https://doi.org/10.1016/j.ejor.2015.05.050

Kaggle (2019). Lending Club Loan Data. Retrieved April 23, 2019 from https://www.kaggle.com/wendykan/lending-club-loan-data/kernels

LendingClub(2019a). How it works. Retrieved April 23, 2019 from https://www.lendingclub.com/investing/alternative-assets/how-it-works

LendingClub(2019b). LendingClub statistics Retrieved April 25, 2019 from https://www.lendingclub.com/info/demand-and-credit-profile.action

Appendix

1) EDA&Visualisation.html
2) ML_Models.html