TREE BASED MODELS

Marco Longo
Francesca Ricci
Maria Rotella

UNIVERSITÀ DELLA CALABRIA

MASTER IN

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

AA 2024/2025

TREE BASED MODELS

RANDOM SURVIVAL FOREST

CASE STUDY: EMPLOYEE CHURN

AGENDA

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

- Dataset Overview
  - Variables involved
- Preliminary Survival Analysis
  - Results with Kaplan-Meier and Cox models
- Tree Based Models
  - Decision Trees
  - Random Forest
- Random Survival Forest
  - Model Overview
  - Parameters Tuning
  - Splitting Methods
  - Application in R and Results
  - Variable Importance
  - Evaluation and Comparison of Models

DATASET OVERVIEW

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

15,000 observations - 10 variables, including:

- Time-to-event variables:
  - **time_spend_company**: follow-up (months)
  - **left**: the event, indicating whether the employee left the company
- Covariates:
  - *satisfaction_level*: Satisfaction index (decimal between 0 and 1)
  - *last_evaluation*: Last evaluation received (decimal between 0 and1)
  - *number_project*: Number of projects assigned to the employee
  - *average_monthly_hours*: Average monthly working hours
  - *work_accident*: Workplace accidents (boolean)
  - ~~*promotion_last_5years*: Promotions in the last 5 years (boolean)~~
  - *department*: Corporate department (e.g., IT, support, HR, sales)
  - *salary*: Salary level (low, medium, high)

PRELIMINARY SURVIVAL ANALYSIS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

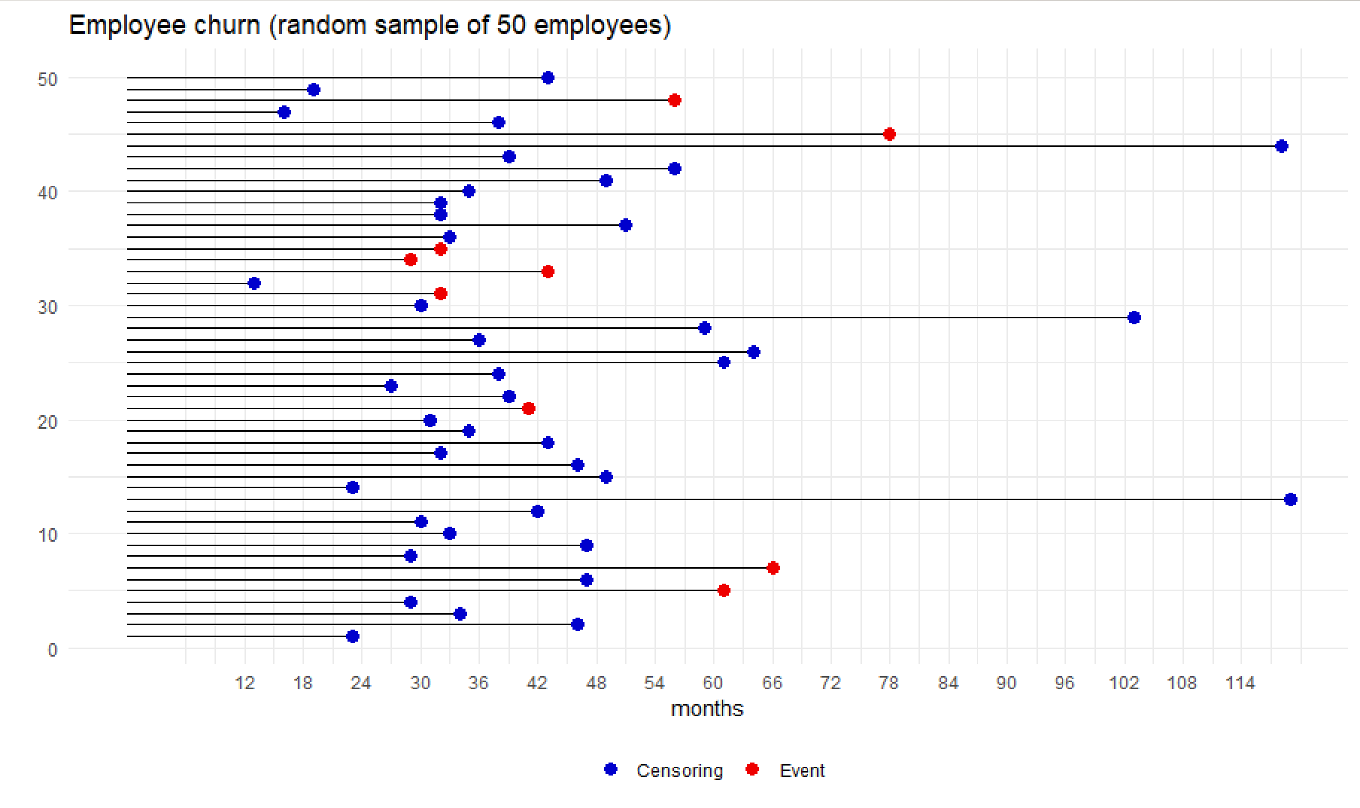
Survival Analysis

- Survival Analysis studies the time until the occurrence of an **event**
- In this case, the event is an **employee** leaving their job
- Observations may be **censored**: for instance, if the employee is still with the company at the end of the observation period (follow up)
- Objective: To analyze and estimate the probability that an individual will remain with the company over a given period of time, using **survival functions** and statistical methods

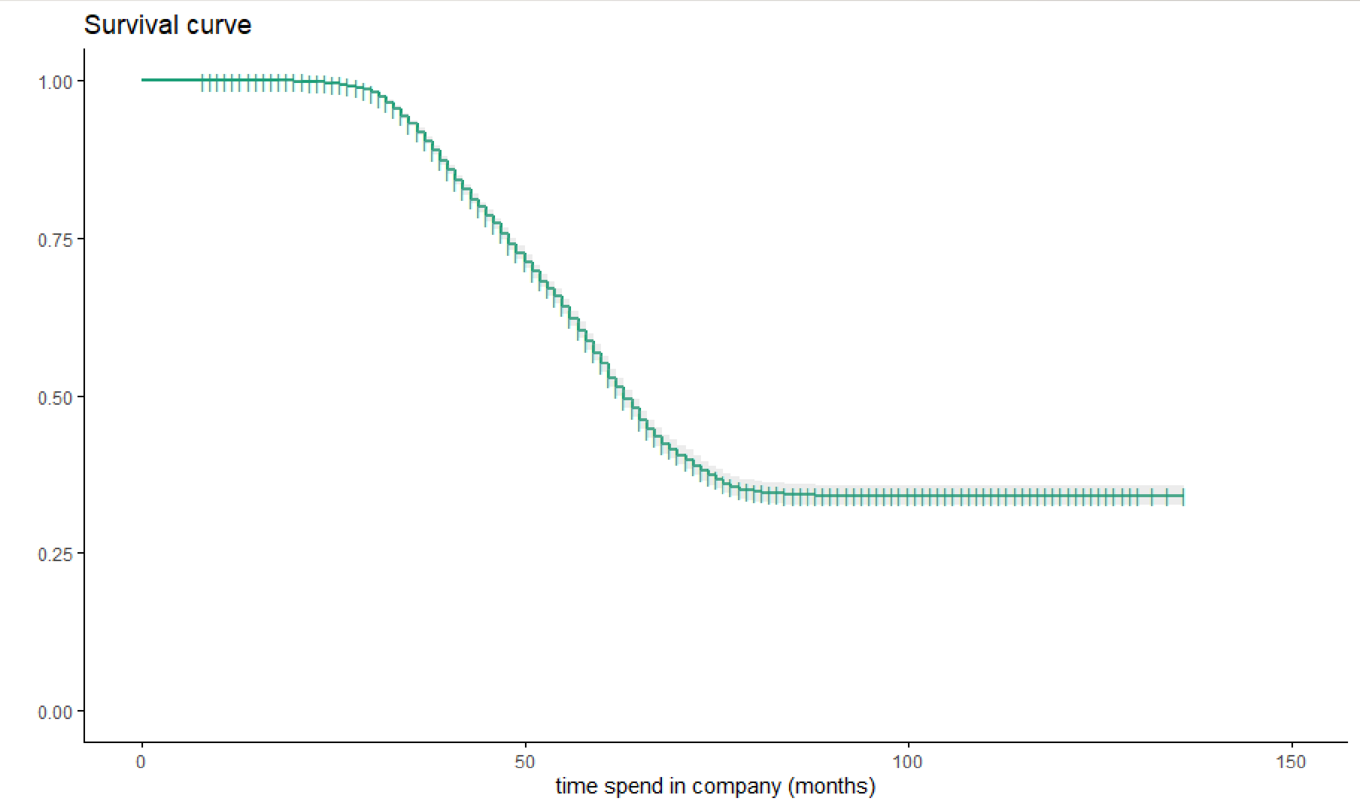
PRELIMINARY SURVIVAL ANALYSIS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN



