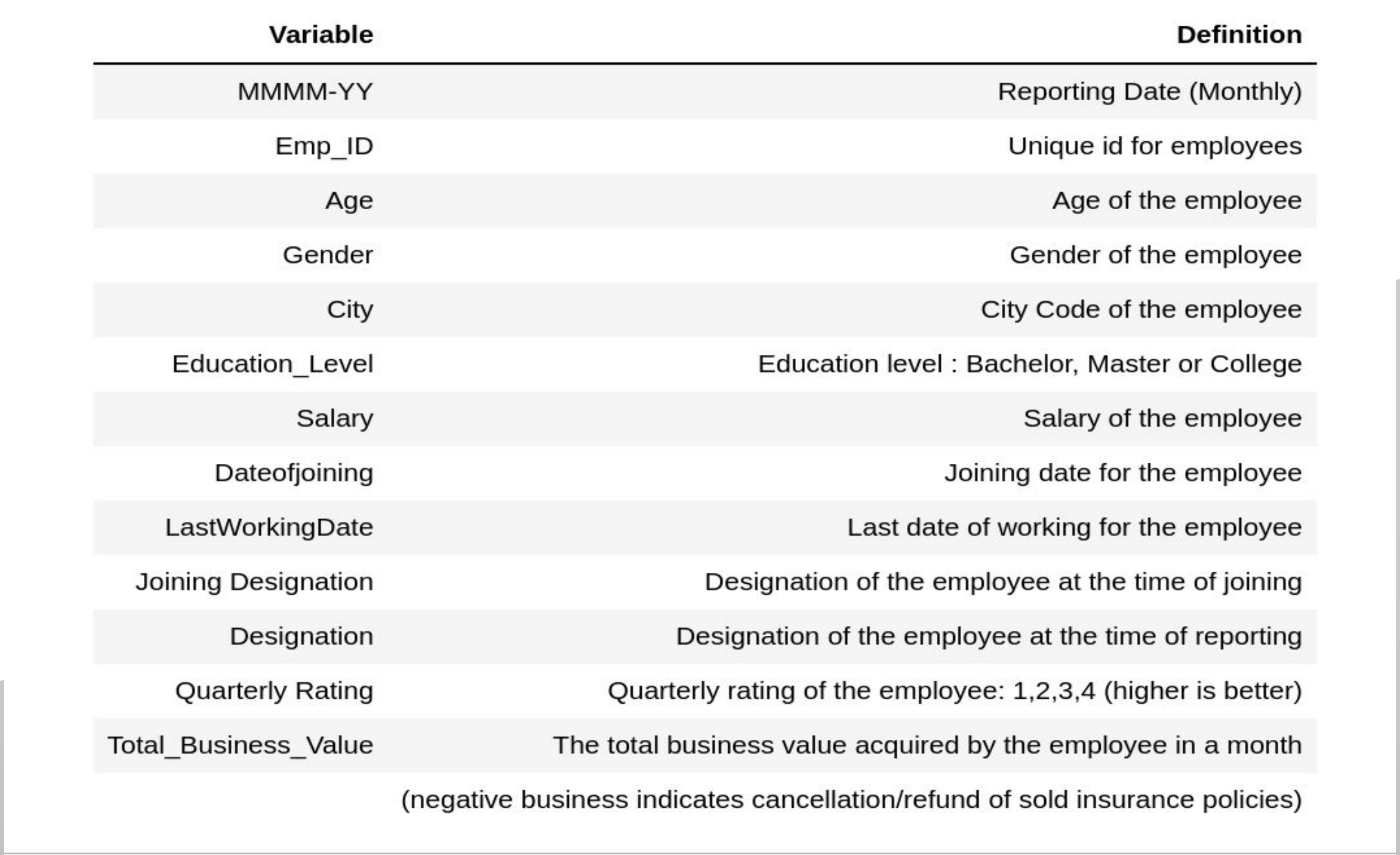
Survival Analysis on Employee Attrition Data

Presented by : **Mahendra Nandi & Sourav Karmakar**
Guided by : **Sudipta Das**

Definition of the variables of the data



Variable	Definition
MMMM-YY	Reporting Date (Monthly)
Emp_ID	Unique id for employees
Age	Age of the employee
Gender	Gender of the employee
City	City Code of the employee
Education_Level	Education level : Bachelor, Master or College
Salary	Salary of the employee
Dateofjoining	Joining date for the employee
LastWorkingDate	Last date of working for the employee
Joining Designation	Designation of the employee at the time of joining
Designation	Designation of the employee at the time of reporting
Quarterly Rating	Quarterly rating of the employee: 1,2,3,4 (higher is better)
Total_Business_Value	The total business value acquired by the employee in a month (negative business indicates cancellation/refund of sold insurance policies)

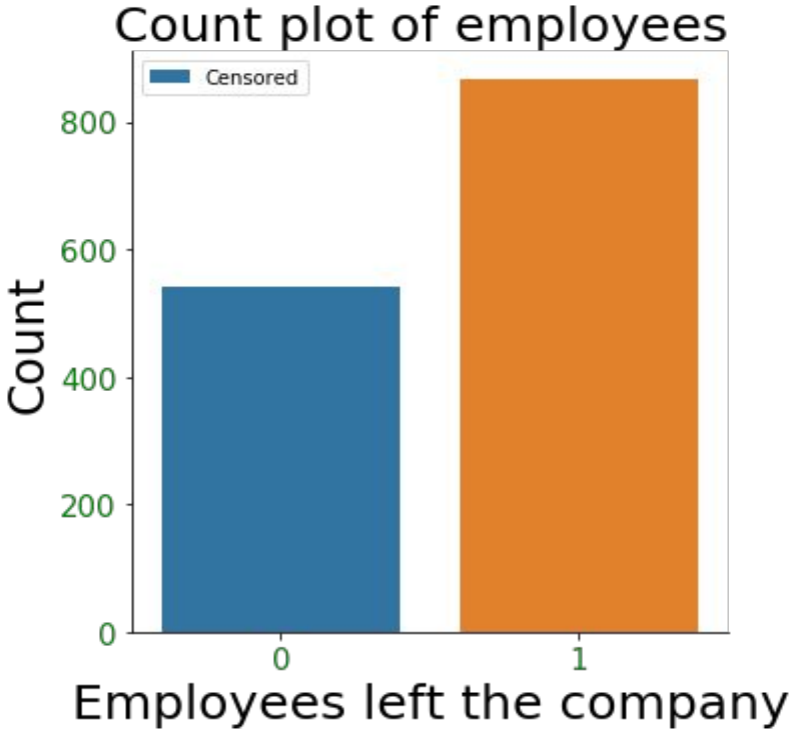
Analysis details

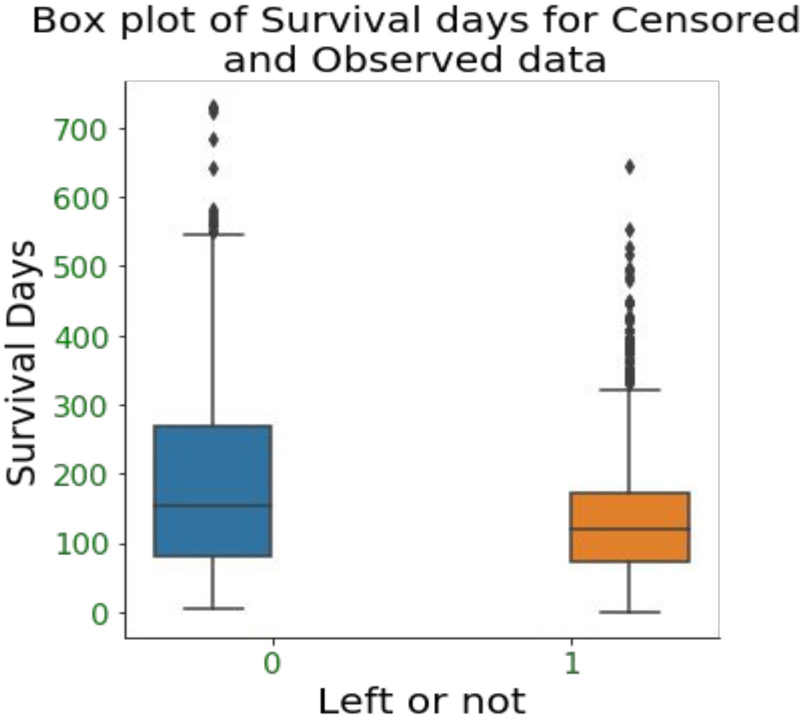
Start of study: 01st January 2016
End of study: 31st December 2017

