GREAT YIELDS
a boutique investment house

Roy Madpis - 319091526
Rawaa Makhoul - 316114370
Alexandra Fatyeyeva - 336540950

Peer to peer lending project – final working paper

Working Paper № 6 – Team A

Executive Summary

Walter, the CIO of Great Yields, considered adding peer lending for the portfolios of some of the company's clients due to the low returns produced from investment in stocks and bonds (of approximately 2%, annualized). We, the Analytics team, were summoned to advise the company on whether to invest in peer-to-peer lending based on insights gained from SoftLending's data (a relatively new US peer-lending platform). In the first meeting, Walter raised five main questions in which we address in the following paper.

The data and the process of building the predicting model.

The data contained 434,407 loans from 2016 with 151 features. It was spread into 8 excel files: 4 were a snapshot of the loans in 2018, and the remaining 4 were snapshots of those loans in 2019. Note that as we weren't given a snapshot of the loans in their initial status in 2016, we had to locate the "leakage" features - meaning features that their given value in 2018 / 2019 wasn't the same in 2016 so that they won't be used in our prediction model. In addition, there were many features with a significant amount of missing values that can't be used. Finally, while optimizing the model, we removed few features that decreased the model's performance[1]. All in all, we started with 151 features and ended up removing 103 of them. At the same time, we created **7** new features (in the feature engineering process) and one new target variable ("realized_return_2%"). The target variable is a feature indicating for each loan whether its realized return was higher or lower than 2%.[2]
Since the company's investments in bonds and stocks yielded 2% annually, we recommend the company to invest in peer-to-peer lending only if it can generate higher returns. That was the reason behind choosing the 2% in our targeted variable.

Through the first four weeks, we explored and cleaned the data in an elaborate EDA process[3].
In the second step, we figured that we couldn't use loans that didn't mature (we don't have their final outcome), i.e., loans with:

a. "Term" higher than three years.
b. A status that is not "Charged Off" or "Fully Paid".

In addition, in further stages, we cleaned missing values and outliers, leaving us with **237,015 loans** and **55 features**.

---

[1] The two main metrics that we used to evaluate the model's performance and fine-tune it were: accuracy and NPV (Negative Predictive Value).

[2] Note that the realized return was calculated according to the formula: $Realized\ Return = \frac{p-f}{f} \times \frac{12}{t}$

p = full amount recovered from the loan = column: total_pymnt_inv (total payment that the investor got)
f = full amount invested in a loan = column: funded_amnt_inv
t = nominal duration of the loan in months = column: term (= 36 months)

[3] The full EDA and cleaning process are presented in Working Paper #2

In the following four weeks, using the filtered loans, we built a model and optimized it: we split those loans into a training set (70% of those loans) and a test set (30%) and trained a classification XGBoost model. The model aims to:

a. Predict whether a given loan will have a realized return that is higher or lower than 2%.
b. Predict the probability ("Score") for a loan to have a realized return that is higher than 2%.

Then, based on a desirable probability[4] - defined as the "**threshold**," the model filters the given loans and outputs the "suggested portfolio," i.e., a list of loans it suggests to invest in. Thus, the number of loans in our "suggested portfolio" is directly influenced by both the number of loans given to the model as well as a chosen threshold.

Note the tradeoff: the higher the threshold, the higher the weighted average realized return[5], however, we get a smaller amount of loans. Meaning, if you chose a very high threshold (0.85 and above), you may receive a narrow "suggested portfolio".

By building a profit curve function and observing the influence of different thresholds on the tradeoff between the weighted average realized return and the number of loans, **we advise using a threshold of 0.75**, and as a result, getting a weighted realized return that is higher than 2% without losing too many loans (The "threshold table" is presented in the appendix in *Question 4 section*).

Answering the company's questions

Now we will answer the five questions that the company has raised at the outset of the project.

