Survival Analysis on employee attrition data

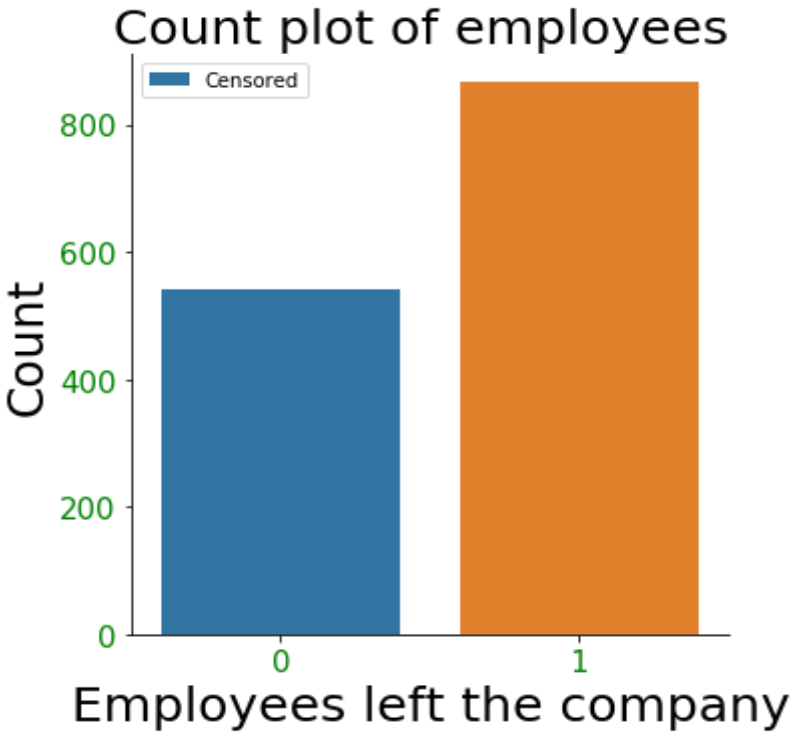
By Mahendra Nandi (B2030032) and Sourav Karmakar(B2030047)

Data description: The dataset is collected during the JOB-A-THON-November 2021 event organized by analytics vidya. The dataset contains the month-wise information of each employee. For this project we took data totally lies between the period of Jan 2016 to Dec 2017. There are 1409 employees in the study and out of them 542 are right-censored and others left the company

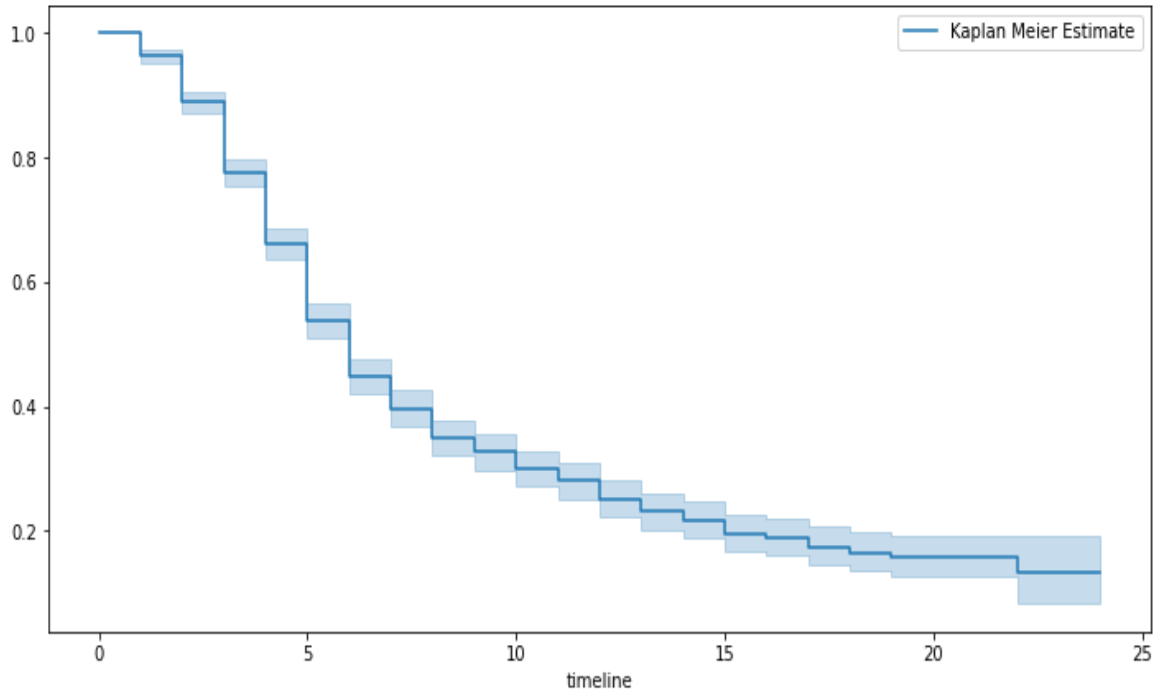
Data Preprocessing:

We have done some preprocessing onto the dataset for better interpretability as follows-

1. We introduced survival months by -
    a. The number of months the employee was in the company from the beginning of the study i.e., Jan 2016 to the month they have left the company or Upto Dec 2017 if they were with the company after the study ends for the employees who have joining date before Dec 2017.
    b. The number of months from the month of the employee starts working to the month of the employee worked or up to Dec 2017 if they sustained during the period for the employees who have joining date after Dec 2017
2. In the joining designation there were very few people with joining designation 4 and 5. So we have taken those as joining designation 3 and above.
3. We transform the age column to categorical column with the transformation young(<30 years of age), mature(30-38years of age), aged(>38years of age).



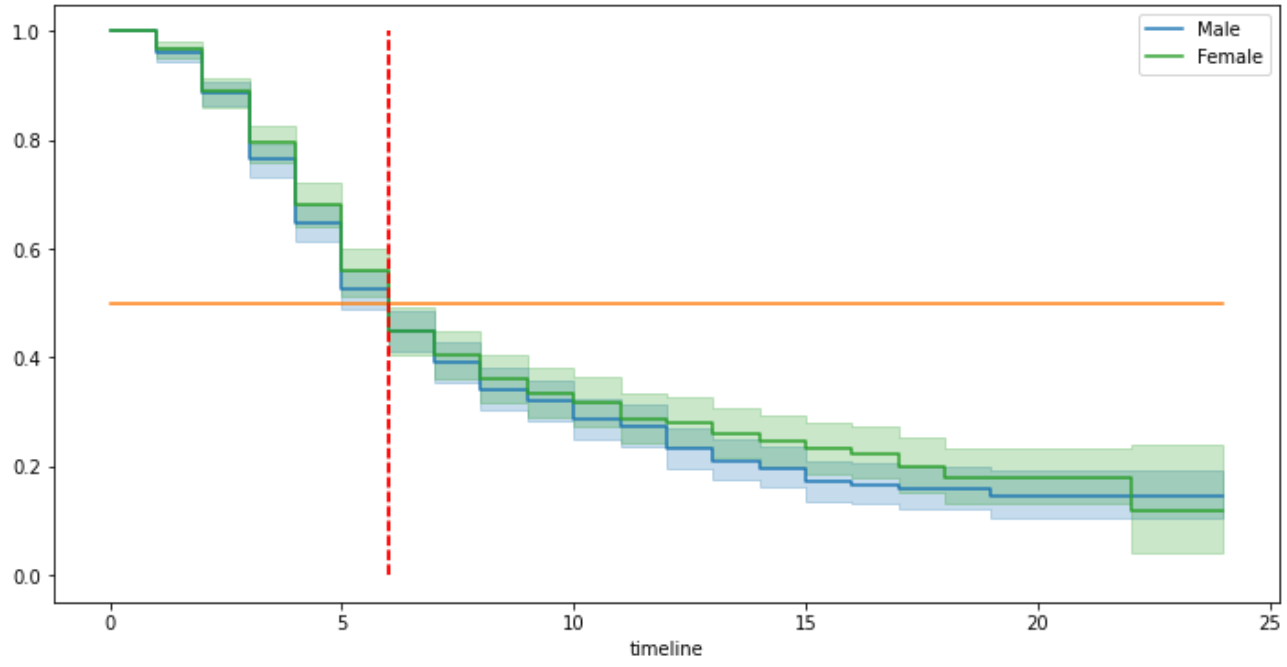
Analysis:
We have fitted the Kaplan-Meier curve for the whole data and for different groups.