Time of event or time of censoring:
“**survival_months**”
Binary event **indicator : “left”**

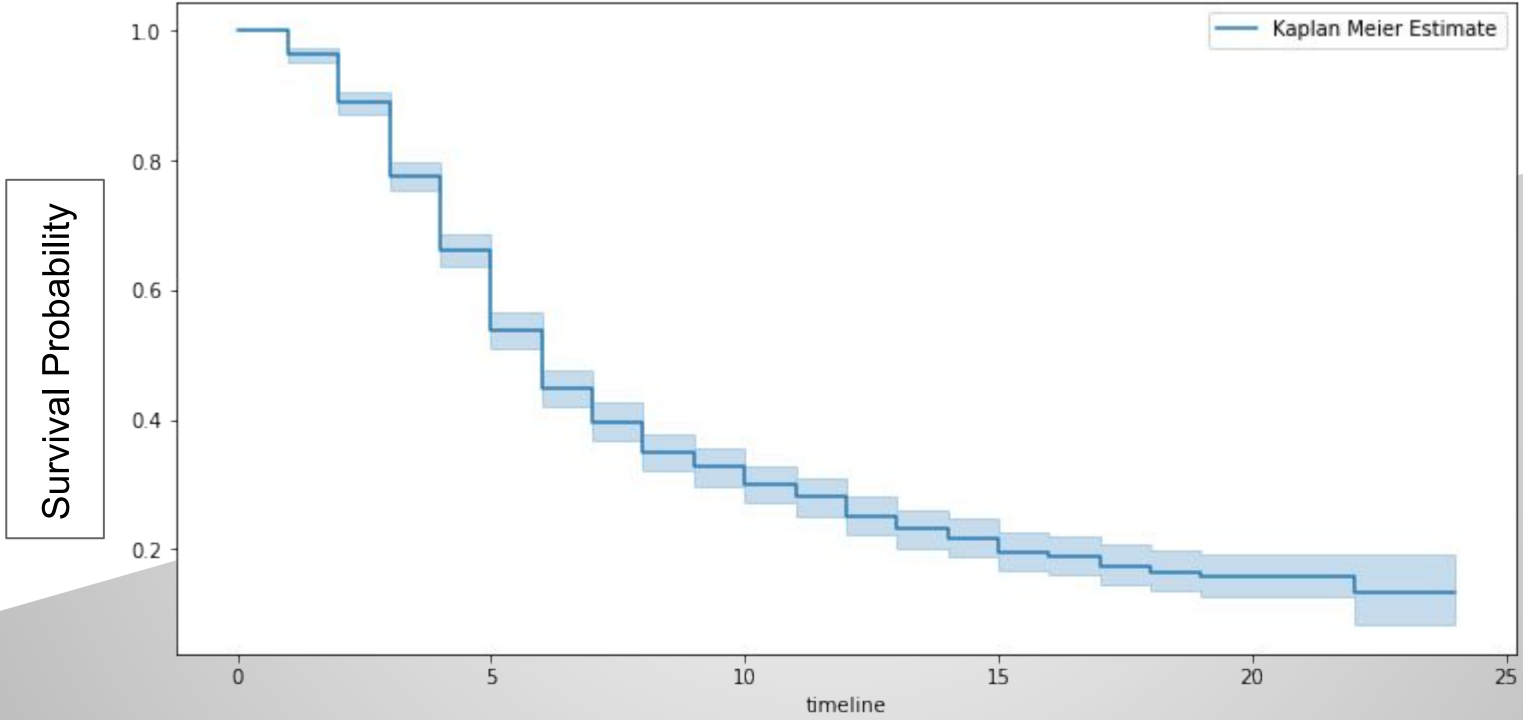
We have 1409 total observations of the employees for each month he/she is in the company,
542 right-censored observation.