PRELIMINARY SURVIVAL ANALYSIS



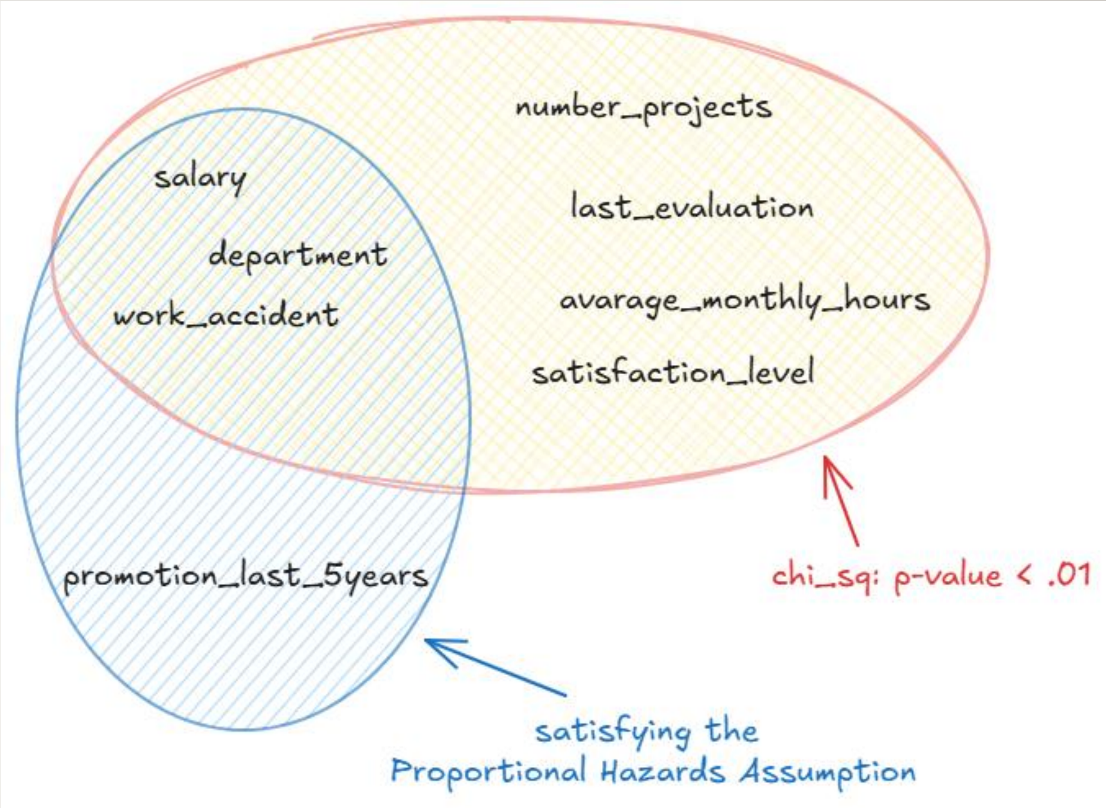
RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

PRELIMINARY SURVIVAL ANALYSIS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN



TREE BASED MODELS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

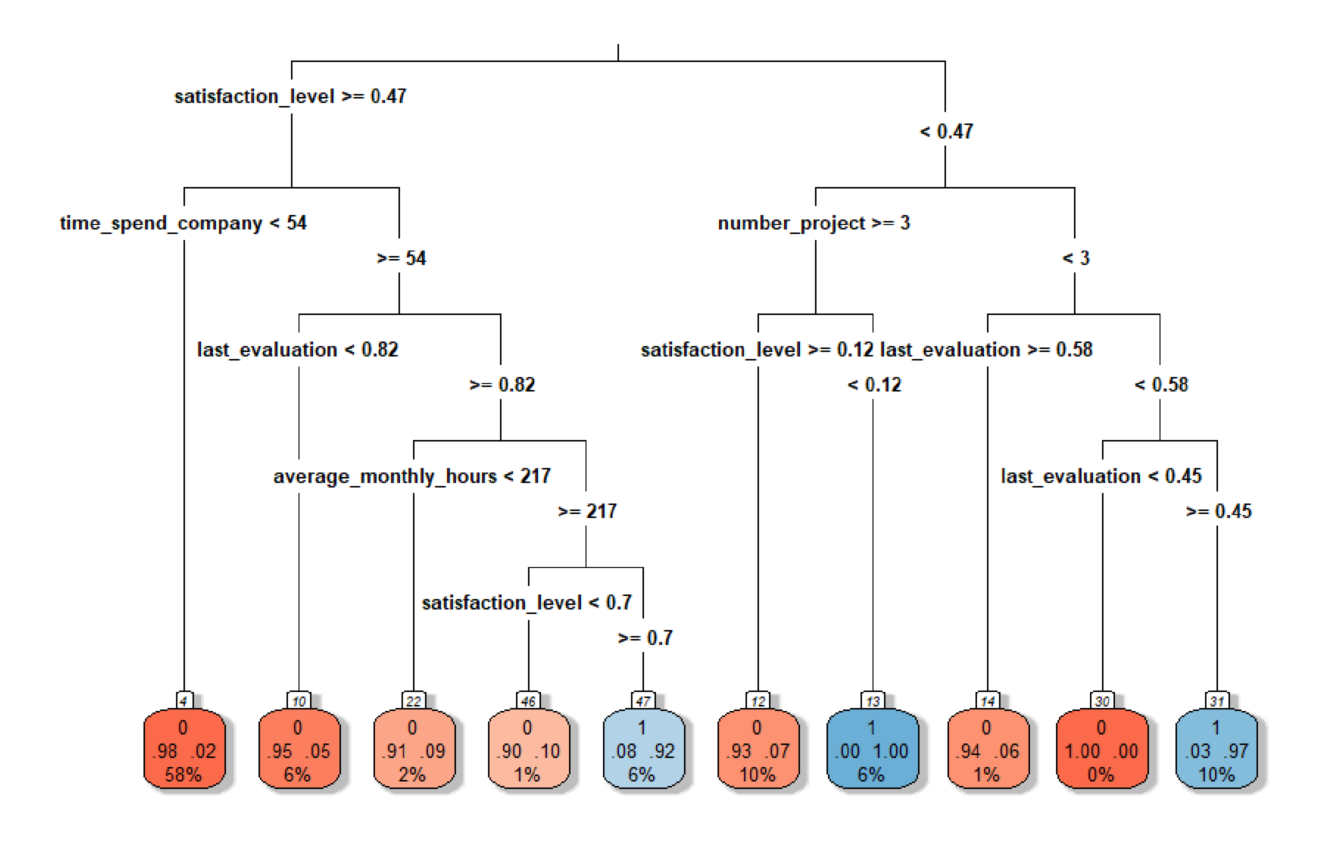
- Tree-based models are widely used for prediction tasks in both **regression** and **classification**.
- Models:
  - Decision Trees
  - Random Forest
- Strengths:
  - High flexibility in **adapting** to different types of data and problems
  - Strong **performance** on real-world datasets
  - Ability to handle both **categorical** and **numerical** variables efficiently

TREE BASED MODELS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Classification - Decision Tree in R with **rpart** (method="**class**")

