



Time to Failure Prediction

Group P6A

Shengxin (Christine) Ding

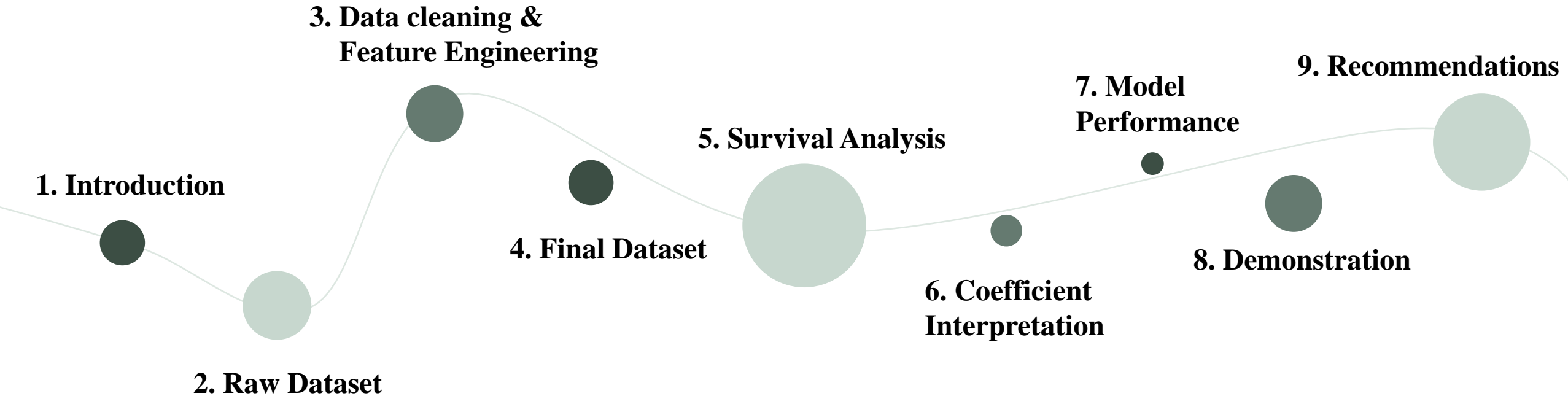
Jishi Liu

Bofu Zhang

Ting Zhang

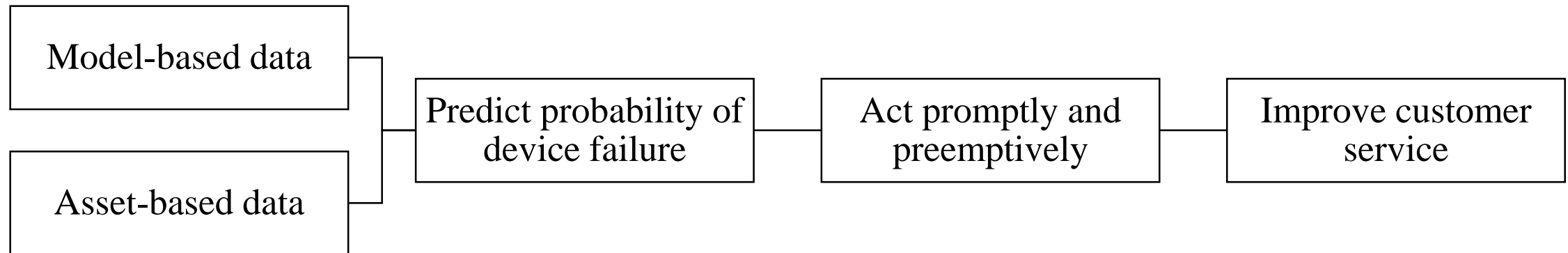
Xiyu (Tracy) Zhang

Agenda



Introduction

Xerox is managing over 1.5 million printing devices for customers by collecting IoT data about devices' characteristics, utilization, failure incidents.



