# Survival and Longitudinal Data Analysis - Project

Data Saiyentist

November 2022

## 1 Introduction

One challenge that large organizations face today is the problem of understanding and predicting which employees are going to leave the business, called **employee turnover prediction**. Indeed, if you have a large workforce, then you may want to be able to **predict which employees are at risk of leaving at any given time, how long they are expected to stay**, and get a hint of which interventions may have a chance of reducing attrition (of valuable employees). Furthermore, **frequent employment turnover can create a major money loss** [1] in the company. So you want to identify and address quickly the issues which cause employees to leave from your company.
Survival Analysis is one of the best approach to predict employee turnover. Indeed, contrary to classification methods, we would be able to predict individual quitting risk.

Here are the first lines of the dataset we will study [2] in R :

```
Rows: 1,129
Columns: 16
$ duration      <dbl> 7.030801, 22.965092, 15.934292, 15.934292, 8.410678,…
$ event         <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
$ gender        <fct> m, m, f, f, m, f, f, f, f, f, f, m, f, f, f, f, f, f…
$ age           <dbl> 35, 33, 35, 35, 32, 42, 42, 28, 29, 30, 40, 23, 22, …
$ industry      <fct> Banks, Banks, PowerGeneration, PowerGeneration, Reta…
$ profession    <fct> HR, HR, HR, HR, Commercial, HR, HR, HR, HR, Marketin…
$ traffic       <fct> rabrecNErab, empjs, rabrecNErab, rabrecNErab, youjs,…
$ coach         <fct> no, no, no, no, yes, yes, yes, no, no, yes, no, yes,…
$ head_gender   <fct> f, m, m, m, f, m, m, m, f, m, m, m, f, f, f, m, m, m…
$ greywage      <fct> white, white, white, white, white, white, white, whi…
$ transport     <fct> bus, bus, bus, bus, bus, bus, bus, bus, bus, bus, ca…
$ extraversion  <dbl> 6.2, 6.2, 6.2, 5.4, 3.0, 6.2, 6.2, 3.8, 8.6, 5.4, 8.…
$ independ      <dbl> 4.1, 4.1, 6.2, 7.6, 4.1, 6.2, 6.2, 5.5, 6.9, 5.5, 4.…
$ selfcontrol   <dbl> 5.7, 5.7, 2.6, 4.9, 8.0, 4.1, 4.1, 8.0, 2.6, 3.3, 1.…
$ anxiety       <dbl> 7.1, 7.1, 4.8, 2.5, 7.1, 5.6, 5.6, 4.0, 4.0, 7.9, 7.…
$ novator       <dbl> 8.3, 8.3, 8.3, 6.7, 3.7, 6.7, 6.7, 4.4, 7.5, 8.3, 6.…
```

Figure 1: First rows of `turnover`

And here are the features' description :