1. What are the expected realized returns for the different loan grades?
How are the returns distributed for each grade?

To answer this question, we built a table summarizing the number of loans, the weighted[5] average realized return, and the weighted standard deviation (of the realized return) an investor would get if he invested in all the loans of a particular grade.

Note: The table is calculated according to the 265,316 loans that were left after filtering loans with a term that is higher than three years and loan status that is not "Charged Off" or "Fully Paid," as well as several more cleanings that are elaborated in *Working Paper #2*.

- See **table 1** in the appendix -

We can see that the realized return distributes differently for the different grades. We can notice that higher grades (**Note**: A is "higher" than B) have both higher weighted average realized return as well as lower weighted standard deviation, meaning that if the company wishes to maximize the realized return and minimize the risk, it can invest in loans that have higher grades (for example - invest only in loans of grade A). However, note that for all the given grades, the weighted realized return is still lower than 2%, and therefore this doesn't meet the company's goal (generating an annual return higher than 2%).
The distribution graphs of the realized return for each grade are presented in the appendix: *Question 1 section*.

[4] Probability of the loans to be with realized return higher than 2%.

[5] The weight is the amount of money invested in the loan (from the column: funded_amnt_inv)

2. Is the available loan data informative, thus can help to select loans to invest in (i.e., can the data help choose loans better than random selection, or selection by simple criteria (e.g., loan grade)?

To determine whether the data is informative, we need to define a criteria/metric at first. Then, we conclude that the data is informative only if it can be used to build a model that can filter the given loans and return a suggested portfolio such that one or both of the following is achieved:

- The **weighted average realized return** of the suggested portfolio is **higher** than the weighted realized return gained by a baseline model[6] (higher return).
- The **weighted standard deviation** of the suggested portfolio is **lower** than the weighted standard deviation gained by a baseline model (lower risk).

The performance of our model on a test set (containing 71,105 loans):
Out of the 71,105 loans given to the model, only 28,970 received a probability[7] higher than the given threshold (0.75). Those 28,970 loans compose the "suggested portfolio." Table **2** presents the weighted realized return and the weighted standard deviation of that "suggested portfolio."

- See **table 2** in the appendix -

Table 3 compares the performance of our model (Table 2) with the performance of a baseline model[6](Table 1)

- See **table 3** in the appendix -

We can see that if the company had invested in all the loans from the suggested portfolio, it would have gained a weighted realized return of 2.34% and a weighted standard deviation of 6.305% (see the last row of tables 2 and 4). In other words, our model succeeded in generating higher returns with lower risk compared to a random model that uses only the grades for investment strategy, **which means that the given data is indeed informative**.

3. If the data is indeed informative, what increased performance can be expected, compared to a baseline of simply selecting loans based on their ratings (grades)?

To state the "**increased performance**" of our model compared to a baseline model, we can calculate the difference between the performance of these two models regarding the weighted realized return and the weighted standard deviation. We present two ways to compare the two models:

- **The basic approach:** Compare our model with the best performance of the baseline model - investing only in loans with grade **A**. In practice, we need to compare the metrics[8] gained by investing in all the loans in the suggested portfolio[9] with the matching metrics achieved by the best performance of the baseline model - investing only in loans with grade A.
- **The two models' approach:** We can measure the performance of our model as if we invest according to the grades: This means to use our model to get a suggested portfolio and then measure the metrics we would get if we invested only in loans of a particular grade (see for example *Table 2*). Then compare the metrics gained in each grade with the matching performance of the baseline model.

[6] The baseline model advises to invest only according to the grades.
[7] The probability of a loan to have a realized return higher than 2%.
[8] The metrics are: The weighted realized return and weighted standard deviation
[9] Our model's output.

Important Note: we can't use the outcome of a specific test set (as presented in *Table 2*) to determine the performance of our model. We instead use cross-validation and use the mean performance gained.