This is the estimate of the survival function using Kaplan Meier Estimate.

For differentiating whether two groups are identical or not we use log rank test for two groups and multivariate log rank test for more than two-groups.

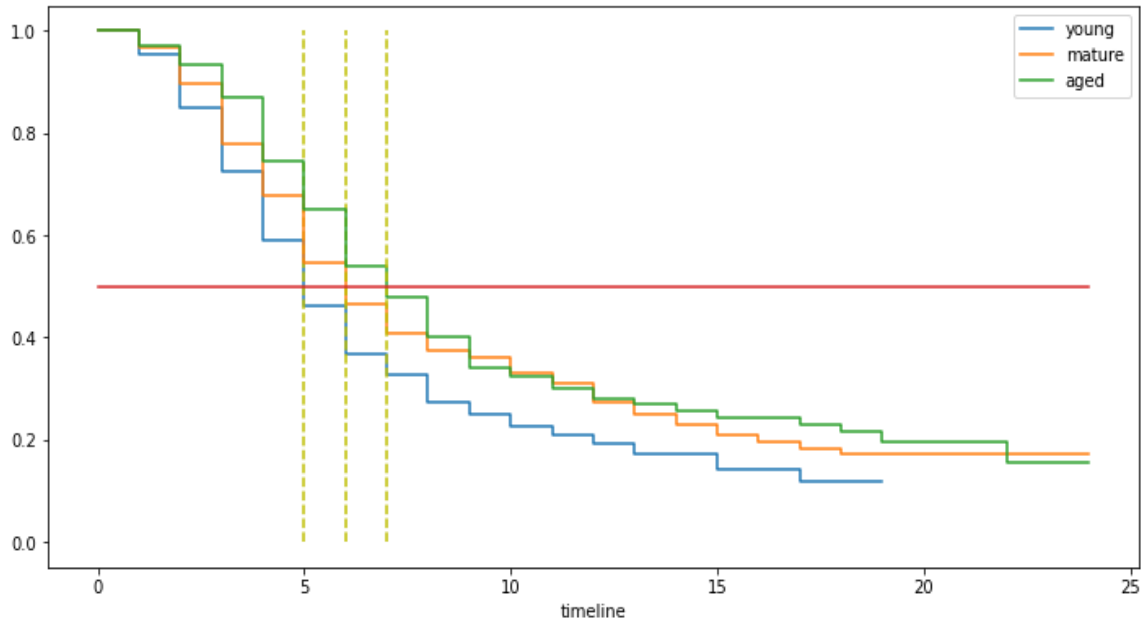
<u>Gender wise survival function using Kaplan-Meier estimate:</u>



Log rank test results:
Test statistics: 1.61
p-value: 0.21

So the p-value is greater than 0.05. So there is no evidence that we can reject the null hypothesis that two groups are identical.
The median lifetime for both men and women is 6 months.

