Turning Loss Into Profit *with Machine Learning*

Muhammad Nanda S.

Table of Contents

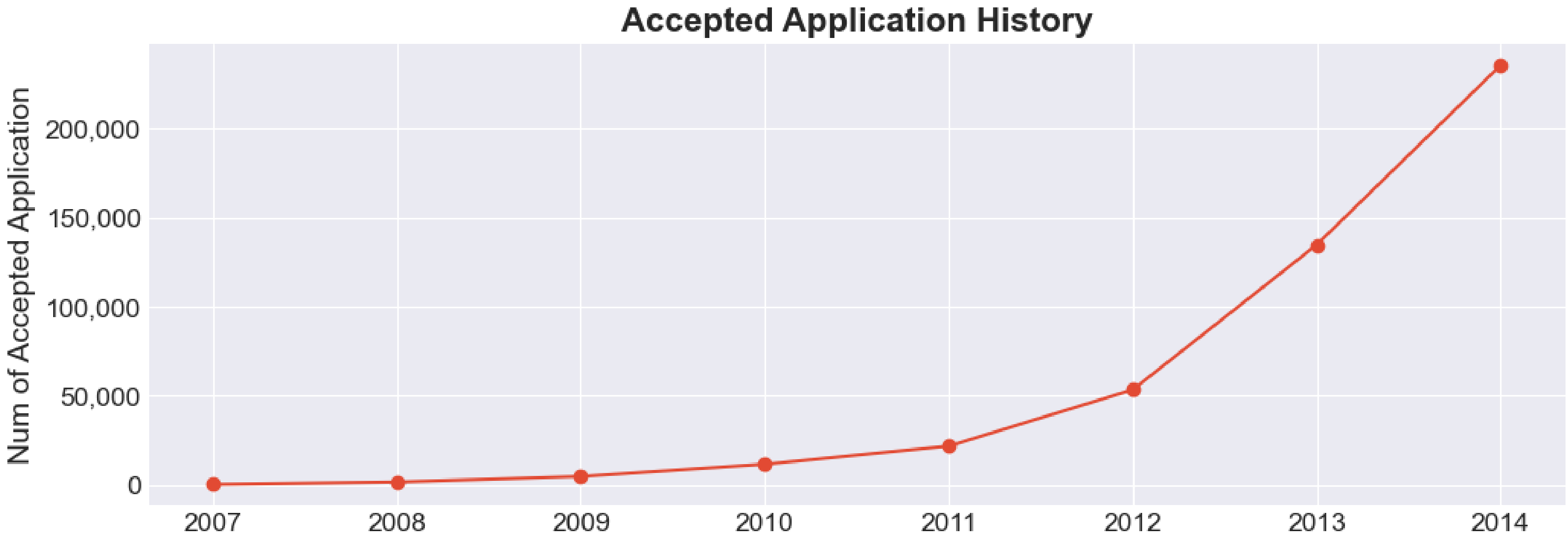
Business and Data Understanding

Company Profile (from dataset observation)

- Sector : Lending Funds (36/60 Months Term)
- Period : About 7 years (2007-2014)
- Total Approval : 463,536 application
- Total Funds Lent : USD $6,651,081,325
- Avg Acceptance : 8,915 application/day

From the profile, it can be seen that the company is **quite mature in running a business in the lending sector**

