Deck Presentation | Bankruptcy Prediction | September 2025

Project Data Science:

Bankruptcy Prediction

by Muhammad Ramadhani

Portfolio Drive Github

Table of Contents

Contents	Slide
Introduction: Companies Bankruptcy	3
Data Preparation: Exploration and Preprocessing	8
Result of Modelling	15
Model Interpretation	19
Demo and Conclustion	22
References	24
Appendix	25

Introduction: Companies Bankruptcy

Bankruptcy: The Cycle within the Company

Definisi

Sebuah kondisi (hukum) di mana sebuah individu atau perusahaan secara resmi dinyatakan tidak mampu membayar utang-utangnya kepada kreditor.

Key Takeaways:

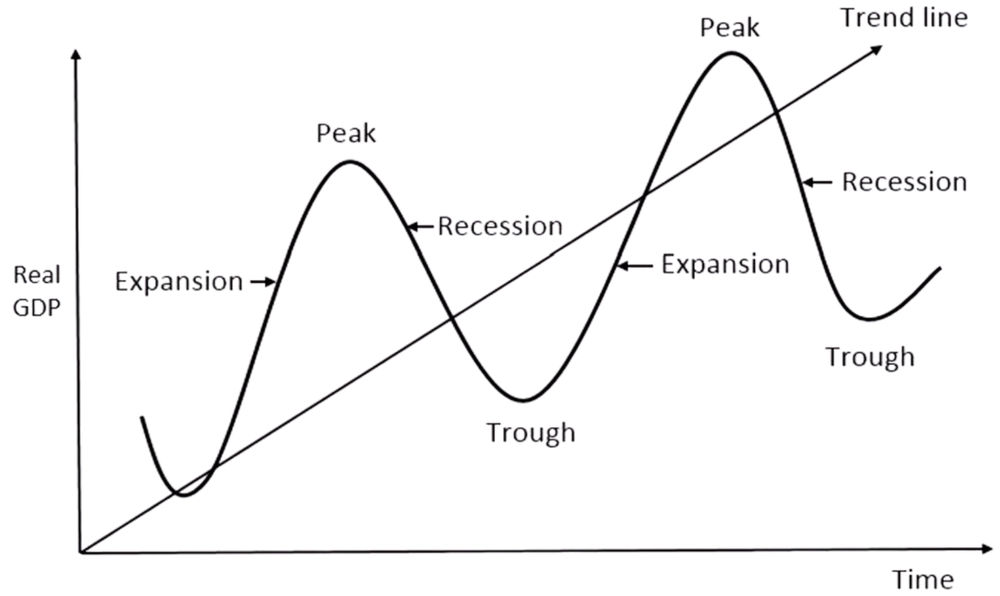
- Kebangkrutan (bankruptcy) adalah **risiko bisnis fundamental** yang (akan) dihadapi setiap perusahaan.
- Risiko kebangkrutan bisa dipicu oleh **faktor eksternal** dengan kondisi perokonomian ataupun **faktor internal** dari kesalahan manajerial (inefisiensi operasional).

Jumlah Perusahaan yang Mengalami Kebangkrutan

Country	Last	Previous	Reference
United States	23043	23309	Jun/25
France	5830	5721	Jun/25
United Kingdom	2081	2125	May/25
Japan	961	848	Jul/25

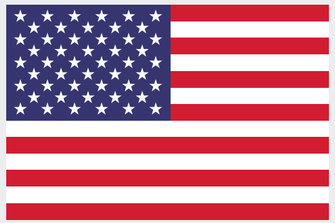
From: Bankruptcy Companies in G20 Countries| Trading Economics

Graph 1 – Phases of a Business Cycle

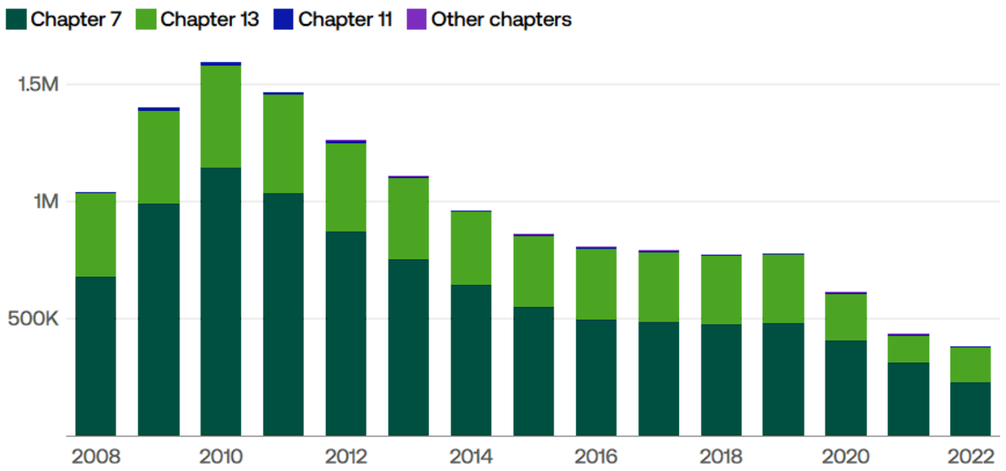


Bankruptcy in Companies: Fact and Aftermath

United States



US Bankruptcy Cases by Chapter (Fiscal Year 2008-2022)

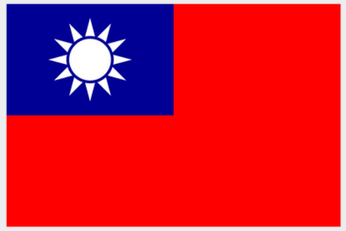


Key Takeaways:

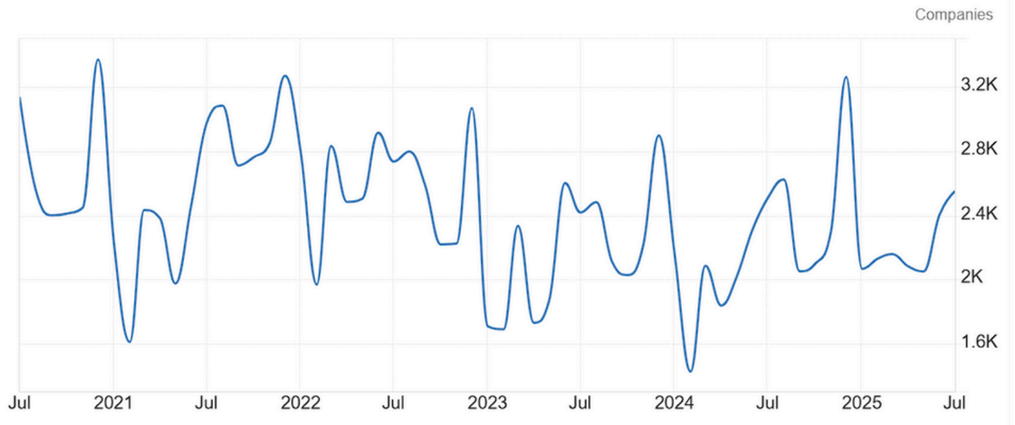
- **Pada puncak Krisis Finansial 2010**, ada lebih dari 1.5 juta kasus kebangkrutan dalam satu tahun.
- Pada tahun 2022, sekitar 380.000 kasus kebangkrutan yang terjadi
- **Chapter 7 (hijau tua) dan Chapter 13 (hijau muda) mendominasi** jumlah kasus di United States

From: USAFacts | Source

Taiwan



Taiwanan Bankruptcies (Jul 2020-Jul 2025)



Key Takeaways:

- Pada Juli 2025, **ada 2.548 perusahaan di Taiwan yang mengajukan kebangkrutan.**
- Angka tersebut **meningkat** jika dibandingkan **dengan bulan sebelumnya** yang mencatat 2.401 perusahaan bangkrut.
- Terlepas dari tren naik atau turun, kebangkrutan adalah risiko fundamental yang kemungkinan akan dialami perusahaan dan jumlahnya tidak sedikit.

From: Trading Economics | Source

Bankruptcy in Companies: Fact and Aftermath

The Aftermath: Efek Domino dari Kebangkrutan

Dampak Finansial

1. Kerugian Kreditor & Bank
2. Hilangnya Model Investor
3. Beban padasistem perbankan

"Pada 2018, PT Saringawi Agricultural Estate Agency (SAEA) dinyatakan pailit dengan total utang ke Bank ICBC mencapai 316 miliar"

From: CNN Indonesia | Source

Dampak Sosial

1. Pemutusan Hubungan Kerja
2. Daya Beli Menurun
3. Peningkatan Beban pemerintah

"PT Sritex dinyatakan pailit dan tutup permanen pada 1 Maret 2025. Sekitar 10 ribu pekerja terkena pemutuhan hubungan kerja (PHK)."

From: Detik Finance | Source

Dampak Operasional

1. Gangguan pada pemasok
2. Kekacauan bagi pelanggan
3. Kehilangan Aset, termasuk aset produktif

"Bowflex, salah satu produsen alat gym, bangkrut pada 2024 dan telah menjual asetnya senilai $37,5 juta untuk menyelamatkan perusahaan"

From: Media Indonesia | Source

Jalur Penyelesaian

Likuidisasi (Pailit)

Pengajuan Permohon Pailit → **Putusan oleh pengadilan** → **Penunjukan kurator** → **Pengambilahina Aset** → **Penjualan Aset (lelang)** → **Pembagian Hasil u/ kreditor**

Key Takeaways: Likuidasi adalah skenario akhir. Perusahaan sudah dipastikan berhenti beroperasi dan asetnya dijual. Tidak ada jalan kembali.