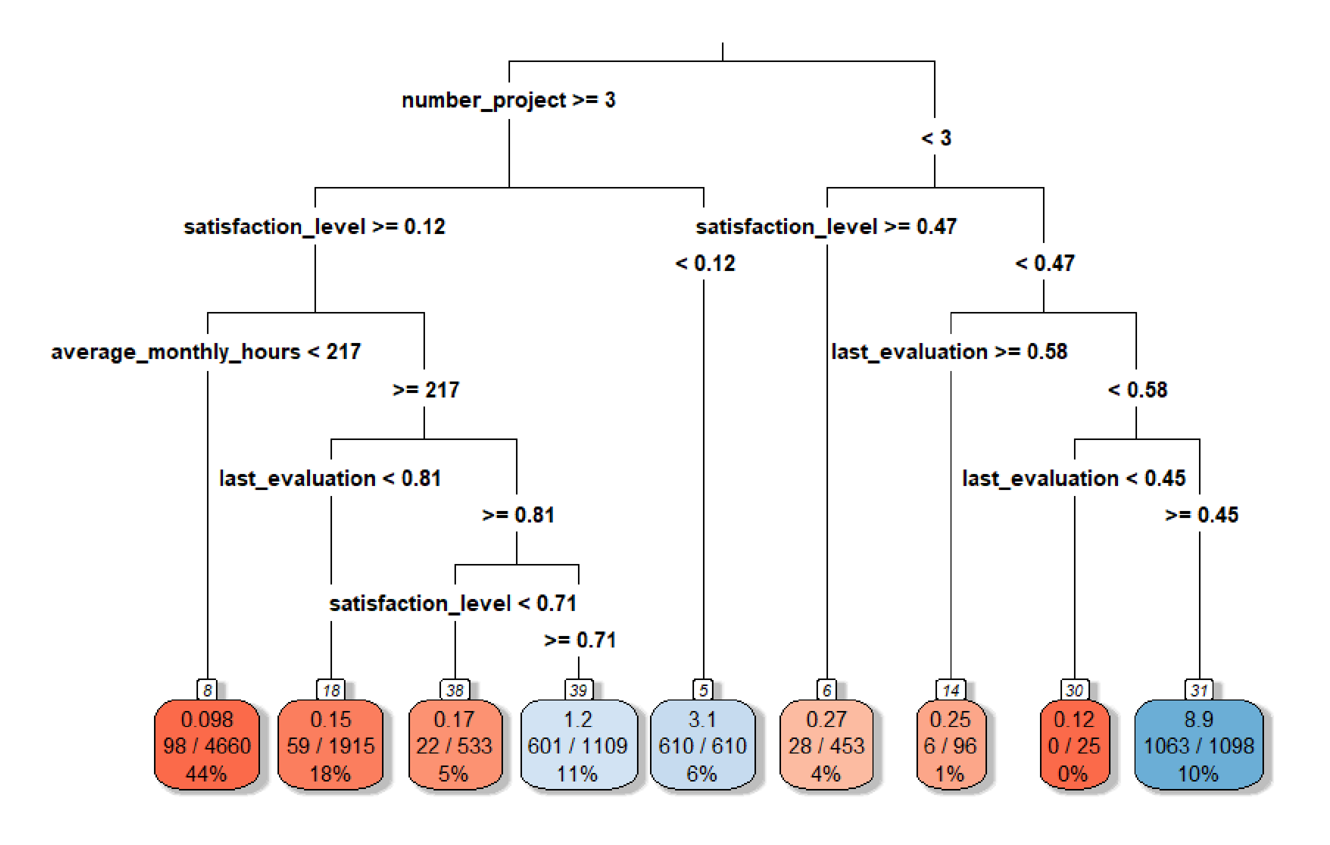


TREE BASED MODELS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Survival - Decision Tree in R with **rpart** (method="**exp**")



RANDOM SURVIVAL FOREST

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

- The Random Survival Forest (**RSF**) algorithm is an extension of Random Forests tailored for **survival** analysis.
- They are **non-parametric** models capable of handling censored data effectively.
- The **log-rank test** is used for splitting nodes during tree construction.

HOW RSF ALGORITHM WORKS

RANDOM SURVIVAL FOREST

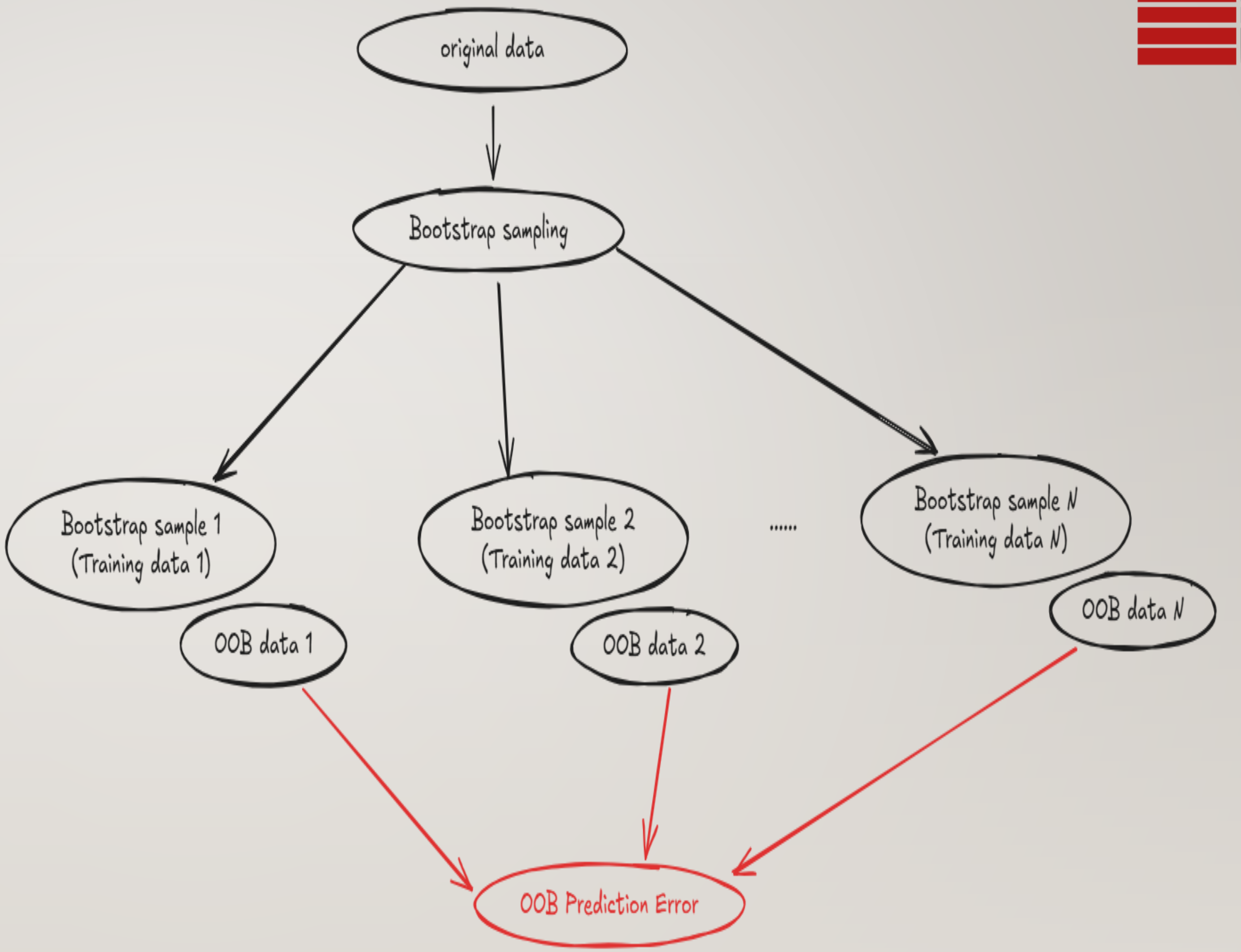
EMPLOYEE CHURN

- RSF builds **multiple decision trees** using bootstrap samples from the dataset.
- At each node, the **log-rank test** is used to determine the best split by comparing survival distributions.
- RSF selects a **subset of covariates** for splitting during the construction of each tree.
- Predictions are aggregated across all trees in the forest to estimate **survival probabilities** or risk.
- RSF handles **censored data** effectively and does not require the proportional hazards assumption.
- Implementation in R with **randomForestSRC**
  - Training: *rfsrc*
  - Survival Curve: *plot.survival*
  - Variable Importance: *vimp*