Total employee attrition and survival days in the company

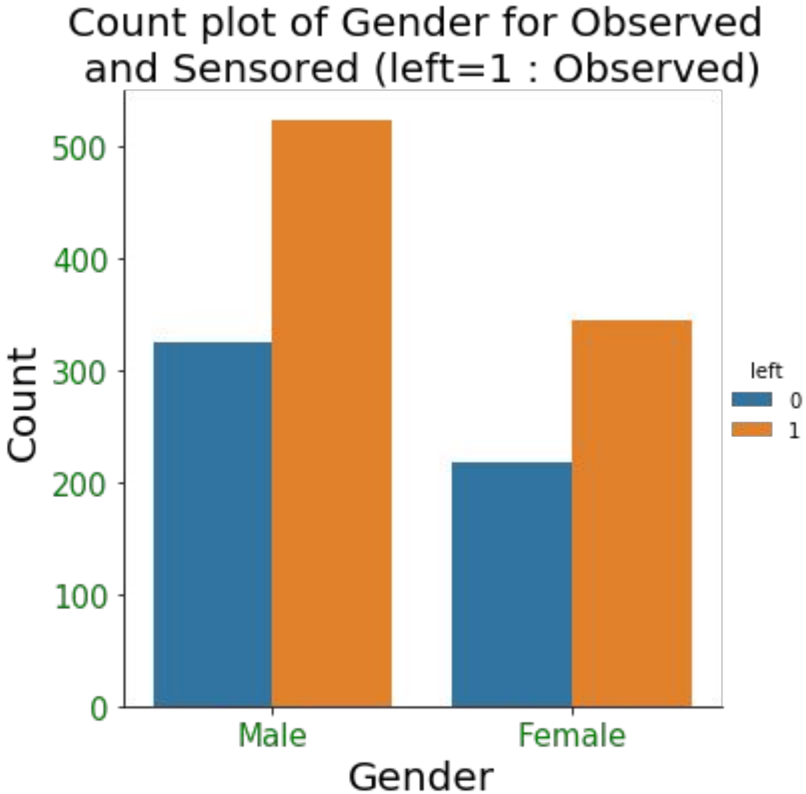


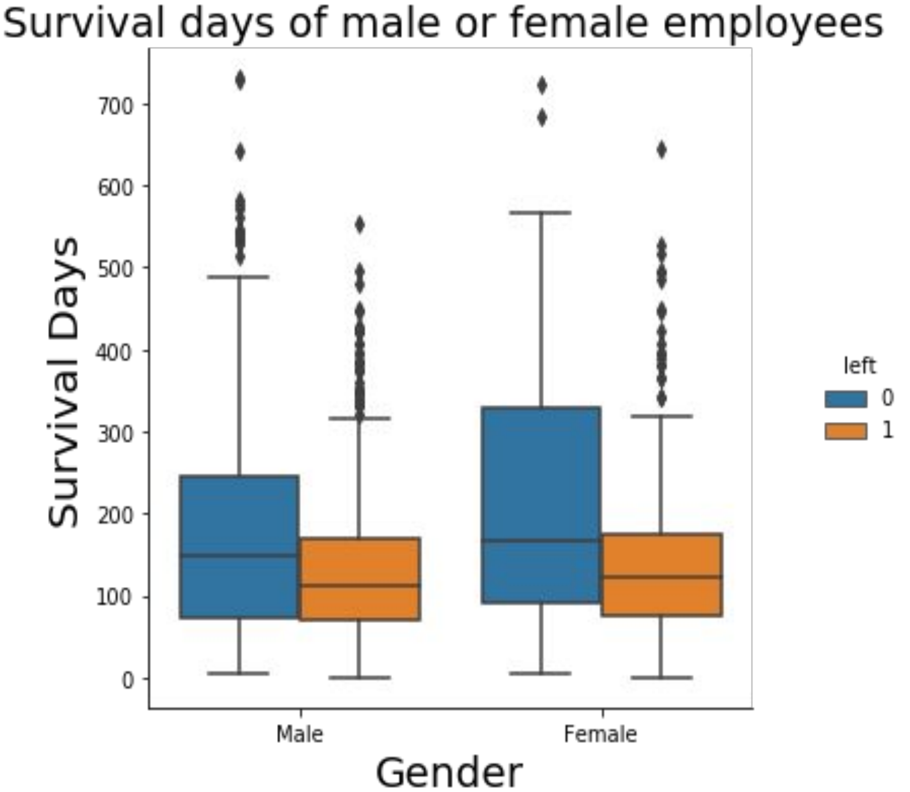


Estimate of the survival function using Kaplan Meier Estimate

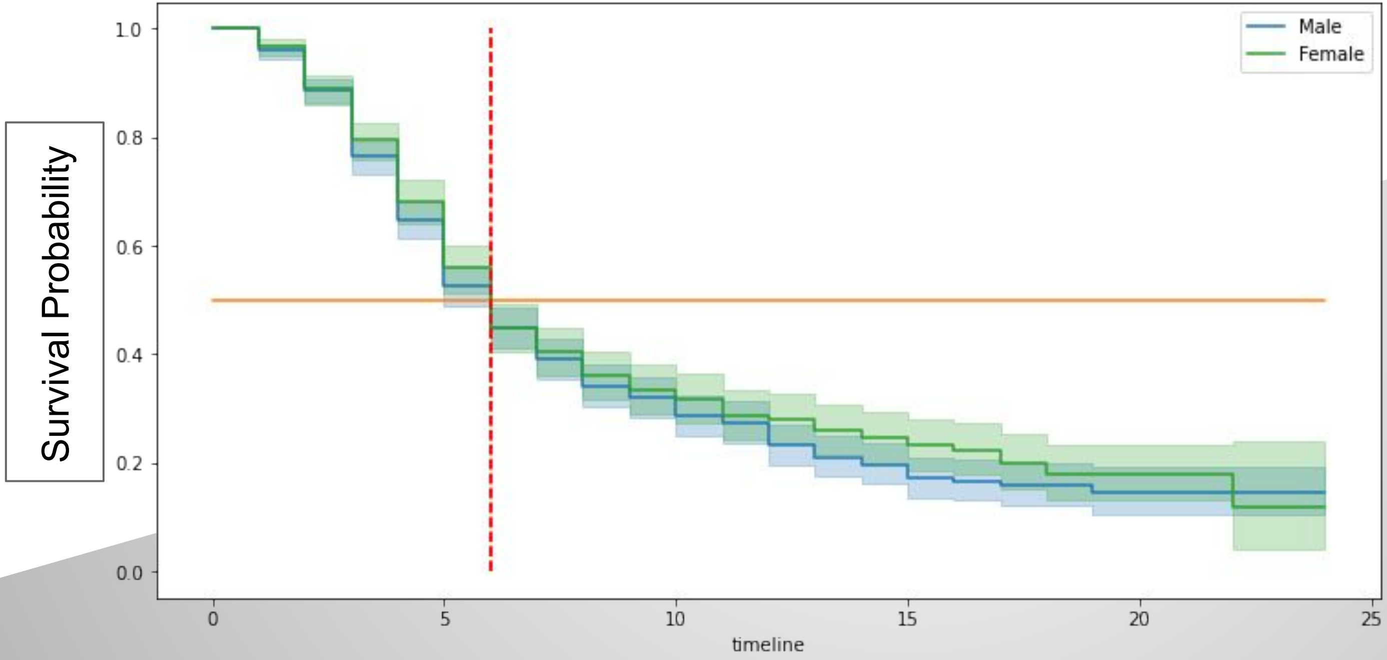


Employee attrition and survival days with **Gender**





Km curve for different Genders:

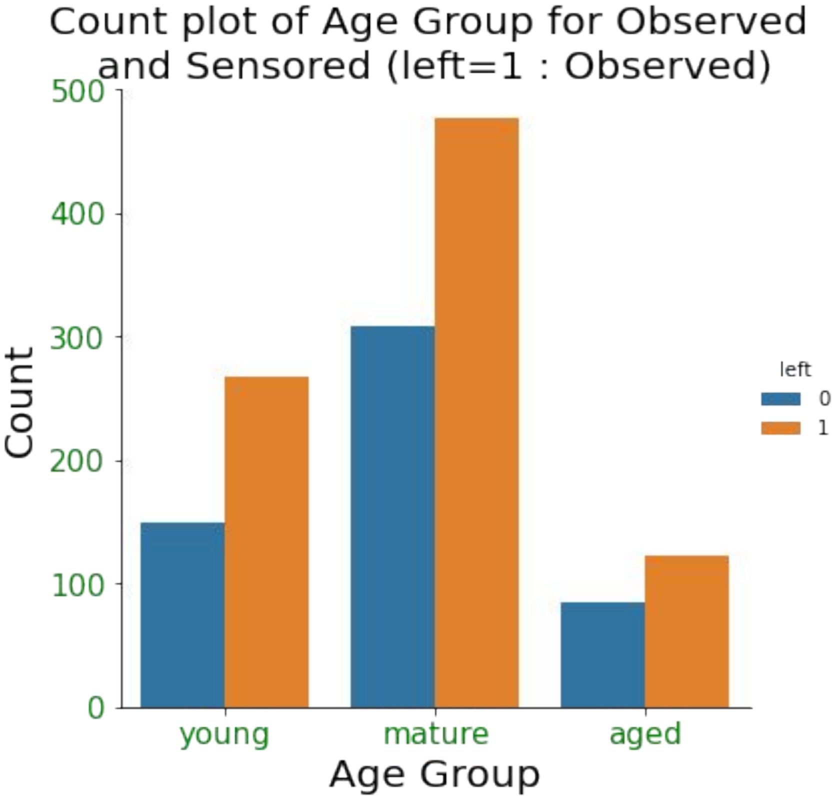


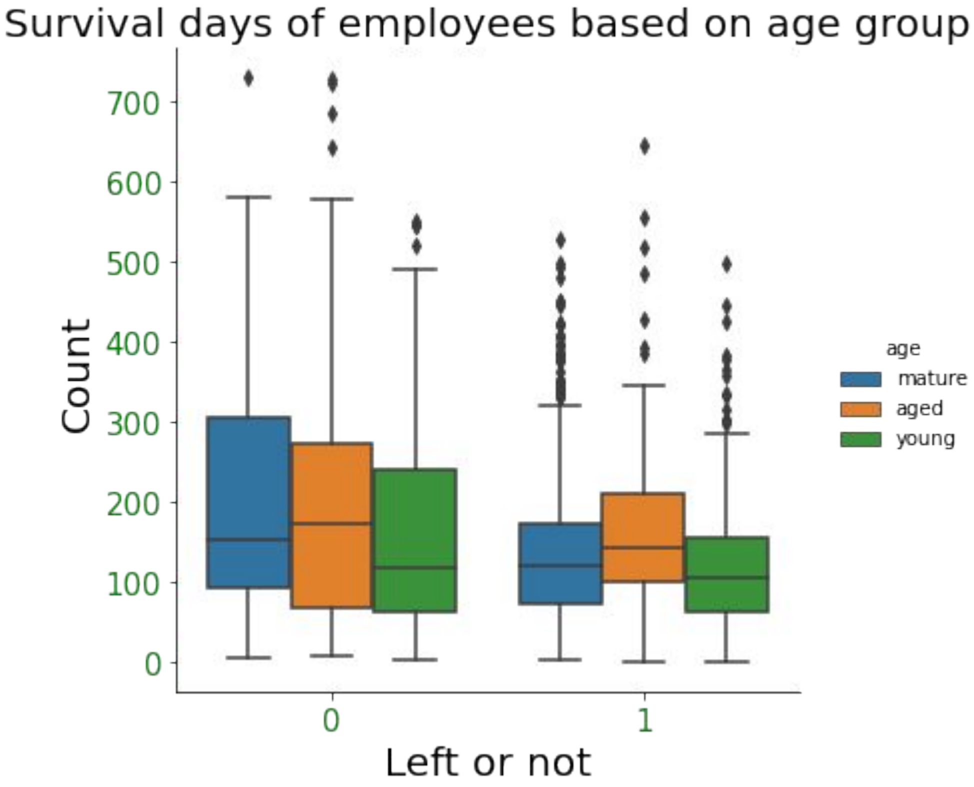
Logrank test
Test statistics: 1.61
P-value: 0.21

