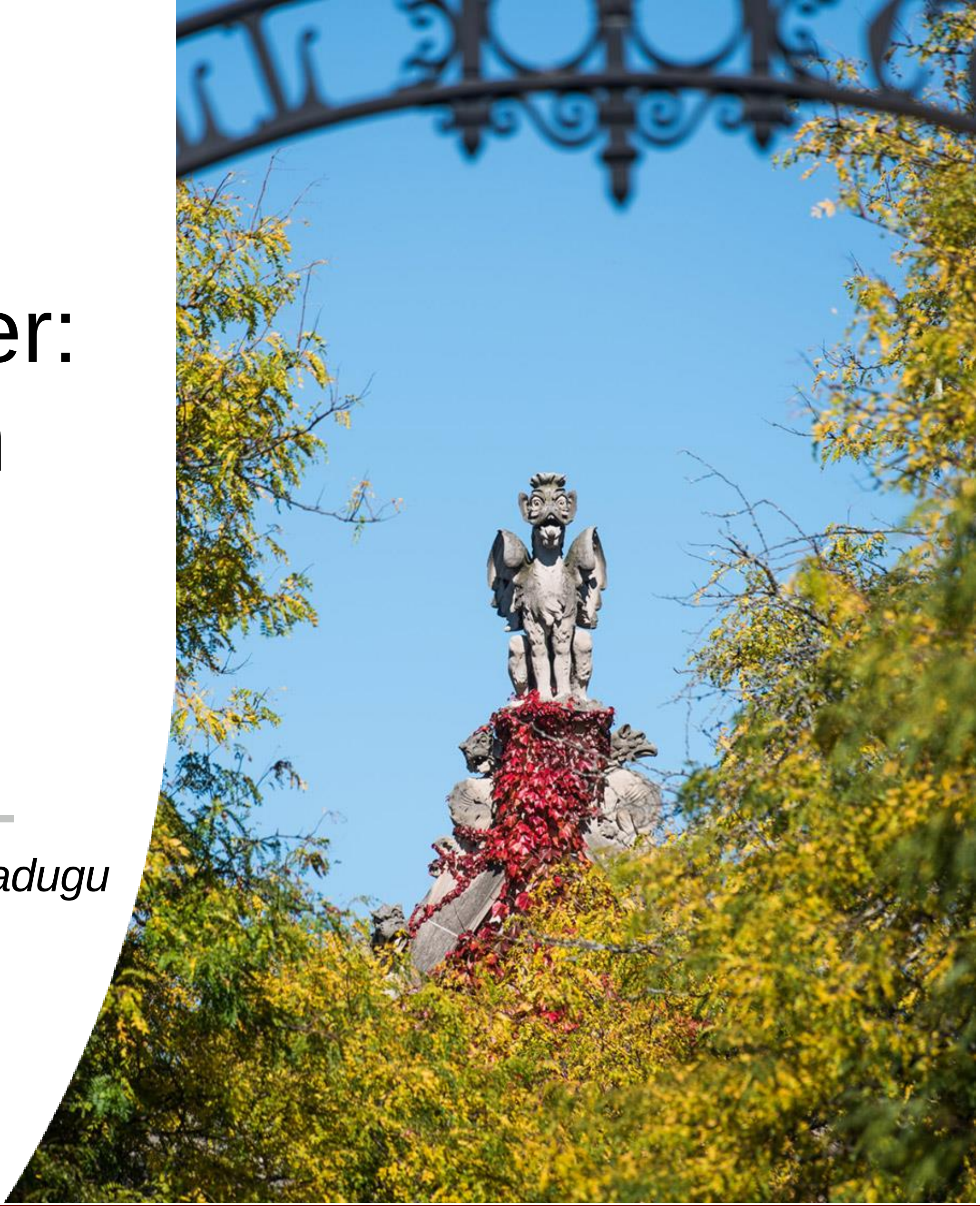
"Forecasting Employee Turnover: A Predictive Analytics Approach for Human Resource Management"

Nicolas Ferreira, Kargil Thakur, Kishor Mannur & Sirivanth Paladugu

MSCA 31010 IP08 Linear and Non-Linear Models
March 5th, 2023



THE UNIVERSITY OF CHICAGO

Agenda

- Problem Statement
- Data Description
- EDA
- Feature Engineering
- Models Selection & Validation
- Interpretations
- Future Scope

Problem Statement

High employee turnover is a major challenge faced by organizations worldwide, as it results in increased recruitment and training costs, decreased productivity, and a negative impact on organizational culture. We aim to develop a predictive model that can identify who, when, and why an employee leaves by answering these:

- **Which** employees are most likely to leave?
- **When** are they likely to leave?
- **Why** are they leaving the organization?

Then, we can take proactive measures to retain valuable employees, improve employee satisfaction, and reduce the costs associated with high turnover rates.

