Donnel K. Cuffie (B.Sc., M.Sc.)

PREDICTING THE RISK OF EMPLOYER INACTIVITY IN A SOCIAL INSURANCE SYSTEM: AN APPLICATION OF SURVIVAL ANALYSIS METHODS

This presentation has been prepared for the 2023 Caribbean Actuarial Association (CAA) Conference.
The CAA wishes it to be understood that opinions put forward herein are not necessarily those of the CAA and the CAA takes no responsibility for those opinions.

Overview

- Context of the research
- (Very) Brief overview of survival analysis
- Research questions
- Data
- Models
- Measures of model performance
- Results
- Conclusion

Context of the Research

- The NIS is a system of social insurance that aims to protect working persons and their dependents against a loss of earnings due to specified contingencies (e.g., sickness, maternity, death, retirement).
- The system works by collecting contributions from working persons and their employers, and paying benefits out of the pool of money collected.
- Employers are responsible for remitting the payments to the NIBTT, typically on a monthly basis.
- An employer who pays contributions in a given period (e.g., a month, year) is deemed to be **'active'**, whereas an employer who makes no payments in the period is deemed to be **'inactive'**.

Context of the Research (Cont'd)

- Because of population ageing, the NIBTT is collecting less in contribution income than it pays in benefits
- Contribution Income vs. Benefit Expenditure (FY2022)

$4.53 Billion	**$5.73 Billion**
Contribution Income	Benefit Expenditure

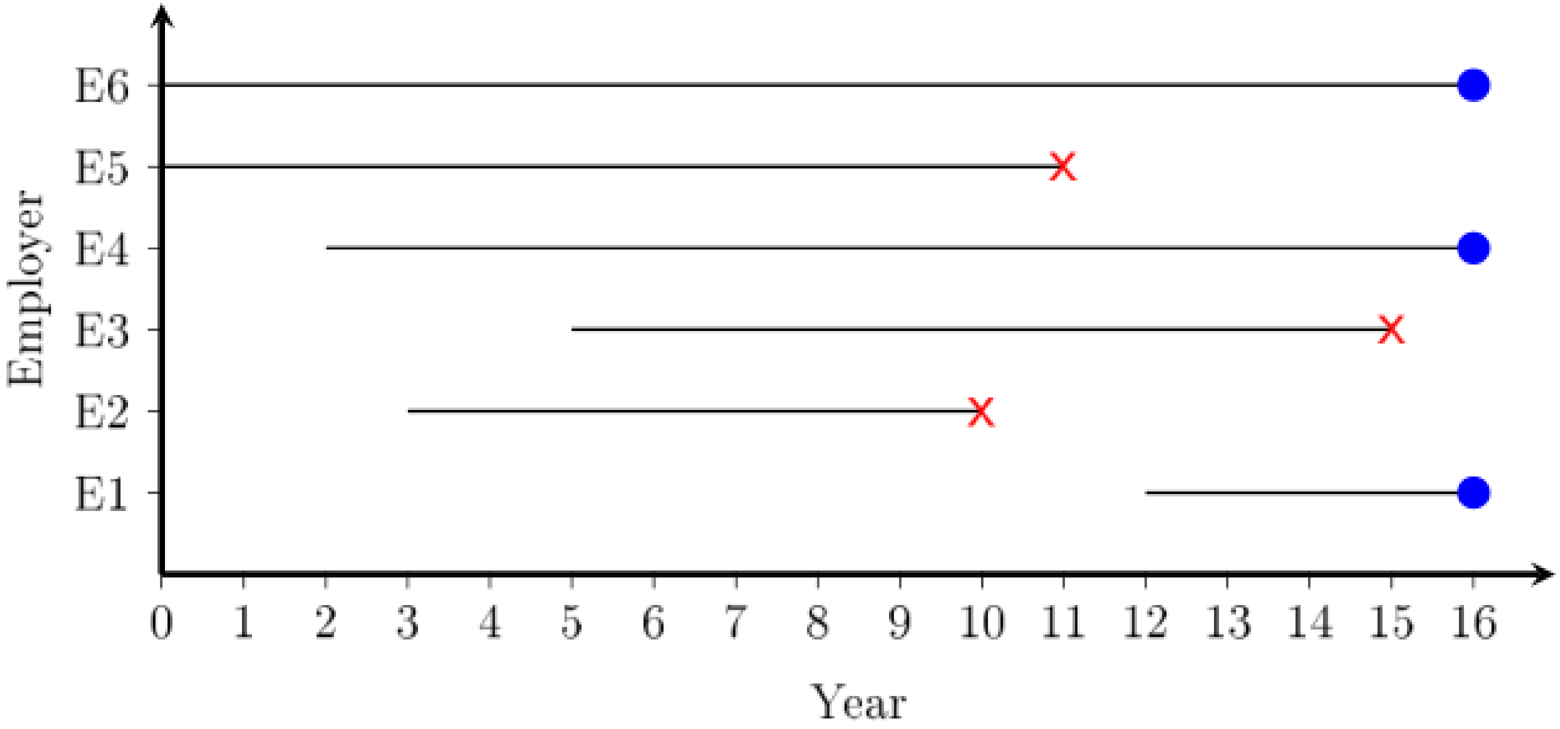
- Hence, the NIBTT can benefit from a system to track employers with a high risk of becoming inactive to ensure that they are not evading their obligations to the NIBTT

Brief Overview of Survival Analysis

- Survival analysis is a subdomain of the discipline of Statistics (and Machine Learning) which deals with the analysis of time to event data.
- The key data analytic issue which survival analysis techniques were developed to overcome is **censoring**.
- A ‘censored’ observation gives us partial information about the survival experience of an individual element in our data.

