Predicting Corporate Bankruptcy Using Financial Ratios via Machine Learning

Nathan Silverglate, Brandon Brooks, Mayuresh Deolekar

Data Science for Business - Technical
Fall 2021

NYU | STERN

Business Case

Problem Statement

- Business customers currently don’t have a tool that allows them to assess whether or not a publicly-traded company is at risk of bankruptcy before buying company stock based on commonly available business financial ratios

Business Case

Our Solution

- A corporate bankruptcy prediction tool that gives customers the insights to properly assess a company's financial health before and while investing their capital

Business Case

Our predictive tool will help customers:

- Collect the necessary data to make better and more informed decisions
- Gain critical insight to at-risk companies in the same industry and notify them of financial risk across that industry
- Highlight critical business issues for at-risk companies that the management team might be withholding
- Maximize their ROI
- Further hold public companies accountable

Business Case

Data mining is extremely useful in discovering patterns and correlations for various players in the finance industry. Here are few alternative use cases for our tool:

- Financial institutions
- Professional investors
- Governments
- Academic researchers

Data

- Data source from Kaggle
- Data collected from *Taiwan Economic Journal* for years 1999-2009
- Bankruptcy is defined by regulations of Taiwanese stock exchange

Data

- Initial inspection yielded data frame of 96 columns with 6,819 entries
- 93 of columns are continuous numerical features corresponding to values including:
    - Operating Gross Margin
    - Realized Sales Gross Margin
    - Operating Profit Rate
    - Pre-tax net Interest Rate
    - After-tax net Interest Rate

Data

- Two of the columns are binary features:
  - Liability-Assets Flag
  - Net Income Flag
  - Both of these features contained same values for all entries and were later dropped
- Final column is a binary classifier column corresponding to Bankruptcy Occurred?
  - Yes = 1
  - No = 0

	A	B	C	D	E	F	G	H	I
1	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate
2	1	0.3705942573	0.4243894461	0.4057497725	0.6014572133	0.6014572133	0.9989692032	0.7968871459	0.8088093609
3	1	0.4642909375	0.53821413	0.5167300177	0.6102350855	0.6102350855	0.9989459782	0.7973801913	0.8093007257
4	1	0.4260712719	0.4990187527	0.4722950907	0.6014500065	0.601363525	0.9988573535	0.7964033693	0.8083875215
5	1	0.3998440014	0.4512647187	0.4577332834	0.5835411292	0.5835411292	0.9986997471	0.7969669683	0.8089655977
6	1	0.4650221811	0.5384321849	0.5222977675	0.5987834936	0.5987834936	0.9989731318	0.7973660806	0.8093037202
7	1	0.388680349	0.4151766245	0.4191337866	0.5901713775	0.5902506522	0.9987580984	0.7969031934	0.8087705621
8	1	0.4161263589	0.4702354993	0.4637828578	0.599115006	0.599115006	0.9989725541	0.7972408824	0.8091758664
9	1	0.4621947058	0.5360335805	0.5144279672	0.5999870278	0.5999870278	0.9989088875	0.7972905467	0.8092226073
10	1	0.453029786	0.5165721762	0.5055945179	0.5969457617	0.5969529685	0.9989399697	0.7970841427	0.8090395498
11	1	0.4720908692	0.5388137811	0.517051234	0.5973925828	0.5973925828	0.9989604216	0.7973659422	0.8092903099
12	1	0.06693316433	0.0571849106	0.05482092189	0.6018607936	0.6018607936	0.9988247692	0.7967792403	0.8087166604
13	1	0.4066689416	0.4672917575	0.4534503988	0.5959944652	0.5959944652	0.9988307776	0.7971905265	0.8091284746
14	1	0.4327499634	0.4927496729	0.4820921891	0.6040156243	0.6040156243	0.9988954841	0.7972899933	0.8092219563
15	1	0.5043630868	0.5622546882	0.5553295144	0.604858819	0.604858819	0.9990597922	0.7974380176	0.8093472061
16	1	0.4706771316	0.5369058003	0.5238503132	0.5976520273	0.5976520273	0.9990096447	0.7973861399	0.8093132246
17	1	0.4701408863	0.5203881378	0.5220300873	0.5973061013	0.5973997896	0.9989036879	0.797310606	0.8092371894
18	1	0.4561010091	0.5207152202	0.5108945875	0.5976304069	0.5978177835	0.9989376588	0.7973330172	0.8092709105
19	1	0.4543460245	0.5137919756	0.5060228064	0.597522305	0.597522305	0.9988994127	0.7972374239	0.8090636363
20	1	0.4681421538	0.5494439599	0.5176401306	0.602055377	0.6022499604	0.9989844554	0.7974057843	0.8093742872
21	1	0.4653146785	0.5265481901	0.5112158038	0.6047507171	0.6047507171	0.999004214	0.7971060005	0.8090455389
22	1	0.4472285867	0.5016354121	0.4926387922	0.5935009153	0.5934792949	0.99890623	0.7971392022	0.809089806
23	1	0.456344757	0.5039249891	0.4985277584	0.5944738321	0.5944738321	0.9988694859	0.7971202495	0.8090623343
24	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Data