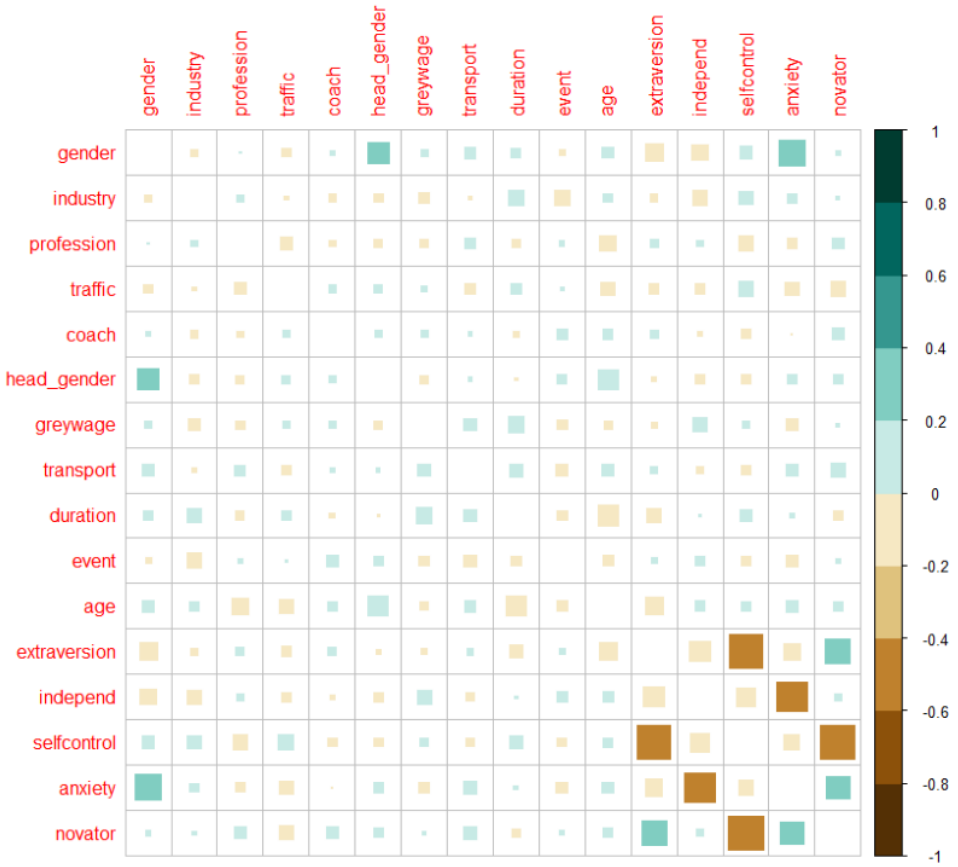
- `duration` : Experience in months.
- `event` : Censorship flag (1 if quit, 0 otherwise).
- `gender` : Gender (`f` for female or `m` for male).
- `age` : Age in years.
- `industry` : Employee's industry (`Agriculture`, `Banks`, `Building`, `Consult`, `HoReCa`, `IT`, `manufacture`, `Mining`, `Pharma`, `PowerGeneration`, `RealEstate`, `Retail`, `State`, `Telecom`, `transport`, `etc`).
- `profession` : Employee's profession (`Accounting`, `BusinessDevelopment`, `Commercial`, `Consult`, `Engineer`, `Finanñe`, `HR`, `IT`, `Law`, `manage`, `Marketing`, `PR`, `Sales`, `Teaching`, `etc`).

- `traffic` : How employee came to the company :
  - `advert` (direct contact of one's own initiative).
  - `recNErab` (direct contact on the recommendation of a friend, not an employ of the company).
  - `referal` (direct contact on the recommendation of a friend, an employee of the company).
  - `youjs` (applied on a job site).
  - `KA` (recruiting agency brought).
  - `rabrecNErab` (employer contacted on the recommendation of a person who knows the employee).
  - `empjs` (employer reached on the job site).
- `coach`: Presence of a coach on probation (`my head`, `yes` or `no`).
- `head_gender`: Gender of the supervisor (`f` for female or `m` for male).
- `greywage` : Whether the salary is fully registered with tax authorities (`white` otherwise `grey`).
- `transport` : Employee's means of transportation (`bus`, `car` or `foot`).
- `extraversion`, `independ`, `selfcontrol`, `anxiety` and `novator` :
  Scores between 1 and 10 given by Big Five Personality Test.

## 2 Data Preprocessing

- We had to convert categorical variables into `factor` for R interpretability.
- There weren't any missing values in the dataset `turnover`.
- There were 13 duplicates in the dataset, so we removed them.

Now let's look at the correlation matrix associated with `turnover` :