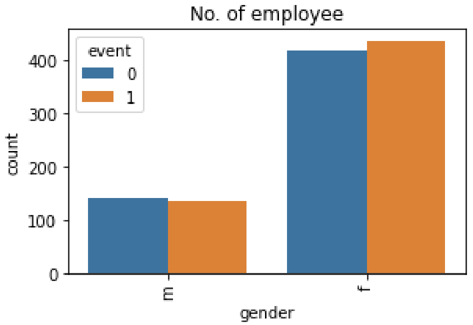
Literature Review

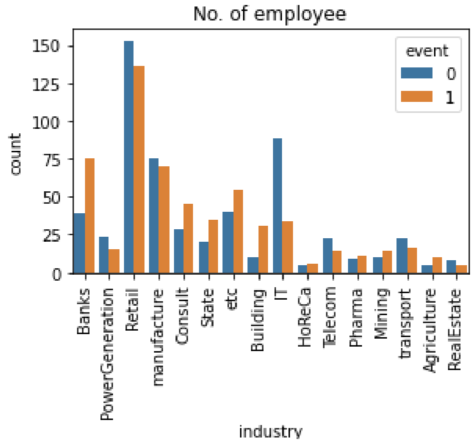
- Our study found that existing machine learning-based methods mainly focus on feature engineering for binary prediction tasks, **ignoring historical events** of turnover behaviors.
- We also study that one of the papers proposes an event-based approach and uses strategies to analyze survival data with censored records for employees with **multiple turnover records**.

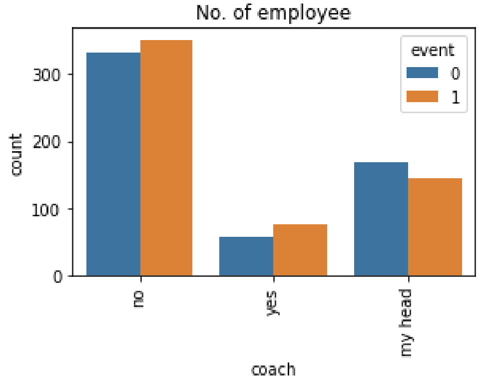
Data Description

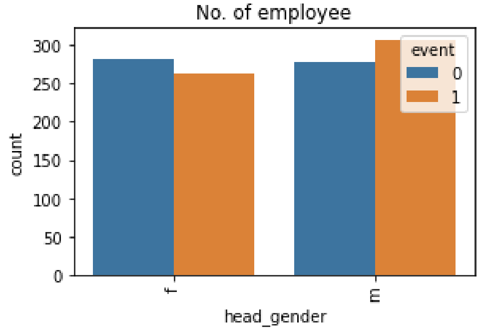
Feature name	Brief description
stag	**Experience (time)**
event	**Employee turnover**
gender	Employee's gender, female (f) or male (m)
age	Employee's age (year)
industry	Employee's Industry
profession	Employee's profession
traffic	**From what pipeline employee came to the company**
coach	Presence of a coach (training) on probation
head_gender	head (supervisor) gender
greywage	Salary Taxes. Greywage in Russia or Ukraine means that the employer (company) pay
way	**Employee's way of transportation**
Extraversion, Independ, Self-control, Anxiety and Novator scores	Scores on various psychological scores

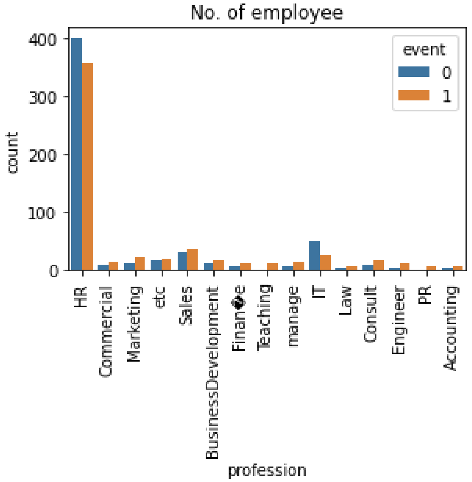
EDA (Exploratory Data Analysis)