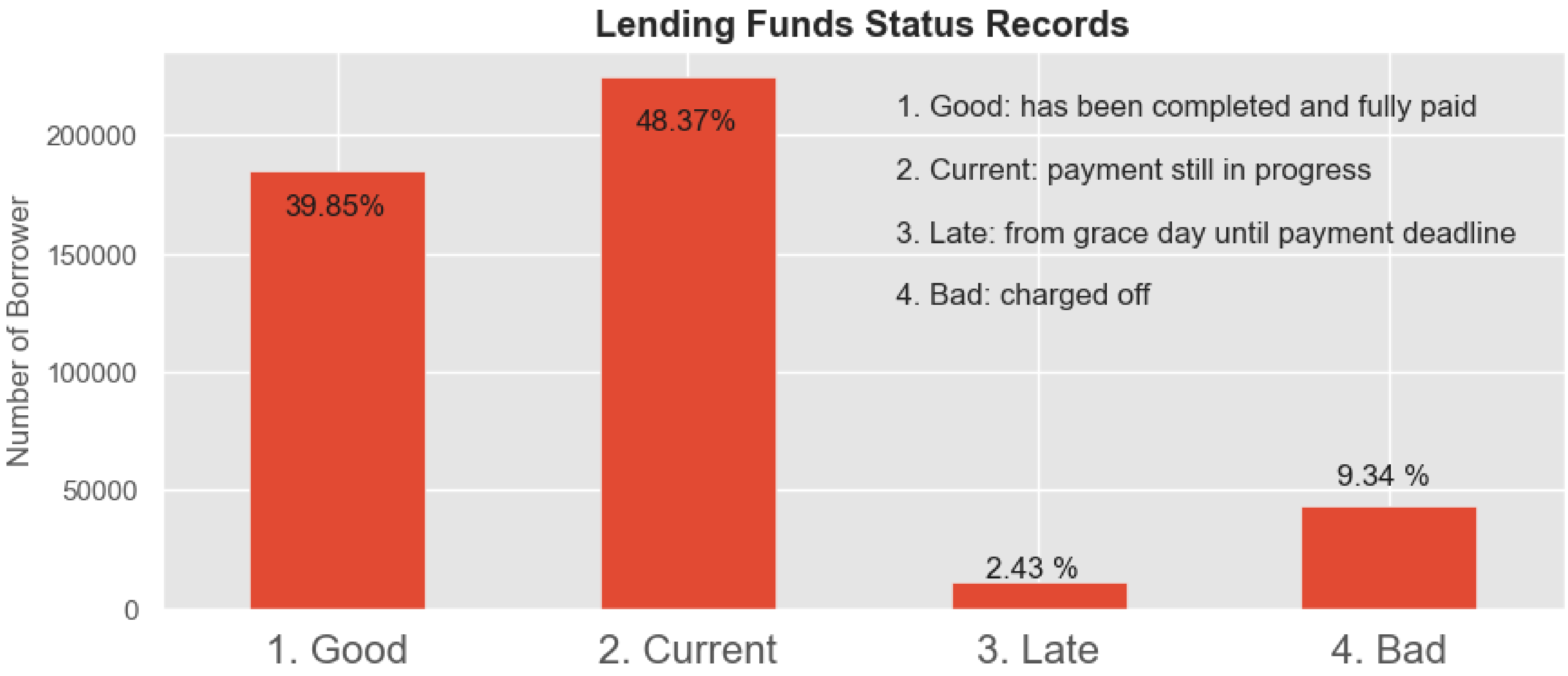
Business Growth

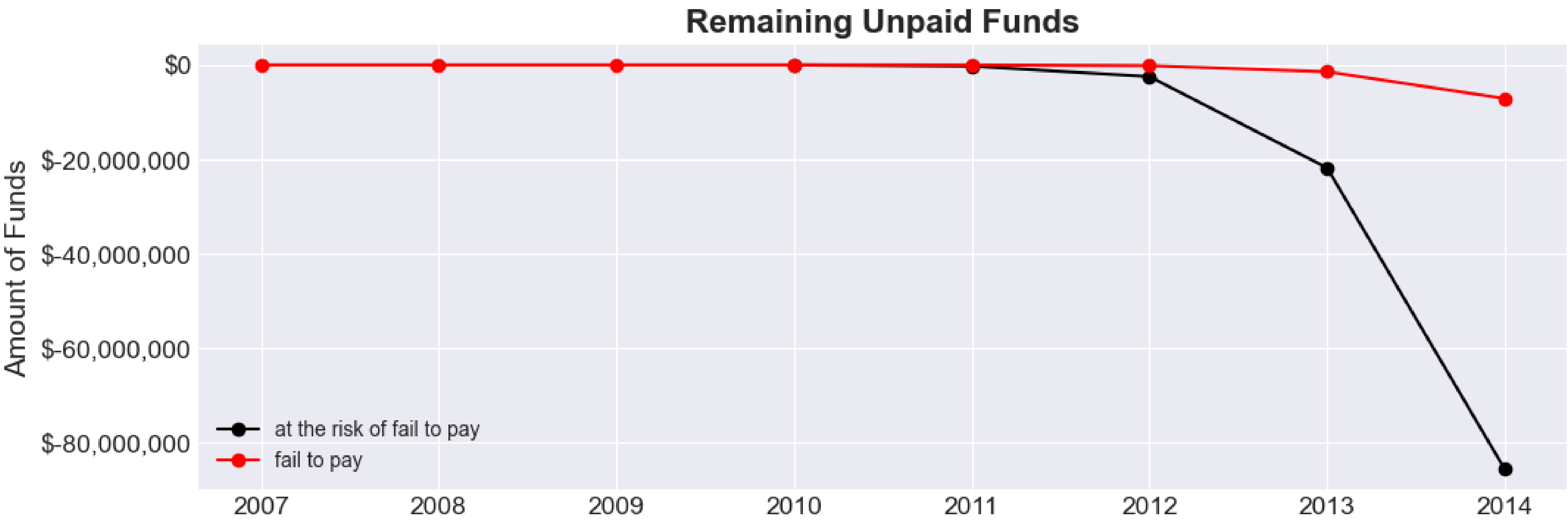


- The number of approved applications and loaned funds **continues to increase every year**.
- The **rapid increase** began to occur since 2013.
- This shows that the business being run has **good prospects** in the future and **gain more customer trust**.