Raw Dataset

Asset Data - device level information

Asset-based

- Asset ID
- Age in month

Model-based

- Model name
- Model class
- Is color
- PPM
- Color PPM
- Is scanner
- Is copier
- Is fax

Volume data – utilization log of devices

- Asset ID
- Read date
- Volume

Incident data – history log of incidents

- Asset ID
- Date
- Problem type


Data Cleaning

Asset Table

Drop Asset ID (no volume):

B69A09A1-EFD4-E211-BA2D-0025B500016E

Volume Table

	ReadDate	Volume 
30C1-0025B500016E	5/21/2014 11:04	-4898105
AD34-0025B500016E	8/6/2012 22:23	-4536319
AD34-0025B500016E	7/30/2012 22:10	-5814894
A66B-0025B500017E	11/13/2012 19:01	-4728894
AEC9-0025B500016E	10/4/2012 6:03	-5272932
B722-0025B500016E	6/28/2014 18:01	8008744

Big Day-by-day table

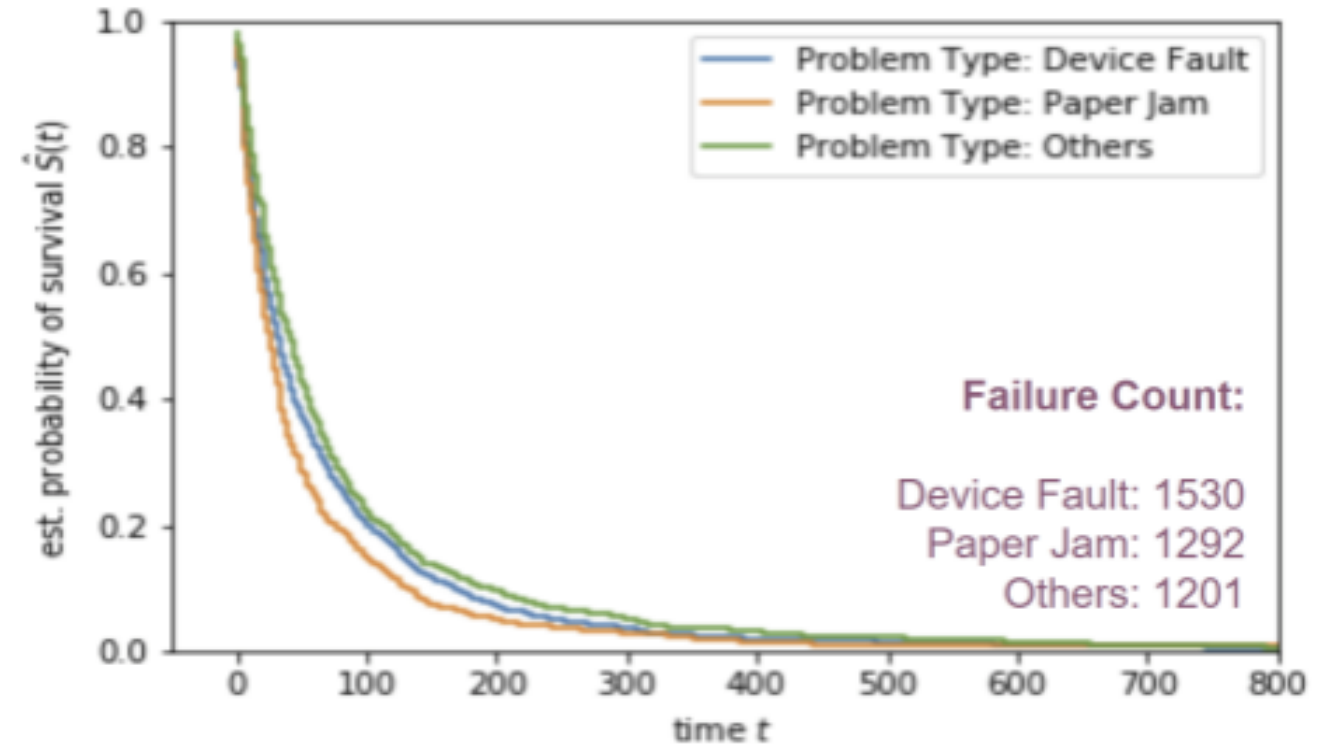
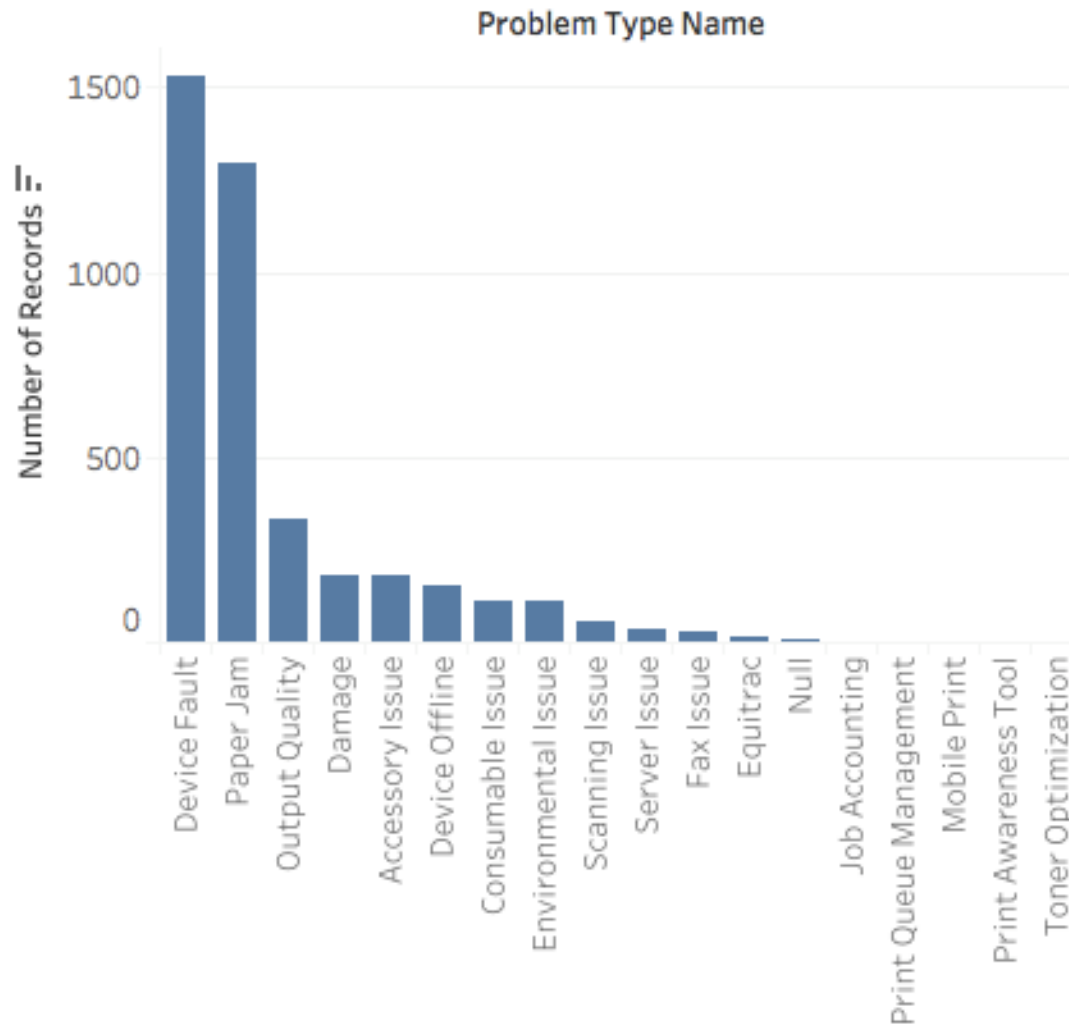
Failure Table

AssetID	ReadDate	ProblemTypeName
-80C2-0025B500016E	2016-01-14	Others,Device Fault
-AD34-0025B500016E	2012-09-25	Paper Jam,Device Fault
-97B2-0025B500016E	2013-10-22	Others,Device Fault
-AD34-0025B500016E	2014-10-03	Paper Jam,Others

Feature Engineering – Age, Volume & Scaling

- ❖ Set age = 1 (month) at the first record for every machine
- ❖ **Survival Day:** cumulative dates since last incident.
- ❖ **We add several new page features that may have effects on the failure probability:**
 - Last 30 Day Volume
 - Last 15 Day Volume
 - Last 7 Day Volume
 - Yesterday Volume
 - Cumulative Pages Printed
- ❖ **We scale the volume-related features at a 100,000 level**