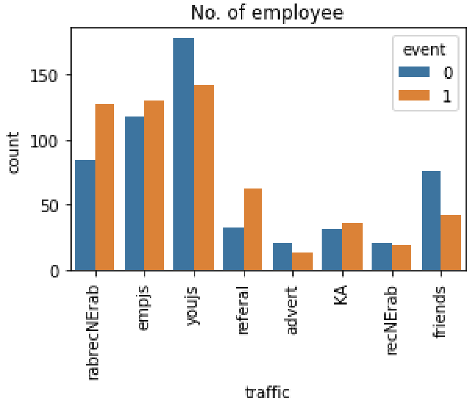


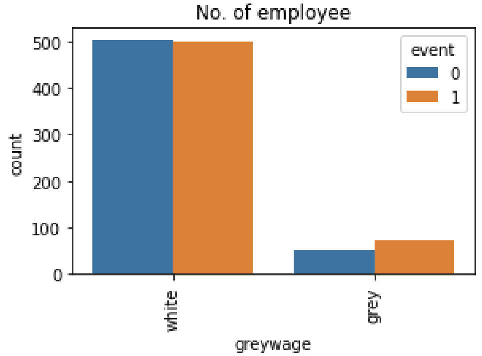


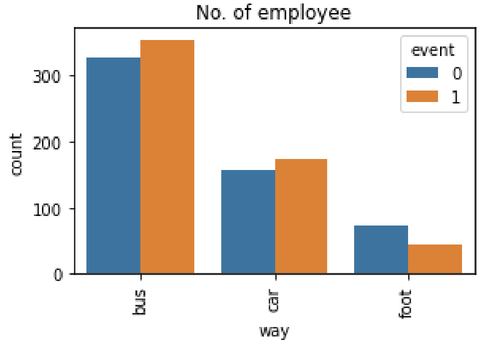












Feature Engineering

We can transform the categorical variables with two values to a binary version of them

Binary values

- gender
- head_gender
- greywage

Categorical variables to one-hot encode

- traffic
- coach
- way

Categorical variables to bin and transform

- industry
- profession

Every other numerical variable will be kept as it is

Take top 5 values and group the others as 'OTHER'

Strip and upper every categorical variable first

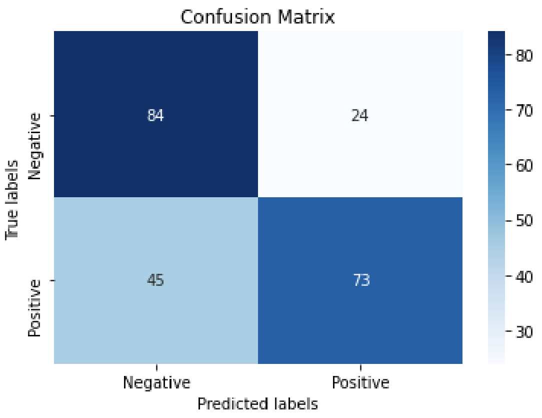
Who will churn? Binary classification problem

We approached this problem like a **binary classification problem.**

Given an 80/20 split of the data. We iterated over different models and found the **Random Forest Classifier** to be the best.

	Model	Score
5	Random_Forest	69.03
6	Naive_Bayes	67.70
0	Logistic_Regression	64.16
4	Decision_Tree	61.95
2	Linear_SVC	58.85
3	KNN	58.85
7	Perceptron	52.65
1	Support_Vector_Machines	51.77
8	Stochastic_Gradient_Decent	47.79

	precision	recall	f1-score	support
0	0.65	0.78	0.71	108
1	0.75	0.62	0.68	118
accuracy			0.69	226
macro avg	0.70	0.70	0.69	226
weighted avg	0.70	0.69	0.69	226



Given our train and test set. We achieve an overall accuracy score of approximately 0.7 with f-1 scores around the same value of 0.7.
Our model does a good job of predicting people that will churn given our set of features.

When will they churn? Time till event prediction

Survival analysis is a statistical technique used to analyze data where the outcome of interest is time until an event occurs, such as death, failure, or in this case, employee churn.

We use this in analyzing when an employee will churn.

Survival analysis will answer a question like:

- How long until the employee churns?

Whereas a logistic regression can only answer a question more like this:

- Will an employee churn or not?

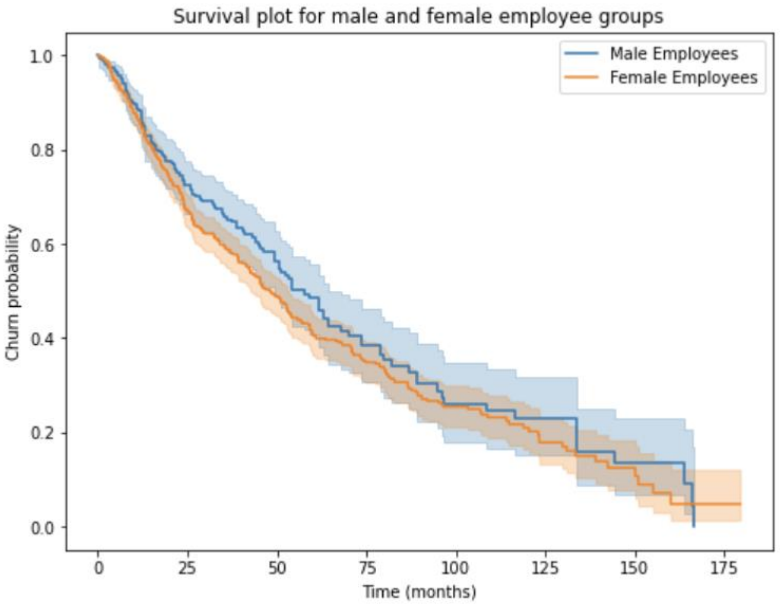
Kaplan-Meier Estimator

A non-parametric method to compute survival probabilities and estimate the survival function

$$\widehat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right),$$

- ti - duration time
- di - number of events that happened at time ti
- ni - number of individuals known to have survived up to time ti