Restrukturisasi (PKPU)

Permohonan PKPU → **Putusan PKPU** → **Penunjukan Pengurus** → **Penyusan & Negoisasi Restrukturisasi** → **Voting Rencana Perdamaian (Keputusan)**

Jika ditolak: Perusahaan Pailit

Jika diterima: Homologasi

Key Takeaways: Restrukturisasi adalah upaya penyelamatan. Ada kemungkinan perusahaan tetap berjalan jika restrukturisasi disetujui (dalam skala/bentuk berrbeda

Problem Statement & Project Goals

"Terlepas dari poin penyelesaian yang ditempuh ataupun tren jumlah kebangkrutan perusahaan, dampak yang ditimbulkan dapat begitu merusak untuk perusahaan maupun pihak yang terkait lainnya. ***Kebangkrutan berkaitan dengan gagal bayar, maka sisi keuangan perusahaan menjadi sorot utama dalam analisis ini. Mengidentifikasi risiko lebih awal menjadi poin penting sebagai alarm agar perusahaan memahami dan mengerti penyebab kemunculan kebangkrutan ataupun potensi kebangkrutan,"***

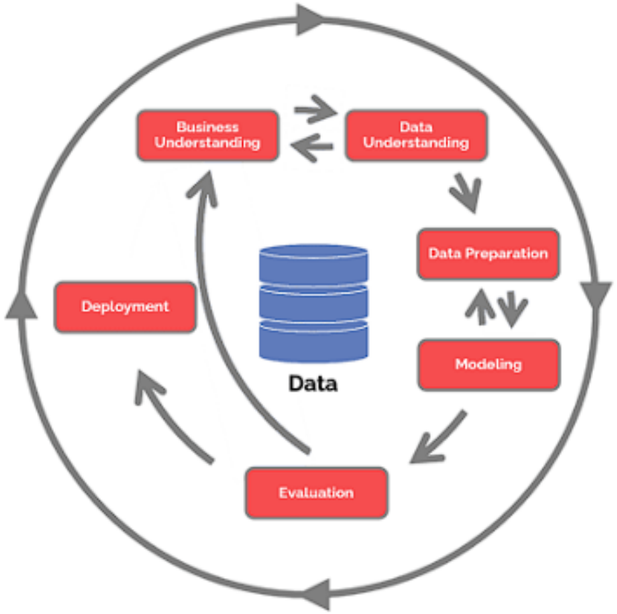
Tujuan

1. Mengindetifikasi faktor-faktor (variabel) yang menyebabkan kebangkrutan dalam perusahaan
2. Membangun dan membandingkan model machine learning yang paling baik dalam memprediksi kebangkrutan
3. Mengidentifikasi dan menganalisis peran masing-masign faktor, seperti likuiditas, profitabilitas, solvabilitas, dsb terhadap prediksi model.

Pertanyaan Analisis

1. Variabel-variabel apa saja yang menjadi penyabab kebangkrutan?
2. Apa model yang terbaik untuk melakukan prediksi? Bagaimana hasilnya?
3. Apakah dampak dari masing-masing variabel yang "penting" terhadap model prediksi? Interpretasikan

Alur Analisis



CRISP-DM Framework

Detail:

- Data diambil dari kaggle dan UCI ML
- Data merupakan **perusahaan Taiwan**
- **Ada 95 fitur di luar dari target**
- Perusahaan didefinisikan berdasarkan peraturan bisnis Bursa Efek Taiwan.

Data Preparation: Exploration and Preprocessing

Data Understanding

Tentang Data

Faktor Keuangan	Variabel Kunci (Contoh)	Dampak Terhadap Potensi Kebangkrutan
Profitabilitas	roa(a) before interest and % after tax; net income to stockholder's equity	**Berbanding Terbalik**: Semakin tinggi nilai profitabilitas (seperti ROA/ROE), semakin rendah kemungkinan perusahaan bangkrut.
Solvabilitas/Leverage	net worth/assets; debt ratio %	**Berbanding Terbalik/Lurus**: Semakin tinggi net worth/assets (porsi modal sendiri), semakin rendah potensi bangkrut. Sebaliknya, semakin tinggi debt ratio % (porsi utang), semakin tinggi potensi bangkrut.
Likuiditas	working capital to total assets; current ratio	**Berbanding Terbalik:** Semakin tinggi rasio likuiditas, semakin besar kemampuan perusahaan membayar utang jangka pendek, sehingga semakin rendah potensi kebangkrutan.
Efisiensi & Aktivitas	total asset turnover	**Berbanding Terbalik**: Semakin tinggi nilai perputaran aset, semakin efisien perusahaan dalam menghasilkan penjualan, sehingga semakin rendah potensi kebangkrutan.

Kemungkinan Informsi yang Redundan

1. Rasio yang (hampir) identik
   - **gross profit to sales** vs. **gross margin**
   - **liquid asset to short term liabilities** vs. **current assets minus inventory to short term liabilities**
2. Metrik dengan denominator yang sama
   - **kelompok aset:** net profit to total asset; ebit_to_total_assets; gross_profit_to_total asset; dst
   - **kelompok penjualan:** gross profit to total sales; net profit to total sales; dst
3. Rasio yang saling berkebalikan
   - **total liabilities to total assets** vs. **total asset to total liabilities**

- Ada 95 fitur data dengan 5455 baris
- Data memiliki *target* yang *imbalanced*, di mana hanya ~4% observasi perusahaan yang bangkrut dari keseluruhan data

Banyaknya fitur bisa menjadi

masalah atau **solusi**

Memilih fitur lebih awal (statistik)

Memilih fitur setelah beberapa eksperimen

Data Cleaning & Manipulating

Assesment	Temuan	Status	(Rencana) Tindakan
Re-organize features order	Terdapat fitur-fitur yang tidak urut	Sudah Dilakukan	• Mengurutkan fitur berdasarkan faktor atau rasio keuangan
Data Duplicates	Terdapat baris yang nilainya sama atau identik	Tidak Ditemukan	• Tidak ada data duplikat
Missing Value	Terdapat nilai yang hilang	Tidak Ditemukan	• Tidak ada *missing value* di data
Encoding	Target masih dalam categoric	Sudah Dilakukan	• Mengubah menjadi 0 dan 1 untuk target yang masih bertipe object
Handling Outlier	Ada fitur yang memiliki nilai ekstrim	Sudah Dilakukan	• Menghapus kolom jika secara mean dan median “rusak” yang terlihat dari nilai yang tidak masuk akal atau wajar • Melakukan Capping berdasarkan persentil ke-99
Scaling	Skala hitung jaraknya masih jauh	Sudah Dilakukan	• Mengubah skala hitung dengan RobustScalar

Outliers: Noise, Valuable Insight, or Anomaly?

Statistik Deskriptif

Sample Variabeles	Mean	Std	Min	25%	50%	75%	max
roa(a) before interest and % after tax	0.557818	0.065355	0.00000	0.535379	0.559365	0.587958	1.00000
cash flow rate	4.672292E-01	1.552481E-02	0.00000	0.461541	0.465076	0.470914	7.432509E-01
cash/current liability	3.704491E+07	5.126982E+08	0.00000	0.001982	0.004840	0.012776	9.65000E+09
research and development expense rate	1.945017E+09	2.597449E+09	0.00000	0.000129	4.92000E+08	3.410000E+09	9.980000E+09
cash turnover rate	2.496308E+09	2.951727E+09	0.00000	0.000276	1.14000E+09	4.550000E+09	1.000000E+10

Fokus: Menghilangkan nilai yang tidak valid atua tidak wajar dalam konteks data (anomaly)

Jika dilihat dari tabel:

1. Terdapat variabel yang tidak ada *noise* atau outlier → **roa (a) before interest and % after tax**
2. Terdapat variabel yang memiliki *noise* atau outlier, namun masih masuk akal → **cash flow rate**
3. Terdapat variabel yang memiliki noise atau outlier dan mulai tidak wajar → **cash/current liabilitiy**
4. Terdapat variabel yang memiliki *noise* atau outlier dan "rusak" → **r&d expense rate; cash turnover rate**

Screening

Mengecek nilai mean dan mediannya

↓

Cek distribusi data menggunakan Boxplot

↓

Cek distribusi data (lagl) menggunakan persentil ke-99 (Lihat Appendix No.1)

↓

Cek persentase nilai outlier dan nilai di luar persentil ke-99

Outliers: Noise, Anomaly, or Valuable Insight?

Detail Proses *Handle* Outlier

Seluruh fitur (Data Asli)

Cek Nilai Mean dan Median

- Jika data ditemukan memiliki mean dan median yang **tidak wajar**, maka fitur atau variabel langsung di-*drop*.
- Namun jika median masih masuk akal, maka masuk ke tahap berikutnya, yakni mengecek distribusi datanya.

Cek Persentil ke-99 dan Persentase Outlier

- Jika setelah pengecekan distribusi dengan persentil ke-99, fitur tidak memiliki pola terhadap kebangkrutan dan boxplot tetap tidak tergambar, fitur akan di-*drop*.
- Jika memiliki pola terhadap target dan persentase nilai di atas persentil ke-99 kecil, maka dilakukan *capping* (pembatasan) dengan nilai persentil ke-99

Siap Dimodelkan

- Data yang telah dibatasi dengan nilai persentil ke-99 adalah fitur yang memiliki anomali sebesar 1% dari keseluruhan data
- Alasan tidak membuang baris data/obeservasi, ditakukan akan menghilangkan target perusahaan yang bangkrut, mengingat secara distribusi target tidak seimbang. Ada 76 variabel (fitur) dan 1 target