- See **table 4** in the appendix -

According to the **basic approach,** our model can generate a higher realized return, however, with a higher risk than the ones obtained by the baseline model.

- See **table 5** in the appendix -

According to the **two models' approach,** our model succeeded in generating a higher weighted realized return with lower risk **for every grade** (compared to the baseline approach of selecting loans only by the grade).

- See **table 6** in the appendix -

To conclude, according to both the basic and the two models approaches, our model **outperforms** the baseline model in terms of **realized return**. The increased performance is explained in tables 5 and 6. Regarding the **standard deviation**, our model **outperforms** the baseline model, i.e., can generate **lower** standard deviation, only according to the two models' approach. The increased performance can be observed in *Table 5*.

4. What "average" returns can GreatYields expect from investing in peer lending loans? Keep in mind that ultimately, the goal is to maximize returns (i.e., make as much money as possible).

As we presented in tables 2 and 4, if the company invests in all the loans in the suggested portfolio, it could generate a weighted realized return of **~ 2.34%.** Continuing the explanation of the tradeoff of our model (as we presented in the executive summary), one can change the threshold in order to increase the weighted realized return at the cost of lowering the amount of 'good'[10] loans (diminishing the suggested portfolio size). As was explained in the executive summary, we advise using a threshold of 0.75, thus getting a weighted realized return that is higher than 2% and at the same time not losing too many loans. See the appendix: "question 4 section" for the "threshold table" presenting the spoken tradeoff. Notice in the "threshold table" that with a threshold of 0.92, you get only one loan of $3600 with a realized return of 4.8%. if you invested in that loan you could get a profit of $173 on the investment. With a lower threshold of 0.88, you invest in 171 loans $3,188,250, with a weighted realized return of 3.5%, you will get a profit of over $100,000. Therefore, maximizing profits directly connects with the amount of money the company has and wants to invest.

5. What is the risk level entailed in such investment (as measured by the volatility)?

We decided to measure the risk level with the **standard deviation of the realized return** in the suggested portfolio[11] as we couldn't calculate the volatility. If the company invests in all the loans in the suggested portfolio, it gets a weighted standard deviation of **~ 6.24%** (see Table 4 - column: mean of the standard deviation of all folds - the last row). Note that if the company invests in all the given loans, without any criteria, the realized return has a weighted standard deviation of **~ 8.41%** (see the last row in table 1 - column weighted SD). Note that although we succeeded in lowering the weighted

[10] Loans predicted to have a realized return that is higher than 2%.

[11] You can see that risk level in *Table 2* - column weighted SD - and in *Table 4* - column Standard deviation all folds.

standard deviation (the risk level), it's still relatively high and thus should be considered carefully when using our model, especially with a low amount of loans.

Conclusion, Suggestions, and Limitations

In this paper, we presented the work we have done and answered the questions that were raised at the outset of the project. We succeeded in building a model that filters given loans and creates a suggested portfolio. When it's wide enough, one can generate a realized return that is higher than 2% (~2.34%). We explained the importance of the threshold[12] and presented the tradeoff between the number of loans and the weighted realized return. We suggest using a threshold of 0.75. If you want to increase the weighted realized return gained, at the cost of investing in fewer loans, you can increase the threshold, otherwise, if you want to increase the number of loans (and increase the amount of money you invest), you can lower the threshold. However, we sincerely suggest not lowering the threshold below 0.7, as you might not be able to generate a realized return higher than 2%. Finally, suppose you wish to minimize the risk, we suggest using our model to generate a suggested portfolio, then invest only in loans of grade A, thus minimizing the risk while still enabling gaining a return higher than 2%.