We use log-rank tests to compare the survival functions of two or more groups



t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	4.07	0.04	4.52

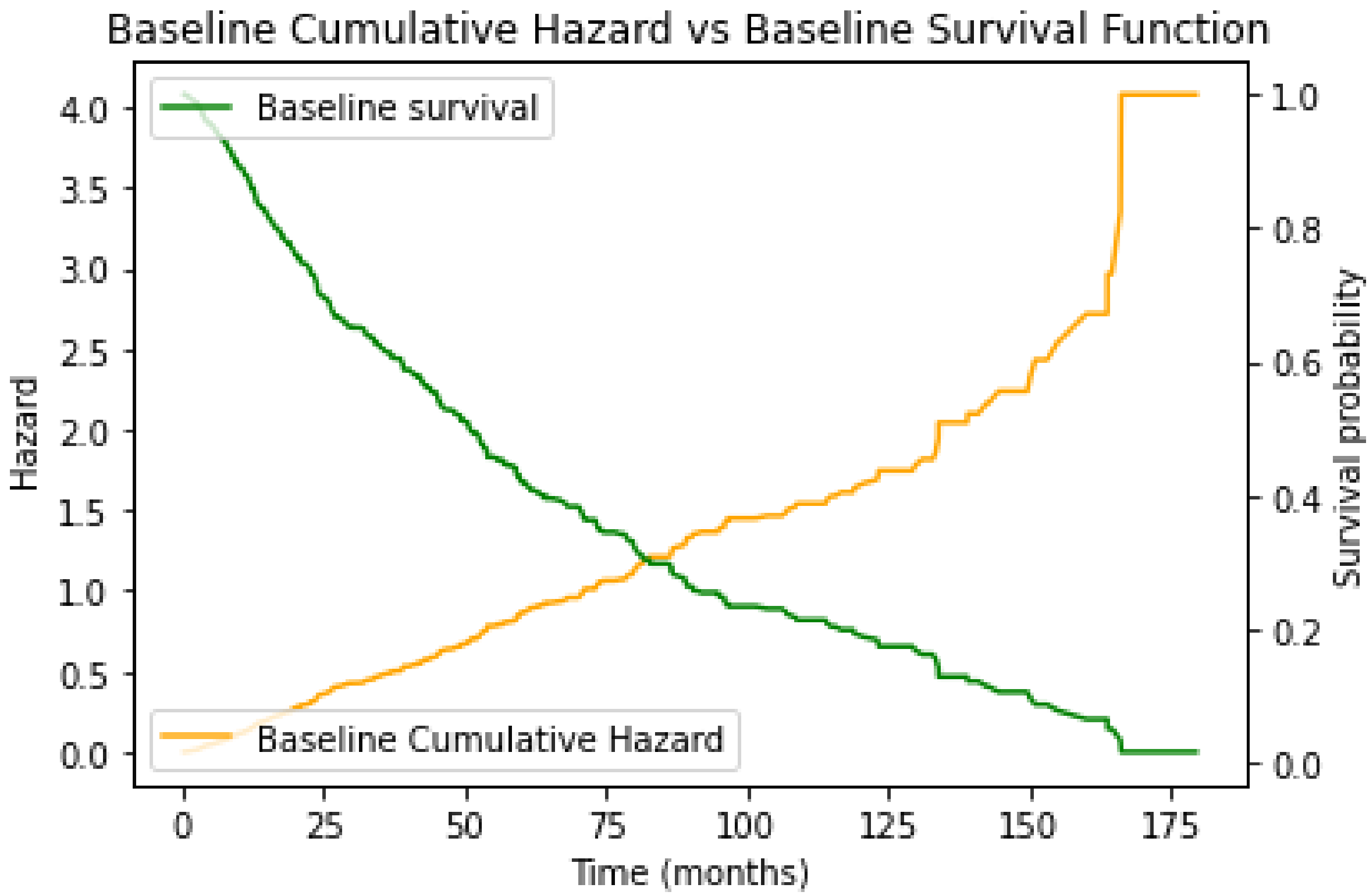
t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	22.34	<0.005	18.74

t_0	-1
null_distribution	chi squared
degrees_of_freedom	1
test_name	logrank_test

	test_statistic	p	-log2(p)
0	2.35	0.13	2.99

Cox Regression



Cox regression (or proportional hazards regression) is method for investigating the effect of several variables upon the time a specified event takes to happen.