Predicting Bankruptcy: Key Variables

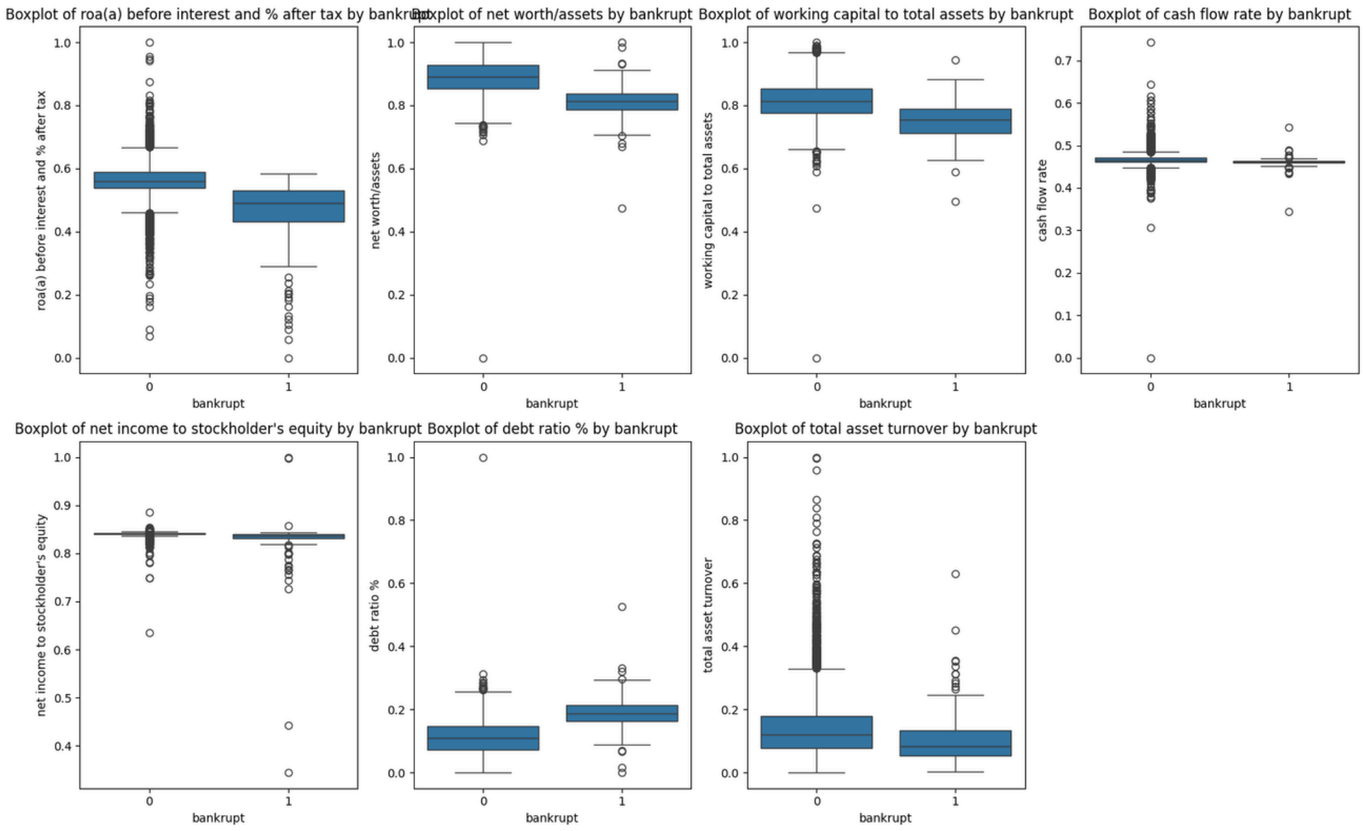
Hipotesis: Variabel yang (Mungkin) Memiliki Pengaruh Terhadap Kebangkrutan Perusahaan

Fitur yang Dicek Distribusinya

1. roa(a) before interest and % after tax (Profitabilitas)
2. net worth/assets (Struktur Modal / Solvabilitas)
3. working capital to total assets (Likuiditas)
4. cash flow rate (Arus Kas)
5. net income to stockholder's equity (ROE - Profitabilitas)
6. debt ratio % (Leverage)
7. total asset turnover (Efisiensi)

Sumber: Chi, 2023; Kücher et al., 2018; Ogachi et al., 2020

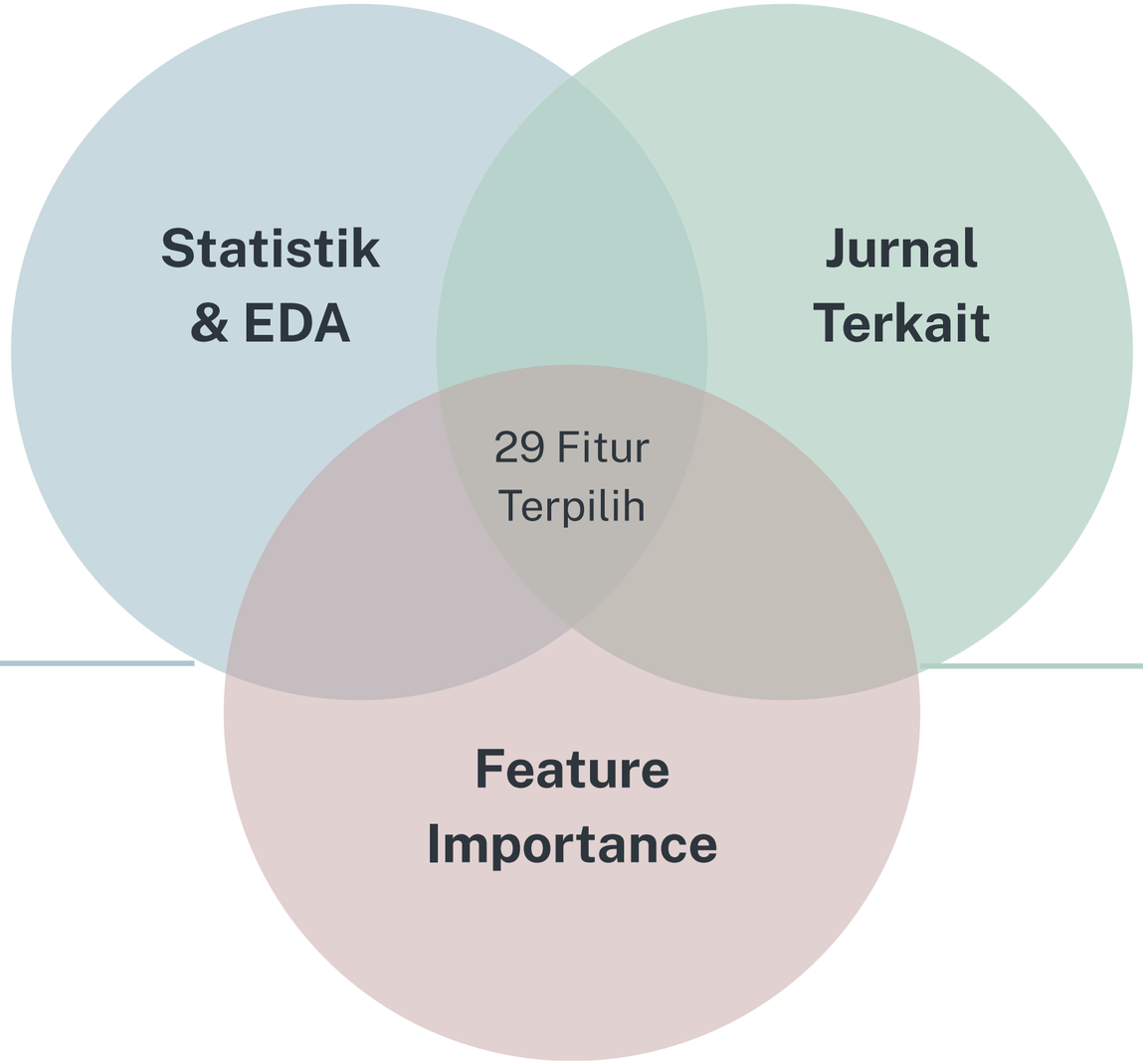
Distribusi Data dengan Boxplot



Feature Selection

Tiga Pendekatan dalam Pemiliha Fitur untuk Dimodelkan

- Menggunakan uji statstik (ANOVA/F-stat)
- Menggunakan nilai korelasi. Jika dalam variabel dalam satu faktor memiliki korelasi kuat, maka akan dipilih salah satu



- Menggunaka beberapa sumber terkait, seperti Chi, 2023; Kücher et al., 2018; Ogachi et al., 2020
- Beberapa variabel yang disebutkan antara lain, ROA, ROE, DER, dan Quick Ratio

Result of Modelling

Modelling Flow

Data Modelling

Membandingkan model pada data latih

80:20

Data Tanpa Scaled → **Split Data** → **Experiment Tanpa Resampling** / **Experiment dengan SMOTE**

80:20

Data dengan Scaled → **Split Data** → **Experiment Tanpa Resampling** / **Experiment dengan SMOTE**

Memilih Salah Satu Model Terbaik

- Berdasarkan metrik f1-score dan recall

Evaluation & Tuning

Model Interpretation & Feature Importance

Re-fit dengan fitur pilihan berdasarkan hasil dari proses *feature selection*

Details of Experiments: The Flow and The Reason for Choosing The Metrics

Detail Pemodelan

- Ada 10 model yang dibandingkan dan terpilih 5 model untuk dievaluasi terhadap data yang tidak pernah dilihat
- Lima model dipilih berdasarkan metrik Recall dan F1-score
- Adapun, 5 model terpilih antara lain, Gradient Boosting (gbc), AdaBoost (ada), XGBoost (xgb), Extra Trees (et), dan Random Forest (rf)
- Model dilatih terpisah untuk masing-masing data scaled dan tidak scaling
- Masing-masing data tidak scaled (lihat appendix no. 3 dan no. 4) dan data scaled dibandingkan (lihat appendix no. 4 dan no. 5).
- Hasil data SMOTE menghasilkan metrik yang dibutuhkan (Recall dan F1-score)

Kenapa Memilij Recall?

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- TP: True Positive
- FN: False Negative

Key Takeaways:

- Ingin memimilkan False Negative karena **biaya kehilangan sangat tinggi**
- **Mengurangi potensi kecolongan** perusahaan yang sebenarnya bankrupt diprediksi tidak bankrupt oleh model
- Sejalan dengan tujuan **untuk pencegahan atau preventif**

Kenapa Memilih F1-score?

$$\text{F1 Score} = \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}$$

- TP: True Positive
- FN: False Negative
- FP: False Positive

Key Takeaways:

- Sebagai ukuran penyeimbang, F1-score **memberikan gambaran** yang **seimbang kinerja model**
- Penting sebagai bahan evaluasi apalagi untuk kasus data tidak seimbang

Result of Candidate of Best Model

Evaluasi Model pada Data Train dan Data Test (SMOTE)

Train Set					
Model	Accuracy	Recall	Prec.	F1	AUC
Gradient Boosting Classifier	0.956242	0.981815	0.934042	0.957333	0.989619
AdaBoost Classifier	0.894582	0.906801	0.88517	0.895855	0.957575
Extra Trees Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
XGBoost Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
Random Forest Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
Test Set					
Model	Accuracy	Recall	Prec.	F1	AUC
Gradient Boosting Classifier	0.907625	0.750000	0.222973	0.343750	0.938602
AdaBoost Classifier	0.883431	0.795455	0.189189	0.305677	0.922435
Extra Trees Classifier	0.955279	0.636364	0.383562	0.478632	0.920506
XGBoost Classifier	0.954545	0.636364	0.378378	0.474576	0.945937
Random Forest Classifier	0.946481	0.590909	0.320988	0.416000	0.935442