- Dataset lacks any temporal features, prevents us from knowing details regarding prediction timing
- Dataset also lacks unique identifiers which prevents us from knowing if multiple instances represent the same company
- The only fixed value columns are the two binary flag features that are to be removed
- All other features are numerical and are actual business financial ratio metrics

Data

- Upon initial inspection, data contained zero NaN values
- Since there are 95 features in dataset, needs to simplify to a more usable number of features
- Performed **Model Based Feature Selection** using **Random Forest Classifier** to judge importance of each feature and how many to use based on the **sensitivity score** of confusion matrix
- Found that **15 features** was optimal

```
Sensitivity Score: 0.164  number of features: 5
Sensitivity Score: 0.230  number of features: 10
Sensitivity Score: 0.262  number of features: 15
Sensitivity Score: 0.262  number of features: 20
Sensitivity Score: 0.213  number of features: 25
Sensitivity Score: 0.230  number of features: 30
Sensitivity Score: 0.230  number of features: 35
Sensitivity Score: 0.246  number of features: 40
```

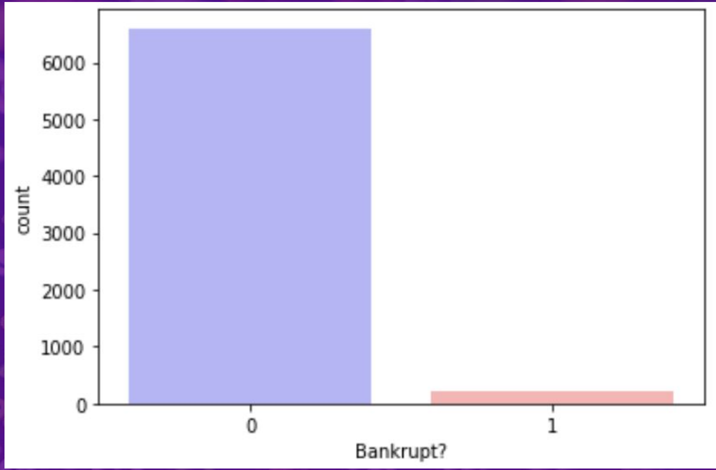
Data

The 15 selected features are as follows:

- Operating Expense Rate
- Interest-Bearing Debt Interest Rate
- Net Value Per Share (B)
- Persistent EPS in the Last Four Seasons
- Net Value Growth Rate
- Total Asset Return Growth Rate
- Interest Expense Ratio
- Borrowing Dependency
- Net Profit Before Tax/Paid-in Capital
- Cash/Total Assets
- Inventory/Working Capital
- Working Capital/Equity
- Total Assets to GNP Price
- Net Income to Stockholder's Equity
- Degree of Financial Leverage (DFL)