There are some important limitations one should consider when using our model. ***Firstly***, the standard deviation, which measures the risk level in such investments, is relatively high (~6.24%). This means that if one inserts a broad portfolio into the model but invests in a low number of loans (given in the suggested portfolio), he might not generate a realized return that is higher than 2%. ***Secondly,*** if the model gets as input a small number of loans, it will output a very narrow suggested portfolio. In such cases, as explained above, one may not get a weighed realized return higher than 2%. ***Thirdly***, the model was trained on a portfolio with the following ratio of grades:

Grade	A	B	C	D	E	F	G
Ratio	0.2194	0.3474	0.2806	0.1126	0.0314	0.0071	0.0016

If one gives the model a new portfolio with a different grades' ratio, he might not get a suggested portfolio that enables generating a weighted realized return higher than 2%. (For example, a portfolio that consists of loans of grade G solely won't fit the model[13]). ***Finally***, note that the model is trained on loans from 2016. Thus, if future loans will "behave" differently than the 2016 loans, our model might not predict the new loans correctly.

[12] The threshold is the desired probability of a loan to be with a realized return that is higher than 2% = to be classified as "0"- and thus to be included in the suggested portfolio.

[13] If the threshold is 0.75 and above, the model would probably filter all those loans.

APPENDIX

Tables:

Table 1: Grades Baseline Table

	weighted_average	weighted SD	Number of Loans
A	1.861573	4.514966	56753
B	1.532950	7.149496	92005
C	0.798874	9.438672	74610
D	-0.194200	11.545009	30543
E	-1.022335	13.340783	8798
F	-2.109506	14.559178	2114
G	-3.628924	15.255247	493
Total All loans Together	1.076391	8.414820	265316

Table 2: The model's grades baseline table on the test set

Base line Grades - Suggested portfolio with threshold of: 0.75
XGBOOST

	weighted_average	weighted SD	Number of Loans
A	2.1611	4.3687	9006
B	2.3072	6.4866	13615
C	2.6130	8.1188	5728
D	3.5900	9.4213	548
E	6.1465	9.0121	70
F	3.2977	10.7355	3
Total All Suggested Portfolio	2.3438	6.3050	28970

Note: the weighted average realized return of the suggested portfolio (of the test set) is 2.34%, and the weighted standard deviation is 6.3%

Table 3: Comparison of the baseline model's performance and our model's performance

	Baseline model		Our model	
	Weighted average realized return	Weighted standard deviation	Weighted average realized return	Weighted standard deviation
Grade A	1.861%	4.514%	2.161% ↑	4.369% ↓
All loans	1.076%	8.414%	2.344% ↑	6.305% ↓

Table 4: Cross-validation (K=5) for the model's grades baseline table

Statistics of Realized return by grade for 5 folds with threshold of 0.75:

	Weighted Average all folds				Standard Deviation all folds				Number of Loans all folds			
	mean	min	max	sd	mean	min	max	sd	mean	min	max	sd
A	2.216684	2.153692	2.294094	0.062531	4.166257	3.939191	4.346233	0.194265	4234.6	4206	4296	39.227541
B	2.328109	2.254273	2.390866	0.059193	6.367769	6.271418	6.548693	0.119580	6299.6	6192	6412	78.697522
C	2.488098	2.280980	2.700446	0.164325	8.185954	7.794123	8.522023	0.274836	2683.2	2610	2764	63.888184
D	3.602997	3.028351	4.061545	0.481267	9.520894	9.091388	10.384430	0.537435	268.6	253	291	15.630099
E	3.800345	1.615217	6.142001	2.118742	11.959265	10.495964	13.778108	1.358434	36.8	28	47	7.049823
F	10.084597	6.949847	12.110476	2.001830	3.560709	1.550174	5.443511	1.761959	3.4	2	5	1.140175
Total All Suggested Portfolio	2.348254	2.301043	2.370801	0.028843	6.241215	6.182338	6.339383	0.059021	13526.4	13398	13656	103.910057

Table 4 presents the mean, minimum, maximum, and standard deviation of the: weighted realized return, the standard deviation of the realized return, and the number of loans. The mean, minimum, maximum, and standard deviation are calculated on the K folds generated in the cross-validation function.