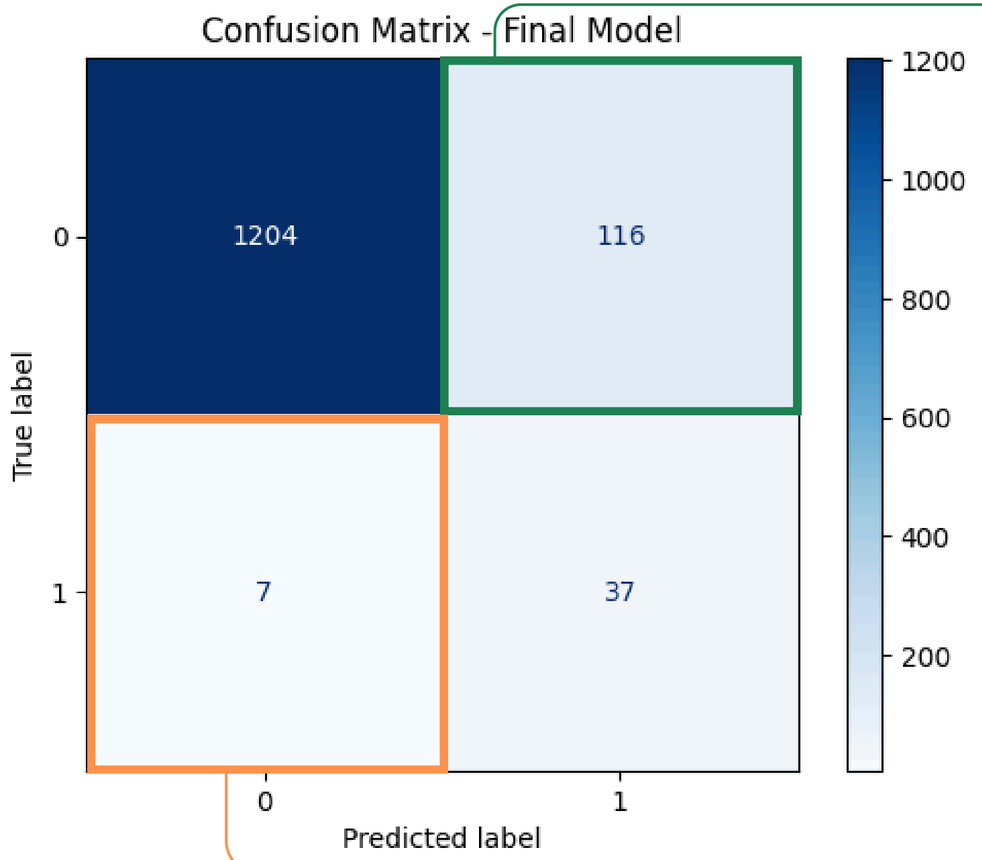
Train Set (Scaled)					
Model	Accuracy	Recall	Prec.	F1	AUC
Gradient Boosting Classifier	0.952927	0.975185	0.933624	0.953952	0.990018
AdaBoost Classifier	0.901686	0.918545	0.888583	0.903316	0.958086
Extra Trees Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
XGBoost Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
Random Forest Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
Test Set (Scaled)					
Model	Accuracy	Recall	Prec.	F1	AUC
Gradient Boosting Classifier	0.910557	0.772727	0.232877	0.357895	0.940255
AdaBoost Classifier	0.878299	0.818182	0.185567	0.302521	0.920136
Extra Trees Classifier	0.953812	0.636364	0.373333	0.470588	0.917588
XGBoost Classifier	0.947214	0.568182	0.320513	0.409836	0.939824
Random Forest Classifier	0.950880	0.613636	0.350649	0.446281	0.929735

Key Takeaways:

- **Model XGBoost Classifier** dengan resampling (tanpa scaled) dipilih sebagai **final model** untruk di-*deploy* atau digunakan.
- XGBoost lebih modern dan lebih cepat dibandingkan AdaBoost, Gradient Boosting, dan Extra Trees, apalagi jika berhadap dengan tuning parameter.
- **Recall dan F1-score tertinggi kedua** dari seluruh model.
- Data tanpa scale memudahkan untuk intepretasi angka dalam model karena masih menggunakan data asli. Selain itu, Nilai F1 score paling tinggi dalam evaluasi menggunakan data resampling (SMOTE) tanpa scaling.
- XGBoost telah dituned dan berhasil meningkatkan Recall, namun ada *trade-off* dengan penuruna Precision yang menyebabkan F1-score juga ikut menurun.

Confusion Matrix and Result After Tuned

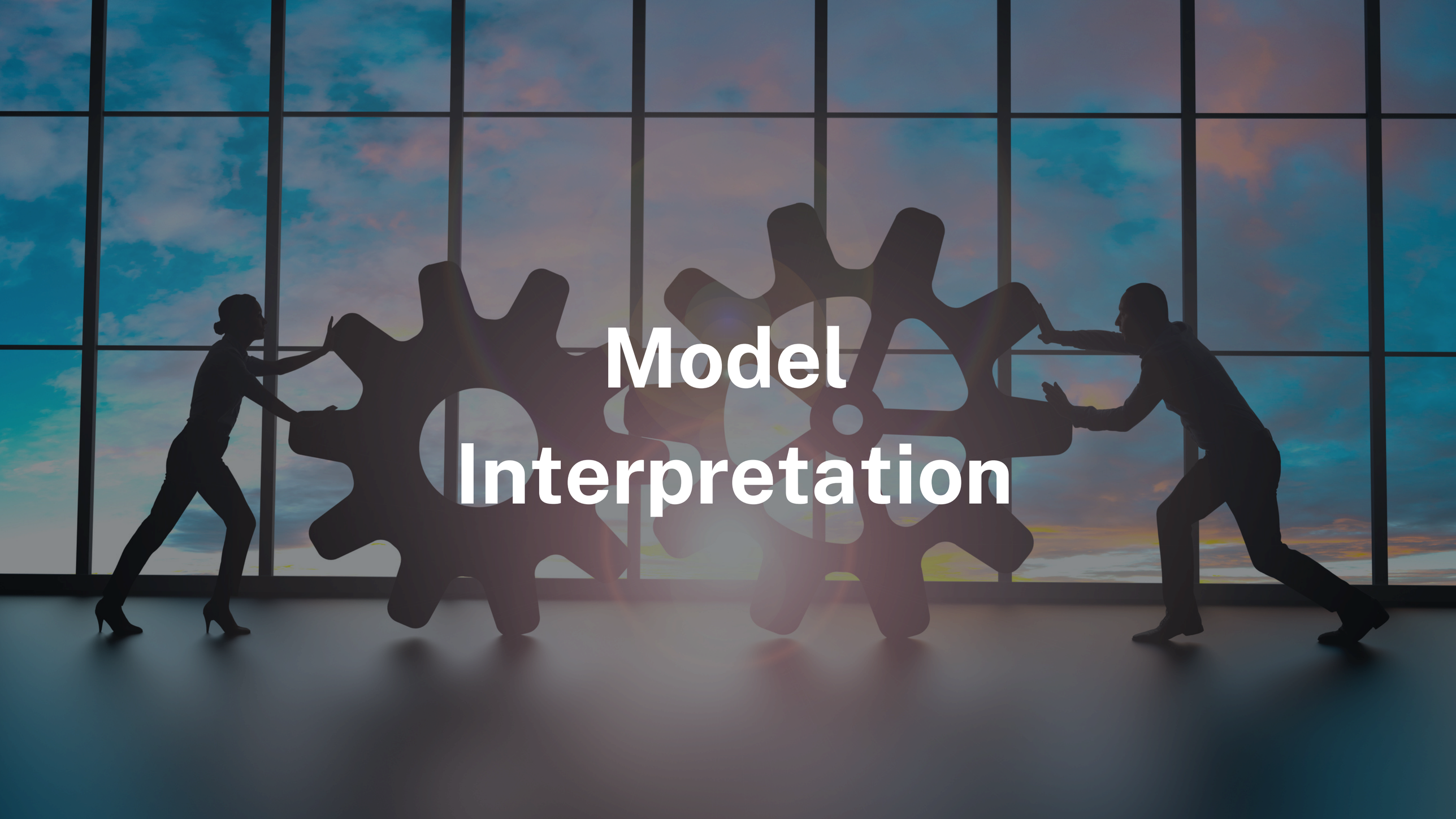
Confusion Matrix of XGBoost Tuned Model



- Ini adalah false positive, nilai 116 adalah nilai yang diprediksi bangkrut, namun sebenarnya tidak tidak bangkrut
- Ini penyebab Precision nilainya rendah dan menyebabkan keseimbangan F1-score nya berkurang
- Hal ini karena model di tuning untuk memperkuat hasil Recall

- Ini adalah false negative, nilai 7 ada nilai yang diprediksi tidak bangkrut padahal sebenarnya bangkrut.
- Nilai 7 sangat rendah sehingga nilai Recall sangat tinggi, berdasarkan hasil evaluasi nilai Recall setelah tuning sebesar 84%.
- Metrik ini coba dimaksimalkan untuk mengurangi kesalahan prediksi

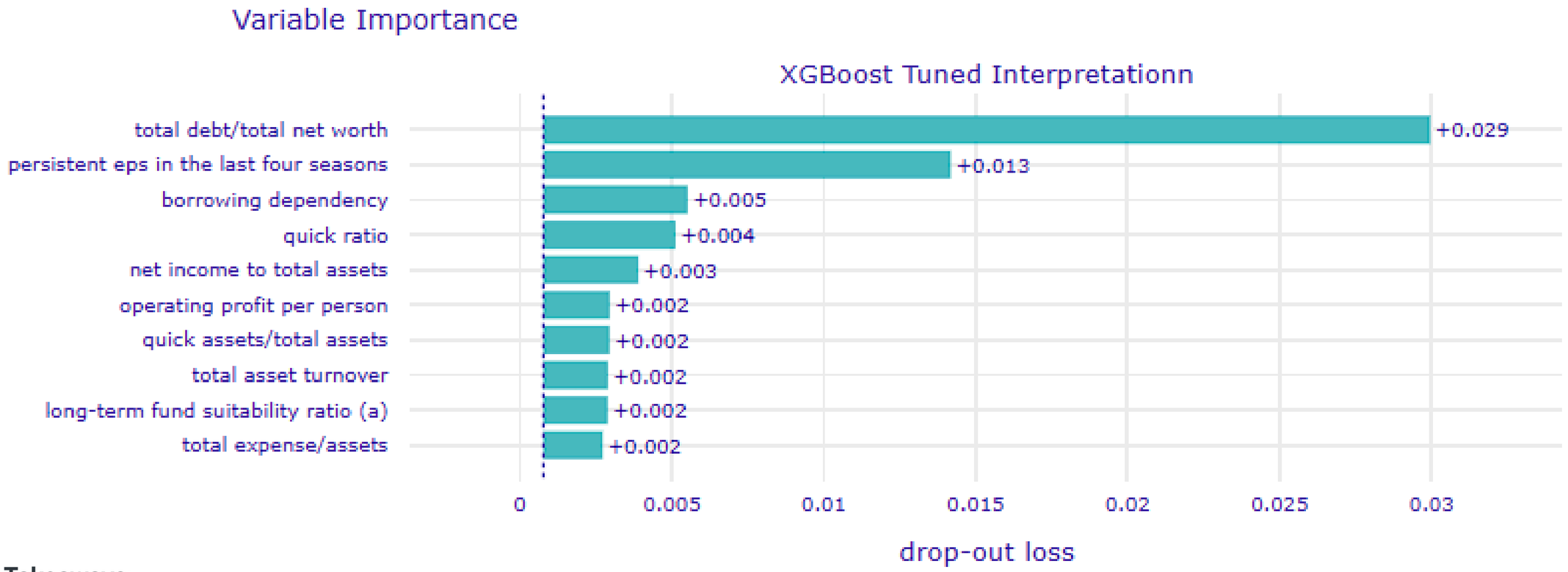
Note: Model Setalah tuned berhasil meningkatkan Recall (semakin baik), namun ada trade-off dengan Precision (lihat appendix no. 5 dan appendix no. 6)