Brief Overview of Survival Analysis (Cont'd)

The diagram below shows fixed right censoring



Research Questions

The research aimed at answering the following research questions:

1. *What is the best model for predicting the risk of an employer becoming inactive?*
2. *Given the best predictive model, what are the important factors in determining the risk of employer inactivity?*

Data

- The data was acquired from the NIBTT and consisted of employers who were active for at least one year during the period from 2000 to 2016.
- For each employer, follow-up began from the start of their first active spell over the 17-year period from 2000 to 2016, and continued until they experienced inactivity or the study period ended in 2016.

Data (Cont’d)

- ***Dependent variables***:
  - The length of the first active spell (in years) over the 17-year observation period (`time`)
  - An indicator of whether or not the employer experienced inactivity (`status`)
- ***Independent variables***:
  - The average number of months per year in which contribution payments were made over the duration of the employer’s active spell (`avgMnth`)
  - The median number of employees on whose behalf contribution payments were made (`medEmp`)
  - The median monthly value of contribution payments per employee (`medContVal`)

Data (Cont’d)

- ***Independent variables (cont’d)***:
  - Whether or not the employer owed a debt to the NIBTT over their active spell (`debt`)
  - The number of compliance audits conducted during the follow-up period (`nAudit`)
  - The type of employer (`type`)
  - The location of the employer (`location`)
  - The number of years the employer has been in the NIS, from registration (or first contribution date), up to December 31st of the first year of their active spell (`nYrs`)

Models

The following six (6) models were compared:

- *Cox Proportional Hazard (PH) Model*
- *Accelerated Failure Time (AFT) Models:*
  - *Weibull*
  - *Log-logistic*
  - *Log-normal*
- *Censored Normal Regression (Tobit) Model*
- *Random Survival Forest*

Cox PH Model

- The Cox PH model is given by

$$h(t,x) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$$

  - Where $h(t,x)$ is the hazard rate at time t for a subject with predictors X
  - $h_0(t)$ is a baseline hazard function
  - X_i $(i = 1, \dots, p)$ is the value of the ith time-independent predictor
  - the β_i's are the model parameters
- No parametric form is assumed for $h_0(t)$

AFT Models

- Let T be a random variable of survival time, the general form of the AFT model is given by

$$\ln T = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X} + \sigma\varepsilon$$

  - Where $\boldsymbol{X}$ is a vector of fixed-time predictors
  - $\boldsymbol{\beta}' = (\beta_1, \dots, \beta_p)$ is a vector of regression coefficients
  - ε is the error distribution
    - **Weibull**: $\varepsilon \sim Gumbel\ (Extreme\ Value)$
    - **Log-logistic**: $\varepsilon \sim Standard\ Logistic$
    - **Log-normal**: $\varepsilon \sim Standard\ Normal$

Tobit Model

- The functional form of the model is

$$T_i^* = \beta_0 + \boldsymbol{\beta}'\boldsymbol{X}_i + \varepsilon_i$$

$$T_i = \begin{cases} T_i^* \; if \; T_i^* < R_i \\ R_i \; if \; T_i^* < R_i \end{cases}$$

  - where T_i^* is the latent survival time for subject i
  - T_i is the observed survival time
  - R_i is the fixed right-censoring time for subject i
  - $\varepsilon_i \sim Normal(0, \sigma^2)$

Random Survival Forest

Random Survival Forest Algorithm

1. Draw B bootstrap samples from the original dataset. Approximately 63% of the observations are sampled (called the in-bag sample) with or without replacement for each bootstrap sample. The remaining 37% of observations are called the out-of-bag (OOB) observations.