RANDOM SURVIVAL FOREST

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN



RANDOM SURVIVAL FOREST

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Analysis: **RSF** - Family: **surv**

Sample size: **10499** – No. of deaths: **2487** – No. of trees: **1000**

Forest terminal node size: **15** - No. of variables tried at each split: **5**

Total no. of variables: **7** - Resampling used to grow trees: **swor**

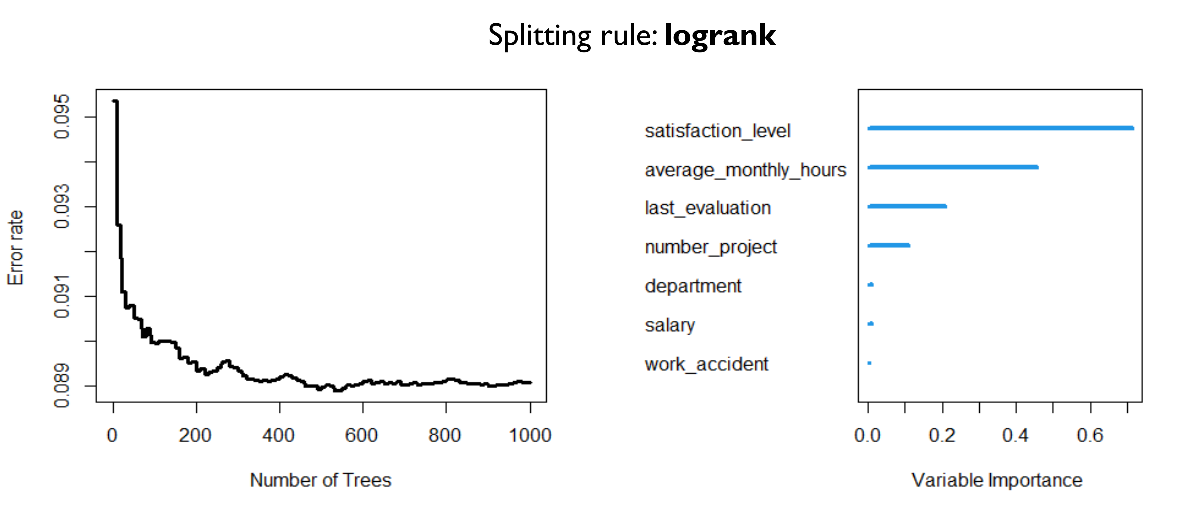
Resample size used to grow trees: **6635 –** No. of random split points: **10**

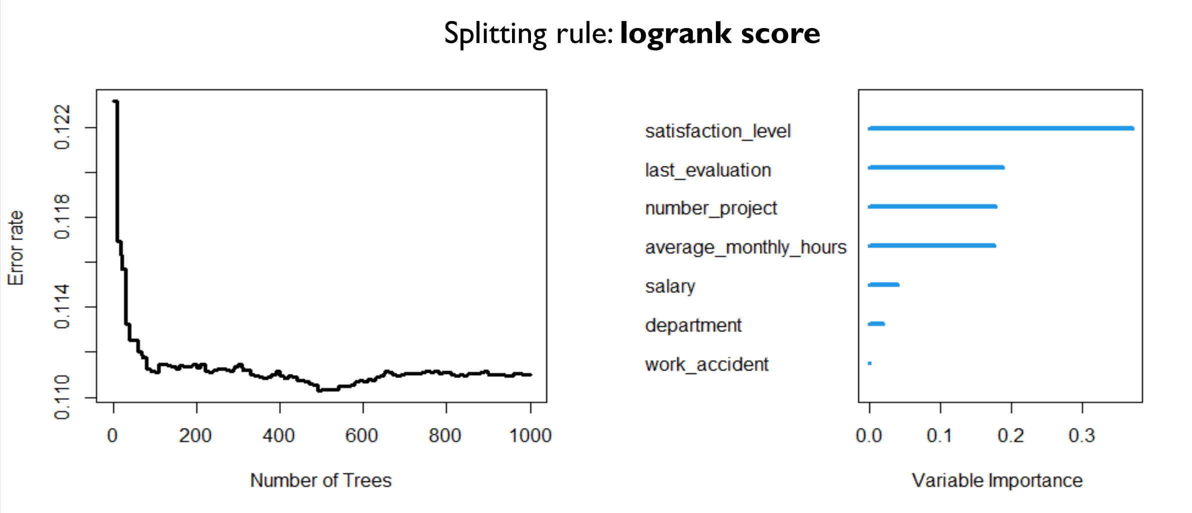
	LOGRANK	LOGRANK SCORE
Average no. of terminal nodes	**249.14**	**395.36**
(OOB) CRPS	**3.1532**	**3.8133**
(OOB) stand. CRPS	**0.0371**	**0.0449**
(OOB) Requested performance error	**0.0891**	**0.1110**
(OOB) C-index	**0.9112**	**0.8900**