The *hazard function* h(t) is defined as the event rate at time *t*

$$h(t) = h_0(t) * \exp(b_1x_1 + b_2x_2 + \cdots + b_nx_n)$$

Where,

$t \quad = \textit{survival time}$

$h(t) \quad = \textit{the hazard function}$

$x_1, x_1, \ldots, x_1 \quad = \textit{covariates}$

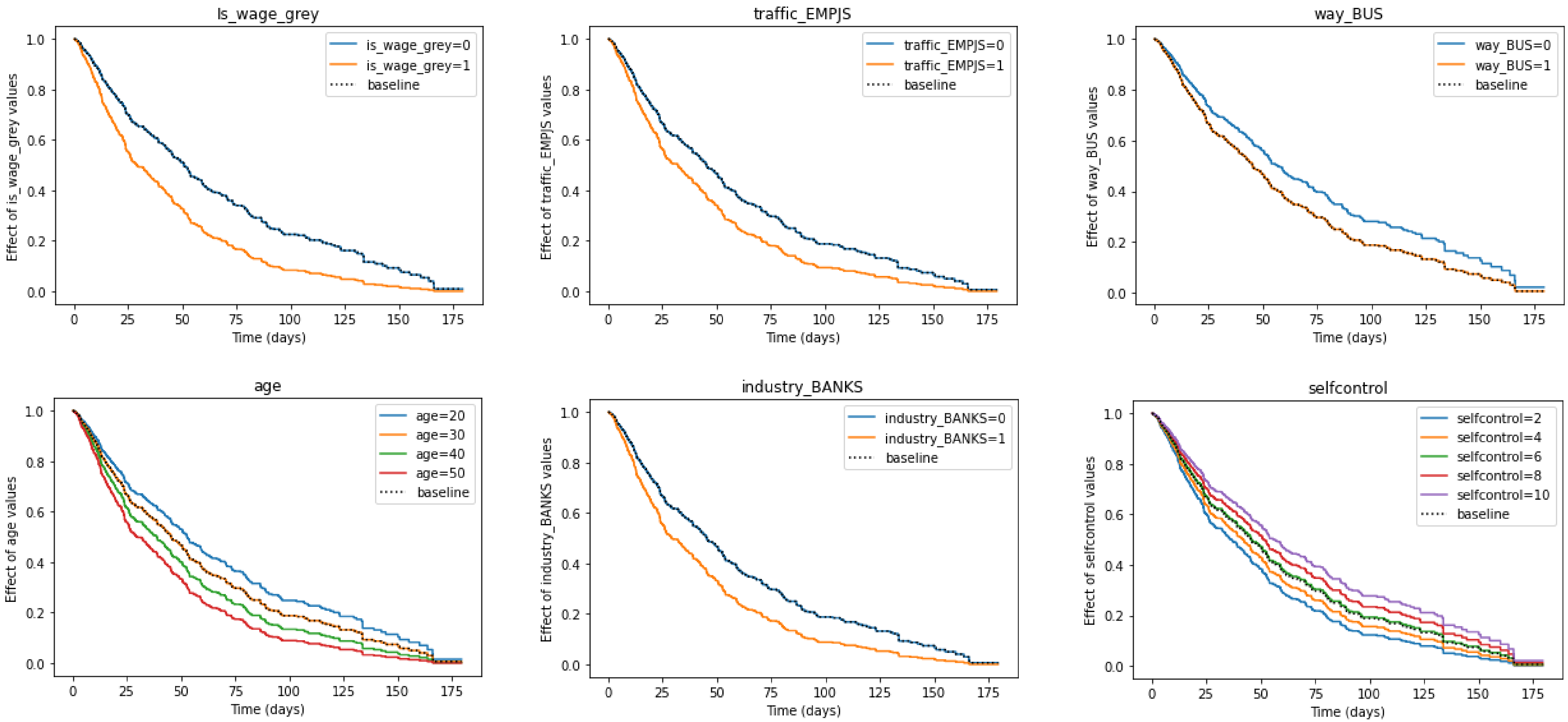
$b_1, b_2, \ldots, b_n \quad = \textit{measures the impact of covariates}$

Cox Regression

Backward Selection

	coef	exp(coef)
age	0.02	1.02
selfcontrol	-0.05	0.95
anxiety	-0.05	0.95
is_wage_grey	0.51	1.66
traffic_EMPJS	0.46	1.59
way_BUS	0.28	1.32
industry_BANKS	0.38	1.46
industry_IT	-0.46	0.63
industry_RETAIL	-0.28	0.76
profession_HR	-0.29	0.74
profession_IT	-0.51	0.60