Feature Engineering - Problem Types



Feature Engineering - K-Means Clustering

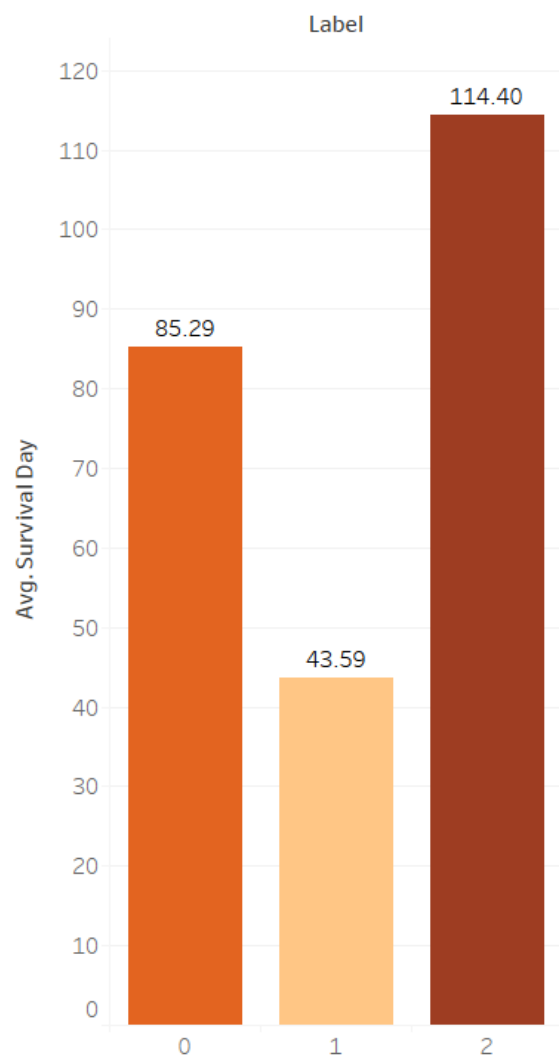
	IsColor	PPM	ColorPPM	IsPrinter	IsScanner	IsCopier	IsFax	Label
0	0	75	0	1	1	1	0	0
1	0	110	0	1	1	1	0	0
2	0	75	0	1	1	1	0	0
3	1	37	32	1	1	1	1	1
4	1	50	50	1	1	1	1	1
5	0	65	0	1	1	1	0	0
6	1	50	50	1	1	1	1	1
7	1	50	50	1	1	1	0	1
8	0	65	0	1	1	1	0	0
9	0	65	0	1	1	1	0	0