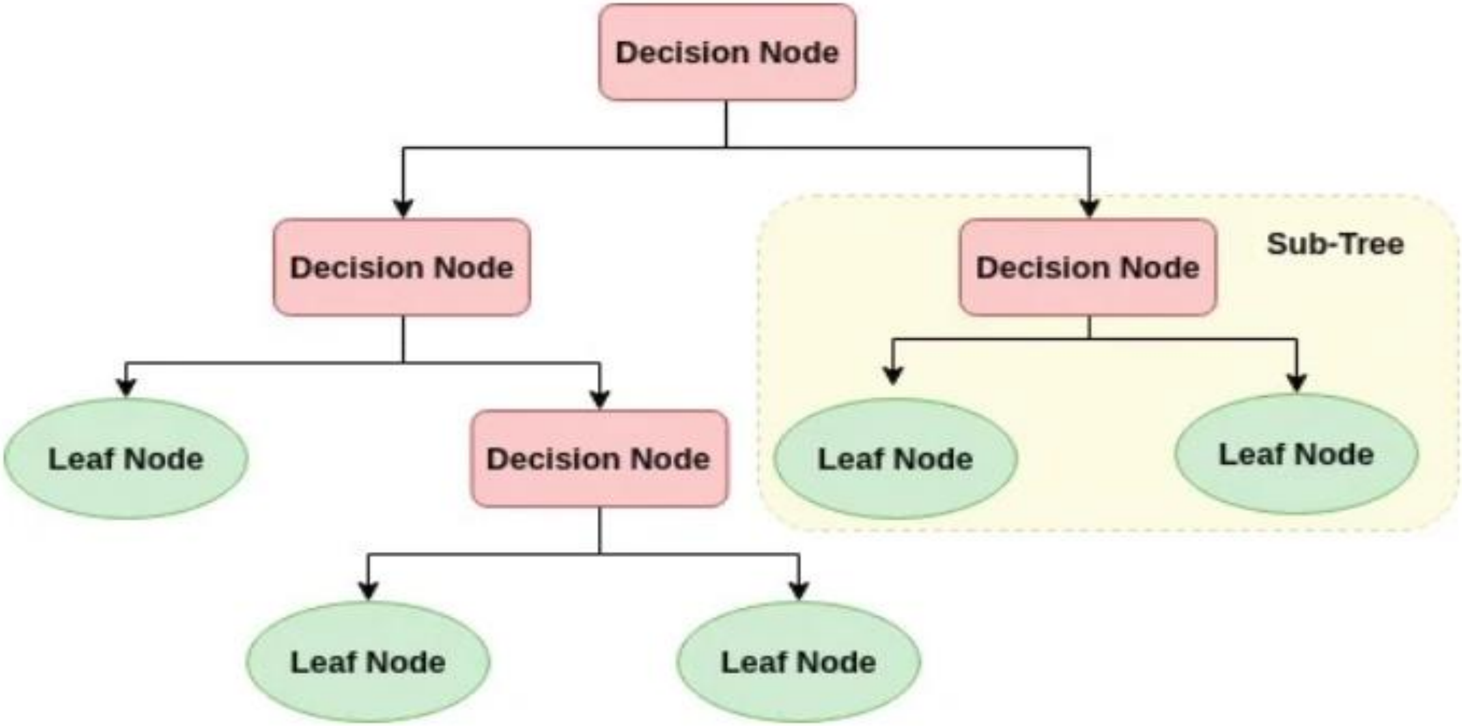
2. Grow the survival tree for each of the $b = 1, \ldots, B$ bootstrap samples, at each node randomly select $\sqrt{p}$ variables. Split the node by selecting the variable that maximises the difference between daughter nodes using a predetermined split rule.

3. Grow the tree to full size under the constraint that a terminal node should have no less than $n_0 > 0$ subjects.

4. Estimate the survival function or the cumulative hazard function (CHF) for each of the B trees.

5. Using the OOB observations, calculate the ensemble survival function or CHF.

Example Decision Tree



Measures of Model Performance

Measures of model performance are divided into 3 categories:

1. *Discrimination* – evaluates the ability of the model to separate subjects' outcomes i.e., separate high-risk employers from low-risk employers.
2. *Calibration* – evaluates the ability of the model to make unbiased estimates of outcomes i.e., to accurately predict the risk of employer inactivity.
3. *Overall* – combines both the *discrimination* and *calibration* aspects of the model into one measure.

Measures of Model Performance (Cont’d)

- ***Discrimination***
  - *Concordance Index (Harrell’s C)* – ranges between 0 and 1, the higher the better.
  - *Somers’ D* – ranges between -1 and 1, the higher the better.
- ***Calibration***
  - Calibration curves at years 5, 9 and 13.
- ***Overall***
  - *Brier Score*
    - It is calculated at a specified time t^*
    - The lower the value of the Brier Score the better

Measures of Model Performance (Cont'd)

- ***Overall***
  - Index of Prediction Accuracy (IPA)