Median Survival Time
Male: 6 months
Female: 6 months

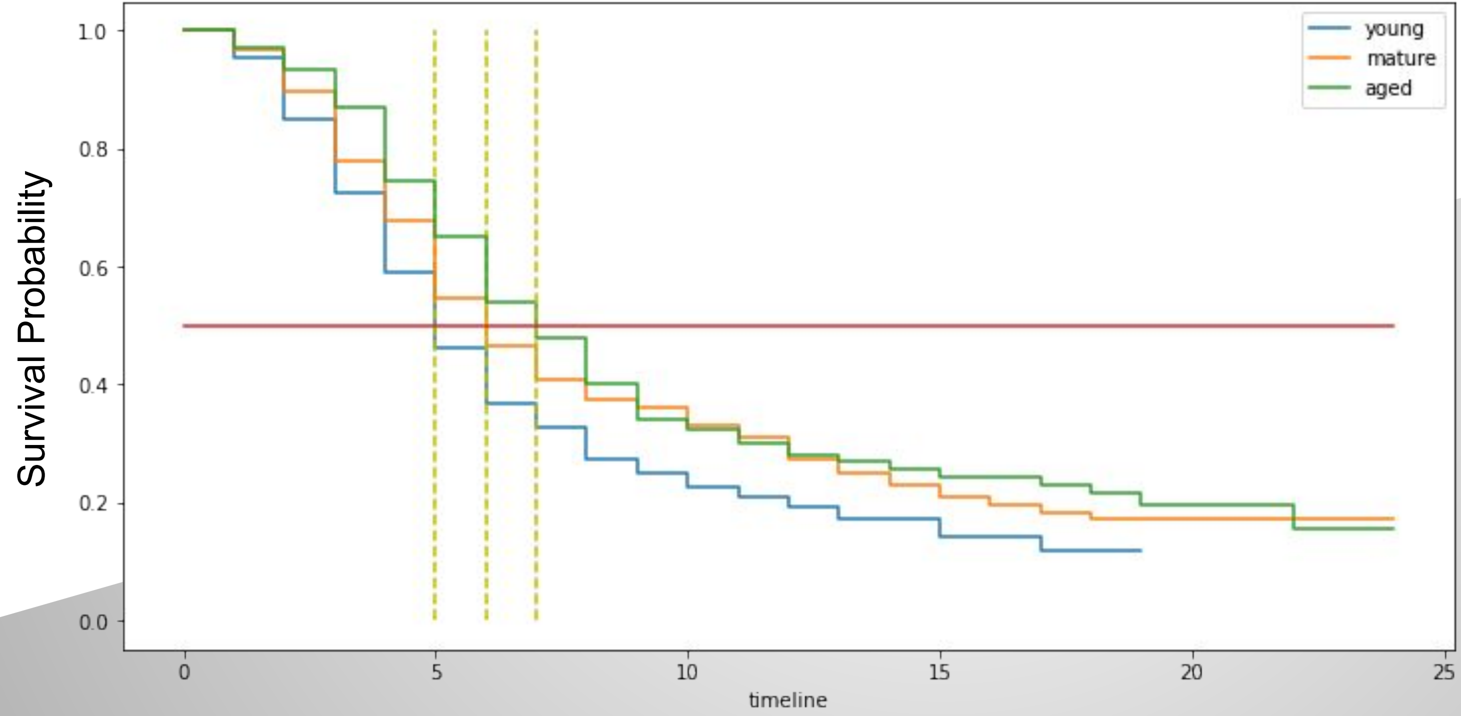
Employee attrition and survival days with age

Age : young(<30),mature(30-38),aged(>38)





Km curve for different Age Group:

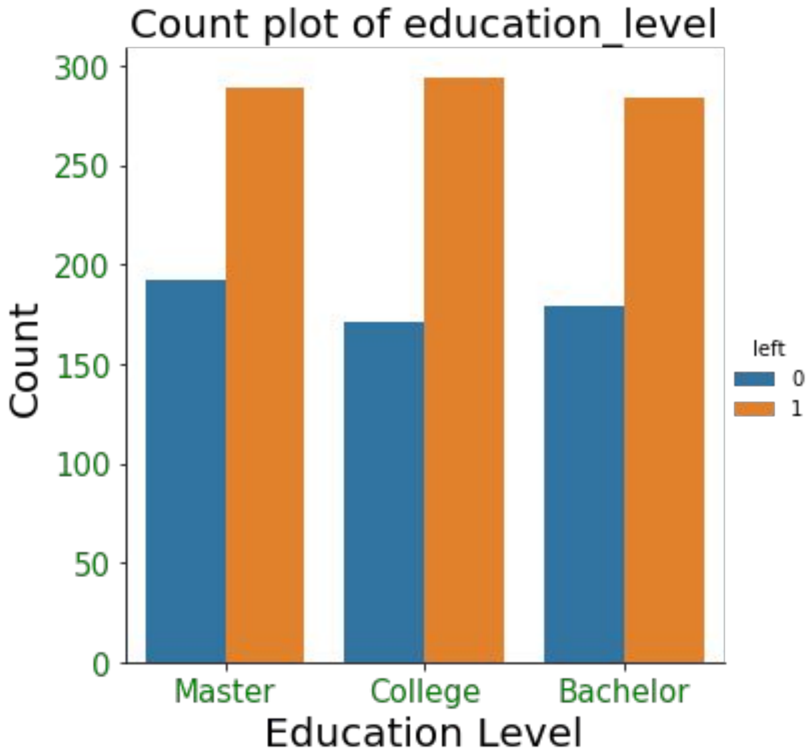


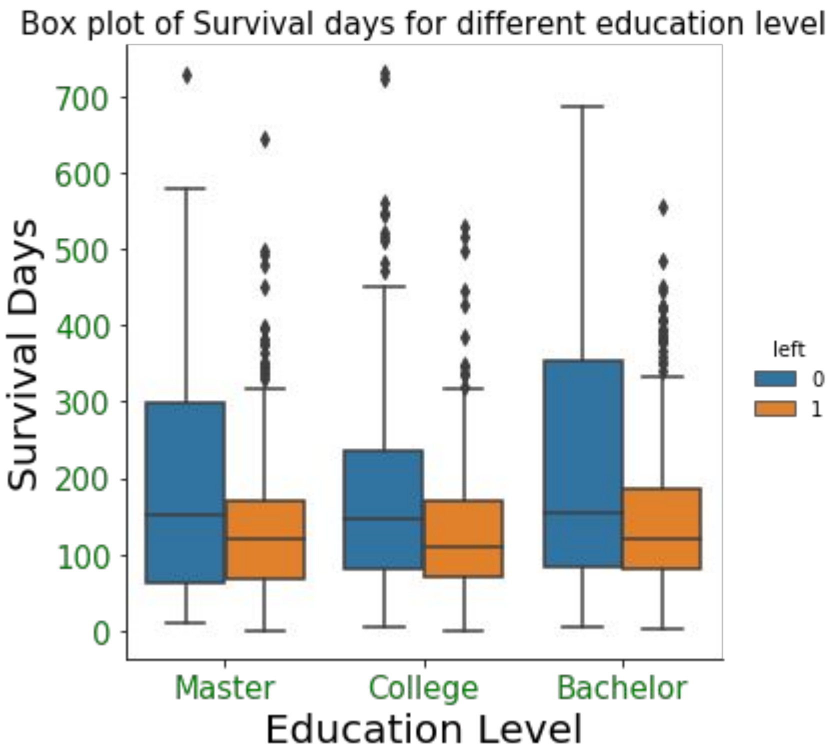
Logrank test

Test statistic:
18.06
P-value: <0.005

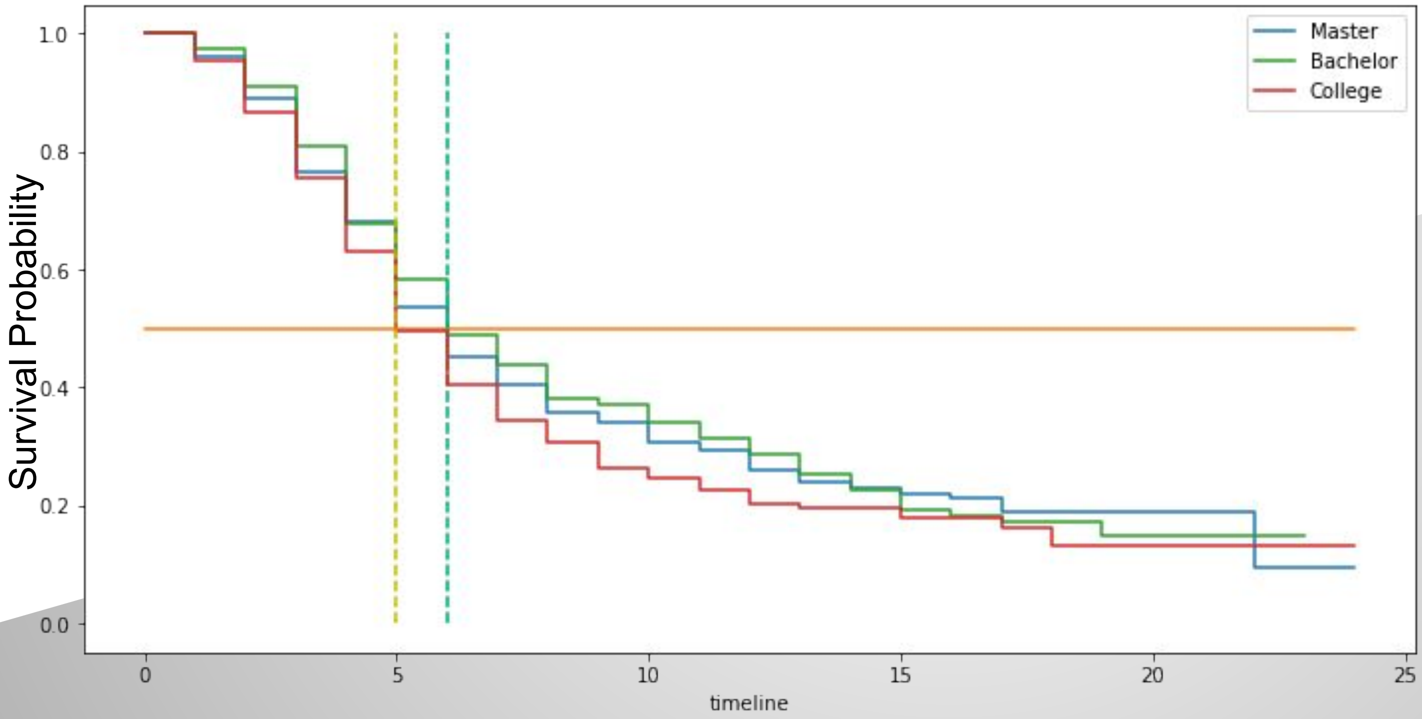
Median Survival Time
young: 5 months
mature: 6 months
aged: 7 months

Employee attrition and survival days with **Education Level**





Km curve for different education level:

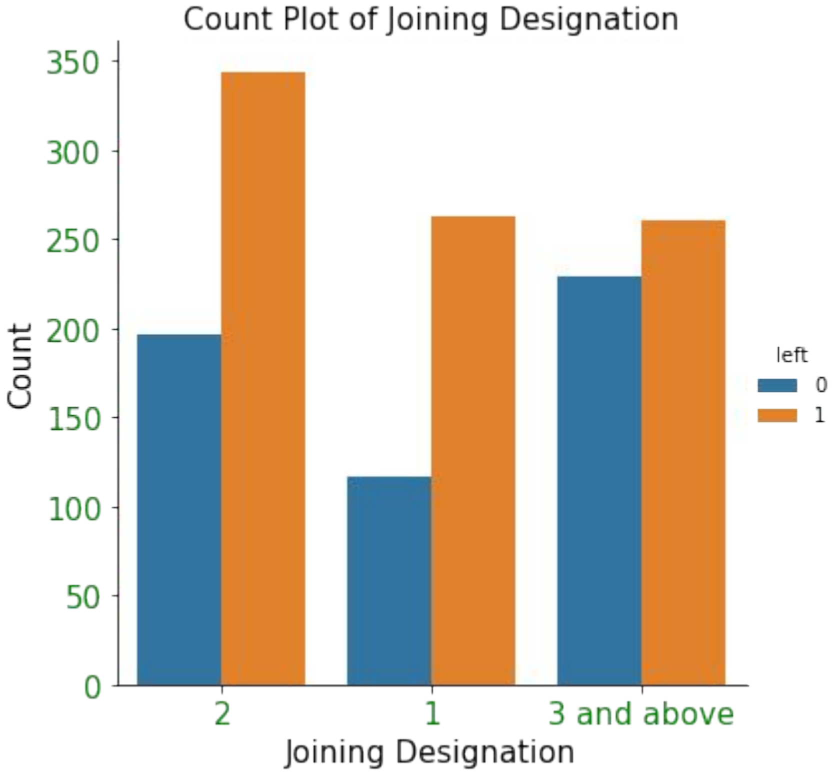


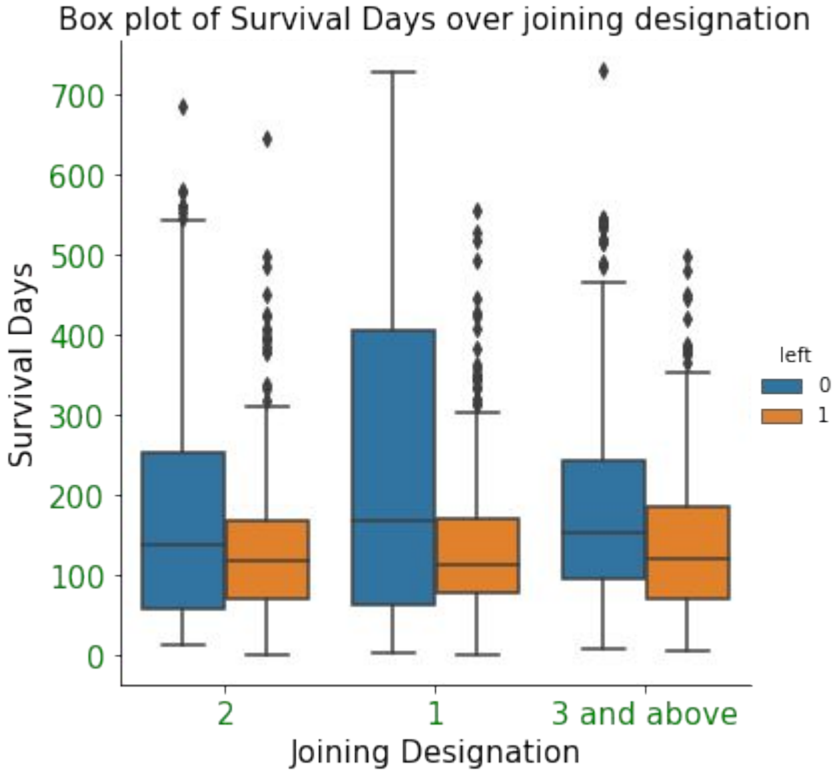
Logrank test

Test statistic: 6.59
P-value: 0.04

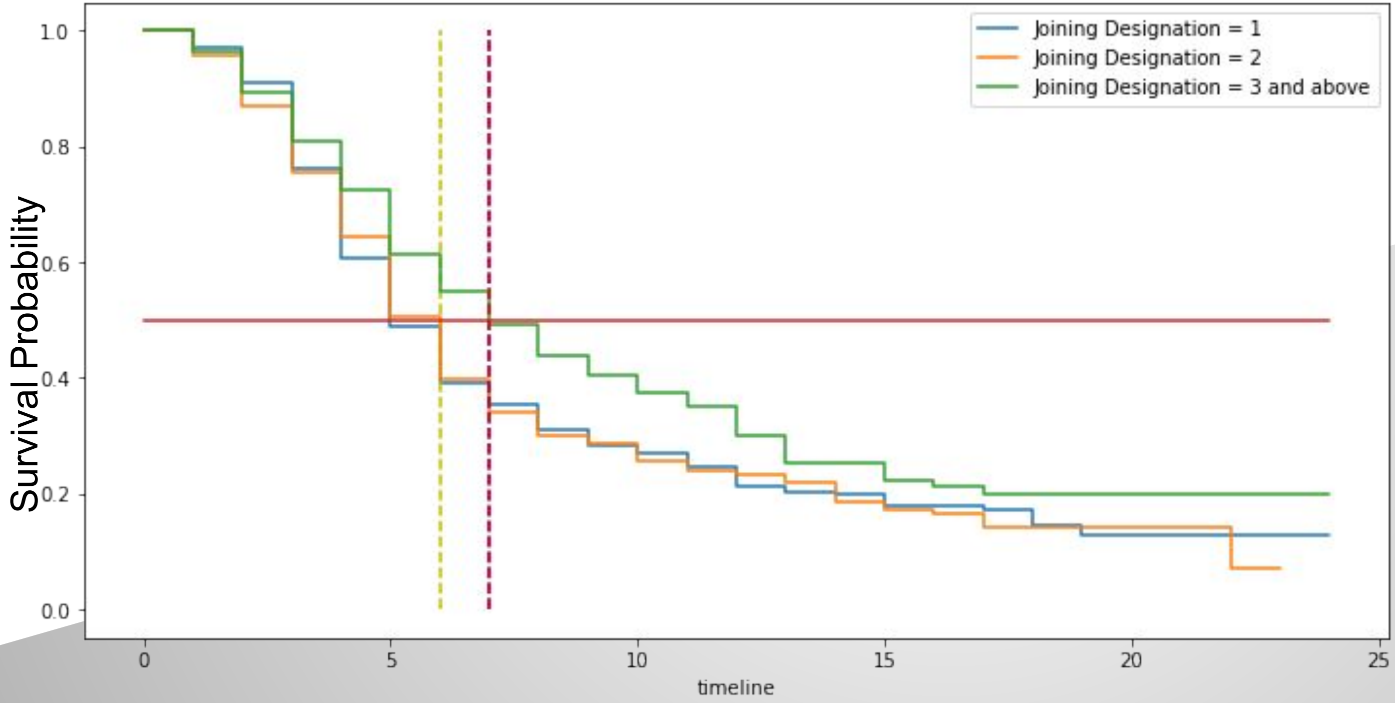
Median Survival Time
College: 5 months
Master: 6 months
Bachelor: 7 months

Employee attrition and survival days with Joining Designation





Km curve for different Joining Designation:



Logrank test

Test statistic: 16.15
P-value: <0.005

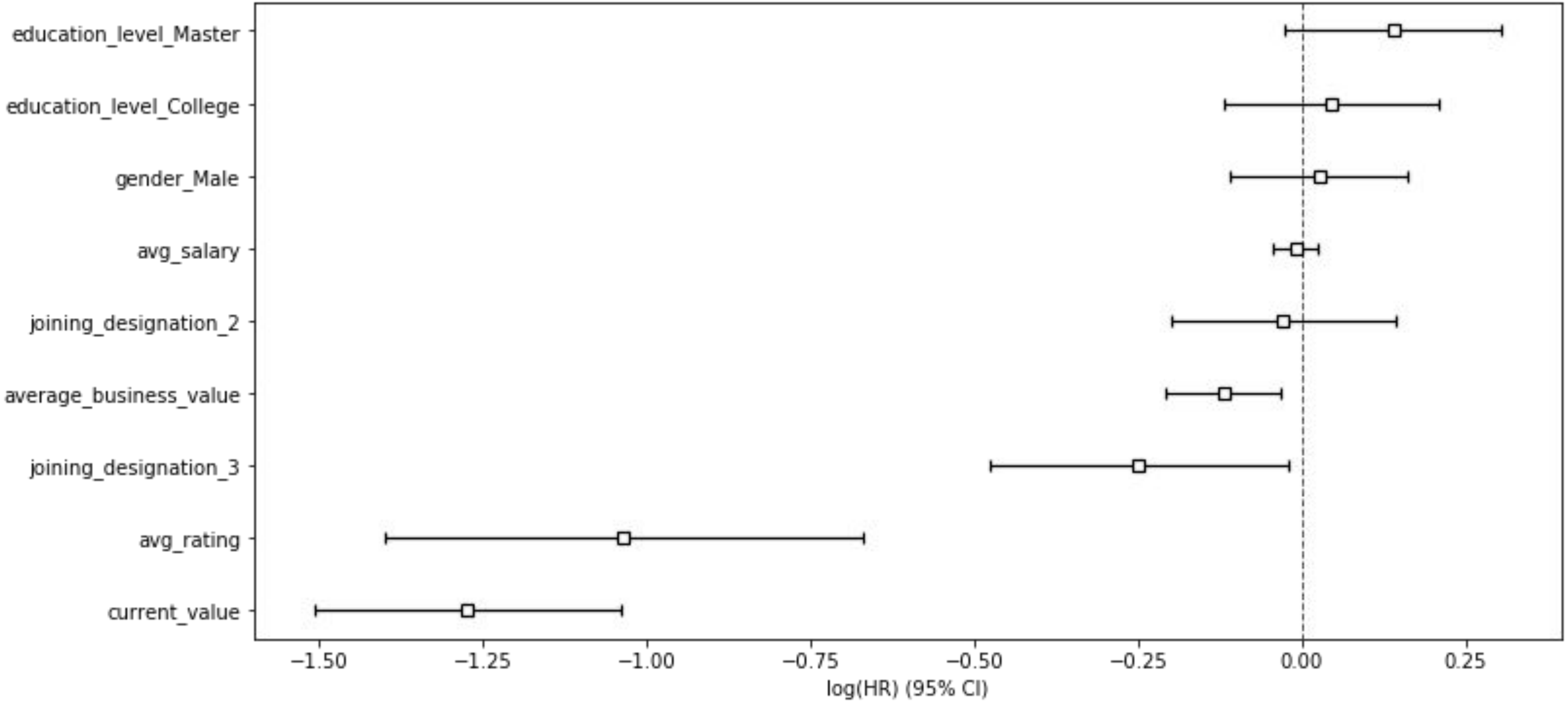
Median Survival Time
Designation 1: 6 months
Designation 2: 7 months
Designation 3 and above : 7 months

Summary of Cox-proportional Hazard Model :

Covariates	coef	exp(coef)	z	p
avg_rating	-1.04	0.35	-5.55	<0.005
avg_salary	-0.01	0.99	-0.59	0.56
average_business_value	-0.12	0.89	-2.72	0.01
current_value	-1.27	0.28	-10.63	<0.005
gender_Male	0.03	1.03	0.36	0.72
education_level_College	0.04	1.05	0.53	0.6
education_level_Master	0.14	1.15	1.64	0.1
joining_designation_2	-0.03	0.97	-0.33	0.74
joining_designation_3	-0.25	0.78	-2.14	0.03

Concordance	0.80
Partial AIC	10439.32

Coefficient values and their 95% confidence interval :

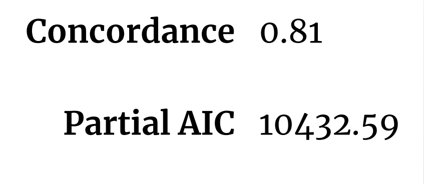


Feature Selection

Name of the Covariate	Concordance values after fitting single covariate
average_business_value	0.813212
avg_rating	0.741244
current_value	0.676032
joining_designation_3	0.562179
Avg_salary	0.537836
education_level_College	0.523045
joining_designation_2	0.522168
gender_Male	0.511090
education_level_Master	0.502353