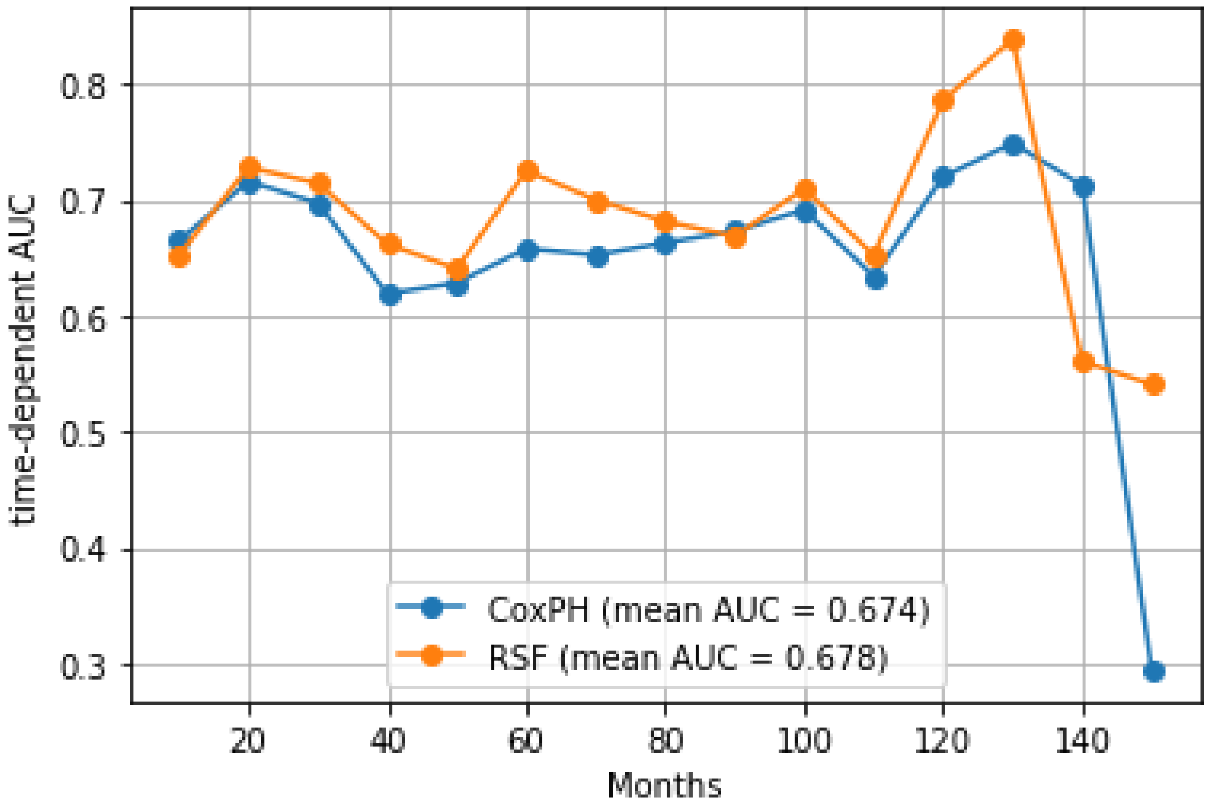
We set the predetermined p-value as 0.05. All the covariates with a p-value greater than 0.05 are eliminated. The remaining covariates are:- age, selfcontrol, anxiety, is_wage_grey, Traffic_EMPJS,way_BUS', industry_BANKS, industry_IT, industry_RETAIL, profession_HR, profession_IT.


Is_wage_grey
is_wage_grey=0
is_wage_grey=1
baseline
Effect of is_wage_grey values
traffic_EMPJS
traffic_EMPJS=0
traffic_EMPJS=1
baseline
Effect of traffic_EMPJS values
way_BUS
way_BUS=0
way_BUS=1
baseline
Effect of way_BUS values
age
age=20
age=30
age=40
age=50
baseline
Effect of age values
industry_BANKS
industry_BANKS=0
industry_BANKS=1
baseline
Effect of industry_BANKS values
selfcontrol
selfcontrol=2
selfcontrol=4
selfcontrol=6
selfcontrol=8
selfcontrol=10
baseline
Effect of selfcontrol values
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125
150
175
Time (days)

Random Survival Forest

- Accuracy / C-Index : 0.658

	importances_mean	importances_std
is_wage_grey	0.027710	0.010983
traffic_EMPJS	0.017719	0.007308
age	0.017550	0.015143
industry_BANKS	0.009552	0.005921
selfcontrol	0.008888	0.007958
way_BUS	0.008295	0.005017



Weibull AFT (Accelerated Failure Time)

Weibull is a parametric method used to model the distribution of survival times.