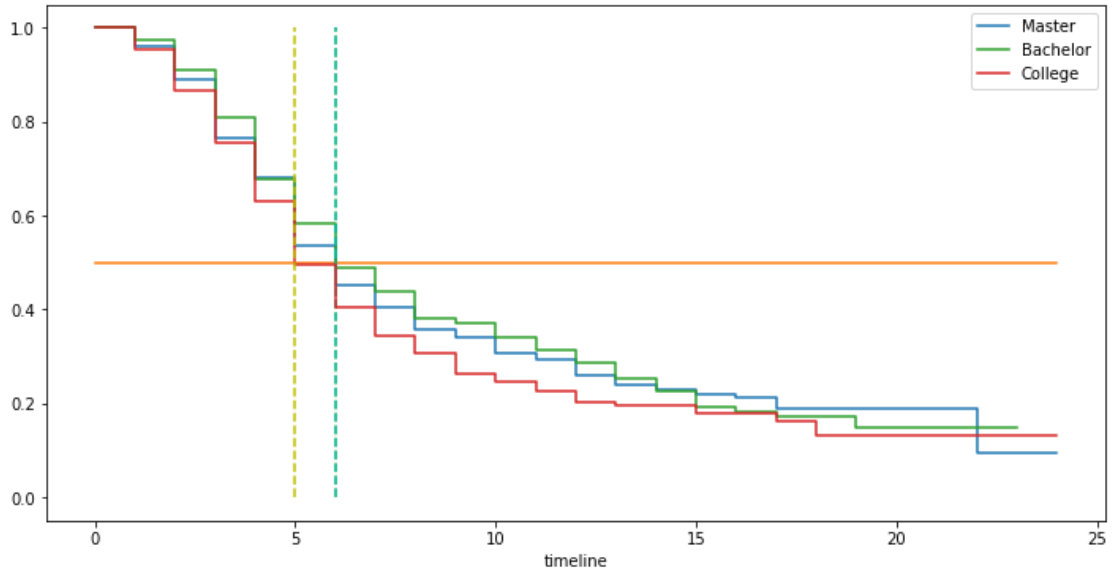
Survival function for different age group:



Log rank test results:
Test statistics: 18.06
p-value: <0.005

So the p-value is less than 0.05. So we can reject the null hypothesis that all the groups are identical.
The median lifetime for young, mature and aged people are 5 months, 6 months and 7 months respectively.

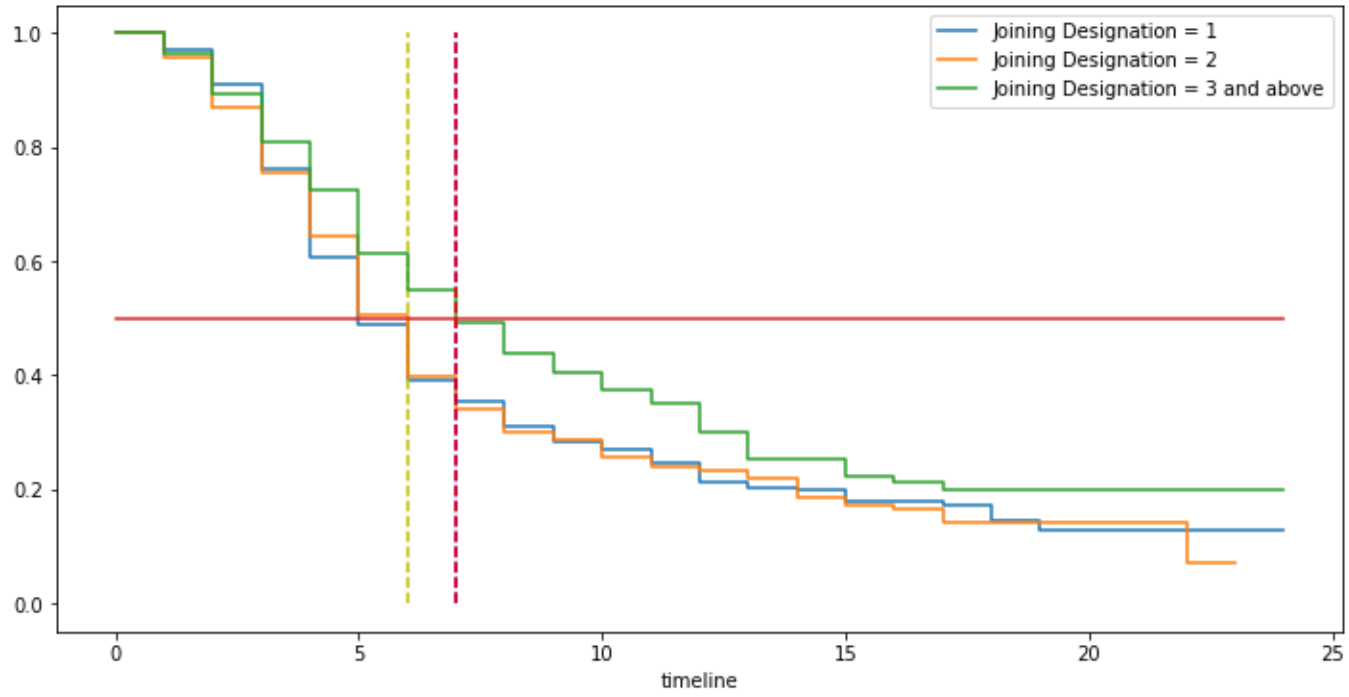
Survival function for different education level:



Log rank test results:
Test statistics: 6.59
p-value: 0.04

So the p-value is less than 0.05. So we can reject the null hypothesis that all the groups are identical. The median lifetime for people with education level as college, master and bachelor are 5 months, 6 months and 7 months respectively.

Survival function for different joining designation:



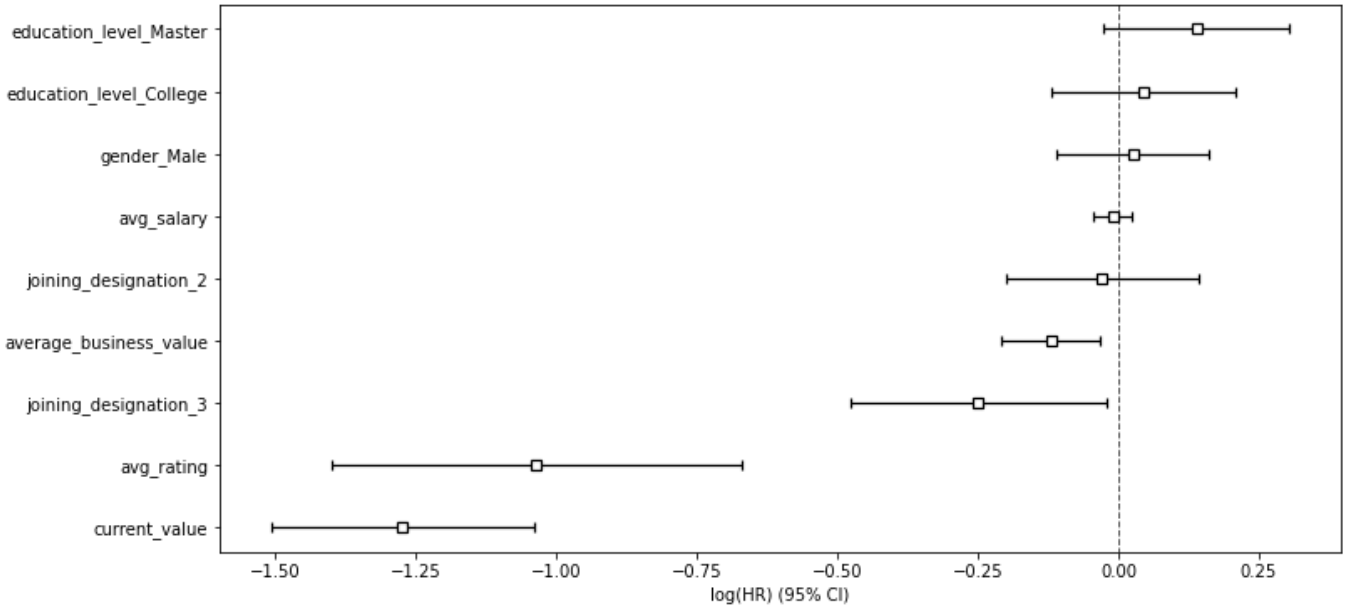
Log rank test results:
Test statistics: 16.15
p-value: <0.005

Here the p-value is less than 0.05. So we can reject the null hypothesis that all groups are identical. The median lifetime for people with joining designation '1', '2' and '3 and above' are 6months, 7 months and 7 months respectively.