Data

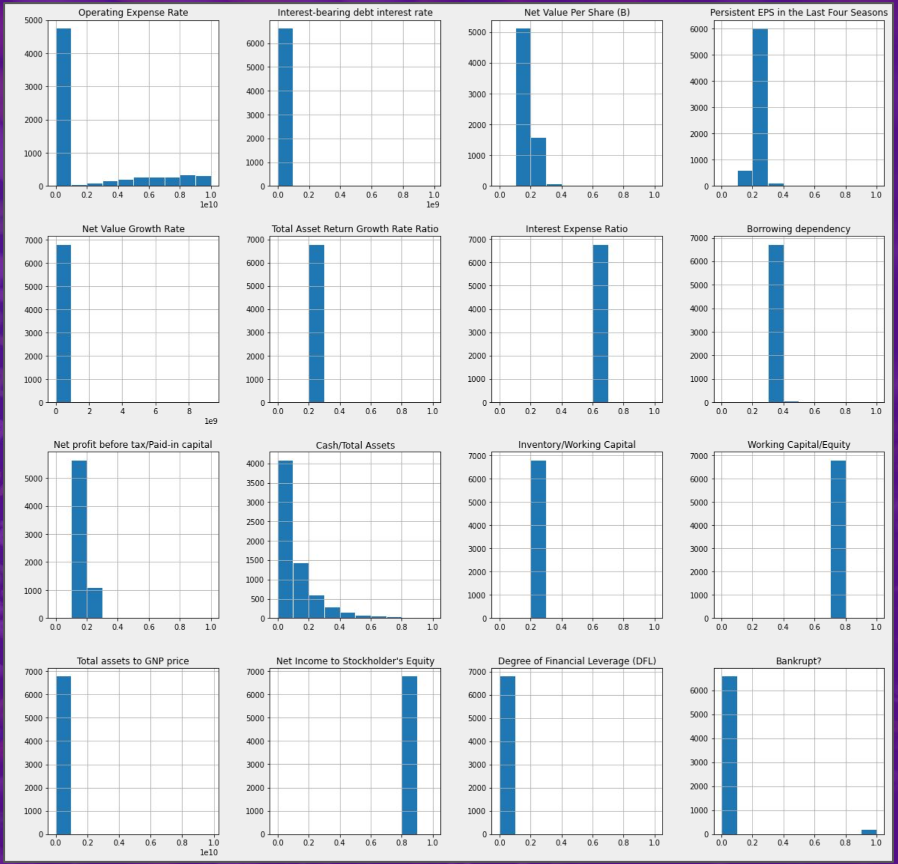
- Next investigated the distribution of target column values
- Distribution was quite skewed
  - Bankrupt = No: 6,599
  - Bankrupt = Yes: 220
- Ratio: 3.23% of entries went bankrupt



Data

Feature Distribution

- Data is both heavy on outliers or contains most values in a single “bin”



NYU | STERN

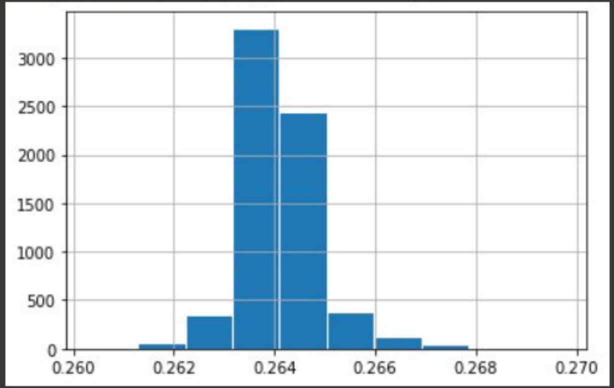
Data

Total Asset Return Growth Rate

- Has normal distribution but influenced by outliers

```
Total Asset Return Growth Rate Ratio
(0.2, 0.3]                              6815
(0.3, 0.4]                                 2
(0.9, 1.0]                                 1
(-0.001, 0.1]                              1
(0.8, 0.9]                                 0
(0.7, 0.8]                                 0
(0.6, 0.7]                                 0
(0.5, 0.6]                                 0
(0.4, 0.5]                                 0
(0.1, 0.2]                                 0
dtype: int64
```

```
Rows with outliers: 6819
Rows withou outliers: 6767
information lost = 52 rows
<matplotlib.axes._subplots.AxesSubplot at 0x7f9e1b736e10>
```