RANDOM SURVIVAL FOREST

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN



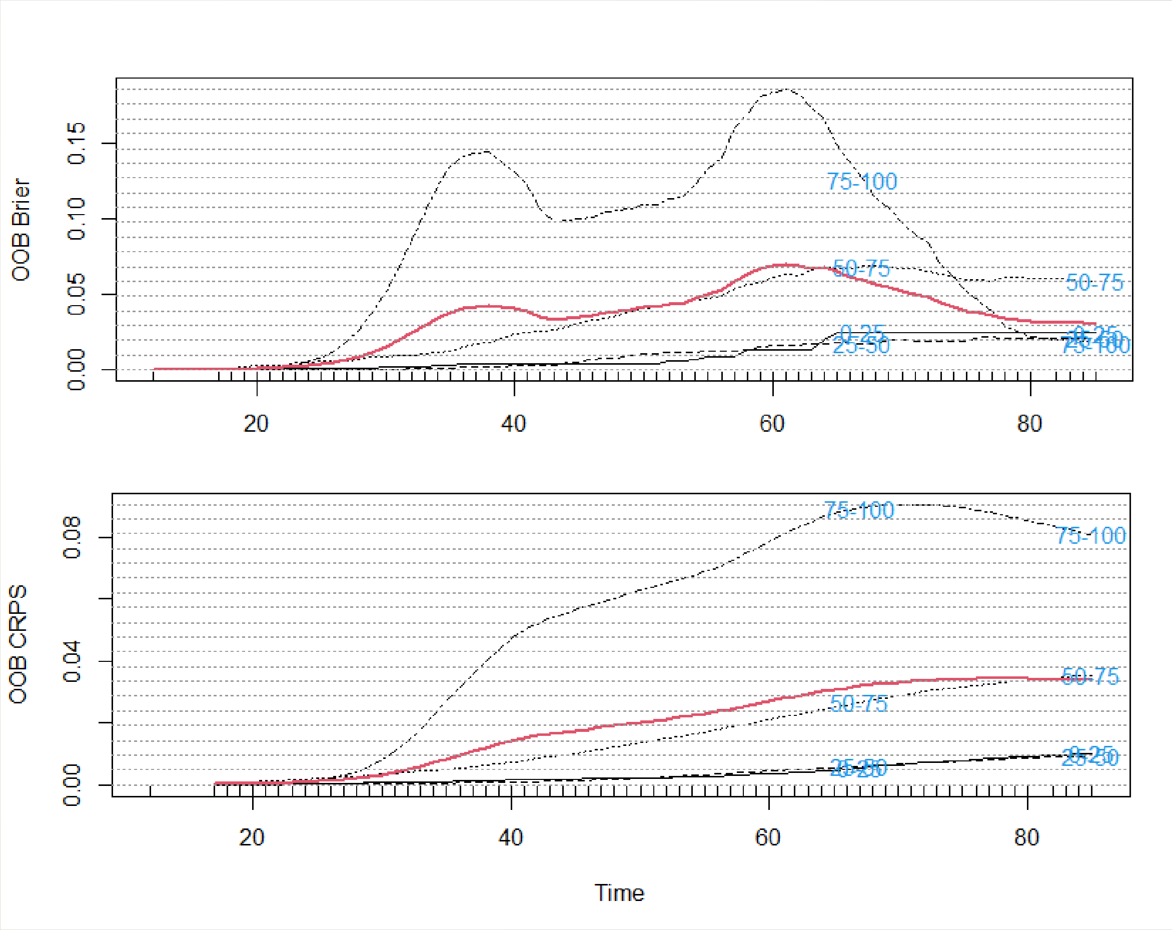


RANDOM SURVIVAL FOREST

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Splitting rule: **logrank**

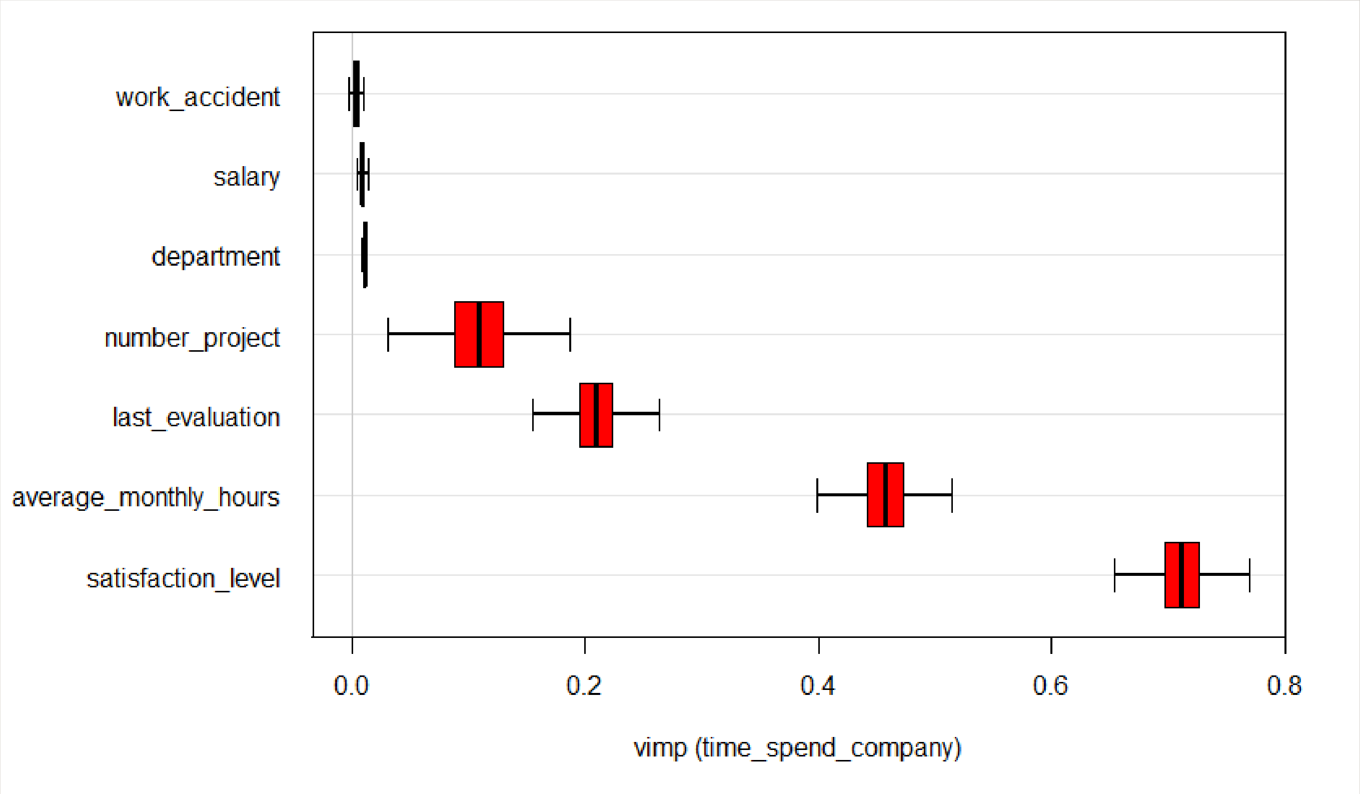


VARIABLE IMPORTANCE

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Variable importance

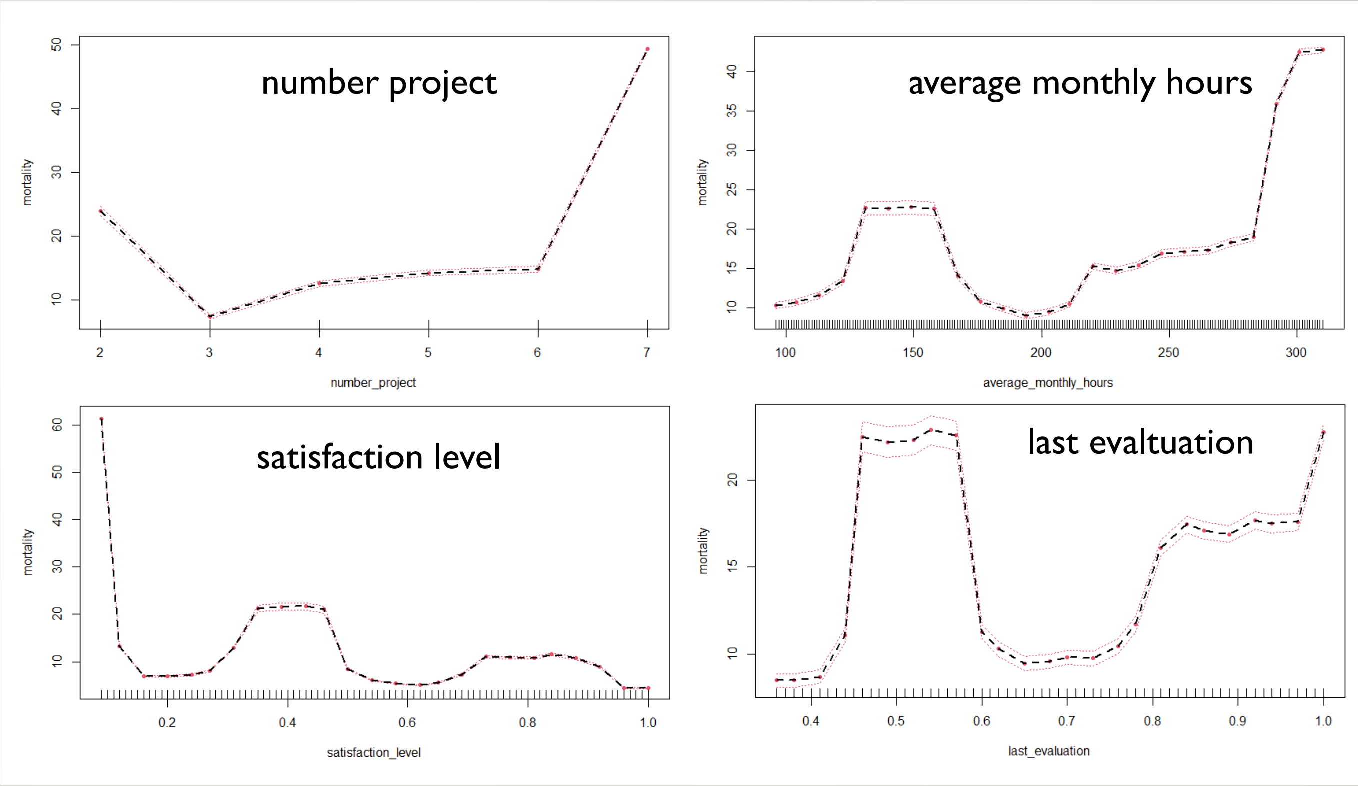


VARIABLE IMPORTANCE