Besides promising growth, there are certainly risks in a business. In the business sector run by the company, the risk comes from the results of the assessment of potential borrowers.

Based on the data, the assessment process that has been carried out by the company is quite good. Why? because there are still **2.43%** of borrowers **who are late** in paying and **9.34%** of borrowers **who fail to pay** as shown in the graph.



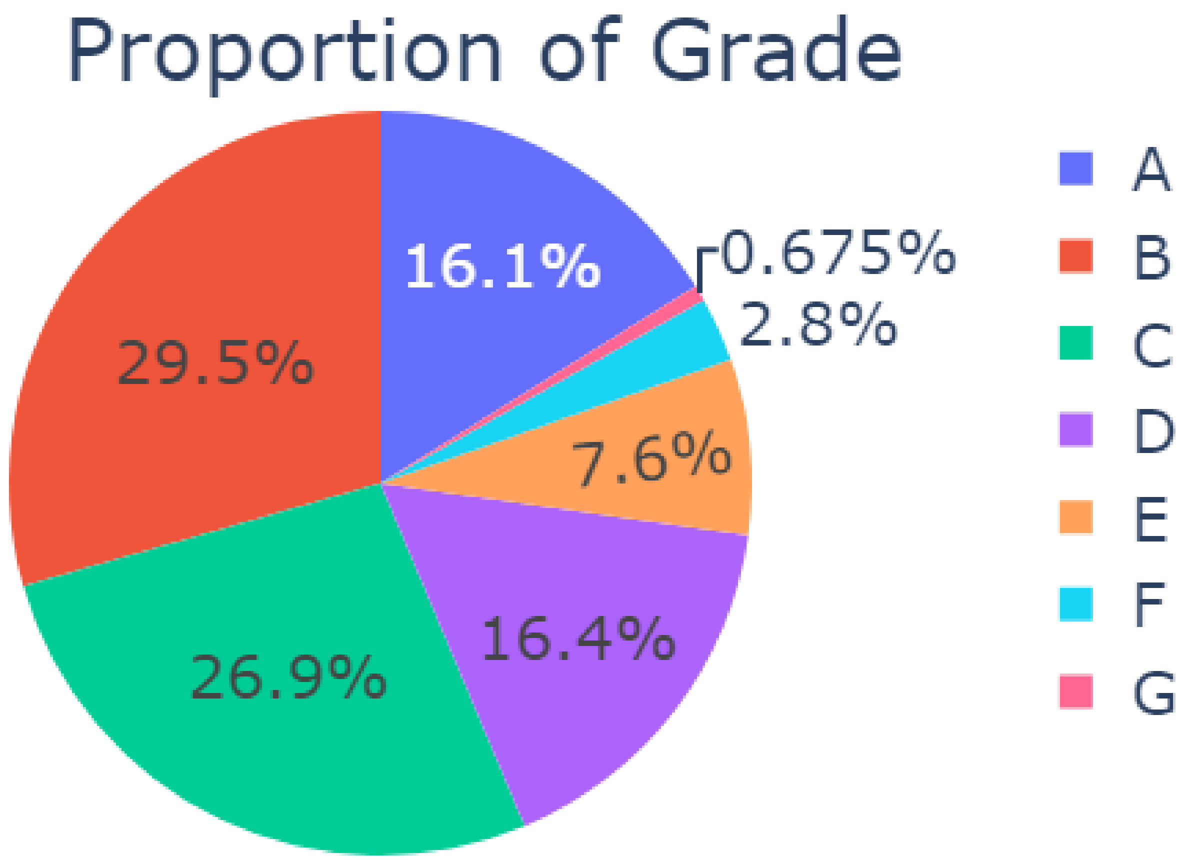


If the risk of loss is converted into money, the company has incurred a loss of around 8 million USD.