Final Dataset

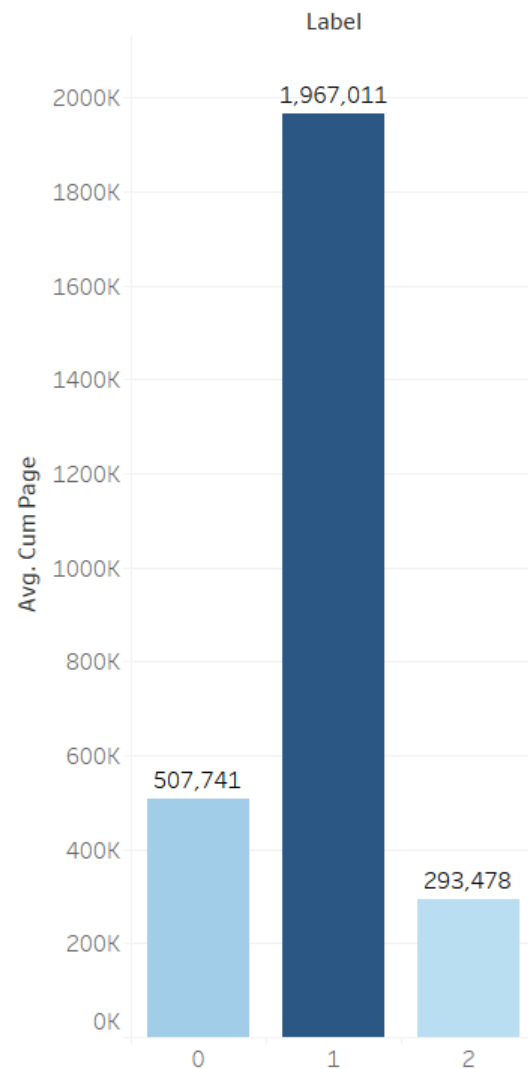
B	C	D	E	F	G	H	I	J	K	L	M	N	O
AssetID	ReadDate	Volume	ProblemOccured	ProblemTypeName	Survival.Day	CumPage	Page.Since.Last.Failure	age	Label	Page.2D	Page.7D	Page.15D	Page.30D
000C52Fi	2012-11-09	0	1	Others	495	15472	15472	17	2	162	341	640	1260
000C52Fi	2013-02-05	81	1	Others	88	20672	5200	20	2	145	385	849	1578
000C52Fi	2013-04-23	0	1	Device Fault	77	23765	3093	22	2	38	160	585	858
000C52Fi	2013-07-16	108	1	Paper Jam	84	27904	4139	25	2	126	355	486	969
000C52Fi	2013-12-06	45	1	Device Fault	143	37969	10065	30	2	100	297	506	1278
000C52Fi	2014-01-29	181	1	Others	54	40739	2770	31	2	223	333	641	1481
000C52Fi	2014-02-25	0	1	Others	27	41755	1016	32	2	0	120	361	1264
000C52Fi	2014-05-28	42	1	Device Fault	92	48975	7220	35	2	139	1583	2010	2941
000C52Fi	2014-06-17	48	1	Device Fault	20	50082	1107	36	2	164	459	860	2844
000C52Fi	2014-08-06	90	1	Device Fault	50	54706	4624	38	2	341	1116	2023	3769
000C52Fi	2014-08-25	106	1	Others	19	57277	2571	38	2	106	1282	2122	3687
000C52Fi	2015-01-08	0	1	Device Fault	136	69318	12041	43	2	0	0	99	769
00A32BA	2012-10-12	1588	1	Paper Jam	54	20849	20849	3	2	1588	2373	5507	12476
00A32BA	2014-09-19	123	1	Paper Jam	707	300716	279866	26	2	402	1468	3856	9912
00A32BA	2015-02-19	6	1	Device Fault	153	324167	23451	31	2	82	240	1320	5110
00A32BA	2015-06-09	0	1	Paper Jam	110	355309	31142	35	2	595	1403	3984	8348
00EAD91	2012-09-10	3183	1	Device Fault	43	22982	22982	3	1	3183	14194	22103	22632
00EAD91	2012-10-23	3042	1	Others	43	94208	71226	4	1	3042	10685	25852	51337
00EAD91	2013-01-02	4200	1	Device Fault	71	221158	126950	7	1	4200	4200	11575	51917
00EAD91	2013-04-09	0	1	Paper Jam	97	370730	149572	10	1	0	288	5652	34502
00EAD91	2013-05-06	5635	1	Others	27	416591	45861	11	1	5635	13365	26681	45861
00EAD91	2013-05-08	1956	1	Device Fault	2	421180	4589	11	1	4589	13602	25606	50450
00EAD91	2013-08-27	0	1	Device Fault	111	482359	61179	14	1	0	0	500	514
00EAD91	2013-09-06	3533	1	Others	10	502889	20530	15	1	8235	20530	20530	21030
00EAD91	2013-09-19	3676	1	Paper Jam	13	525239	22350	15	1	5257	12654	30585	42880
00EAD91	2013-10-01	2489	1	Paper Jam	12	547816	22577	16	1	4138	14194	31495	65457