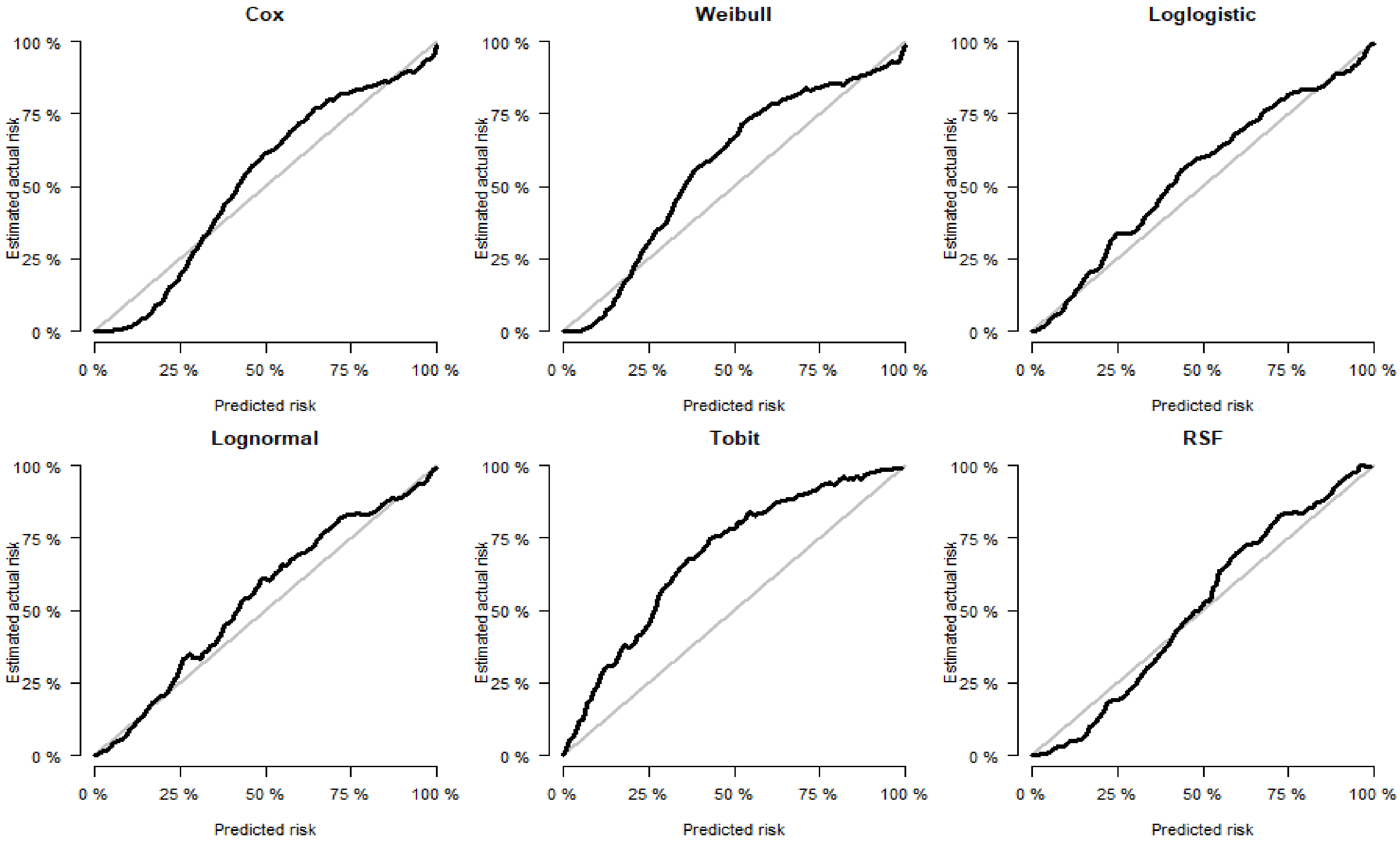
$$IPA = 1 - \left(\frac{model\ Brier\ Score}{null\ model\ Brier\ Score}\right)$$

- The null model contains no predictors i.e., the Kaplan-Meier estimator.
- A IPA value of 100% indicates a perfect model and a value of ≤ 0 indicates a useless or harmful model.

Results: Discrimination

Model	Harrell's *C*	Somers' *D*
Cox PH	87.646	75.293
Log-logistic	87.638	75.276
Log-normal	87.621	75.241
Weibull	87.609	75.218
Tobit	87.088	74.176
RSF	85.904	71.808