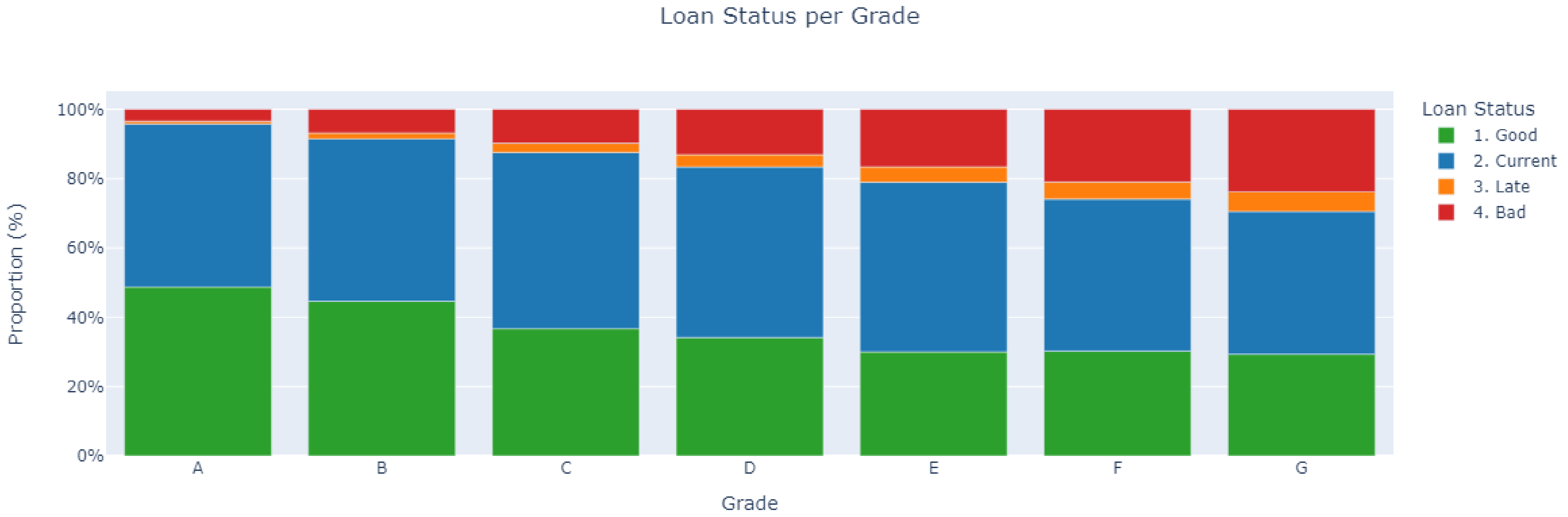
This risk of loss has the **potential to increase by around 110 Million USD** if those who are classified as late to pay become unable to pay.

Then, *how is the assessment process by the company?*

- The company **assigns a 'grade'** to those who are approved. There are *7 grades starting from A to G*, from the best to the best to the minimum qualification threshold (the origin of this assumption will be discussed in the next slide).
- From the available grades, **grade A** is not the majority and **grade G is very small, not even reaching 1%**.



let's see in more detail the effect of grade on loan status ...



- From the proportion of loan status per grade, **a pattern formed**. From grade A to G, the proportion of the good applicants is decreasing. On the other hand, the proportion of bad applicants is getting bigger. This is why, it is assumed that group A is the best and G is the minimum qualification.
- In addition, the plot shows that **the current assessment process still has gaps for groups who cannot repay loans**, because in every grade there is always such a group.

Initial Summary

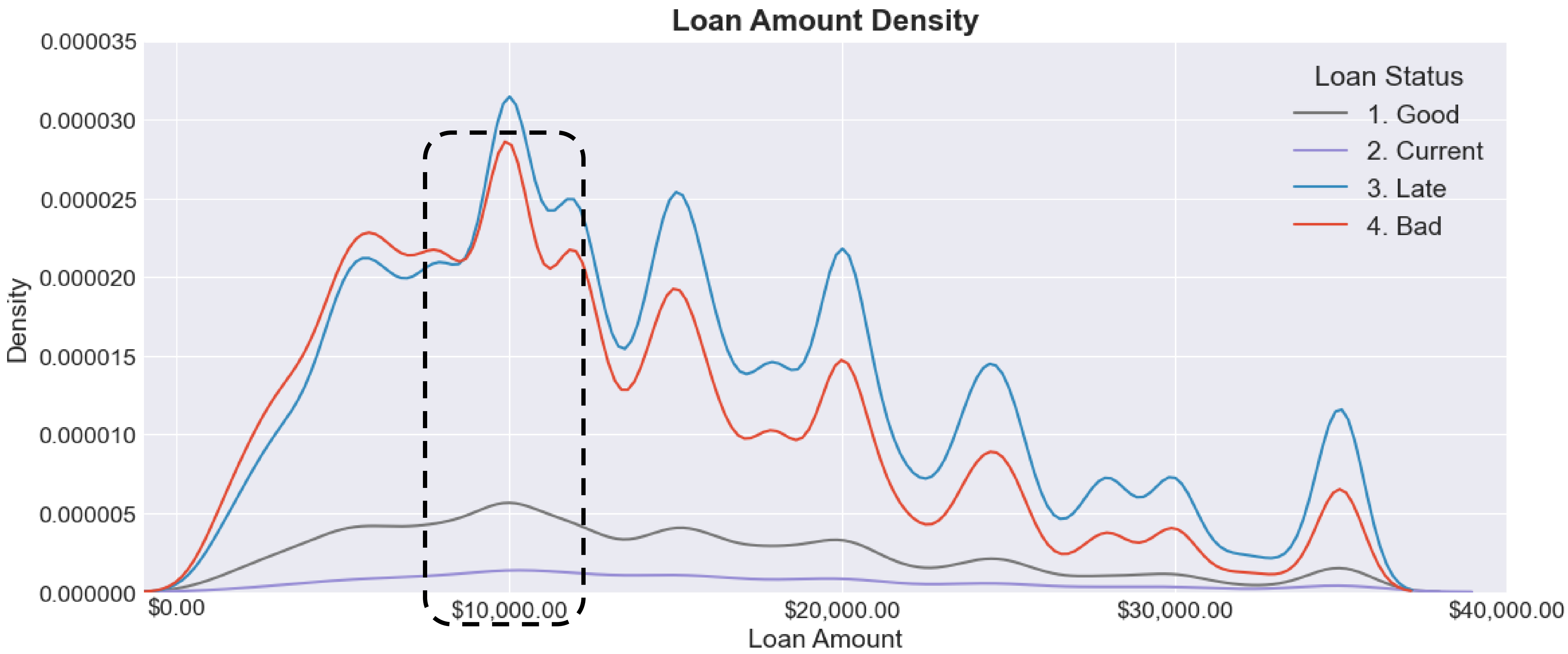
The company **has good prospects in the coming years** and their **customers have a rapid increase**. On the other hand, there is **still a risk from the assessment** process which still has loopholes that can result in losses to the company. Trying practical solutions like upgrading (in other words, eliminating some of the bottom grades) won't be enough. This is because **bad applicants are always present in each grade**.