Data

	Operating Expense Rate	Interest-bearing debt interest rate	Net Value Per Share (B)	Persistent EPS in the Last Four Seasons	Net Value Growth Rate	Total Asset Return Growth Rate Ratio	Interest Expense Ratio	Borrowing dependency	Net profit before tax/Paid-in capital	Cash/Total Assets	Inventory/Working Capital	Working Capital/Equity	Total assets to GNP price	Net Income to Stockholder's Equity	Degree of Financial Leverage (DFL)	Bankrupt?
count	6.819000e+03	6.819000e+03	6819.000000	6819.000000	6.819000e+03	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6819.000000	6.819000e+03	6819.000000	6819.000000	6819.000000
mean	1.995347e+09	1.644801e+07	0.190661	0.228813	1.566212e+06	0.264248	0.630991	0.374654	0.182715	0.124095	0.277395	0.735817	1.862942e+07	0.840402	0.027541	0.032263
std	3.237684e+09	1.082750e+08	0.033390	0.033263	1.141594e+08	0.009634	0.011238	0.016286	0.030785	0.139251	0.010469	0.011678	3.764501e+08	0.014523	0.015668	0.176710
min	0.000000e+00	0.000000e+00	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000
25%	1.566874e-04	2.030203e-04	0.173613	0.214711	4.409689e-04	0.263759	0.630612	0.370168	0.169376	0.033543	0.277034	0.733612	9.036205e-04	0.840115	0.026791	0.000000
50%	2.777589e-04	3.210321e-04	0.184400	0.224544	4.619555e-04	0.264050	0.630698	0.372624	0.178456	0.074887	0.277178	0.736013	2.085213e-03	0.841179	0.026808	0.000000
75%	4.145000e+09	5.325533e-04	0.199570	0.238820	4.993621e-04	0.264388	0.631125	0.376271	0.191607	0.161073	0.277429	0.738560	5.269777e-03	0.842357	0.026913	0.000000
max	9.990000e+09	9.900000e+08	1.000000	1.000000	9.330000e+09	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	9.820000e+09	1.000000	1.000000	1.000000

(6819, 16)

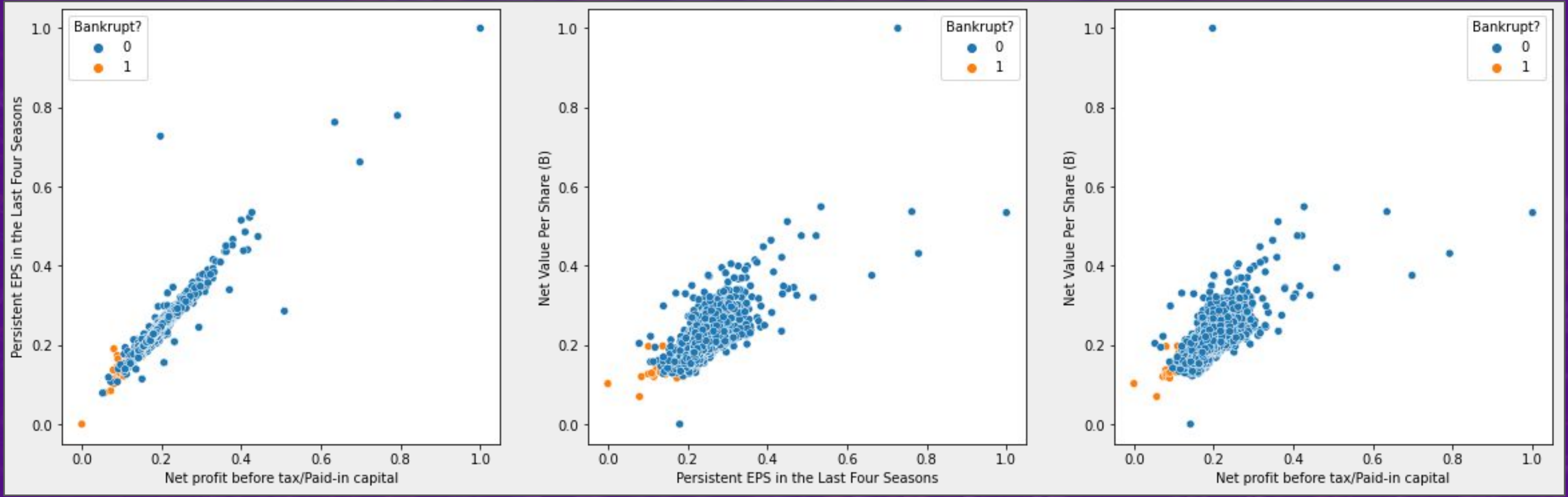
Data

Looking for Correlations

- Find that top correlations are:
  - *Net profit before tax/Paid-in capital vs. Persistent EPS in the Last Four Seasons*
  - *Persistent EPS in the Last Four Seasons vs. Net Value Per Share (B)*
  - *Net Profit Before Tax/Paid-In Capital vs. Net Value Per Share (B)*