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Variable importance

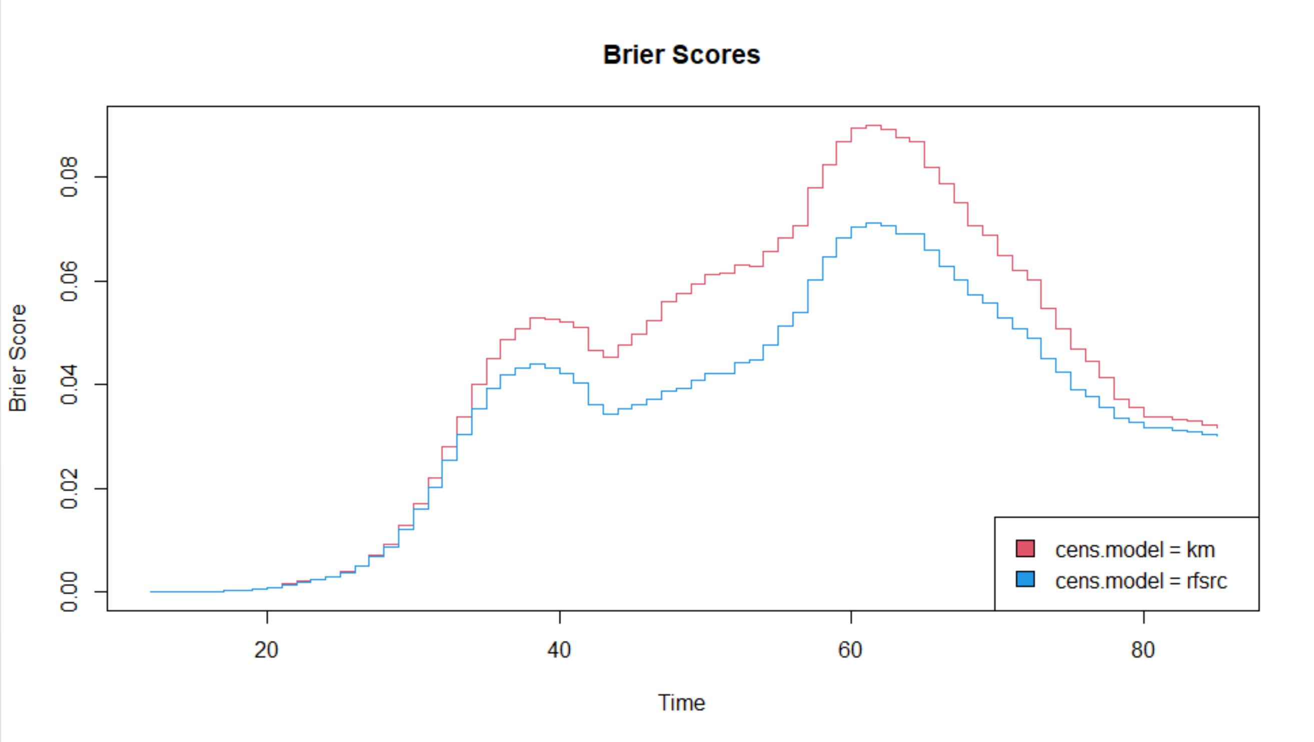


RSF RESULTS EVALUATION

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

RSF and Kaplan-Meier Brier Scores

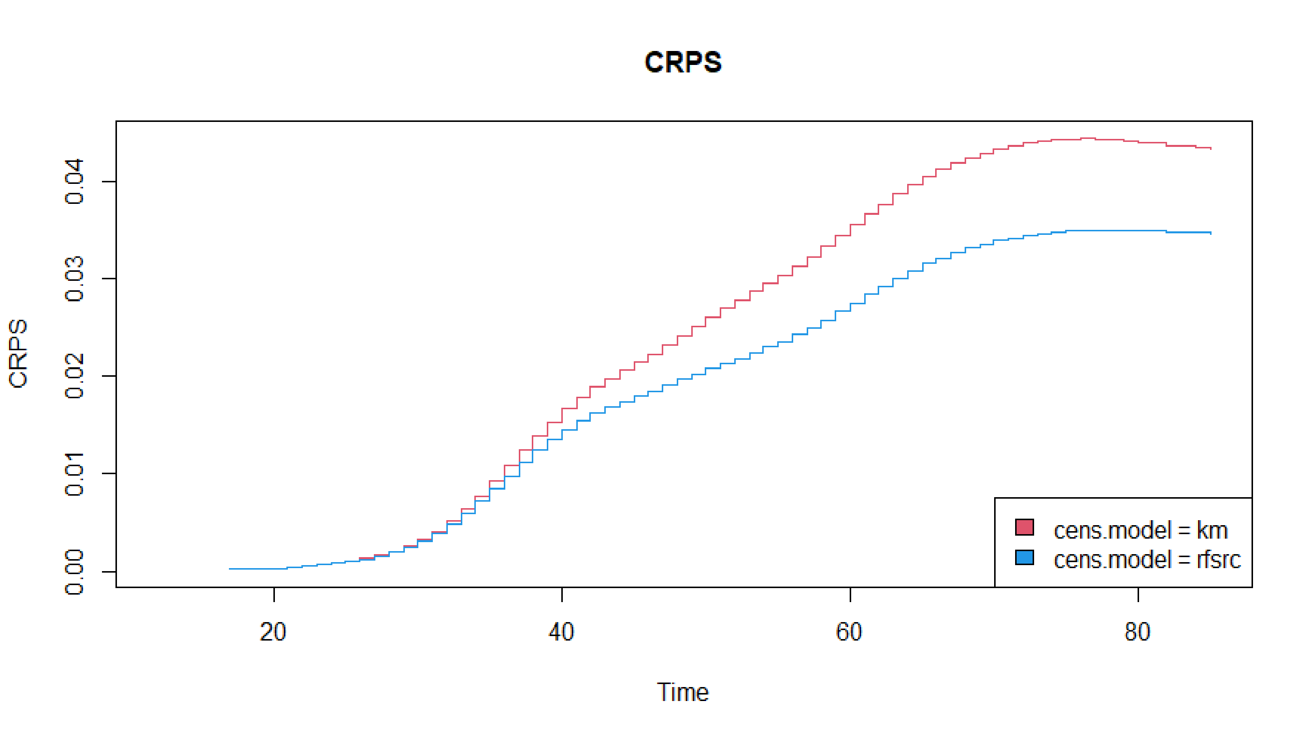


RSF RESULTS EVALUATION

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

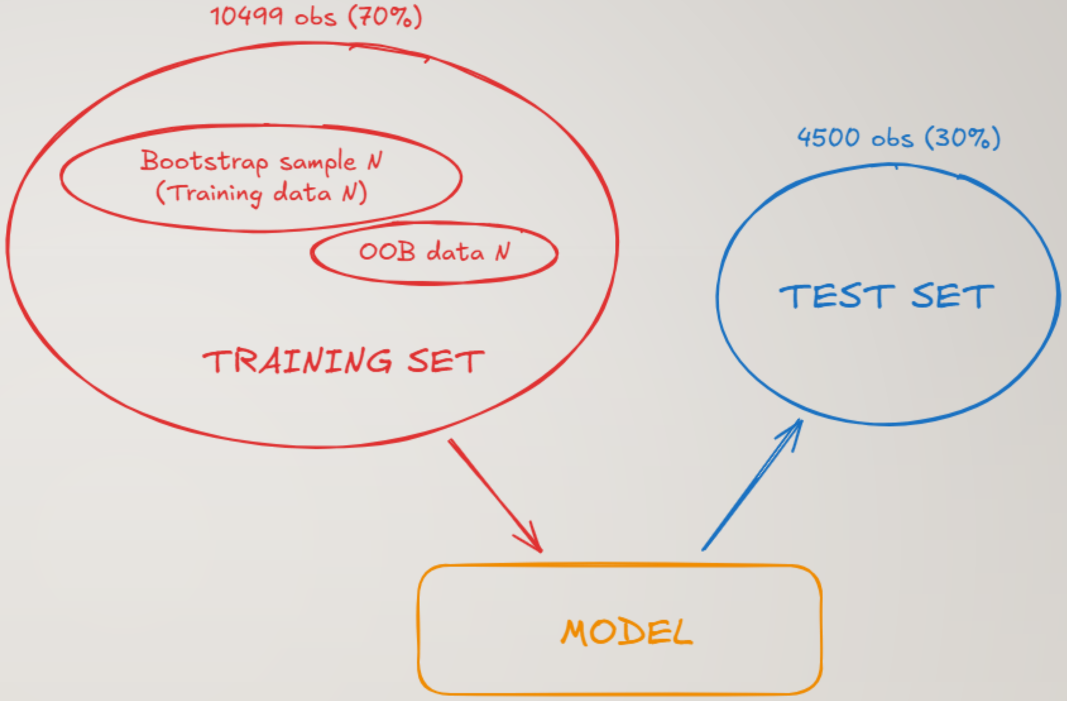
RSF and Kaplan-Meier CPRS Scores



RSF RESULTS EVALUATION