Now, Let's take look at the effect of the bad applicants on the company's profit growth . . .

*will only try some features from data

From the loans that have been completed, it turns out that **the company has suffered losses since 2013**. This is in ***contrast*** to the growth of customers and loan funds approved by the company. This means that *small group of bad applicants has a big impact on the company's profit*.

Then, do bad applicants always borrow large amounts?



The answer is NO. Notice the peak of the red line,

that means *bad applicants tend to borrow relatively small funds*, about $10,000

Another question is how the income of the applicants.
Are bad applicants only those with low incomes?

Annual Income Statistics

average	$73.288,39
std_dev	$54.871,71
min	$3.000,00
1%	$18.960,70
10%	$34.000,00
25%	$45.000,00
50%	$63.000,00
75%	$89.000,00
90%	$120.000,00
99%	$242.000,00
max	$7.500.000,00

Lowest income group, **14,62%** are bad applicants

Very high income group, **5,62%** are bad applicants

From the statistics, it shows that *those with high incomes also make loans and **do not guarantee** that they are good applicants*.

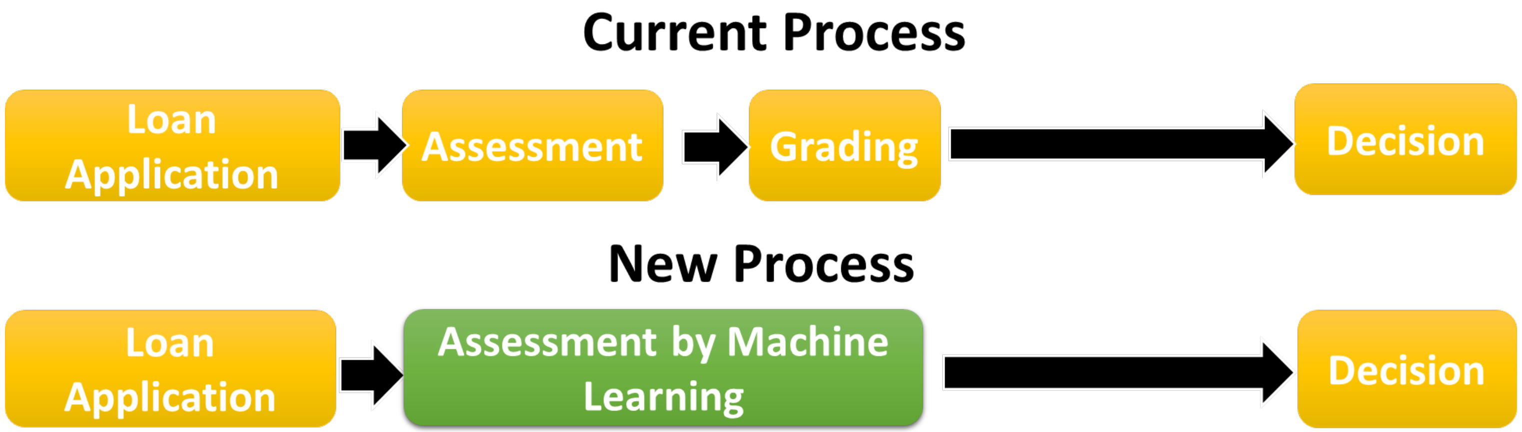
Methodology

What we already know …

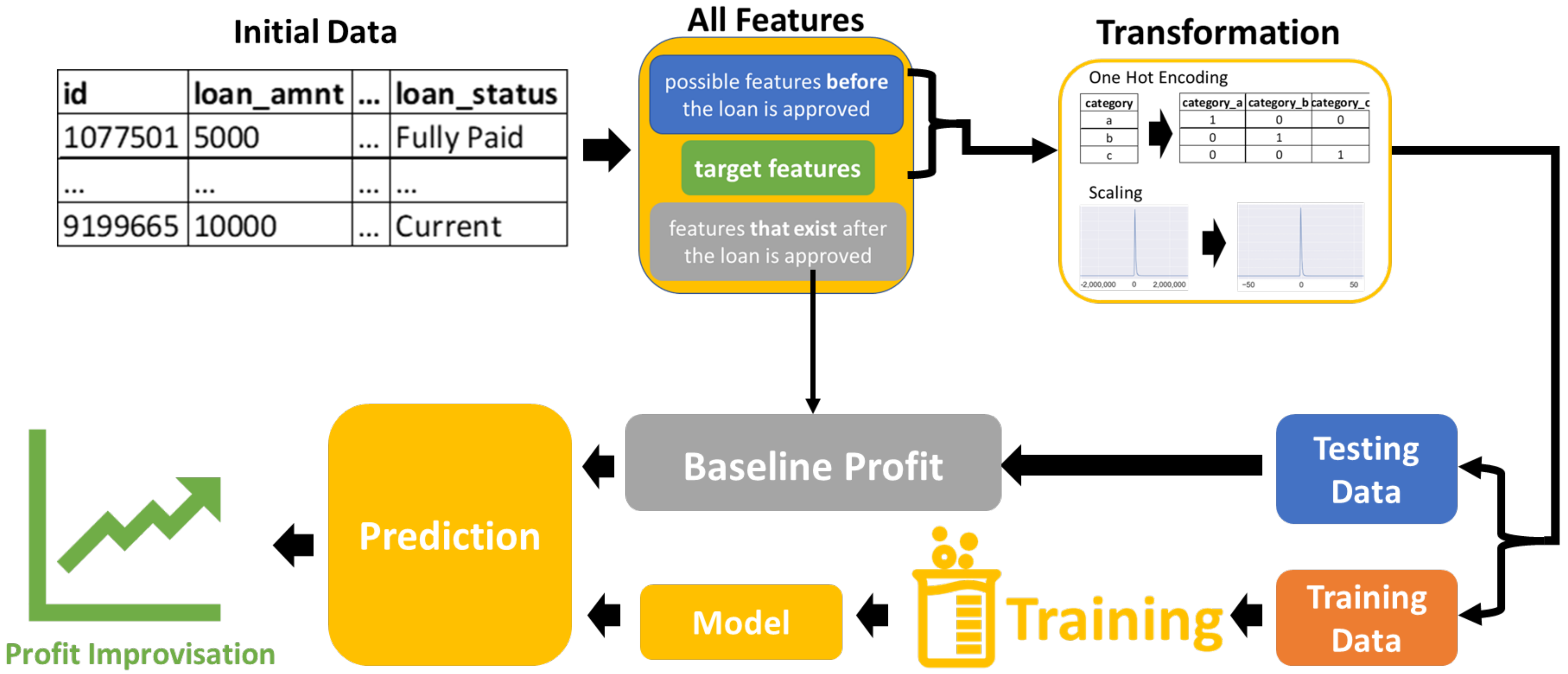
- The company has already lost more than US$65 million and this could add up from ongoing and future borrowings.
- The data source has a description of applicant grading, but there is no description of how it is grading and there may be data used other than the currently available data

Solution

- I offer a solution by applying machine learning *to reduce bad applicants by replacing the grading assessment.*
- The main metric as a measure of success is ***profit.***



Framework



Initial Data
id | loan_amnt | ... | loan_status
1077501 | 5000 | ... | Fully Paid
... | ... | ... | ...
9199665 | 10000 | ... | Current
All Features
possible features before the loan is approved
target features
features that exist after the loan is approved
Transformation
One Hot Encoding
category
a
b
c
category_a
category_b
category_c
1 0 0
0 1 0
0 0 1
Scaling
-2,000,000 0 2,000,000
-50 0 50
Testing Data
Training Data
Baseline Profit
Training
Model
Prediction
Profit Improvisation