Summary of Cox-proportional Hazard Model after fitting with the best covariates

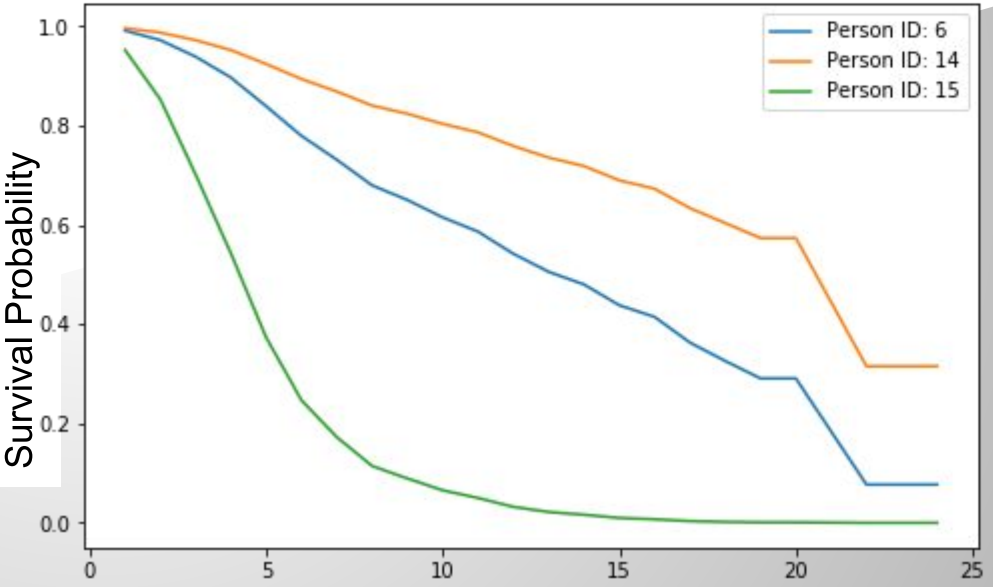
Covariates	coef	exp(coef)	z	p
average_business_value	-0.13	0.88	-2.91	<0.005
current_value	-1.27	0.28	-10.61	<0.005
joining_designation_3	-0.26	0.77	-3.45	<0.005
avg_rating	-1.01	0.36	-5.45	<0.005



Concordance	0.81
Partial AIC	10432.59

Prediction of survival function of different Employee

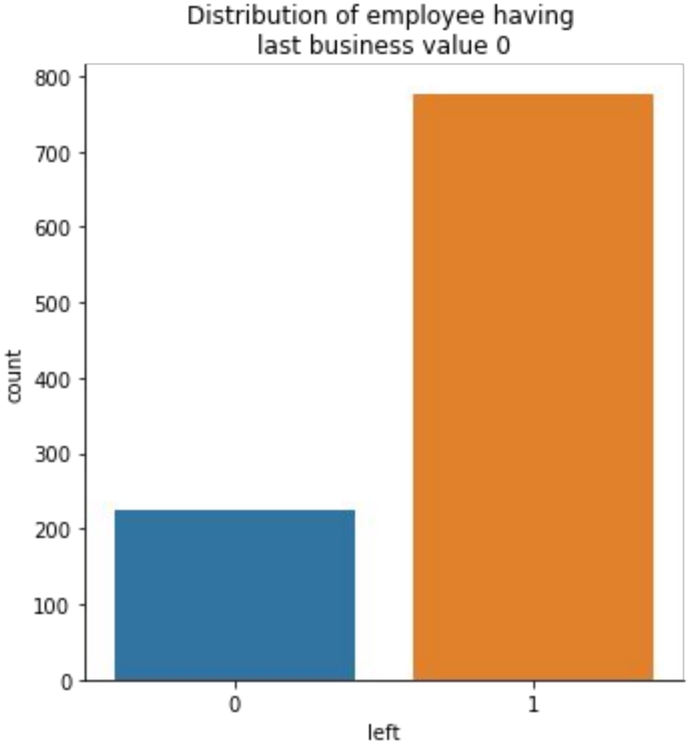
Id	avg_rating	average_business_value	current_value
6	2.5	4.345300	0
14	2.0	3.055521	1
15	1.0	1.083367	0

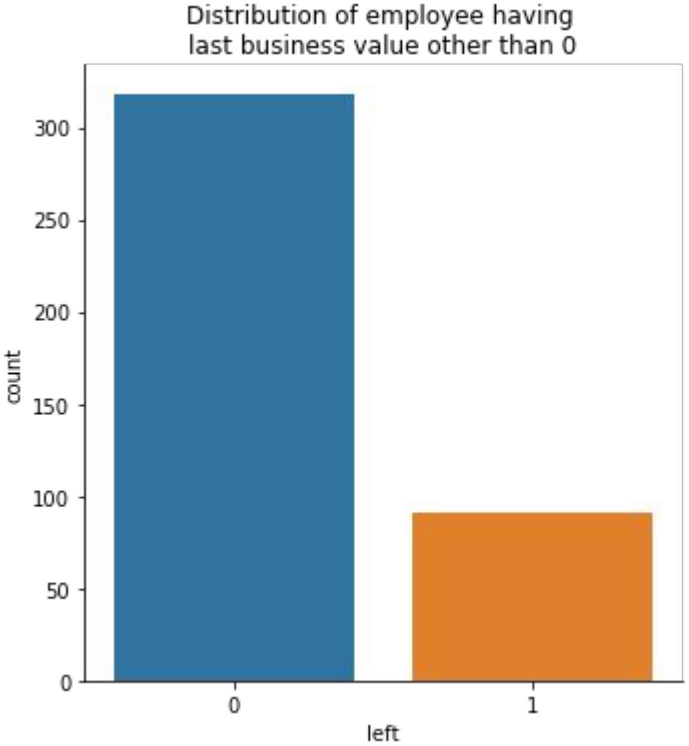


Thank You

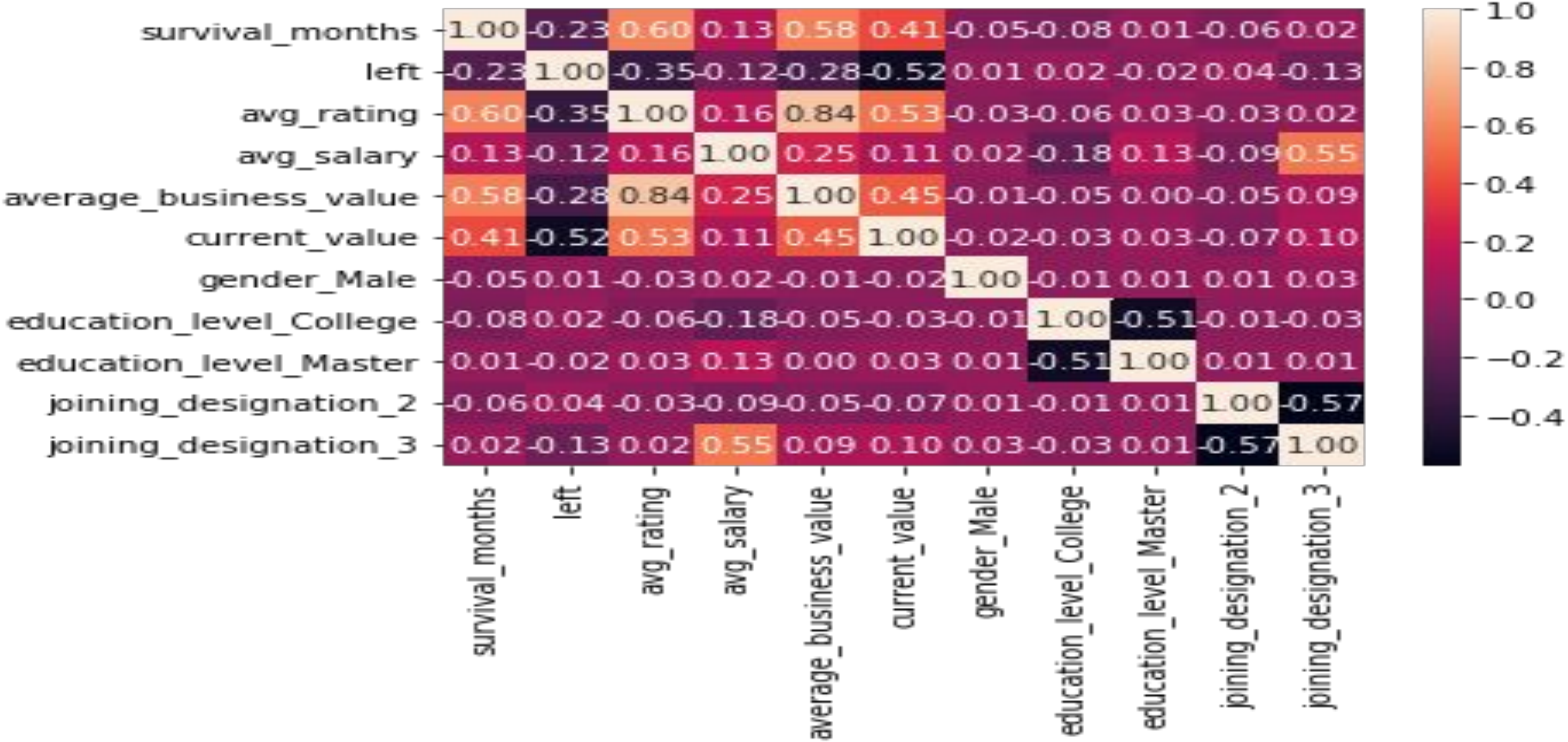
The interpretation of concordance value is identical to the traditional area under the ROC curve metric for binary classification: - a value of 0.5 denotes a random model, - a value of 1.0 denotes a perfect model, - a value of 0.0 denotes a perfectly wrong model.