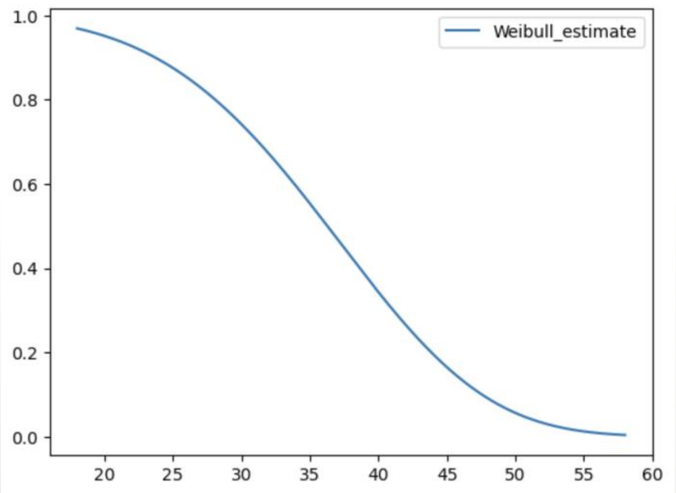
The probability density function of a Weibull random variable is ;

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

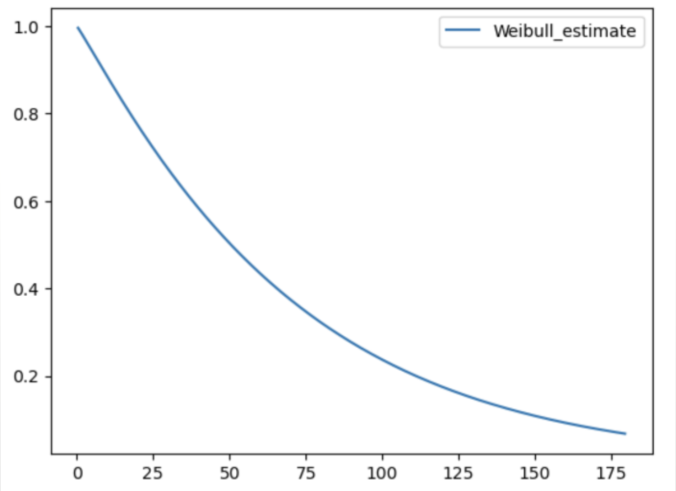
The cumulative density function of a Weibull random variable is ;

$$F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k}$$

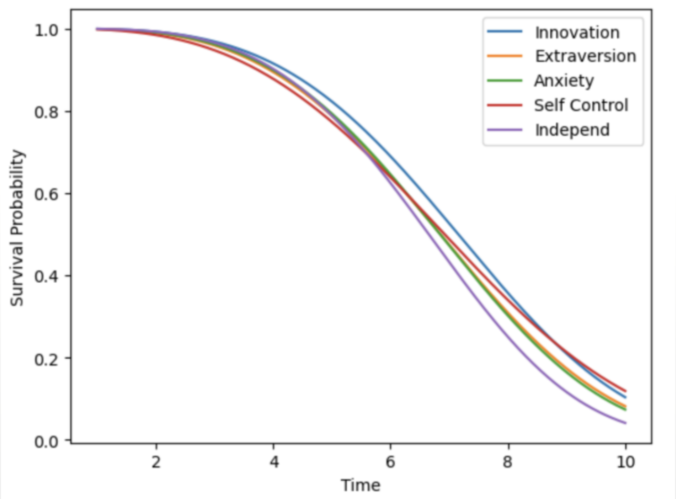
Here k is the shape parameter, λ is the scale parameter and x is time-to-failure.



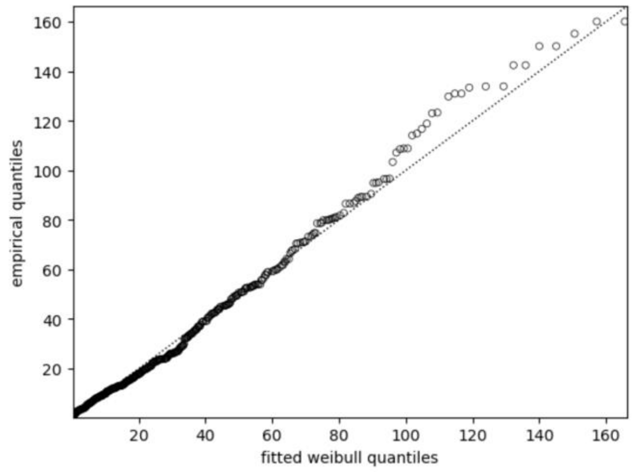
Weibull estimate for
Age vs Event

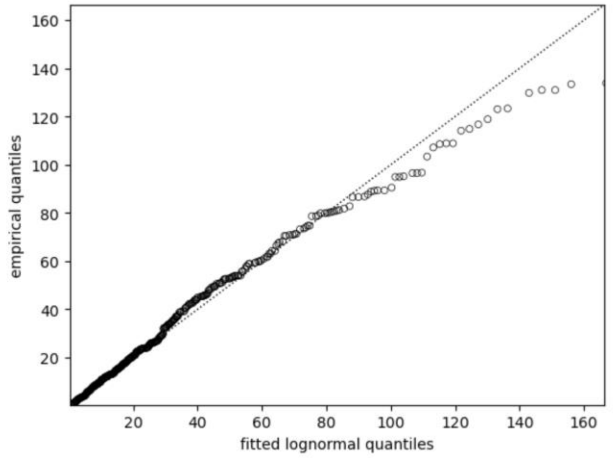


Weibull estimate for
Stag vs Event



Weibull estimate for
5 additional parameters vs Event






```
wb.predict(10)
```

