Pension Survival Analysis: Handling Censored Data Using Survival Models and IPCW

Aman Shah

November 6, 2025

1. Project Overview

This project explores the impact of different censoring strategies on survival analysis models when predicting pension-related events such as retirement, withdrawal or death. It was completed as part of my third-year Survival Modelling coursework and focuses on addressing the challenges of censored data using traditional and modern statistical techniques.

2. Motivation

Why this project matters:

- **Academic:** Part of University coursework on survival modelling.
- **Real-World Relevance:** Accurate pension liability predictions are crucial for actuarial science and financial planning.
- **Technical Challenge:** Censored data introduces bias if ignored or incorrectly handled.

Examples of censoring in pension data:

- Individuals leave the pension scheme before the event occurs.
- Administrative censoring at the end of the observation period.
- Members transfer to another scheme.

3. Methodology

Censoring Methods Compared

Method	Description	Pros	Cons
ZERO Method	Treats all censored cases as non-events	Simple	Biased; underestimates risk
DISCARD Method	Removes censored cases from training	Avoids censoring bias	Reduces sample size
IPCW	Weights data by inverse probability of censoring	Unbiased under MAR assumption	Computationally complex

Models Evaluated

Survival Models:

- Cox Proportional Hazards (Cox PH)
- Weibull Accelerated Failure Time (AFT)

Machine Learning Classifiers:

- Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbours (KNN)

Evaluation Metrics

Survival Models:

- Concordance Index (C-index)
- Integrated Brier Score (IBS)
- AUC at 15 years (AUC@T∗)

Classification Models:

- Accuracy, AUC, F1-Score
- Net Reclassification Improvement (NRI)

4. Project Structure

```
pension-survival-analysis/
 notebooks/
    Data generation/
       synthetic_survival_data.csv
    ipcw_and_other_censoring/
       ipcw_and_other_censoring.ipynb
       data/censoring_methods/
    model_eval/
        model_evak.ipynb
results.csv
README.tex
```

5. How to Run

Step 1: Install Dependencies

```
pip install pandas numpy matplotlib scikit-learn scikit-survival lifelines
```

Step 2: Generate Data (Optional)

Run the data generation notebook to create `synthetic_survival_data.csv`.

Step 3: Create Censored Datasets

```
cd notebooks/ipcw_and_other_censoring/
jupyter notebook ipcw_and_other_censoring.ipynb
```

Step 4: Run Model Evaluation

```
cd notebooks/model_eval/
jupyter notebook model_evak.ipynb
```

Step 5: View Results

```
import pandas as pd
df = pd.read_csv('results.csv')
print(df.groupby(['Method','Model_Type']).mean())
```

6. Key Parameters

```
T_STAR = 15.0                          # Prediction horizon (years)
TIMES = np.array([1,5,10,15,17])  # Evaluation times
MAX_WEIGHT = 20.0                      # IPCW weight cap
```

7. IPCW Stabilisation

```
G_hat_clipped = np.clip(G_hat, 0.05, 1.0)
ipcw_capped = np.clip(1/G_hat_clipped, 0, MAX_WEIGHT)
```

8. References

- Klein & Moeschberger (2003) — *Survival Analysis*
- Robins & Rotnitzky (1992) — Inverse Probability Weighting
- scikit-survival & lifelines documentation