Business value





Correlation among the final Covariates



	survival_months	left	avg_rating	avg_salary	average_business_value	current_value	gender_Male	education_level_College	education_level_Master	joining_designation_2	joining_designation_3
survival_months	1.00	-0.23	0.60	0.13	0.58	0.41	-0.05	-0.08	0.01	-0.06	0.02
left	-0.23	1.00	-0.35	-0.12	-0.28	-0.52	0.01	0.02	-0.02	0.04	-0.13
avg_rating	0.60	-0.35	1.00	0.16	0.84	0.53	-0.03	-0.06	0.03	-0.03	0.02
avg_salary	0.13	-0.12	0.16	1.00	0.25	0.11	0.02	-0.18	0.13	-0.09	0.55
average_business_value	0.58	-0.28	0.84	0.25	1.00	0.45	-0.01	-0.05	0.00	-0.05	0.09
current_value	0.41	-0.52	0.53	0.11	0.45	1.00	-0.02	-0.03	0.03	-0.07	0.10
gender_Male	-0.05	0.01	-0.03	0.02	-0.01	-0.02	1.00	-0.01	0.01	0.01	0.03
education_level_College	-0.08	0.02	-0.06	-0.18	-0.05	-0.03	-0.01	1.00	-0.51	-0.01	-0.03
education_level_Master	0.01	-0.02	0.03	0.13	0.00	0.03	0.01	-0.51	1.00	0.01	0.01
joining_designation_2	-0.06	0.04	-0.03	-0.09	-0.05	-0.07	0.01	-0.01	0.01	1.00	-0.57
joining_designation_3	0.02	-0.13	0.02	0.55	0.09	0.10	0.03	-0.03	0.01	-0.57	1.00

Physical Interpretation of Coding covariates

$Z_1 = 1$ if the subject is male, 0 otherwise
$Z_2 = 1$ white

$h(t|z_1=1,z_2=0)=h_o(t) \exp(beta1)$
$h(t|z_1=0,z_2=1)=h_o(t) \exp(beta2)$
$h(t|z_1=0,z_2=0)=h_o(t)$

The risk of the events occurring among male relative to the risk of the events occurring among whites is $\exp(beta1-beta2)$

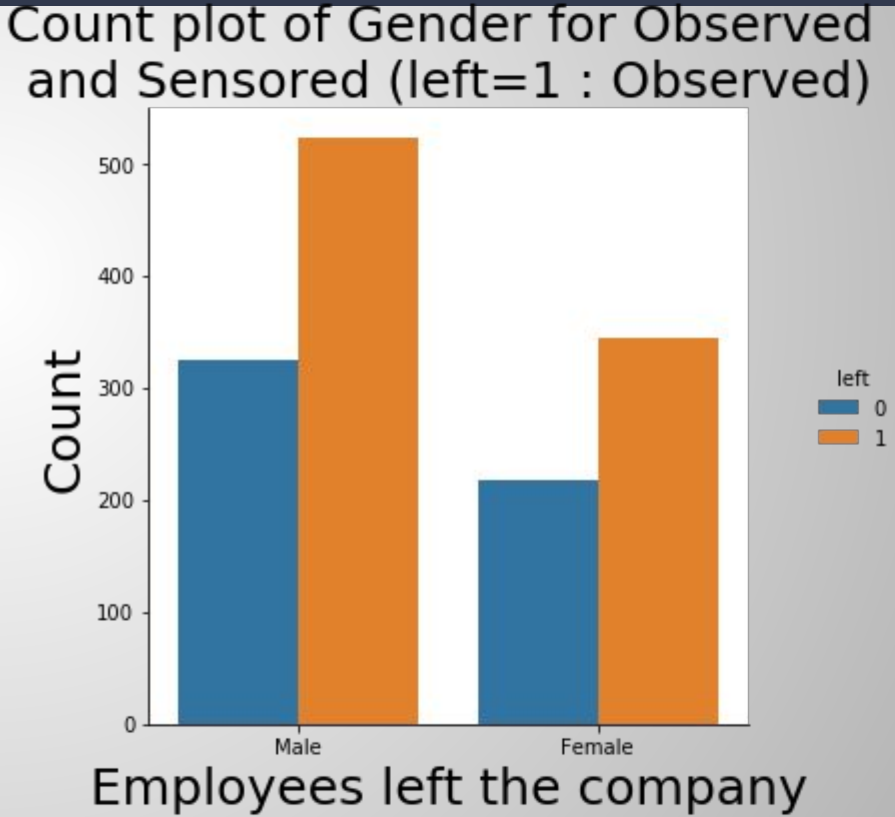
White pink $\exp(b_2)$
Black Pink $\exp(b_1)$

Two samples are concordant if the one with a higher estimated risk score has a shorter actual survival time.
Two samples are comparable if (i) both of them experienced an event (at different times), or (ii) the one with a shorter observed survival time experienced an event, in which case the event-free subject “outlived” the other.

Observed Ratio:

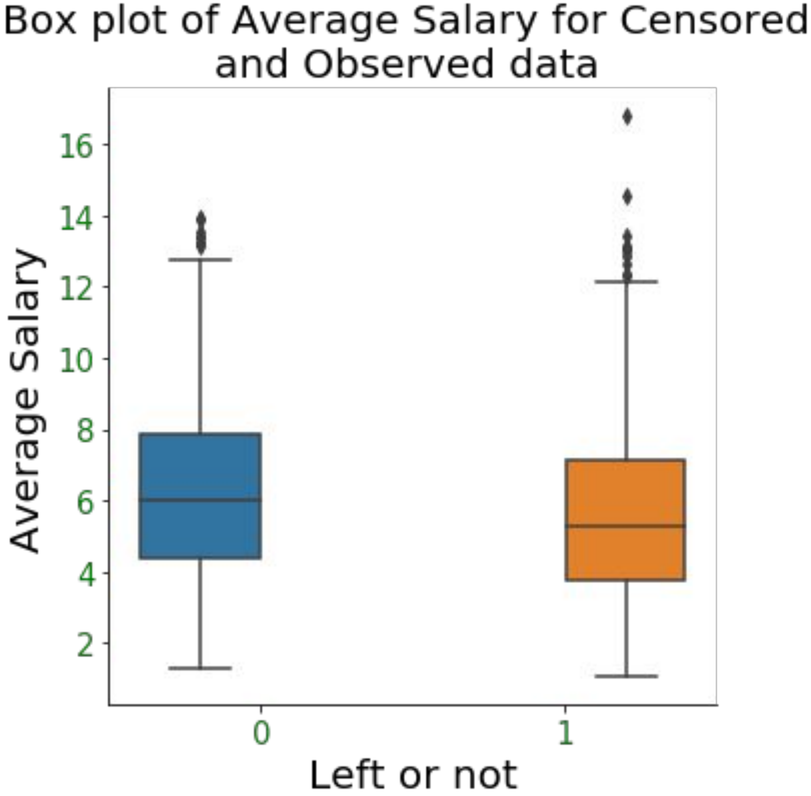
Above all: 1409
867/542 =1.60
Male: 847
523/324 = 1.61

Female: 562
344/218 = 1.58



Count plot of Gender for Observed and Sensored (left=1 : Observed)

Salary



Box plot of Average Salary for Censored and Observed data
Average Salary
16
14
12
10
8
6
4
2
0
1
Left or not