Table 5: Comparison between a baseline model and our model – according to the **basic approach**:

	Baseline model[14]	**Our model**	**Difference** between Our model and the **Baseline** model
Weighted realized return	1.861%	2.348% ↑	**Higher** in 0.487% than the baseline model
Weighted standard deviation	4.515%	6.241% ↑	**Higher** in 1.726% than the baseline model

Table 6: Comparison between a baseline model and our model – according to the **two models** approach:

	Baseline model	Our model	Baseline model	Our model	**Difference** between Our model to the baseline model	
Grade	Weighted **Realized return**	Weighted **Realized return**	Weighted **Standard Deviation**	Weighted **Standard Deviation**	**Difference in** Weighted **Realized return**	**Difference in** Weighted **Standard Deviation**
A	1.862%	2.161%	4.515%	4.166%	0.299% ↑	-0.349% ↓
B	1.533%	2.328%	7.149%	6.367%	0.795% ↑	-0.782% ↓
C	0.799%	2.488%	9.439%	8.186%	1.689% ↑	-1.253% ↓
D	-0.194%	3.603%	11.545%	9.52%	3.797% ↑	-2.025% ↓
E	-1.022%	3.8%	13.341%	11.959%	4.822% ↑	-1.382% ↓

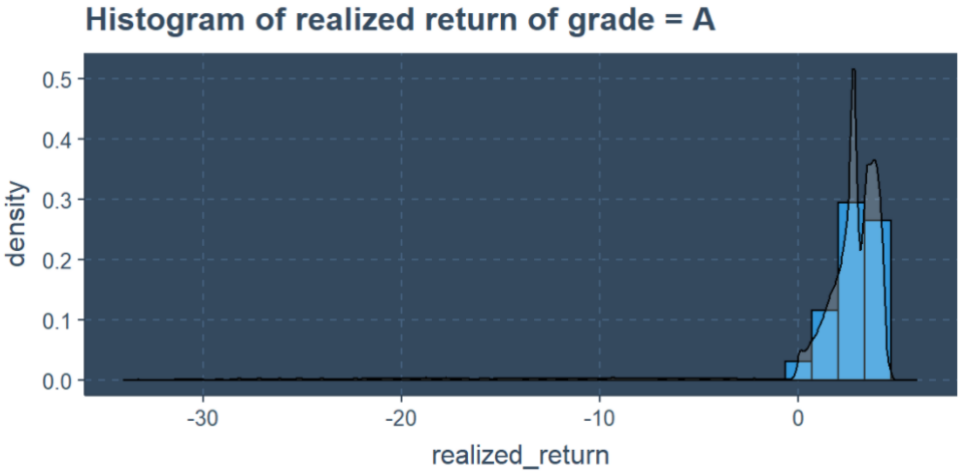
*We do not compare the F and G grades as there are only a few loans with the F grade and none of the loans in grade G entered the suggested portfolio.

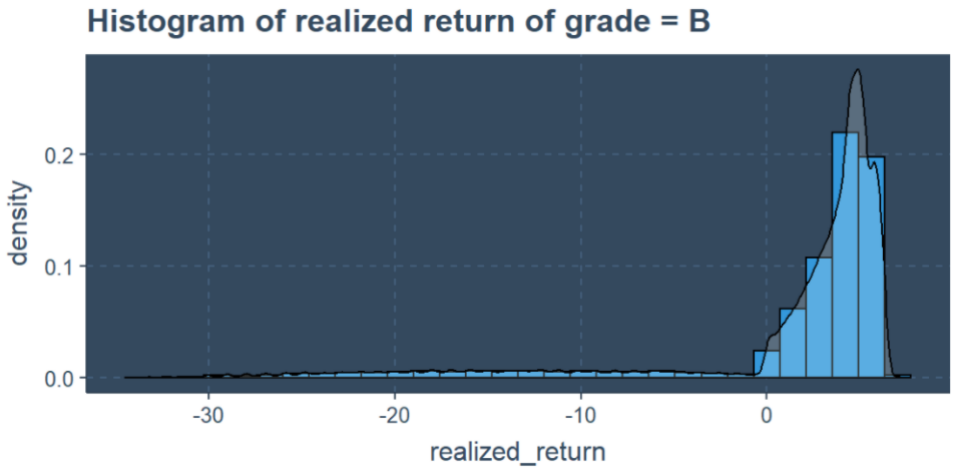
*The difference in the weighted realized return and in the weighted standard deviation is simply the **subtraction** of the metrics of our model and matching metrics of the baseline model.