Notes

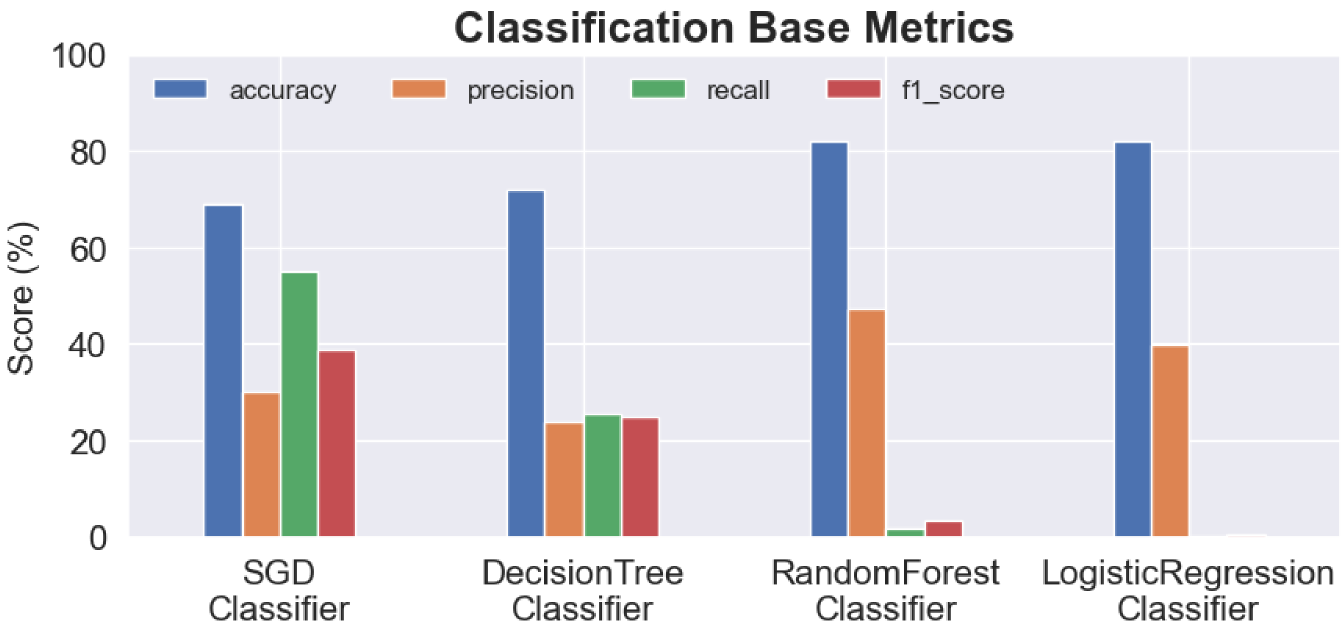
- **Restriction**: only use data with good and bad loan status
- In the modeling section, several machine learning models are used as follows:
    1. **LogisticRegression** (simple and light model)
    2. **DecisionTreeClassifier** (good for non-parametric model)
    3. **RandomForestClassifier** (more robust for non-parametric model)
    4. **SGDClassifier** (good for large dataset)

Result

Base Metrics

From a machine learning perspective, predictive results from models are generally measured by:

- **Accuracy:** measure of correctly predicted applicants to the total applicants.
- **Precision:** measure how many retrieved bad applicants are relevant
- **Recall:** measure how many relevant bad applicants are retrieved.
- **F1-Score**: measure weighted average of Precision and Recall



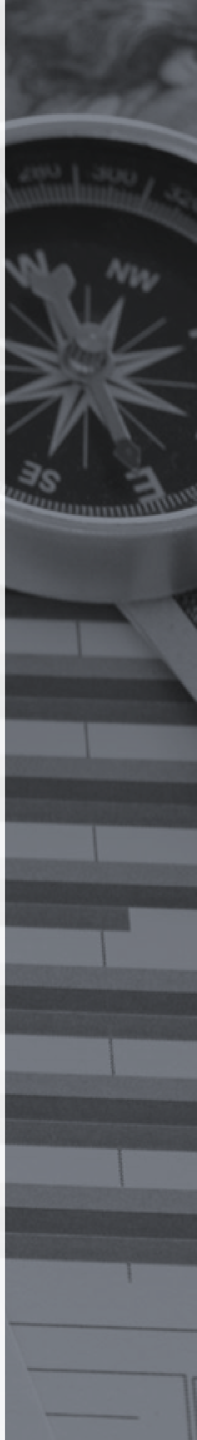
Result from 47,862 test data (8,661 of them are bad applicants):

- Highest Accuracy : 81.88% (LogisticRegressionClassifier)
- Highest Precision : 47.35% (RandomForestClassifier)
- Highest Recall : 54.91% (SGDClassifier)
- Highest F1-Score : 38.95% (SGDClassifier)

Profit Metric

It has been stated in the methodology that the main metric is **Profit**. So here need to do the conversion. The reasons behind that are:

- Decision makers or other divisions may need a *metric that is more general and easy to understand* in general, especially from business perspective.
- A low base metric value **doesn't mean** machine learning isn't working well.



