Survivorship Bias in Data Science

Abstract

Survivorship bias—focusing only on entities that "survive" a selection process—can catastrophically distort conclusions. This project revisits Abraham Wald's WWII bomber study, where accounting for non-returning aircraft led to optimal armor placement, and extends its lessons to three modern domains: business dynamics, higher-education outcomes, and financial markets. Historical bomber records alongside BLS turnover data, public university graduation rates, and S&P 500 survivorship lists with price histories were compiled. Through exploratory data analysis, hypothesis testing, supervised learning, and clustering, it was demonstrated that including the "failures" yields robust, generalizable insights, whereas analysis limited to survivors produce over-optimistic narratives. This work combines concrete code pipelines, data visualization, and interpretable models to offer a blueprint for detecting and correcting survivorship bias in diverse data-science contexts.

---

1. Introduction

Survivorship bias arises whenever analysis exclude non-survivors—leading to inflated success metrics. Abraham Wald's 1943 Statistical Research Group memoranda used damage patterns on returning bombers to infer vulnerabilities on the non-returning ones, overturning conventional reinforcement strategies. Inspired by this, the project asks:

1. **How does survivorship bias influence decision-making in historical and modern datasets?**
2. **What key patterns emerge when comparing the bomber study with modern business and education datasets?**
3. **How does the exclusion of failure cases lead to skewed conclusions in areas such as career success and business survival?**
4. **Can identifying survivorship bias improve the accuracy of predictive models and policy-making?**

---

2. Data Collection & Features

Dataset	Source & Period	Key Features
Bomber Study	SRG memoranda (Wald, 1943)	Aircraft ID, damage counts by region, survival status
Business Turnover	BLS annual establishment data (1994–2024)	% Opening, % Closing, % Expanding, % Contracting, NetChange
Graduation Rates	IPEDS public universities (2020–2022)	4-yr, 5-yr, 6-yr cohort graduation rates
S&P 500 Survivorship	Current & delisted tickers (2014–2024) + Yahoo Finance	10-yr cumulative return, annualized Sharpe ratio

- **Business**: NetChange > 0 → "growth" class; else "decline."
- **Graduation**: Predict 6-yr rate from 4- and 5-yr rates.
- **Finance**: Label survivors (1) vs. delisted (0); features capture return and risk.

All numeric features were standardized before modeling.

---

3. Exploratory Analysis & Hypothesis Testing

3.1 Business Turnover

- **Opening vs. Closing** (paired t-test): $t = 6.19$, $p < 0.001$ → openings significantly exceed closings.
- **Expanding vs. Contracting**: $t = 2.22$, $p \approx 0.034$ → expansions exceed contractions.
- **NetChange vs. 0** (one-sample t-test): $t = 6.19$, $p < 0.001$ → positive net growth.

Implication: Restricting analysis to "growth years" would overlook downturns, leading to over-optimistic business climate assessments.

3.2 Graduation Rates

- **4-yr vs. 6-yr** (paired t-test): $t \approx -35.11$, $p \approx 2\times10^{-58}$.
- **Mean gap** ≈ 0.223 (22.3%).

Implication: A 4-year–only lens underestimates total graduates by ~22%—late completers are crucial "survivors."

3.3 S&P 500 Survivorship

- **Cumulative return comparison** (unpaired t-test): $t \approx 4.72$, $p \approx 3.3\times10^{-6}$; survivors outperform delisted firms by an average of ~23 pp (95 % CI: [14.7%, 33.7%]).

Implication: Ignoring delisted firms drastically inflates index performance—survivorship bias in financial back-tests.

---

4. Machine Learning Analyses

4.1 Business Classification

- **Model**: Logistic Regression, Random Forest
- **Evaluation**: 70/30 split and 5-fold CV yielded
  - **Accuracy** ≈ 1.0
  - **AUC** ≈ 1.0

Interpretation: Turnover features perfectly separate growth vs. decline when both classes are included. Coefficients highlight that larger opening-closing and expansion-contraction gaps drive growth predictions.

4.2 Graduation Regression

- **Model**: Gradient Boosting Regressor (200 trees, LR = 0.1)
- **Evaluation**: $R^2 \approx 0.857$, RMSE ≈ 0.069
- **Scatter plot**: Actual vs. predicted cluster tightly about the 45° line.
- **Residual histogram**: Centered near zero, ±5 pp spread, few outliers.

Interpretation: Including both on-time and delayed rates yields a model that explains ~86 % of variance in eventual graduation, quantitatively correcting 4-yr–only bias.

4.3 S&P 500 Classification

- **Model**: Random Forest (200 trees)
- **Evaluation**: Accuracy ≈ 0.95, AUC ≈ 0.93
- **Confusion matrix**: High true positive (survivors) and moderate true negative (delisted) rates.
- **ROC curve**: Strong separation across thresholds.

Interpretation: Firms with higher returns and Sharpe are far more likely to survive. Excluding delisted data misrepresents true market risk.

---

5. Unsupervised Learning: Clustering

5.1 Hierarchical Clustering (Finance)

- **Dendrogram** (truncated) reveals three natural groups:
    1. **Moderate performers** (n = 159)
    2. **High-performing survivors** (n = 401)
    3. **Extreme outlier** (NVDA alone)

Insight: Survivorship bias emerges as the high-performing cluster dominates index returns, while moderate under-performers (including delisted firms) are marginalized.

---

6. Discussion

1. **Generalizing Wald's Insight**: Each domain showed that including failures unveils the true decision boundary—armoring, business health, degree completion, or firm survival.
2. **Policy Implications**:
    - **Business strategy**: Monitor contraction signals; don't assume perpetual growth.
    - **Higher-ed policy**: Support late completers; target resources beyond 4 years.
    - **Investment back-tests**: Include delisted firms for realistic performance estimates.
3. **Modeling Best Practices**:
    - Always stratify and include all outcomes in training and evaluation.
    - Use both supervised and unsupervised methods to understand class separation and latent segments.
4. **Limitations**:
    - **Data breadth**: Business data annual; graduation data limited to 3 years; finance data excludes mergers.
    - **Feature scope**: Additional covariates (sector, demographics) could refine models.

---

7. Conclusion

Through rigorous statistical testing, machine learning, and clustering, this project extends Abraham Wald's principle to modern datasets. By deliberately modeling both survivors and non-survivors, the magnitude of survivorship bias—20–30% distortions— was uncovered and practical analysis to avoid it were provided. The accompanying GitHub repository includes all code, data pipelines, and visualizations, serving as a reproducible guide for bias-aware data science.

References

- Mangel, M., & Samaniego, F. J. (1984). Abraham Wald's Work on Aircraft Survivability. *J. Amer. Statist. Assoc.* 79(386):259–267.
- U.S. Bureau of Labor Statistics, Business Employment Dynamics (1994–2024).

- National Center for Education Statistics, IPEDS Graduation Rates (2020–2022).
- Yahoo Finance via `yfinance` Python package for market data (2014–2024).