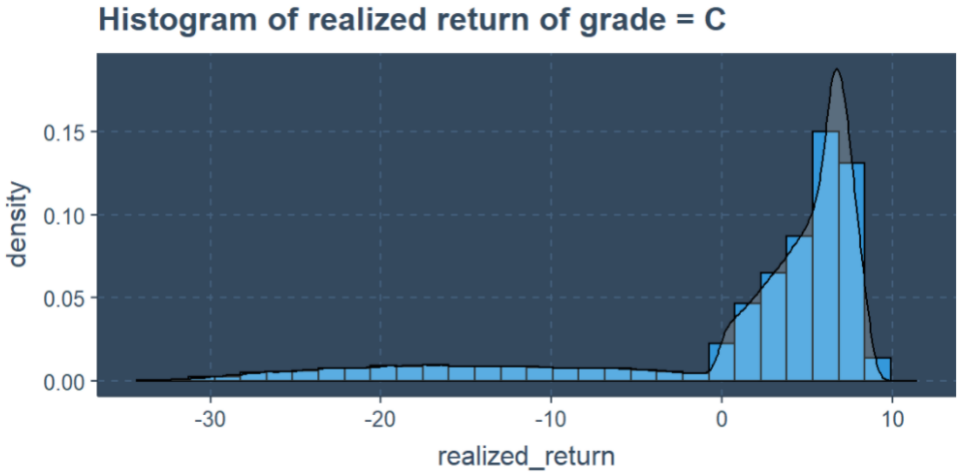
[14] A baseline model that suggests investing only in loans of grade A.