Model Interpretation

Feature Importance

Top 10 Features

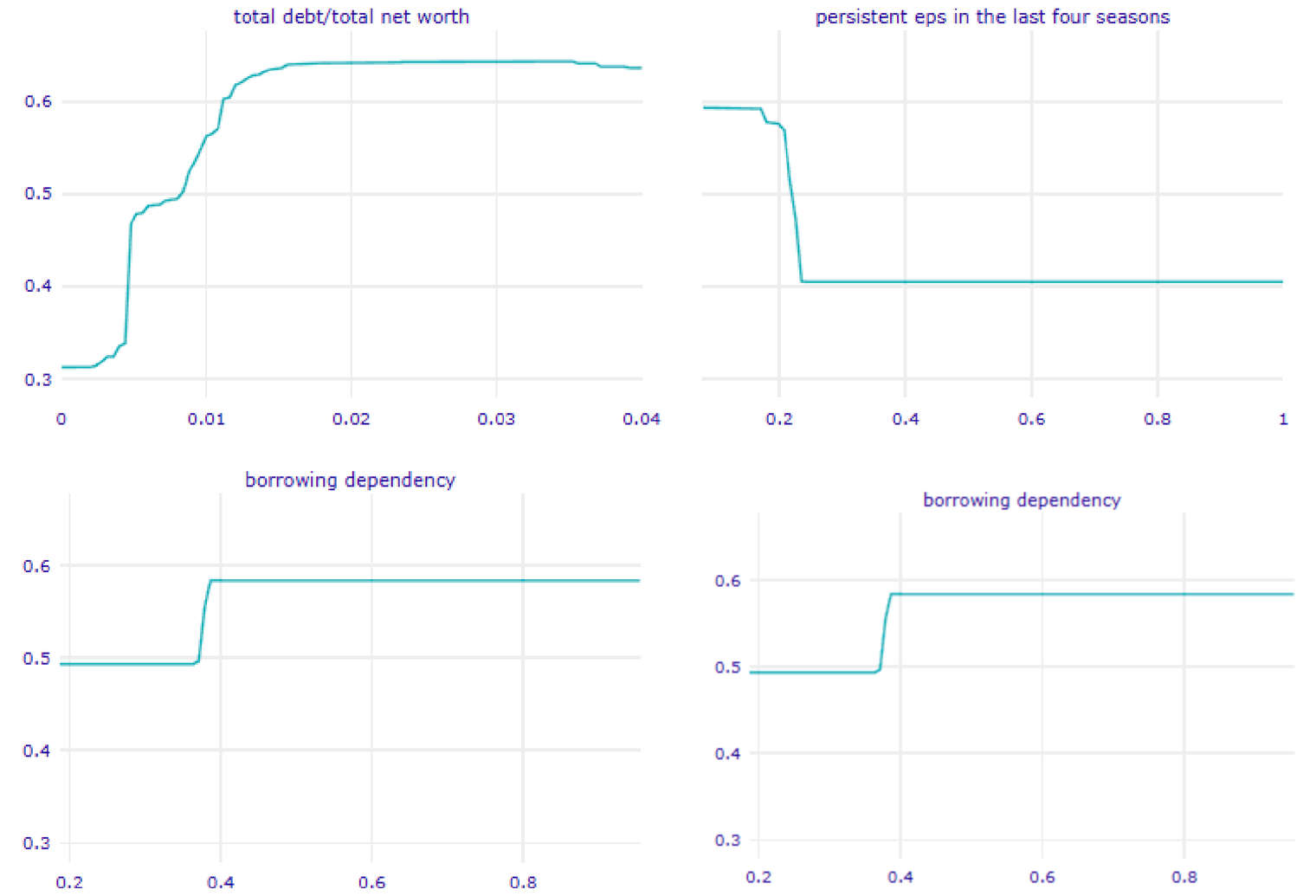


Key Takeaways:

- .**Variabel total debt/total net worth terbukti menjadi fitur yang paling penting**, di mana pengacakan nilainya menyebabkan performa model menurun paling signifikan (peningkatan loss sebesar 0.029).
- **Hasil ini mengonfirmasi bahwa faktor leverage atau solvabilitas adalah salah satu prediktor kunci** dalam menentukan risiko kebangkrutan perusahaan.
- Dalam 10 fitur paling penting ini, terdapat faktor profitabilitas seperti ROA (net income to total assets), faktor likuiditas seperti quick ratio, dan lain sebagainya

Interpretation Model and Recommendation for Business

Partial Dependence Plot



Tota debt / total net worth (Leverage):

- Risiko kebangkrutan **meningkat tajam saat rasio utang terhadap modal bersih naik dari 0% ke 1.5%**. Setelah melewati 1.5%, risiko tetap sangat tinggi (di atas 60%).

persistent eps in the last four seasons (Profitabilitas):

- Terdapat ambang batas tajam di sekitar nilai 0.2. Di atas nilai EPS ini, risiko bangkrut dianggap sangat rendah (sekitar 40%). Di bawah nilai ini, risiko melonjak signifikan (mendekati 60%).

net income to total assets (Profitabilitas):

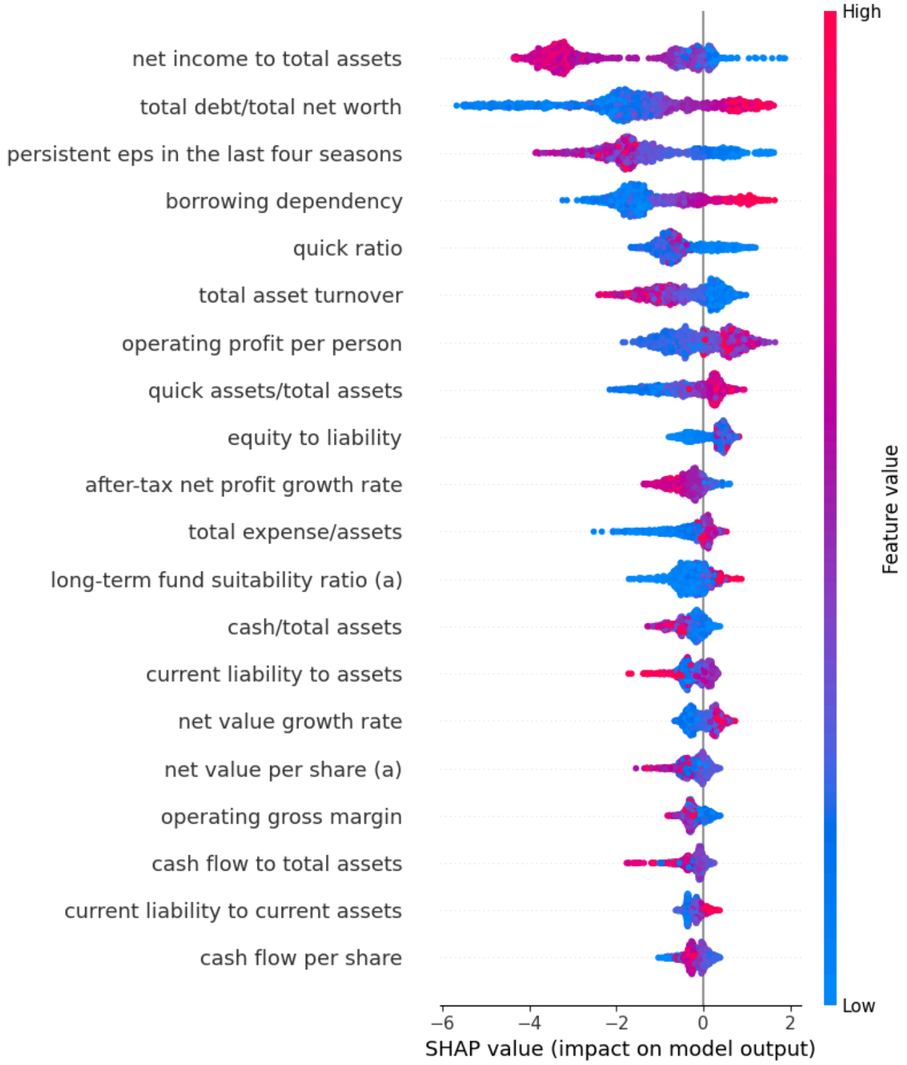
- Mirip dengan EPS, namun penurunannya lebih bertahap. **Risiko mulai meningkat saat rasio laba bersih terhadap aset turun di bawah 0.8** dan menjadi sangat tinggi saat turun di bawah 0.7.

borrowing dependency (Leverage):

- Ada "lompatan" risiko yang jelas di sekitar nilai 0.4. Perusahaan dengan **tingkat ketergantungan pada pinjaman di atas 40%** dianggap **memiliki probabilitas bangkrut yang lebih tinggi** (mendekati 60%) oleh model.

Interpretation Model and Recommendation for Business

Shapley Value Graph (beeswarm plot)



How to read a beeswarm plot:

- Sumbu-x memberikan dampak prediksi.
  - Nilai Positif (>0): Mendorong prediksi ke arah bangkrut.
  - Nilai Negatif (<0): Mendorong prediksi ke arah tidak bangkrut.
- Warna titik
  - Merah: Nilai fitur tinggi.
  - Biru: Nilai fitur rendah.

Interpretasi net income to total assets (Profitabilitas)

- Profitabilitas yang tinggi adalah tanda kuat perusahaan sehat. Ini mengonfirmasi bahwa **semakin tinggi profitabilitas, semakin berkurang potensi kebangkrutan** sebuah perusahaan.
- Titik-titik biru (nilai rendah) tersebar luas di sisi kanan (mendorong ke arah bangkrut), sementara titik merah (nilai tinggi) menumpuk di sisi kiri (mendorong ke arah tidak bangkrut).