NYU STERN

Data



NYU | STERN

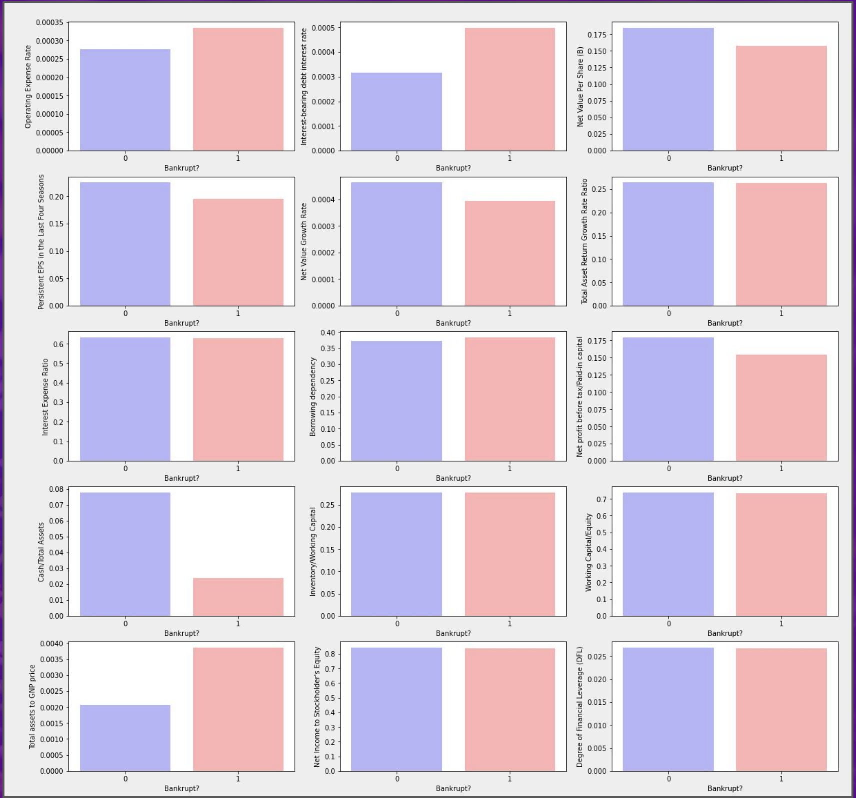
Data

Patterns

- Companies with a low 'Net profit before tax/Paid-in capital', 'Persistent EPS in the Last Four Seasons' and 'Net Value Per Share (B)' tend to go bankrupt

Data

- Comparing median of each feature for bankrupt vs non-bankrupt companies



Bankrupt?	Operating Expense Rate	Interest-bearing debt interest rate	Net Value Per Share (B)	Persistent EPS in the Last Four Seasons	Net Value Growth Rate	Total Asset Return Growth Rate Ratio	Interest Expense Ratio	Borrowing dependency	Net profit before tax/Paid-in capital	Cash/Total Assets	Inventory/Working Capital	Working Capital/Equity	Total assets to GNP price	Net Income to Stockholder's Equity	Degree of Financial Leverage (DFL)
0	0.000276	0.000317	0.185074	0.225111	0.000463	0.264057	0.630703	0.372474	0.179021	0.077684	0.277179	0.736072	0.002063	0.841232	0.026810
1	0.000335	0.000499	0.158021	0.195944	0.000396	0.263724	0.630283	0.382655	0.154012	0.023755	0.276981	0.732669	0.003853	0.836707	0.026689

Data

Conclusions of Data Analysis

- Since companies with a low 'Net profit before tax/Paid-in capital', 'Persistent EPS in the Last Four Seasons' and 'Net Value Per Share (B)' are more likely to go bankrupt, we will want to use a KNN model as we have clear clustering
- **Interest-bearing Debt Interest Rate:** When high, tends to bankruptcy
- **Total Assets to GNP Price:** When high, tends to bankruptcy
- **Cash/Total Assets:** When low, tends to bankruptcy

Models & Evaluation

According to the Data analysis, for this prediction we would train our data through different machine learning algorithms such as:

1. **K-Nearest Neighbors**
2. **Gradient Boosting**

Models & Evaluation

K-Nearest Neighbor

- This model is one of the most commonly used supervised machine learning algorithms.
- It is mostly used in solving classification problems and is based on the nearest neighbor principle.
- This KNN model was trained with the features identified as most predictive during data analysis process: **Net profit before tax/Paid-in capital, Persistent EPS in the Last Four Seasons, Net Value Per Share (B), Interest-bearing debt interest rate, Total assets to GNP price, Cash/Total Assets**
- Evaluated by creating confusion matrix and finding sensitivity scores

Models & Evaluation

K-Nearest Neighbor

```
best number of neighbors: 1
best training set sensitivtiy score : 1.000
best test set sensitivity score: 0.226
```

```
training set score : 0.97
test set score: 0.96
Time:  0.31626288700044825
```