Summary of the cox proportional hazard model with the covariates:

Covariates	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	z	p	-log2(p)
avg_rating	-1.04	0.35	0.19	-1.40	-0.67	-5.55	<0.005	25.08
avg_salary	-0.01	0.99	0.02	-0.04	0.02	-0.59	0.56	0.85
average_business_value	-0.12	0.89	0.04	-0.21	-0.03	-2.72	0.01	7.24
current_value	-1.27	0.28	0.12	-1.51	-1.04	-10.63	<0.005	85.26
gender_Male	0.03	1.03	0.07	-0.11	0.16	0.36	0.72	0.48
education_level_College	0.04	1.05	0.08	-0.12	0.21	0.53	0.60	0.74
education_level_Master	0.14	1.15	0.08	-0.03	0.30	1.64	0.10	3.31
joining_designation_2	-0.03	0.97	0.09	-0.20	0.14	-0.33	0.74	0.43
joining_designation_3	-0.25	0.78	0.12	-0.48	-0.02	-2.14	0.03	4.93

Concordance: 0.80
Partial Aic: 10439.32



From this plot we can see that some of them are important covariates and taking only those if we fit again we get:

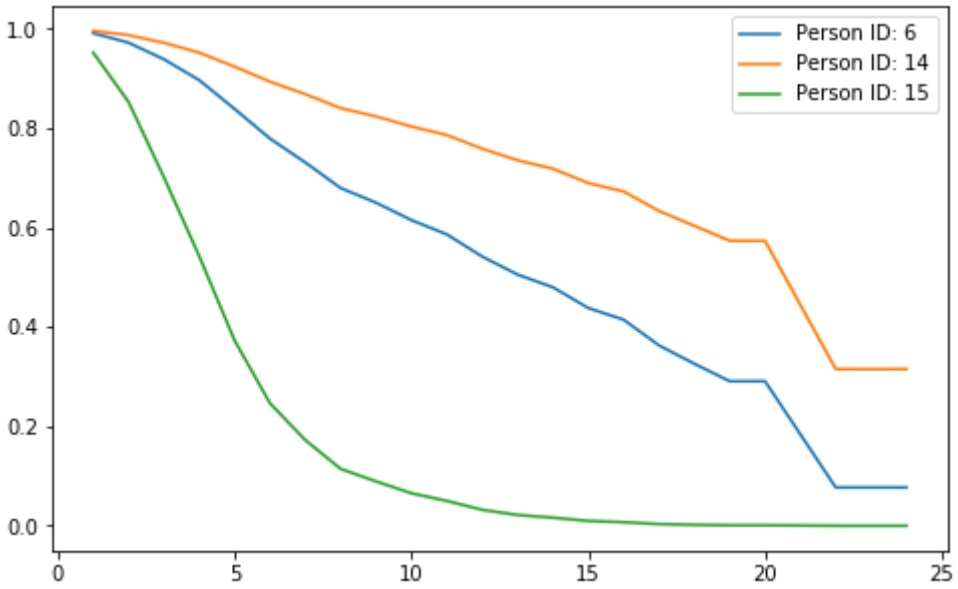
Covariates	coef	exp (coef)	ee (coef)	coef lower 95%	coef upper 95%	z	p	-log2(p)
average_business_value	-0.13	0.88	0.04	-0.22	-0.04	-2.91	<0.005	8.12
current_value	-1.27	0.28	0.12	-1.50	-1.03	-10.61	<0.005	84.94
joining_designation_3	-0.26	0.77	0.08	-0.41	-0.11	-3.45	<0.005	10.79
avg_rating	-1.01	0.36	0.19	-1.37	-0.65	-5.45	<0.005	24.23

Concordance: 0.81
Partial Aic: 10432.59

Also if we want to visualize the effect in the survival function, that can be directly reflect in the next plot, where 3 such employee with very different covariates are taken, as we have shown some of the covariates below

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

Corresponding survival functions are:



Survival probability is in y direction and time in months is in x direction

–o–