Interpretasi total debt/total net worth (Leverage)

- Tingkat utang yang tinggi secara konsisten diidentifikasi oleh model sebagai faktor risiko utama kebangkrutan. Ini juga mengkonfirmasi, **bahwa semakin tinggi fitur ini, maka mendorong potensi kebangkrutan perusahaan.**
- Titik merah (nilai tinggi) tersebar di sisi kanan (mendorong ke arah bangkrut), dan titik biru (nilai rendah) menumpuk di sisi kiri (menjauhi kebangkrutan).

Demo and Conclusion

Recommendation and Room for Imporvement

Recommendation

1. Nilai atau variabel pada fa**ktor leverage sangat mempengaruhi potensi kebangkuratan perusahaan**, kemudian **dilengkapi dengan variabel** yang ada di faktor **profitabilitas dan likuiditas.**
2. Hipotesis terkait faktor atau variabel yang termasuk dalam faktor terbukti, misalnya **jika profitabilitas naik maka kemungkinan potensi kebangkrutan akan turun**. Kemudian, jika leverage naik atau ketergantungan utang naik maka tingkat potensi kebangkrutan juga akan ikut naik.
3. **Menjaga tingkat ketergantungan utang di bawah 40**%, pun **jika harus melebihi nilai tersebut** pastikan **profitabilitas seperti ROA mencukupi setidaknya di angka 0.6-0.7 poin.**
4. Dengan kata lain, perusahaan mesti hati-hati jika dalam melakukan utang-piutang. Perlu melihat kemampuan perusahaan dalam menghasilkan profit, kondisi aset, dan kondisi likuiditas.

Room for Improvement

1. **Mengeskplorasi dan memilih variabel yang penting mungkin terlewat** atau variabel yang kurang penting, namun terlanjur dimasukkan dalam pemodaln.
2. **Mengeskplorasi teknik *preprocessing*** untuk dilakukkan eksperimen, misalnya mencona menghapus atau mentransformasi nilai outlier
3. **Mencoba model lain**, utamanya dalam deep learning, seperti Neural Network dengan Tensorflow dan sebagainya
4. Mencari atau menambah observasi untuk model lebih banyak belajar, utamanya untuk observasi perusahaan yang dikatakan bankrut agar ketidakseimbangan kelas berkurang.
5. **Mengaitkan atau melibatkan expert atau bahan penelitian yang lebih beragaman di masa depa**n.

Model yang dideploy bisa dikases atau digunakan di link berikut:
Streamlit App

References

References

Chi, K. (2023). Company bankruptcy prediction. Advances in Economics, Management and Political Sciences, 22, Article 20230282. https://doi.org/10.54254/2754-1169/22/20230282

Kücher, A., Mayr, S., Mitter, C., Duller, C., & Feldbauer-Durstmüller, B. (2018). Firm age dynamics and causes of corporate bankruptcy: age dependent explanations for business failure. Review of Managerial Science, 14, 633–661. https://doi.org/10.1007/S11846-018-0303-2

Ogachi, D., Ndege, R., Gaturu, P., & Zoltan, Z. (2020). Corporate bankruptcy prediction model, a special focus on listed companies in Kenya. Journal of Risk and Financial Management, 13(3), 47. https://doi.org/10.3390/jrfm13030047

USAFacts team. (2023, June 8). Bankruptcy in the US: how many companies file for Chapter 7, Chapter 11, and Chapter 13? USAFacts. https://usafacts.org/articles/bankruptcy-in-the-us-how-many-companies-file-for-chapter-7-chapter-11-and-chapter-13/

CNN Indonesia. (2025, January 30). Daftar perusahaan terkenal yang bangkrut di Indonesia. CNN Indonesia. https://www.cnnindonesia.com/ekonomi/20250130072420-92-1192579/daftar-perusahaan-terkenal-yang-bangkrut-di-indonesia

Khairally, E. T. (2025, March 15). 8 perusahan yang bangkrut hingga mengakibatkan PHK massal di Indonesia. detikFinance. https://finance.detik.com/berita-ekonomi-bisnis/d-7823172/8-perusahan-yang-bangkrut-hingga-mengakibatkan-phk-massal-di-indonesia

Subari, W. A. (2025, August 19). 12 perusahaan dunia bangkrut di 2024: Kisah & penyebabnya. Media Indonesia. https://mediaindonesia.com/ekonomi/802635/12-perusahaan-dunia-bangkrut-di-2024-kisah--penyebabnya

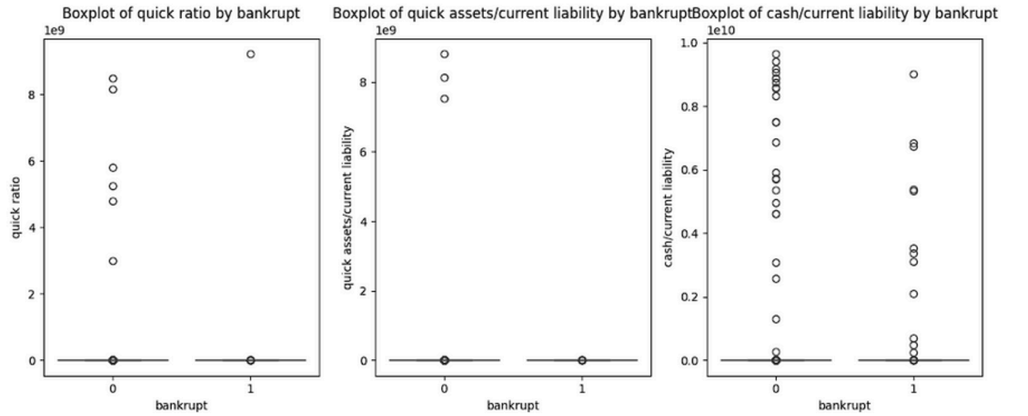
Appendix

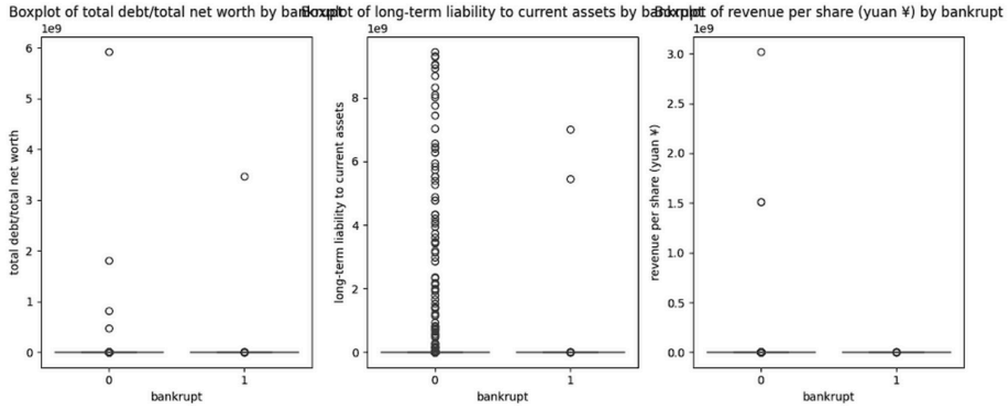
Appendix No. 1

Outliers: Noise, Valuable Insight, or Anomaly?

Mengapa Distribusi Data Dilihat Dua kali?