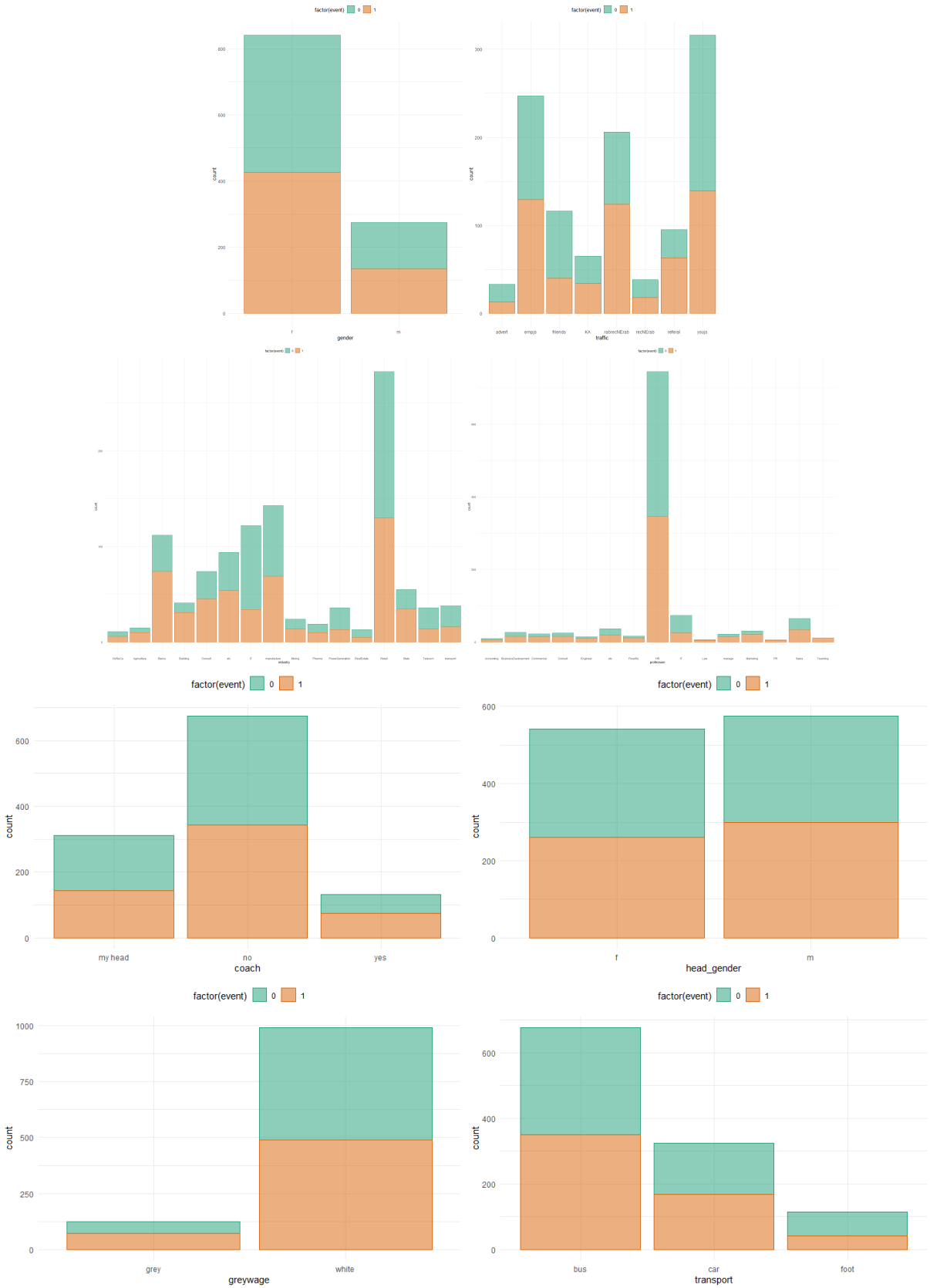



`Figure 2:` Correlation matrix

We notice that there aren't covariates that are correlated significantly (ie. values close to -1 or 1). Consequently, we decided to conserve all features in `turnover`.