0.9977167788510253

```
wb.predict(50)
```

0.05627439383943856 6

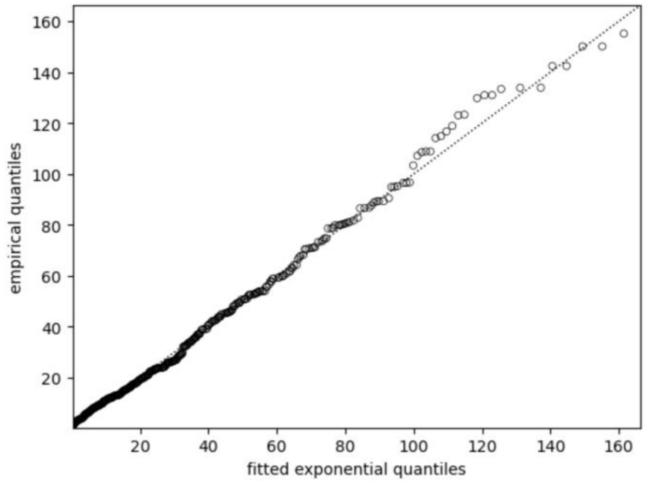
```
wb.predict(75)
```

2.8375963329618498e-08

```
wb.predict(100)
```

9.276725365714685e-28

Survival Rates at different months



Business Impact

Based on our analysis of employee churn, we identified that certain factors such as **age**, **grey wage**, **traffic**, and **way of travel** (especially **by bus**) have a significant impact on employees' leaving over time

These findings suggest that the company should focus on improving:

- **Job satisfaction** by timely promotions, and other benefits to people of higher age
- Provisioning a **transportation service** for those coming by bus (like the **UChicago Downtown Campus Connector** XD)
- Consider **implementing retention strategies** for employees who have been with the company for a certain number of years, as they are at a higher risk of leaving

For better modeling and accuracy, having more predictors like **salary**, **time off**, **promotion level, bonus** etc., could be beneficial.

Summary

- Binary Classification approach answers our question about **who** (employees) would churn
- The Kaplan-Meier Estimator helped visualize the significance between different groups.
- Cox Proportional Hazards model explained the effect of covariates with a unit increase in time thus answering **when** an employee would churn
- Weibull AFT backward selection technique was also used to identify predictors with the highest effect on employee churn.
- The Random Survival Forest approach also was helpful in determining **why** an employee would churn from the importance of each predictor (covariate)

References

- https://www.researchgate.net/publication/324055697_Balanced_Random_Survival_Forests_for_Extremely_Unbalanced_Right_Censored_Data
- https://sassofia.com/wp-content/uploads/2016/05/An-Overview-of-Weibull-Analysis.pdf
- Frierson, Jessica, and Dong Si. "Who's next: Evaluating attrition with machine learning algorithms and survival analysis." In Big Data–BigData 2018: 7th International Congress, Held as Part of the Services Conference Federation, SCF 2018, Seattle, WA, USA, June 25–30, 2018, Proceedings, 251–259. Springer, 2018
- Gemar, German, Ismael P Soler, and Vanesa F Guzman-Parra. "Predicting bankruptcy in resort hotels: a survival analysis." International Journal of Contemporary Hospitality Management 31, no. 4 (2019): 1546–1566
- Jin, Ziwei, Jiaxing Shang, Qianwen Zhu, Chen Ling, Wu Xie, and Baohua Qiang. "RFRSF: Employee turnover prediction based on random forests and survival analysis." In Web Information Systems Engineering–WISE 2020: 21st International Conference, Amsterdam, The Netherlands, October 20–24, 2020, Proceedings, Part II 21, 503–515. Springer, 2020.
- Logubayom, Ida Anuwoje, and Albert Luguterah. "Survival analysis of time to first birth after marriage." Survival 3, no. 12 (2013). Zhu, Qianwen, Jiaxing Shang, Xinjun Cai, Linli Jiang, Feiyi Liu, and Baohua Qiang. "CoxRF: Employee turnover prediction based on survival analysis." In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 1123–1130. IEEE, 2019.4

Thank you!