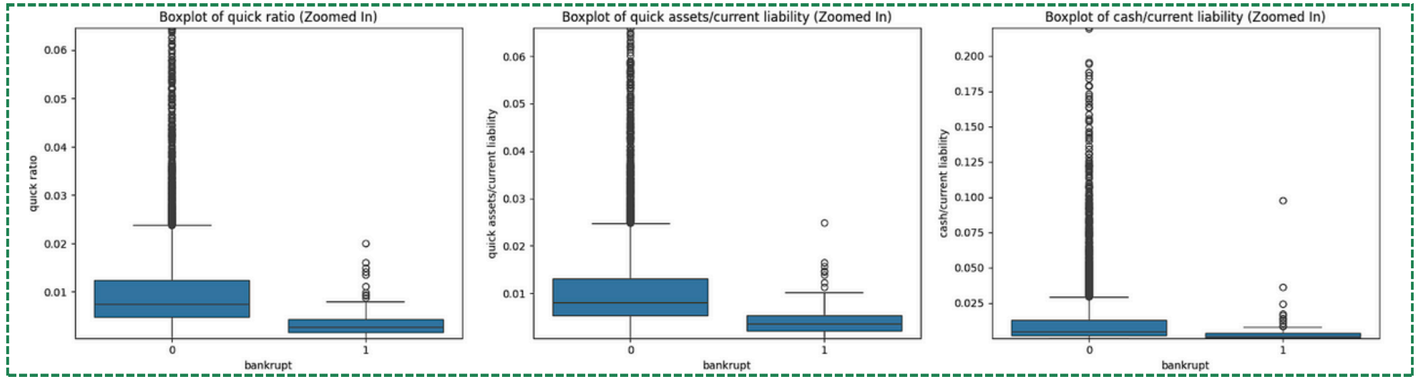
Menggunakan seluruh data

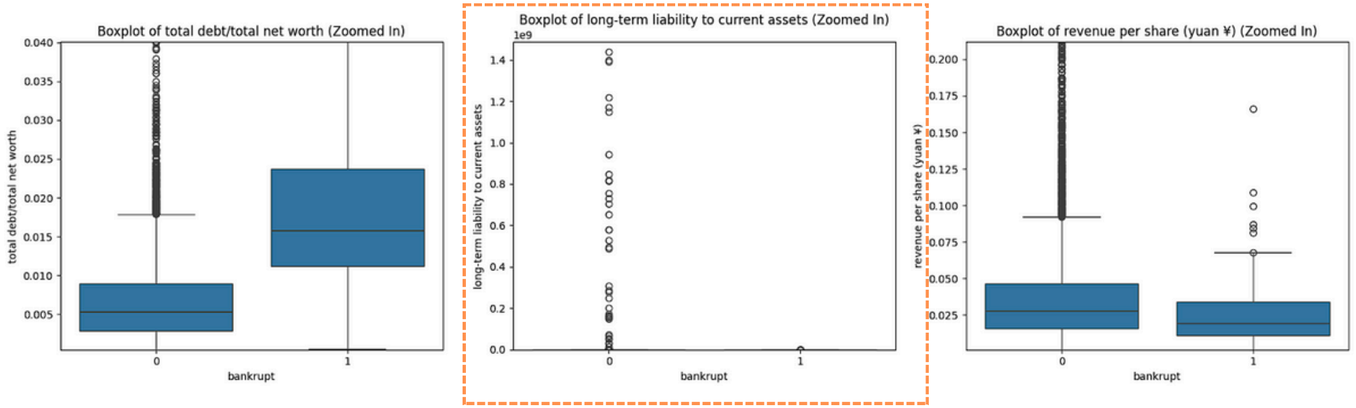




Note: Seluruh Data

Menggunakan data di persentil ke-99





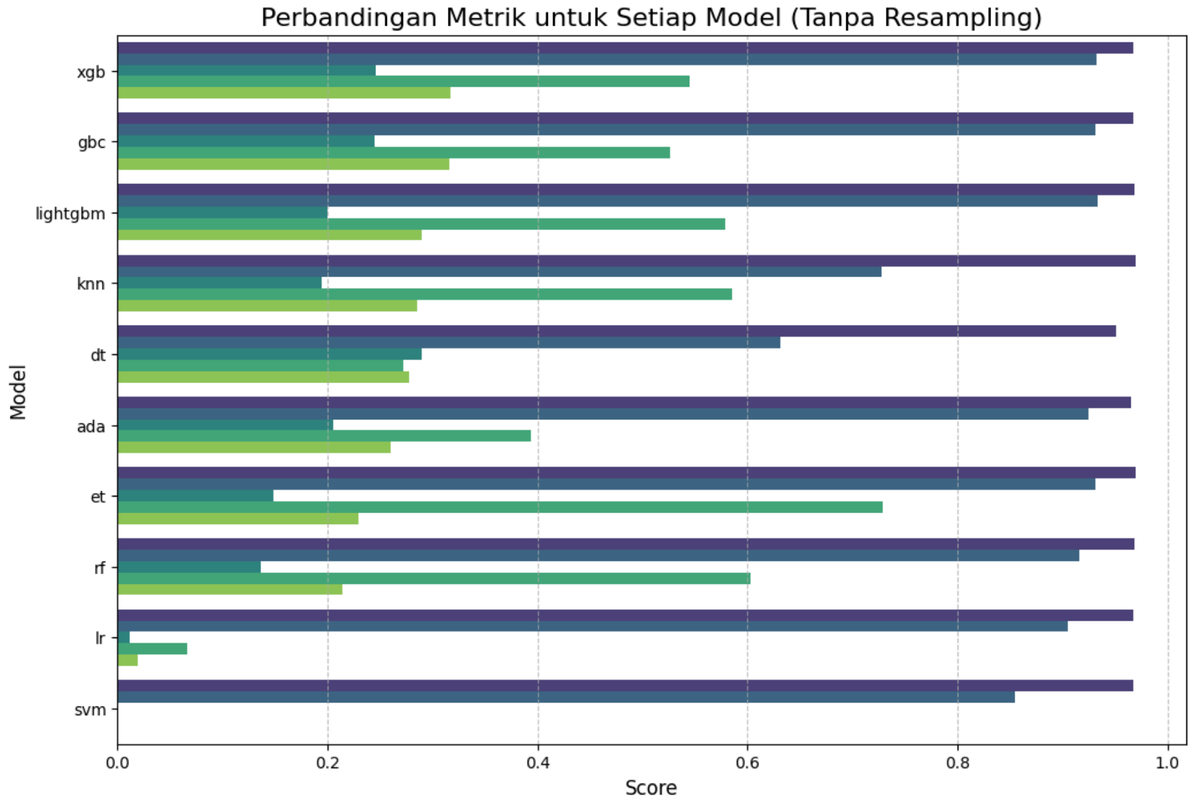
Note: Data dari persentil ke-1st sampai ke-99th

Appendix No. 2

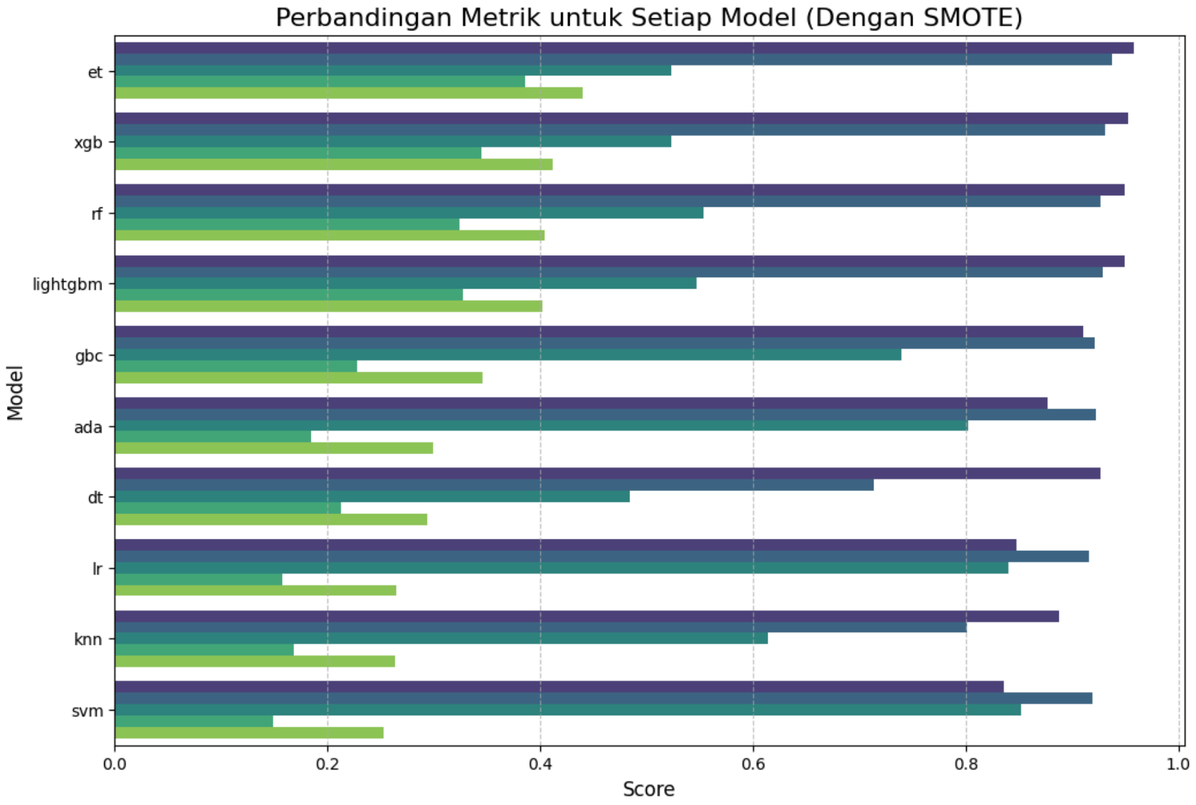
Compare the Result of the Experiment on Not Scaled Data

Hasil pada Data Tanpa Scale

Eksperimen Tanpa *Resampling* (SMOTE)



Eksperimen dengan SMOTE



Metric
- Accuracy
- AUC
- Recall
- Prec.
- F1

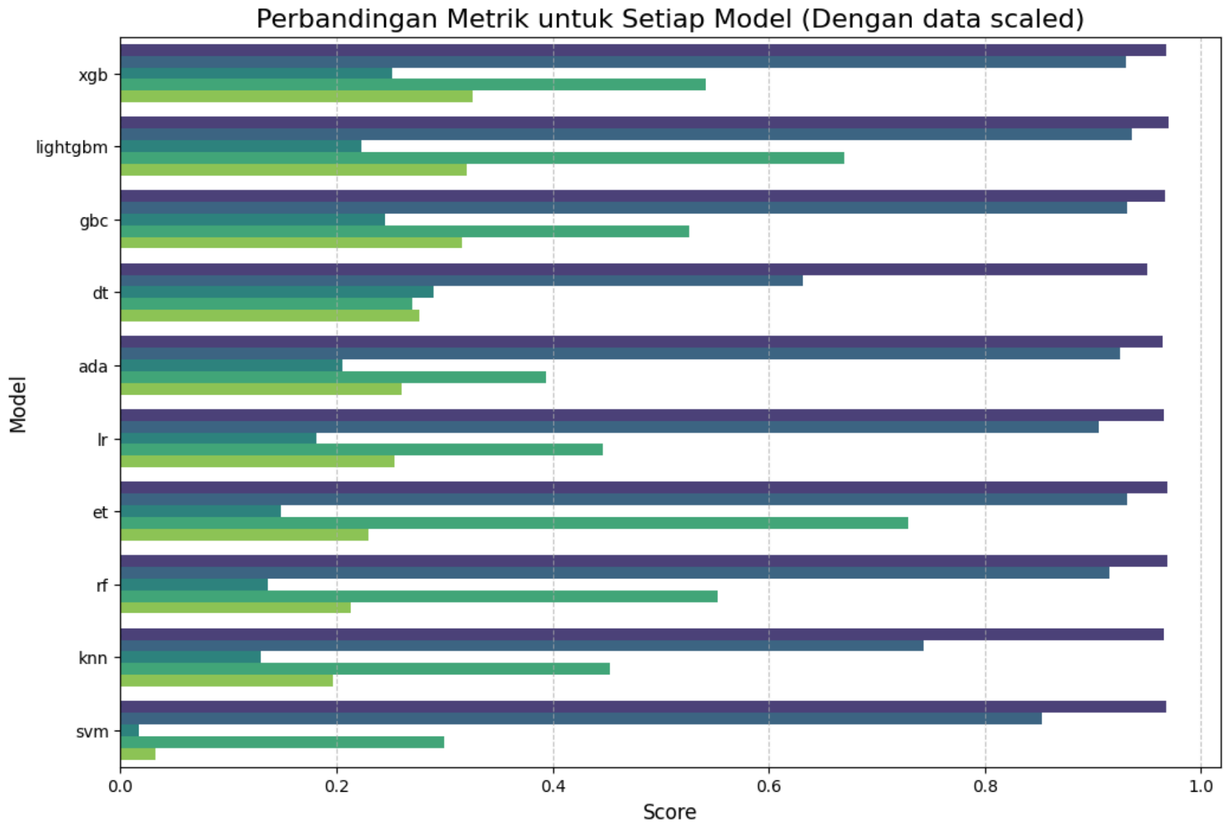
Note:
- et: Extra Trees Classifier
- xgb: XGBoost Classifier
- lightgbm: LightGBM Classifier
- rf: Random Forest Classifier
- gbc: Gradient Boosting Classifier
- ada: AdaBoost Classifier
- lr: Logistic Regression
- knn: K-Nearest Neighbors
- dt: Decision Tree Classifier
- svm: Support Vector Machine