# 3 Data visualization

We plotted histograms for continuous covariates and bar charts for discrete ones by coloring according to the value of `event` as follows :



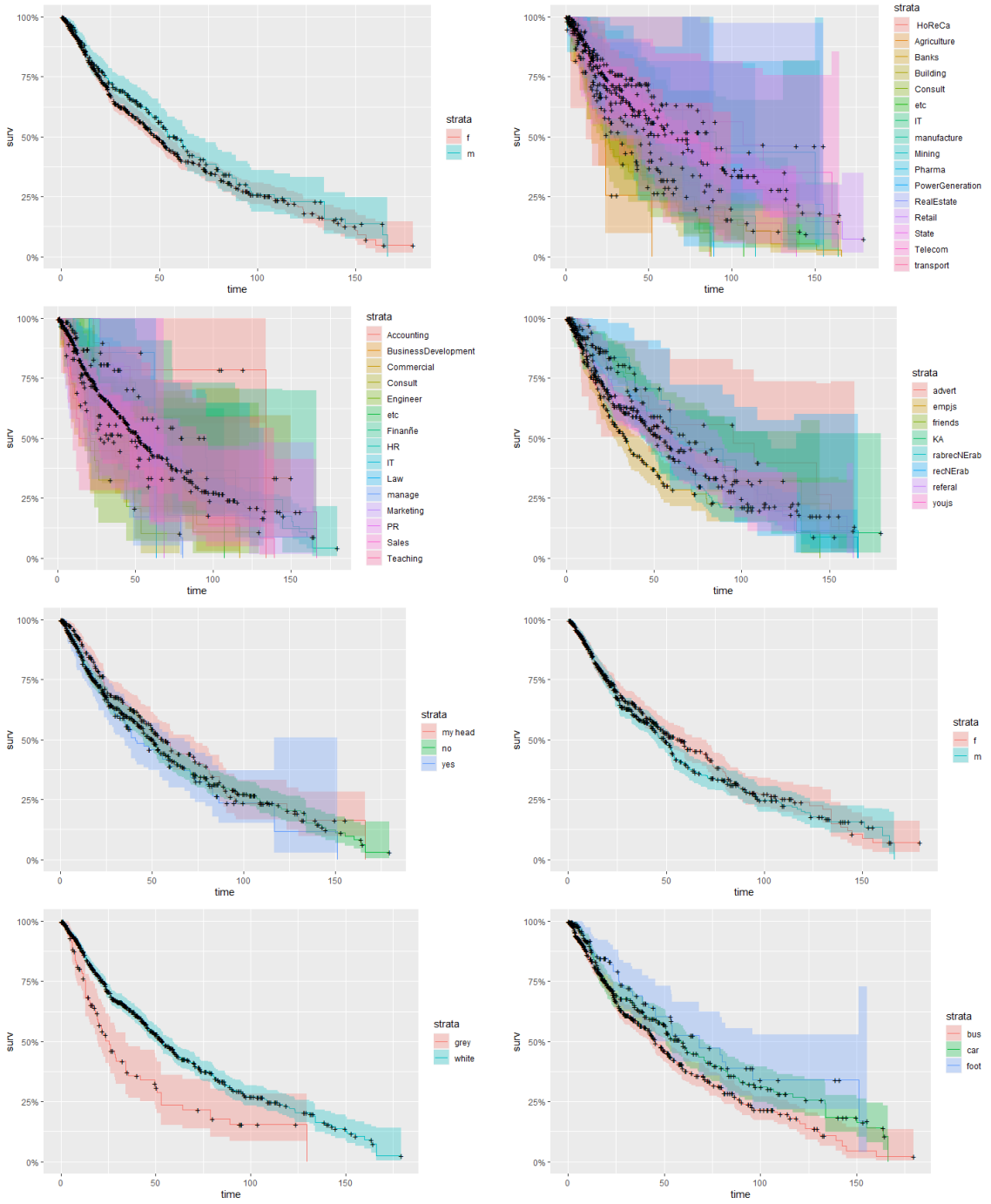

`Figure 3:` Histograms & Bar charts of `turnover` features colored according to `event`

For now, we can only observe that almost half of samples is censored (more precisely, 49.82079 %).