Results: Calibration at Year 5



Cox

Weibull

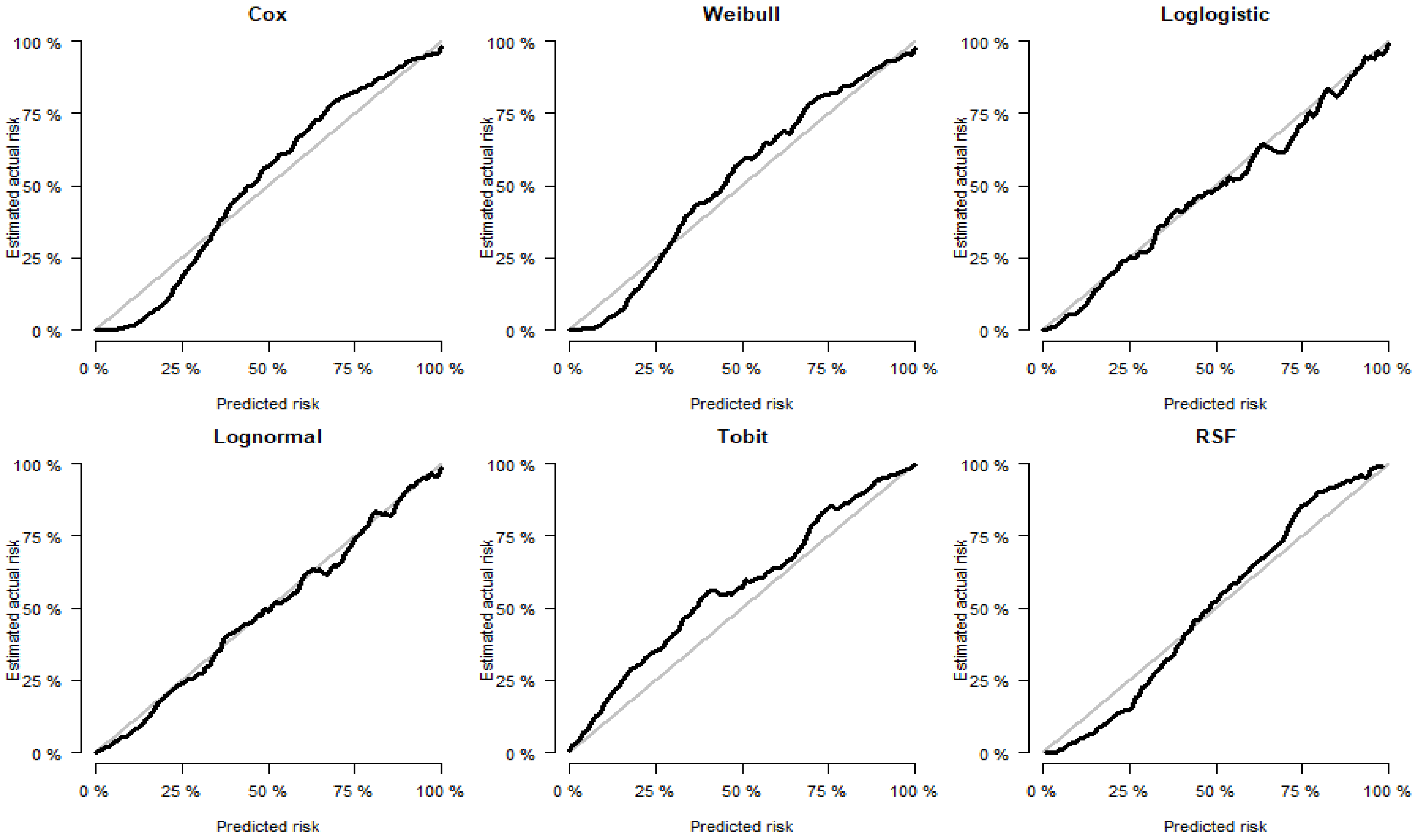
Loglogistic

Lognormal

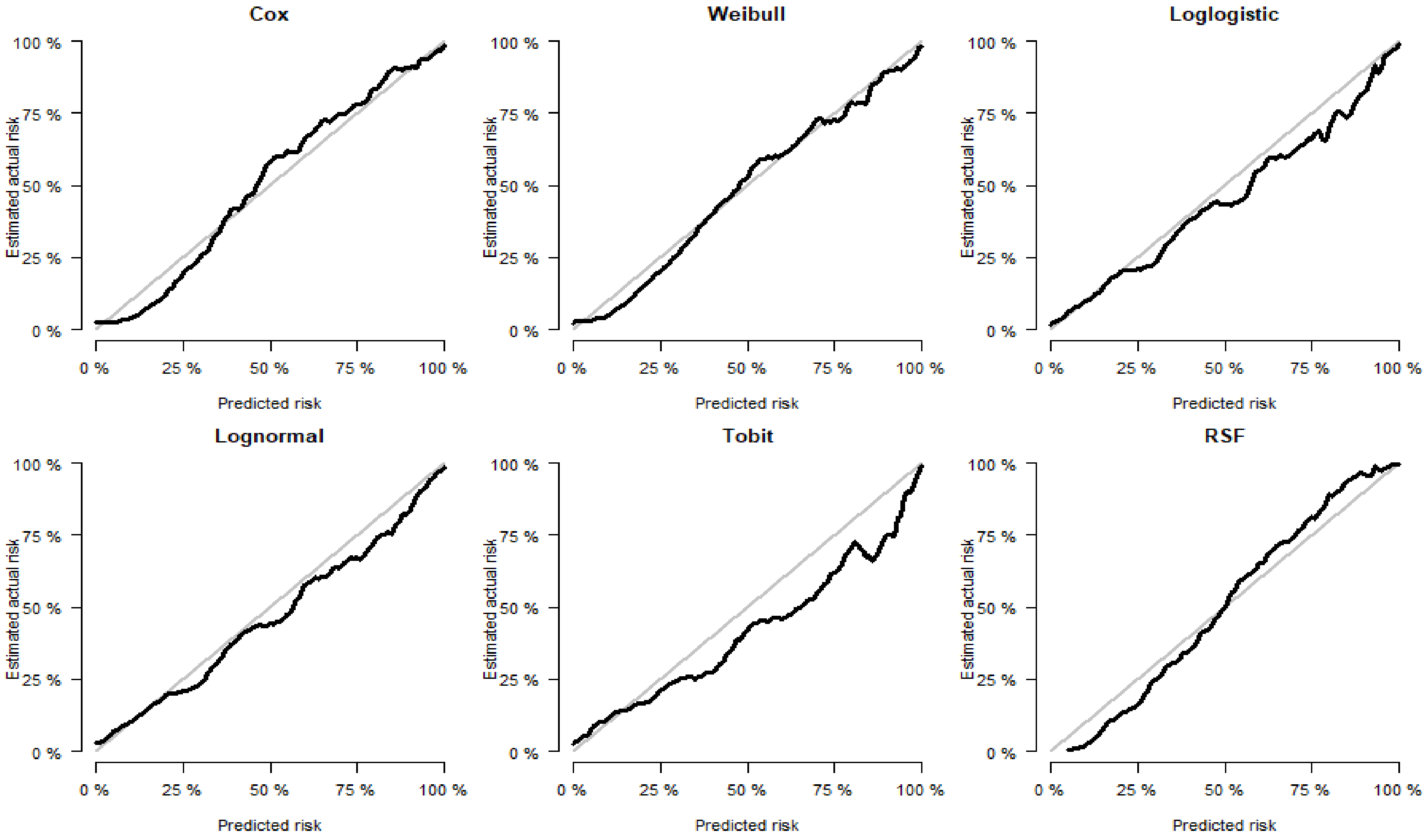
Tobit

RSF

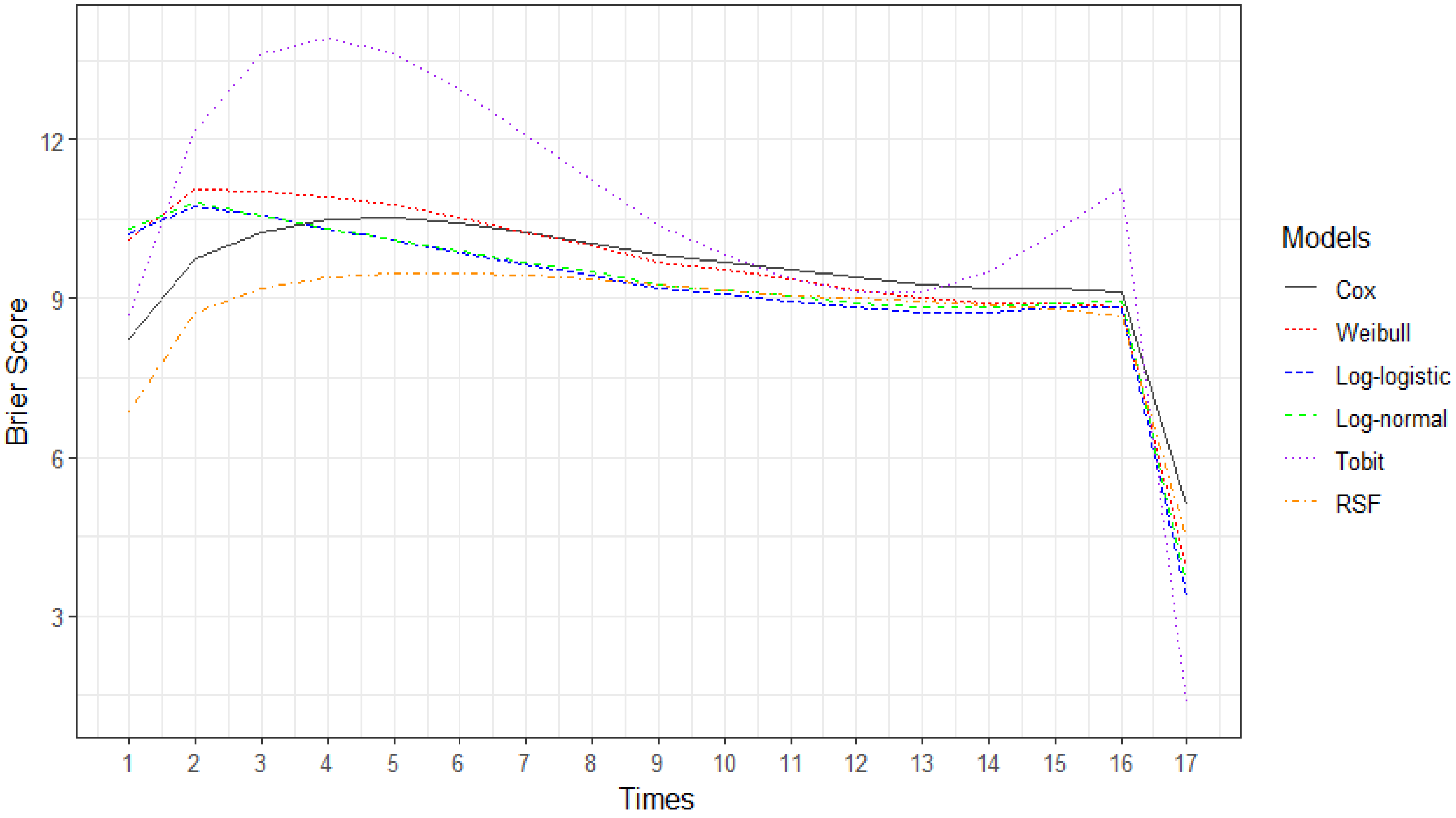
Results: Calibration at Year 9



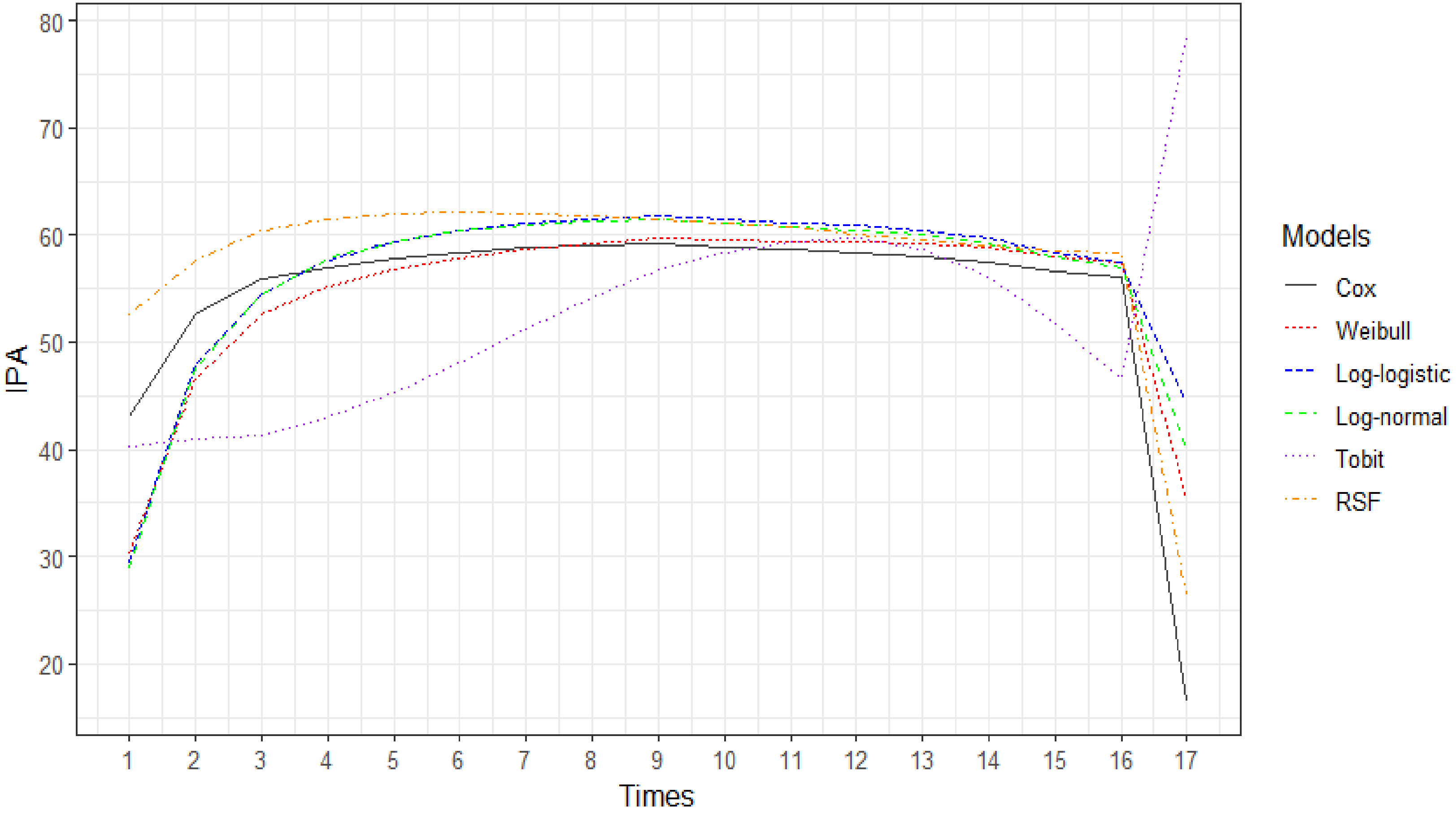
Results: Calibration at Year 13



Results: Overall – Brier Score



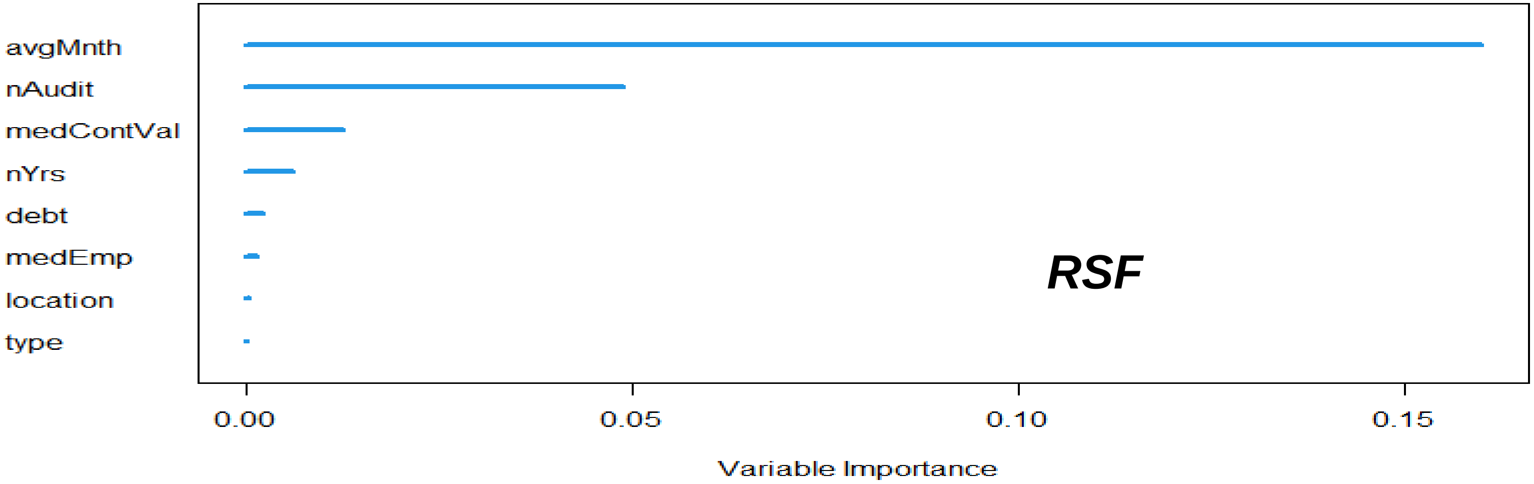
Results: Overall – IPA

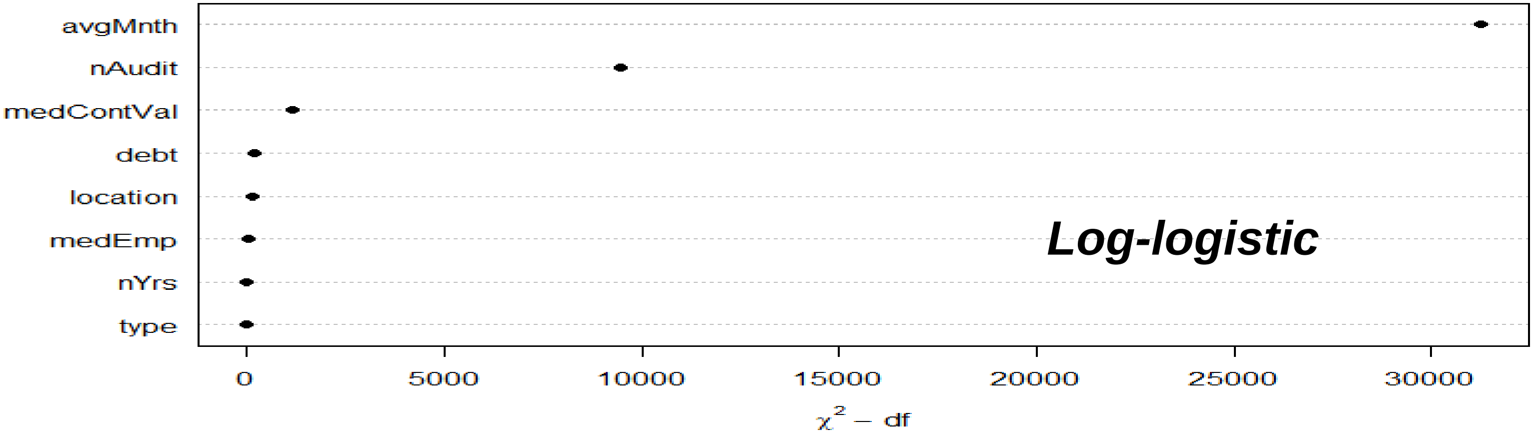


Results: Overall

Model	Mean Brier	Mean IPA	Mean Rank
RSF	8.72	57.9	1.94
Log-logistic	9.15	56.3	2.06
Log-normal	9.22	55.8	2.88
Cox PH	9.43	54.2	4.47
Weibull	9.53	54.3	4.35
Tobit	10.5	52.3	5.29

Results: Most Important Predictors





Conclusion

Best model for predicting the risk of inactivity

The Random Survival Forest was the best model overall for predicting the risk of an employer becoming inactive, followed closely by the Log-logistic model.

Model important predictors

1. The average number of months per year in which contribution payments are made.

2. The number of compliance audits conducted over the duration of becoming inactive.

3. The median value of contribution payments made on behalf of employees.

Conclusion (Cont'd)

- The present study represents the first formal investigation on the risk of employer inactivity within the local NIS.
- The results of the analysis are promising and demonstrate clearly that there are patters in employer activity/inactivity within the NIS that can be detected.
- These patters can be used to predict the risk of inactivity with a reasonably high level of accuracy using the appropriate statistical or machine learning model.
- This work can be used by the NIBTT and built upon to develop a robust system for deciding which employers to target for compliance audits, based on their predicted risk of inactivity.

QUESTIONS