Models & Evaluation

Gradient Boosting

- This gives us prediction model in the method of transformation from weak learners into strong learners.
- First, it creates a starting leaf and then creates new trees by taking into account the errors that occur.
- This process is continued until a better result cannot be obtained.
- Here, we will apply classifier on our reduced data and then on the whole dataset.
- Doing two gradient boosting models and evaluating based on sensitivity scores and time to completion

Models & Evaluation

Gradient Boosting

```
training set sensitivity score : 0.30
test set sensitivity score: 0.19
Time:  0.3831532900003367
```

Gradient boosting model performance with reduced (n=6) feature set

```
training set sensitivity score : 0.30
test set sensitivity score: 0.23
Time:  5.036059335999198
```

Gradient boosting model performance with total original (n=95) feature set

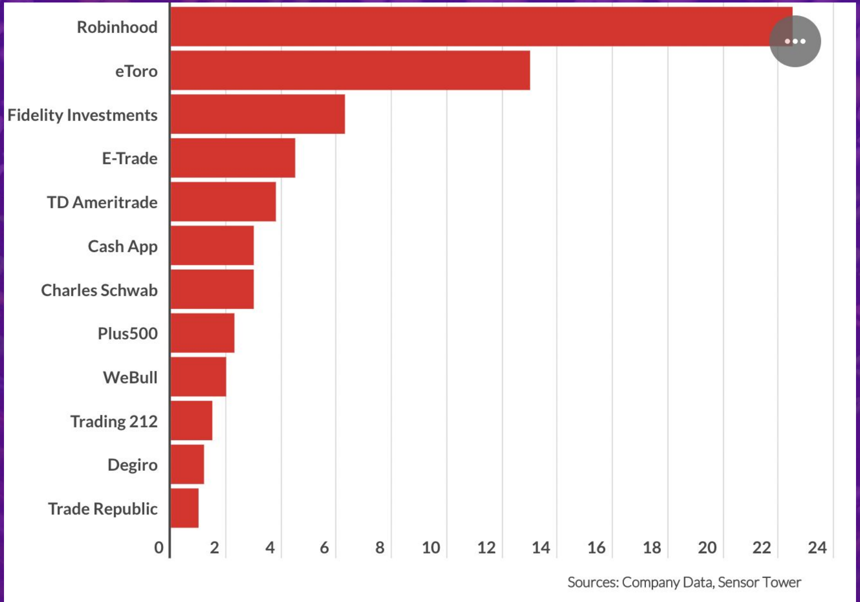
Deployment

Business Model

- Our predictive tool will be packaged as a B2B SaaS product
- We'll license the software to consumer-focused stock investment apps (e.g. TD Ameritrade, Robinhood) and charge them based on the number of users

Deployment

Stock Trading App Users by App (in millions)



Deployment

Issues and Risks to be Aware of

- SEC regulations
  - Since our initial target customer will be investment apps, our product has to be in compliance with SEC regulations
- Consumer volatility
  - Investment apps, like Robinhood, tend to encourage emotionally-based reactions instead of logic-based decisions so consumers turn to social media or forums for advice (e.g. Wallstreetbets vs. Robinhood)
- Possible backlash from companies
  - If our tool works as planned, there will be easier access to indicators and information that predict a company's risk of bankruptcy
- Payment-for-order-flow (PFOF) model
  - Many zero-commission investment apps use this process to generate income. The PFOF model bundles trades and sends them to a third-party market maker which then compensates the stockbroker for making the trade. Ultimately, investment apps are selling the users data and monetizing it.

Deployment

Conclusion

- As we can see, there are inherent risks with selling financial services products, specifically when selling directly to consumers
- However, we believe that because we have a B2B model and by aligning ourselves with the SEC's regulations, as well federal and GDRP privacy and security laws, we're able to mitigate some of the associated risks

Appendix

Section Authors

- **Business Case:** Brandon Brooks, Nathan Silverglate
- **Data:** Nathan Silverglate
- **Models:** Mayuresh Deolekar, Nathan Silverglate
- **Evaluation:** Mayuresh Deolekar, Nathan Silverglate
- **Deployment:** Brandon Brooks

Appendix

Sources

1. https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction
2. Deron Liang, Chia-Chi Lu, Chih-Fong Tsai, Guan-An Shih, Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study, European Journal of Operational Research, Volume 252, Issue 2, 2016, Pages 561-572