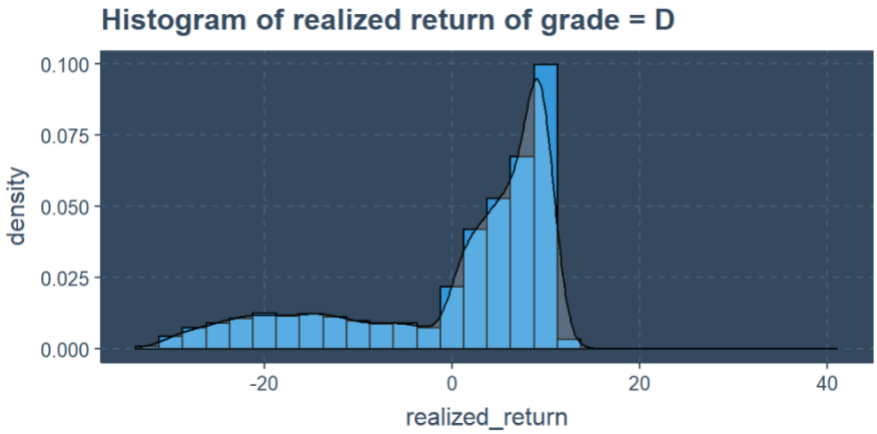
Question 1 section

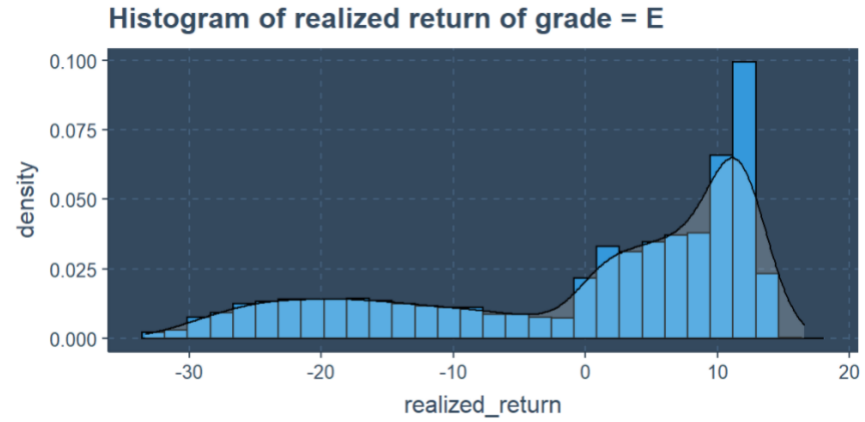
The realized return distribution for each grade graph:

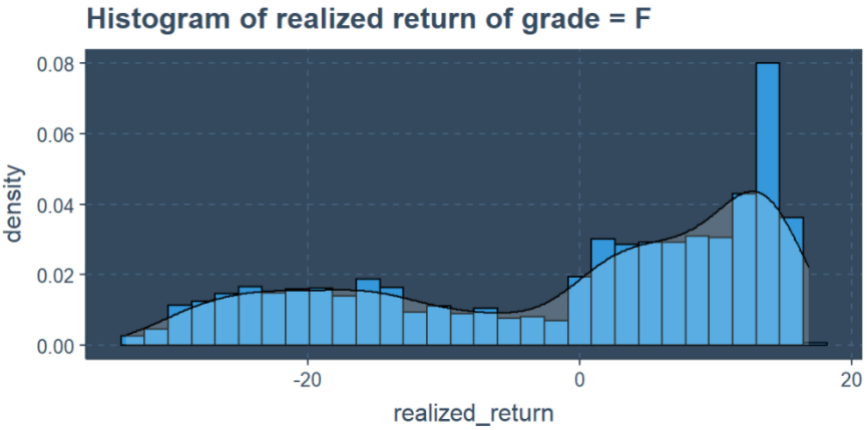


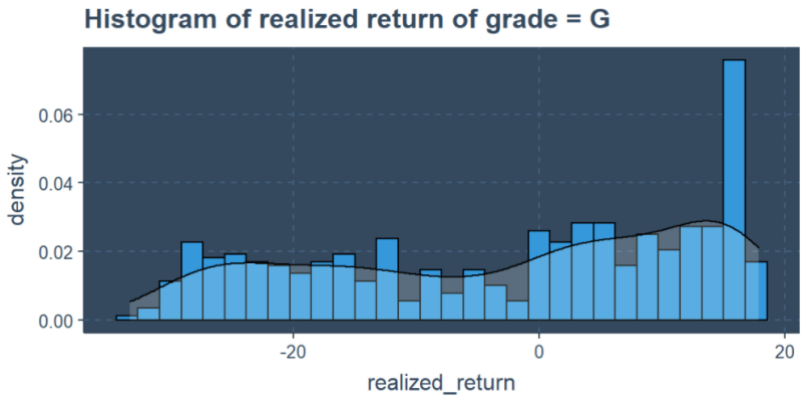




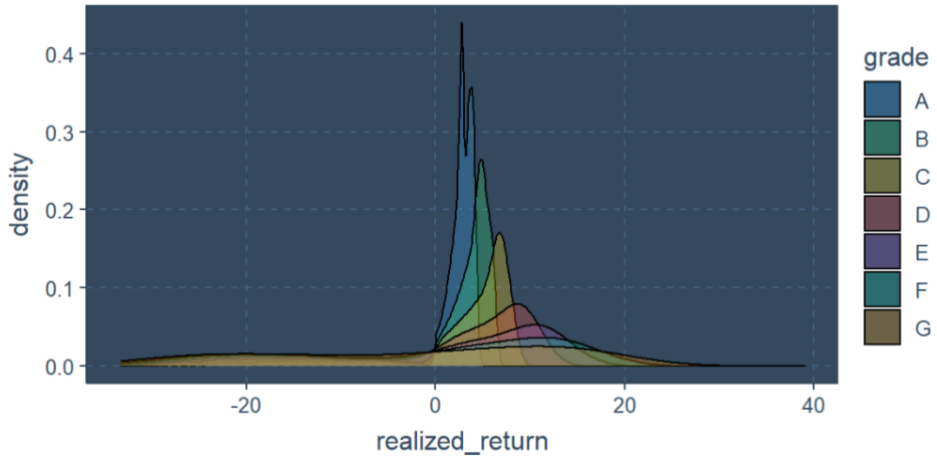

Histogram of realized return of grade = D
density
0.100
0.075
0.050
0.025
0.000
-20
0
20
40
realized_return


Histogram of realized return of grade = E
density
0.100
0.075
0.050
0.025
0.000
-30
-20
-10
0
10
20
realized_return


Histogram of realized return of grade = F
density
0.08
0.06
0.04
0.02
0.00
-20
0
20
realized_return


Histogram of realized return of grade = G
density
0.06
0.04
0.02
0.00
-20
0
20
realized_return

All the grades together:



Question 4 section

The "threshold table" (gained from the profit curve function): the following table presents for each threshold the weighted average realized return of the suggested portfolio as well as the number of loans, the total investment in thousands of dollars, the number of FN and TN yield:

	thresholds	WA_Realized_return_SP	number of loans	Total investment in thousands of $	Number of FN	Number of TN
0	0.5000	1.1894	69615.0000	910831.3570	19261.0000	50354.0000
1	0.5200	1.2195	68971.0000	901522.3070	18941.0000	50030.0000
2	0.5400	1.2448	68185.0000	890849.4820	18583.0000	49602.0000
3	0.5600	1.3007	67115.0000	876351.4820	18065.0000	49050.0000
4	0.5800	1.3513	65700.0000	857311.7570	17454.0000	48246.0000
5	0.6000	1.4405	63637.0000	830272.0250	16598.0000	47039.0000
6	0.6200	1.5194	60890.0000	795966.0000	15524.0000	45366.0000
7	0.6400	1.6236	57495.0000	754610.9750	14227.0000	43268.0000
8	0.6600	1.7364	53248.0000	702290.8000	12740.0000	40508.0000
9	0.6800	1.8846	48309.0000	641297.0250	11071.0000	37238.0000
10	0.7000	2.0301	42790.0000	572988.9250	9354.0000	33436.0000
11	0.7200	2.1649	37243.0000	503835.7750	7705.0000	29538.0000
12	0.7400	2.2829	31742.0000	434507.1750	6187.0000	25555.0000
13	0.7600	2.3703	26231.0000	365302.9500	4831.0000	21400.0000
14	0.7800	2.4988	20601.0000	292780.3500	3562.0000	17039.0000
15	0.8000	2.6326	14845.0000	216308.8250	2348.0000	12497.0000
16	0.8200	2.7738	9181.0000	139178.3750	1310.0000	7871.0000
17	0.8400	2.8396	4352.0000	70129.3750	539.0000	3813.0000
18	0.8600	3.0607	1336.0000	23543.5000	139.0000	1197.0000
19	0.8800	3.4959	171.0000	3188.2500	12.0000	159.0000
20	0.9000	4.0558	6.0000	125.8000	0.0000	6.0000
21	0.9200	4.8641	1.0000	3.6000	0.0000	1.0000

We can observe the **tradeoff** between the weighted average realized return and the number of loans: as we increase the threshold, there are fewer loans to invest in but higher weighted realized return.