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN



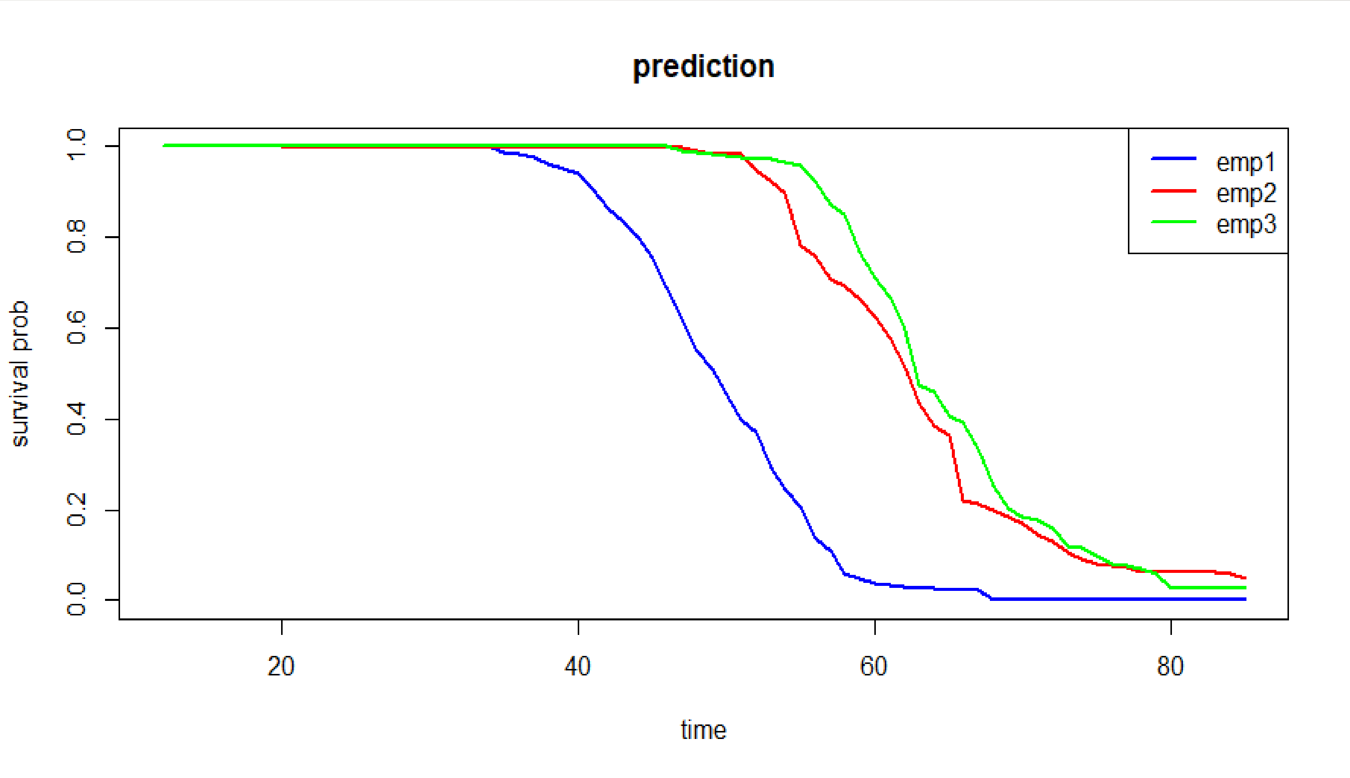
	LOGRANK
(OOB) C-index	**0.9112**
(TEST SET) C-index	**0.9138**

PREDICTION ON TEST SET

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Prediction on a sample of employees (TEST SET)

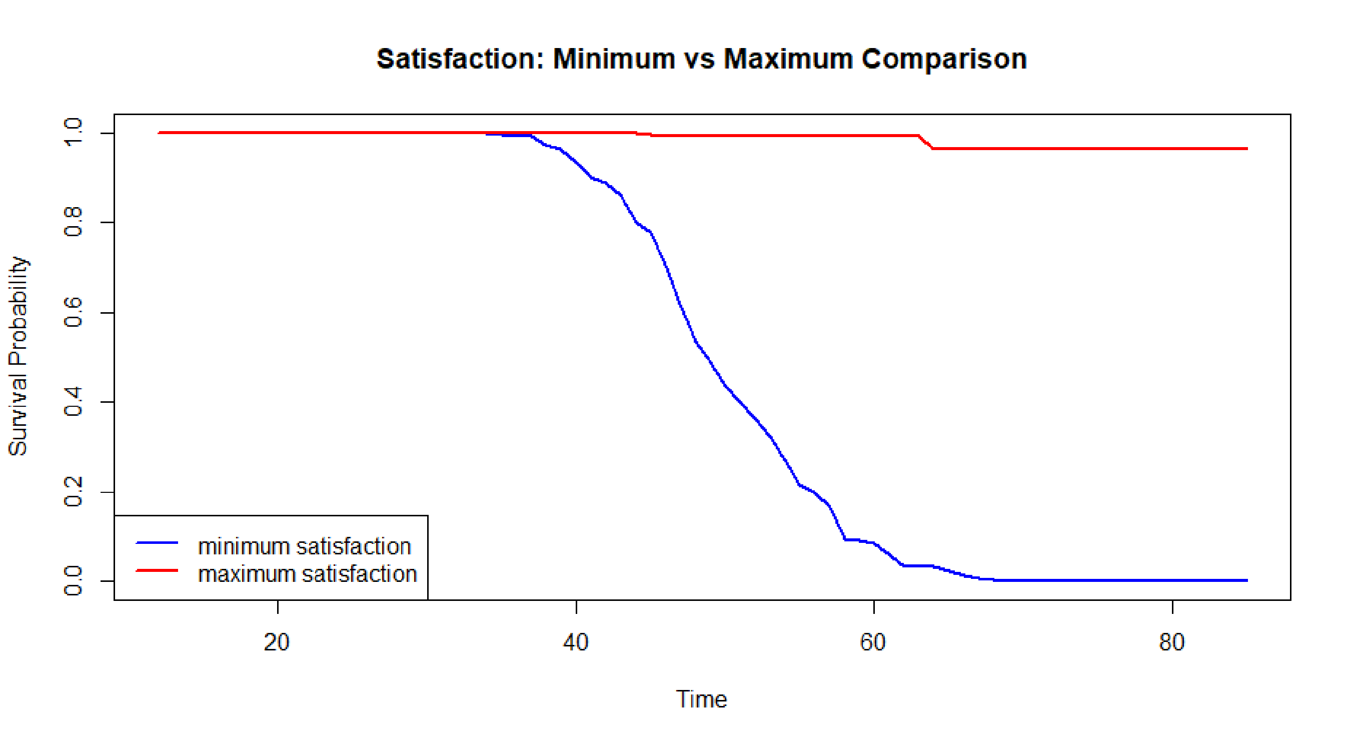


PREDICTION ON TEST SET

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Prediction for two employees with different satisfaction level