# 4 Survival & Longitudinal Data Analysis

## 4.1 Comparing survival distributions

Let's graphically represent the survival functions in the subgroups defined by the categorical variables :




`Figure 4:` Survival functions for each categorical variables

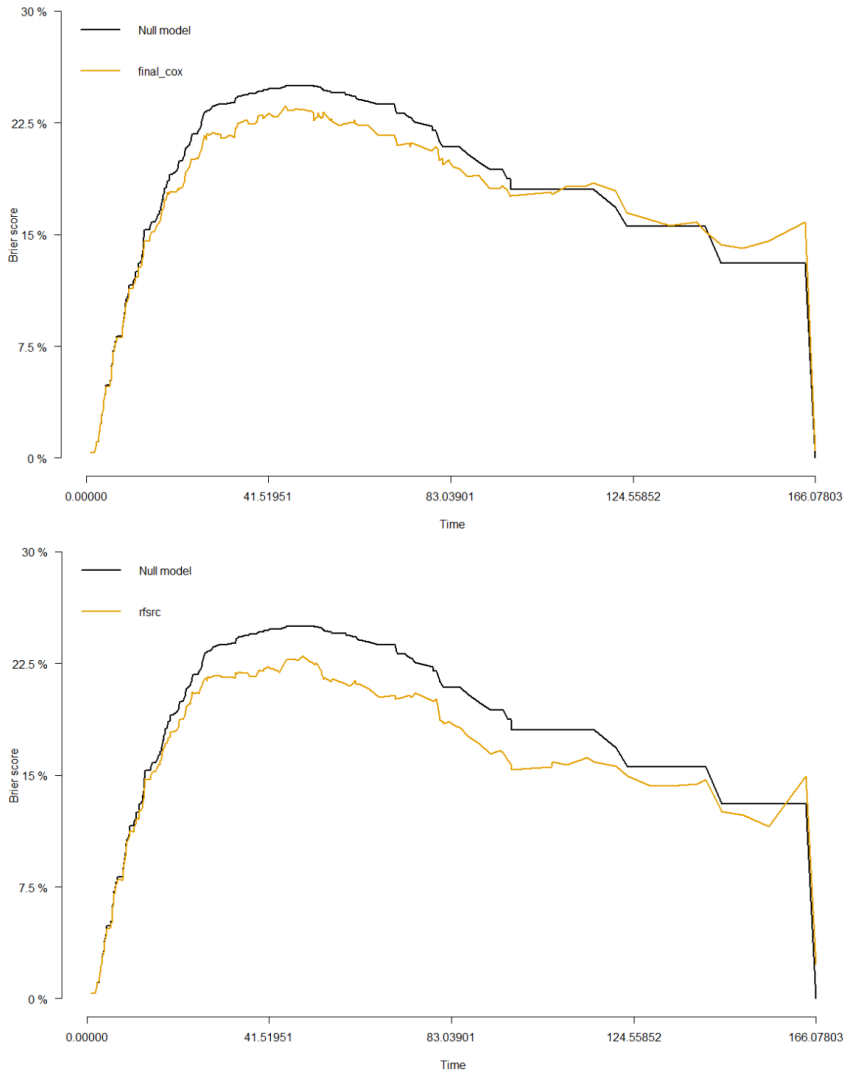
We see that some variables might have an influence on employee turnover (like `industry`), because survival curves of subgroups seems to be different.
However, we can do more than just analyze survival curves. Analysis like that represent a limited use case of the potential of survival analysis for turnover modeling, because we are using it for the aggregated level of the data. Instead of that, **we can create survival curves for individual cases, based on the effects of covariates and try to predict when employees leave**.

## 4.2 Hazard modeling

Now, we will compare survival analysis methods, especially **Cox model** and **survival Random Forests** (from the library `randomForestSRC` [3]). And to compare their performances, we will create a 75/25 partition of data in `train` and `test` samples via the `caret` library (Yet, the partitions are stratified well such that there is approximately the same percentage of censorship in `train` and `test`).
Here are the model used (and trained with `train`) :

- **Cox model** : `coxph(Surv(duration, event) ~ profession + industry + coach + independ + head_gender + anxiety + transport + gender + novator + extraversion + selfcontrol + traffic, data = train)`
  *The sparse subset of covariates was given by a forward procedure thanks to the BIC criterion.*
- **survival Random Forest** : `rfsrc(Surv(duration, event) ~ ., data = train)`