Appendix No. 3

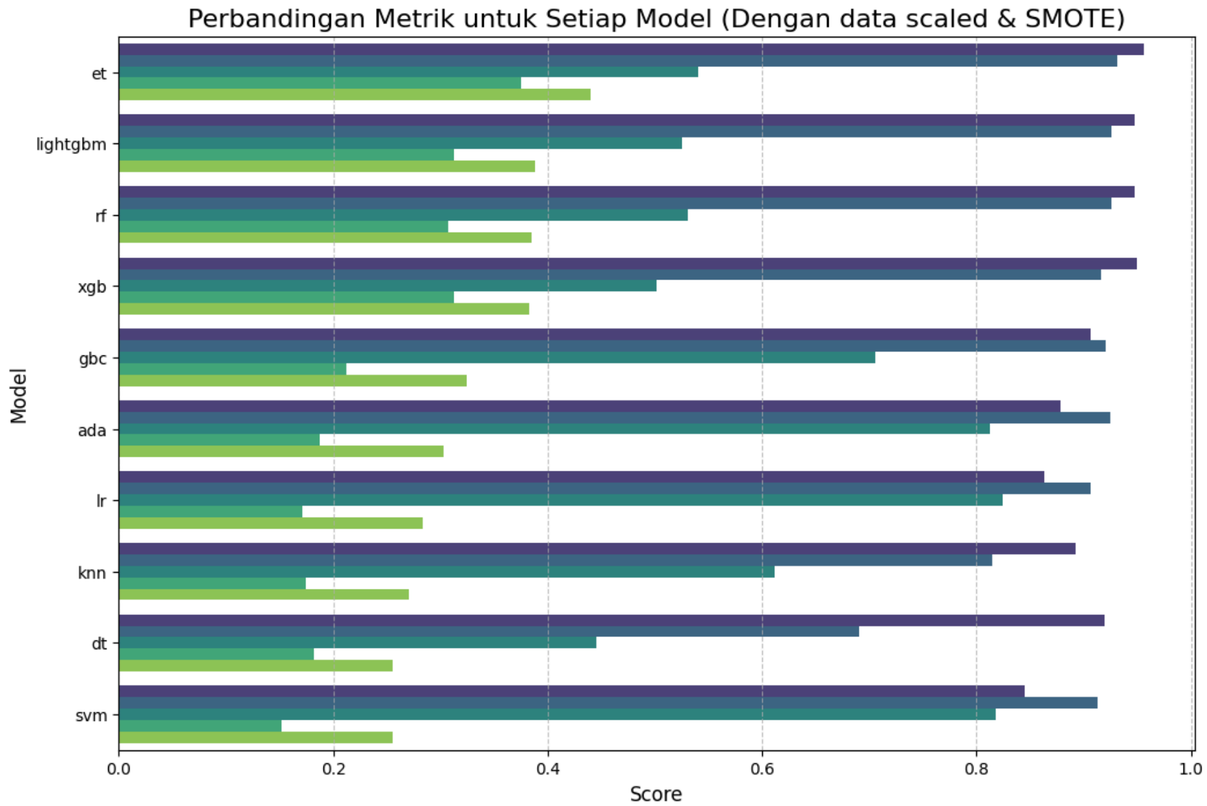
Compare the Result of the Experiment on Scaled Data

Data Scale

Eksperimen Tanpa *Resampling* (SMOTE)



Eksperimen dengan SMOTE



Metric
- Accuracy
- AUC
- Recall
- Prec.
- F1

Note:

- et: Extra Trees Classifier
- xgb: XGBoost Classifier
- lightgbm: LightGBM Classifier
- rf: Random Forest Classifier
- gbc: Gradient Boosting Classifier
- ada: AdaBoost Classifier
- lr: Logistic Regression
- knn: K-Nearest Neighbors
- dt: Decision Tree Classifier
- svm: Support Vector Machine

Appendix No. 4

Table Scores or Metrics Models

Data Tanpa Scale

Hasil Fit Cross Validation Data No Scaling and No Resampling					
Model	Accuracy	AUC	Recall	Prec. (Precision)	F1
xgb	0.967551	0.932098	0.245752	0.544268	0.317322
gbc	0.967186	0.931589	0.245098	0.526105	0.316124
lightgbm	0.969020	0.933023	0.200000	0.578929	0.290050
knn	0.969753	0.728056	0.194118	0.585238	0.284886
dt	0.951051	0.631322	0.289542	0.272415	0.278116
ada	0.964801	0.924793	0.205229	0.393864	0.260349
et	0.969385	0.931721	0.148039	0.729048	0.229706
rf	0.969019	0.915630	0.136928	0.602857	0.213886
lr	0.967553	0.904693	0.011111	0.066667	0.019048
svm	0.967737	0.854458	0.000000	0.000000	0.000000
Hasil Fit Cross Validation Data No Scaling and Resampling (SMOTE)					
Model	Accuracy	AUC	Recall	Prec. (Precision)	F1
et	0.958019	0.937187	0.523203	0.386139	0.440031
xgb	0.952519	0.930655	0.523529	0.344317	0.412044
rf	0.949034	0.92654	0.552941	0.323652	0.404423
lightgbm	0.949034	0.928237	0.547386	0.326954	0.401777
gbc	0.909808	0.92094	0.739869	0.227279	0.346037
ada	0.876987	0.921988	0.801634	0.185056	0.299273
dt	0.926485	0.71296	0.484641	0.212942	0.293528
lr	0.847101	0.91612	0.840196	0.157469	0.26478
knn	0.887983	0.801577	0.613725	0.16858	0.263462
svm	0.835922	0.918619	0.852288	0.148516	0.252635

Data dengan Scaling

Hasil Fit Cross Validation Data Scaling and No Resampling					
Model	Accuracy	AUC	Recall	Prec. (Precision)	F1
xgb	0.967734	0.930367	0.250980	0.541761	0.325685
lightgbm	0.970118	0.935871	0.222549	0.669859	0.320118
gbc	0.967186	0.931589	0.245098	0.526105	0.316124
dt	0.950684	0.631133	0.289542	0.270515	0.277096
ada	0.964801	0.924793	0.205229	0.393864	0.260349
lr	0.966085	0.905861	0.181699	0.446667	0.253193
et	0.969385	0.931748	0.148039	0.729048	0.229706
rf	0.968836	0.915689	0.136928	0.552857	0.213360
knn	0.965718	0.743625	0.129412	0.452619	0.196860
svm	0.968287	0.852687	0.017320	0.300000	0.032749
Hasil Fit Cross Validation Data Scaling and Resampling (SMOTE)					
Model	Accuracy	AUC	Recall	Prec. (Precision)	F1
et	0.956186	0.931659	0.54085	0.374864	0.440362
lightgbm	0.947935	0.925939	0.524837	0.312291	0.387951
rf	0.94757	0.925475	0.530719	0.30767	0.385418
xgb	0.94977	0.91604	0.501961	0.312583	0.382611
gbc	0.906137	0.919991	0.705556	0.212386	0.324861
ada	0.87864	0.924814	0.812418	0.18705	0.303278
lr	0.863602	0.906309	0.823856	0.171407	0.283107
knn	0.892755	0.81435	0.611765	0.174504	0.270641
dt	0.91897	0.689968	0.445098	0.181808	0.255739
svm	0.844357	0.912768	0.818301	0.151267	0.254894

Appendix No. 5

Metrics After Tuning

Metrics and Score XGBoost Tuned

Train Set					
Model	Accuracy	Recall	Prec.	F1	AUC
Gradient Boosting Classifier	0.956242	0.981815	0.934042	0.957333	0.989619
AdaBoost Classifier	0.894582	0.906801	0.88517	0.895855	0.957575
Extra Trees Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
XGBoost Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
Random Forest Classifier	1.000000	1.000000	1.000000	1.000000	1.000000
Test Set					
Model	Accuracy	Recall	Prec.	F1	AUC
Gradient Boosting Classifier	0.907625	0.750000	0.222973	0.343750	0.938602
AdaBoost Classifier	0.883431	0.795455	0.189189	0.305677	0.922435
Extra Trees Classifier	0.955279	0.636364	0.383562	0.478632	0.920506
XGBoost Classifier	0.954545	0.636364	0.378378	0.474576	0.945937
Random Forest Classifier	0.946481	0.590909	0.320988	0.416000	0.935442

Result After Tuned

Trial	Recall	F1-Score	AUC
Trial 8	0.8409	0.3756	0.9418
Trial 79	0.7954	0.3414	0.9417
Trial 99	0.8181	0.3529	0.9449

Appendix No. 6

Set of Hyperparameters Tuning


```
{
    "trial_8": {
        "recall": 0.7841269841269841,
        "f1_score": 0.347634196181214,
        "params": {
            "n_estimators": 581,
            "learning_rate": 0.03330107024564265,
            "max_depth": 6,
            "subsample": 0.60050570929574,
            "colsample_bytree": 0.9640340910014457,
            "gamma": 0.09433674579200935,
            "reg_alpha": 0.18088228614280707,
            "reg_lambda": 13.23592116016826,
            "min_child_weight": 10,
            "scale_pos_weight": 49.74867788658356
        }
    },
    "trial_79": {
        "recall": 0.7953968253968254,
        "f1_score": 0.33196496712061113,
        "params": {
            "n_estimators": 632,
            "learning_rate": 0.027634773618341508,
            "max_depth": 7,
            "subsample": 0.7201626572287096,
            "colsample_bytree": 0.7280835752953629,
            "gamma": 0.4852182650160797,
            "reg_alpha": 13.089096821800045,
            "reg_lambda": 0.006940281348331692,
            "min_child_weight": 2,
            "scale_pos_weight": 71.46628773344582
        }
    },
    "trial_99": {
        "recall": 0.8180952380952382,
        "f1_score": 0.3310842253496428,
        "params": {
            "n_estimators": 383,
            "learning_rate": 0.07802498659583937,
            "max_depth": 9,
            "subsample": 0.9276250041336165,
            "colsample_bytree": 0.8654476657562518,
            "gamma": 0.5419571076372606,
            "reg_alpha": 35.202909359800984,
            "reg_lambda": 4.5648180388438593e-07,
            "min_child_weight": 8,
            "scale_pos_weight": 1.9465854298118055
        }
    }
}
```