PREDICTION ON TEST SET

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

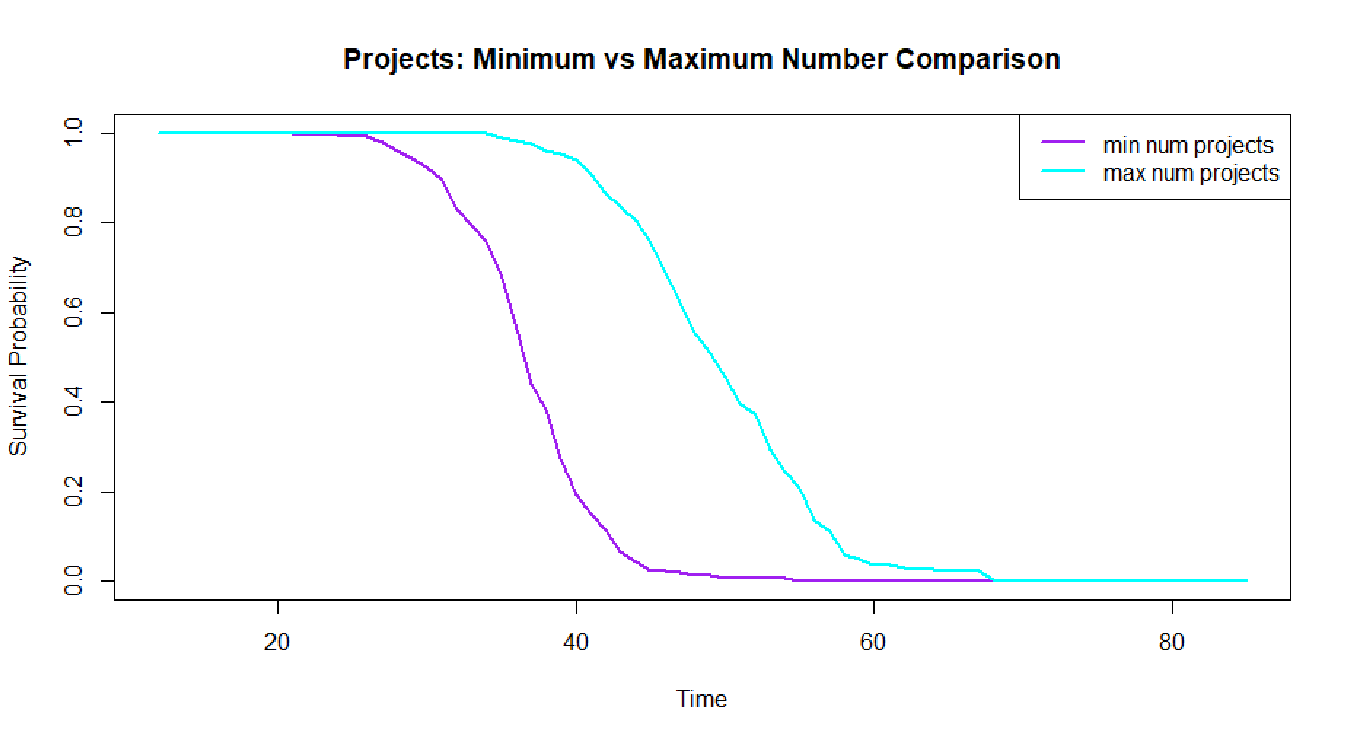
Prediction for two employees with different salary level

PREDICTION ON TEST SET

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

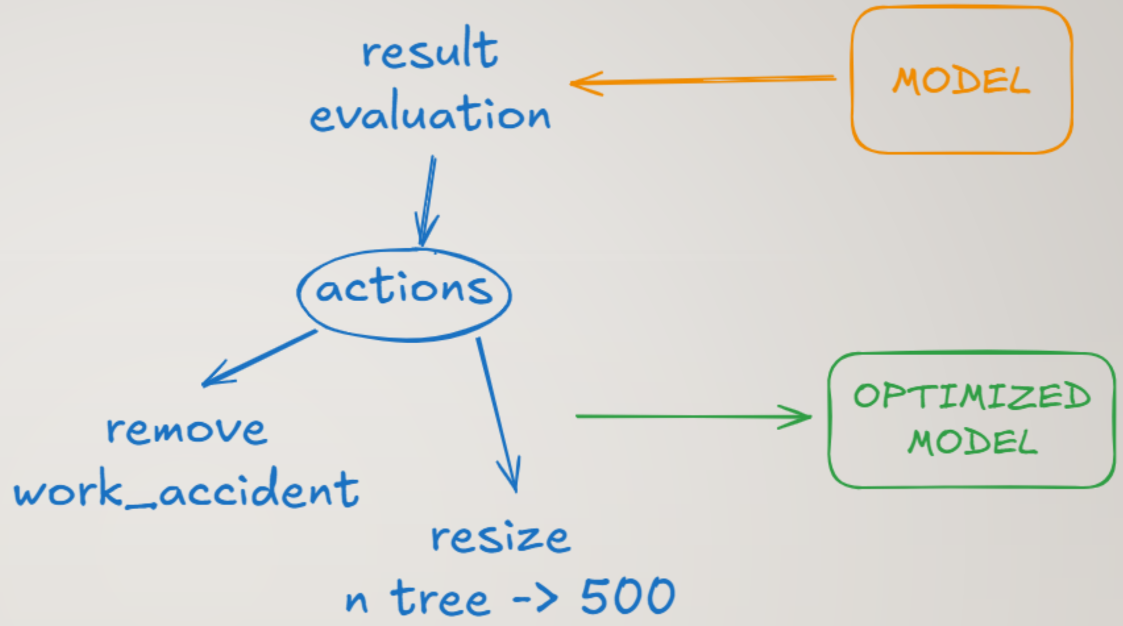
Prediction for two employees with different no. of projects



OPTIMIZATION

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN



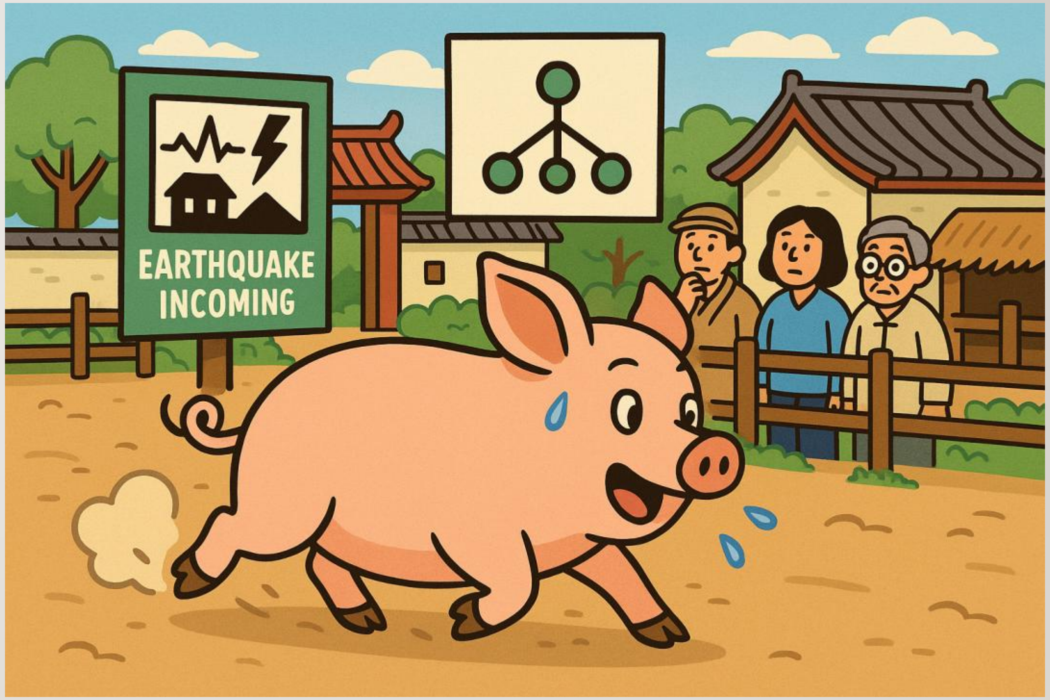
	Model	Optimized Model
C-index	**0.9138**	**0.9261**

WARNING

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

Chinese Pig Prediction Problem



CONCLUSIONS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

- Random Survival Forests excelled in handling **censored data** and non-proportional hazards
- Variable importance and parameter tuning proved essential for **optimization**
- The availability of a large dataset (15000 observations) positively impacted **sampling** and model **robustness**
- Having 7 covariates allowed clearer identification of key variables, enabling a more robust evaluation of variable **importance**.
- Made effective use of numerical covariates without the need for transformations, addressing traditional **limitations** in survival models.
- Practical implementation in R confirmed their real-world **applicability** and reliability.

TREE BASED MODELS

RANDOM SURVIVAL FOREST

EMPLOYEE CHURN

THANKS