For model comparison, we computed the Brier score [4] (from `riskRegression` library [5]) as a function of time :

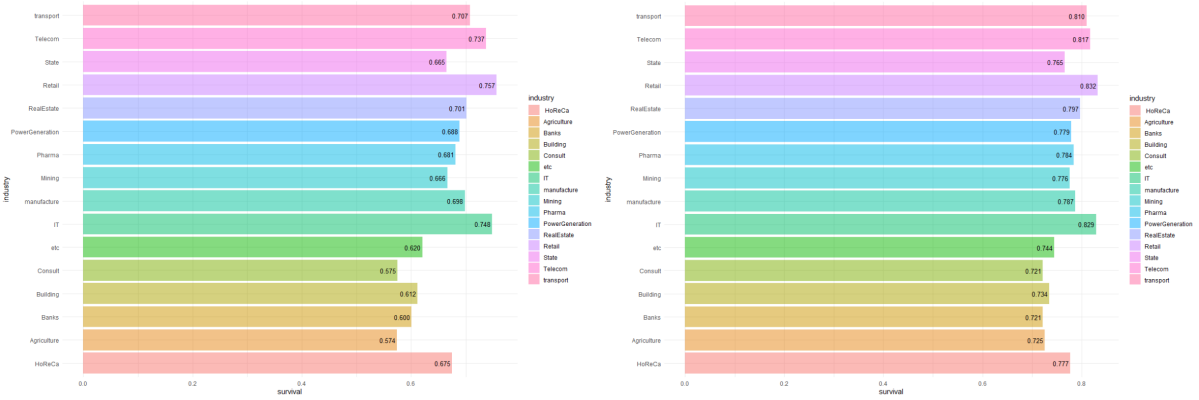



`Figure 5:` Brier scores over time (given by `test`)

But, we prefer to compute the Integrated Brier Score (IBS) [4] that provides a better overview of model performances. Thus, we obtained $IBS_{Cox} = 17.69507 > \boldsymbol{IBS_{RandomForest}} = \mathbf{16.51271}$.
So, we decided to take the model given by survival Random Forest, because it gave us the lowest IBS.
*The IBS was computed using the formula in [4], especially using the trapezoid method [6] to approximate the integral of Brier score (ie. the area under the curves in* ***Figure 5****).*

Finally, let's test that model (with different `industry`) on the following profiles :

- Considering an employee whose features are Female of age 30, referred by an employee of the company (referral), profession HR, commuting by bus, having a coach during the probation, with male supervisor, whose characteristic scores are 5 for all categories, let's give an estimate of the probability that this employee will stay for longer than 3 years.
- Considering another employee with the same profile as above but who has already worked for one year, let's give an estimate of the probability that this employee will stay for another 2 years.




`Figure 6:` Estimated Probabilities of respective profiles (with different `industry`)

We observe that, for both profiles, the `industry` has an effect on the probabilities. More generally, this is the main reason why we should not decide beforehand (like in section **4.4**) whether a covariate has an influence or not on a survival problem.
Now let's fix their `industry` profile as `IT` for instance. In theory, we might think that nowadays, the longer you stay in a company, the higher chances of leaving are (for example, you want to move elsewhere, to work for another company, aso.). Yet, we notice the inverse in practice (with our case), because the second probability is higher than the first one ($0.829 > 0.748$). In other words, a person that has already worked one year has a high chance to stay two years within the same company (ie. three years in total).

# 5 Conclusion

The benefits of applying an employee turnover prediction model like this extend beyond its pure prediction capabilities towards insights that can modify the operations of the organization as a whole. The cost savings to the organization are two-fold as HR professionals can use the model's explanations to develop retention policies across the business and also target high risk individuals with retention initiatives.

# 6 References


[1] Work Institute's 2019 retention report
[2] Employee turnover dataset shared from Edward Babushkin's blog
[3] randomForestSRC documentation
[4] Brier score and IBS documentation from Python library `pysurvival`
[5] riskRegression documentation
[6] A brief explanation of trapezoid method