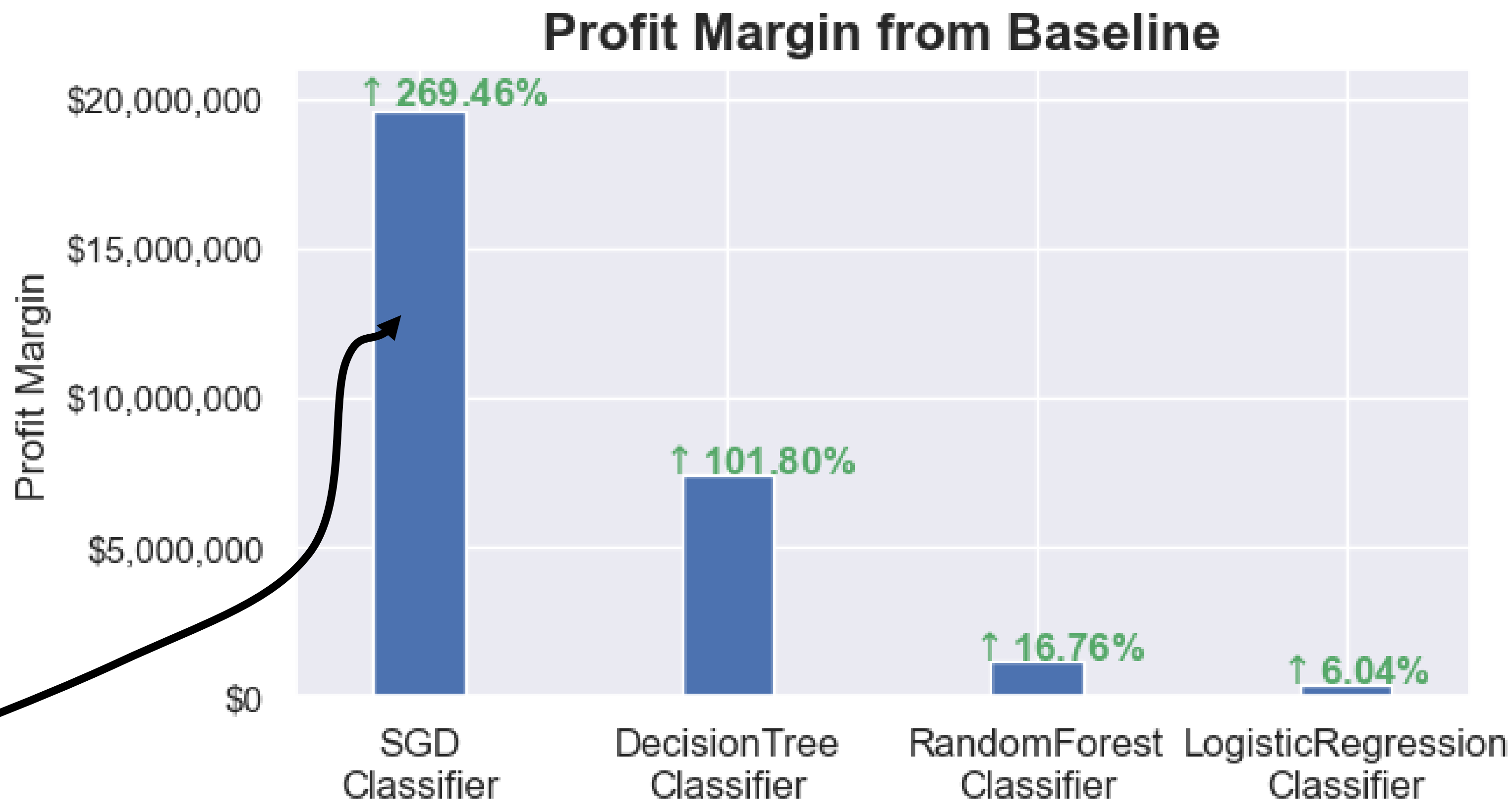
Conversion Result

Conversion Idea:

Only calculates profit/loss available in the baseline data for those who pass as applicants

Baseline Profit (on data test): $-7,272,745.26 (suffered loss)

The Improved Profit (on data test) up to **$12,324,133.22 / ↑269.46%** from baseline



Bad Applicants Reduction

Model	Detected Bad Applicants	Ratio
SGD Classifier	4756	↓ **53.67%**
DecisionTree Classifier	2207	↓ 24.9%
RandomForest Classifier	152	↓ 0.017%
LogisticRegression Classifier	23	↓ 0.003%

- By reducing the bad applicants correctly by 24.9%, at least it could avoid the company from not losing money.
- If the correct decline reaches 53.67%, the company could get a profit of more than 269% from the baseline

Conclusion

Conclusion

- The company has good prospects for growth, but the risk of losses that will be faced will also increase.
- The grading process in the assessment of loan funds still has gaps for groups that fail to pay. For this reason, it is proposed to use ***machine learning as a new assessment process to replace the current one***.
- The simulation results show that *the solution the solution works well, it could reduce bad applicants by 53.67% to increase profit up to 269%*