Final Dataset – Statistics

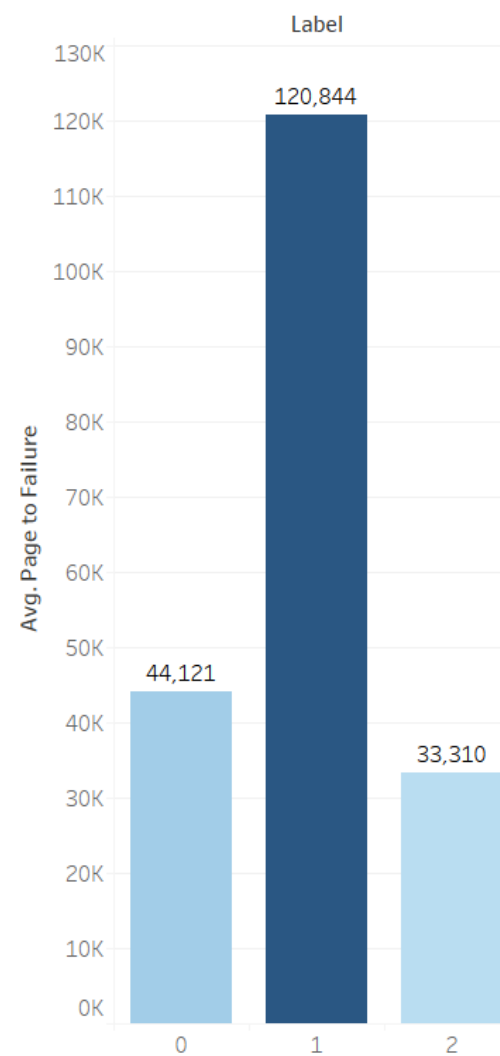
Average Survival Day by Label



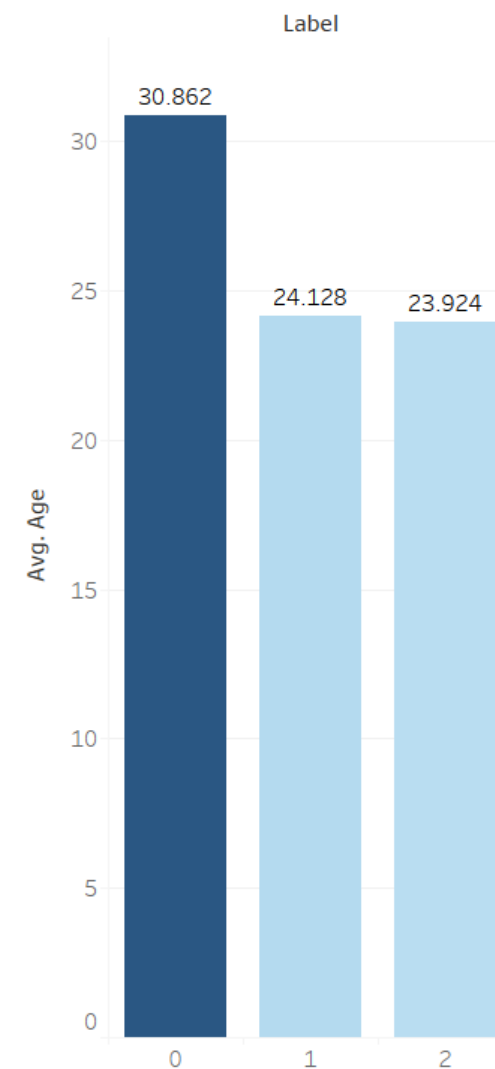
Average Cum Page by Label



Average Page since failure by Label



Average Age by Label



Introduction to Survival Analysis

Statistical methods for analyzing longitudinal data on the occurrence of events

Events may include death, injury, onset of illness, recovery from illness (binary variables) or transition above or below the clinical threshold of a meaningful continuous variable

- ❖ **Estimate time-to-event for a group of individuals**
Eg. time until second heart-attack for a group of MI patients
- ❖ **To compare time-to-event between two or more groups**
Eg. treated vs. placebo MI patients in a randomized controlled trial
- ❖ **To assess the relationship of co-variables to time-to-event**
Eg. does weight, insulin resistance, or cholesterol influence survival time of MI patients?



Introduction to Multivariate Cox Regression

The most popular techniques for survival analysis is Cox proportional hazards regression

- ❖ The objective is to assess simultaneously the effect of several risk factors on survival time. It allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time.
- ❖ The measure of effect is the hazard rate.
- ❖ Hazard rate: the risk of failure given that the participant has survived up to a specific time

Variable coefficients summary

	Univariate Regressions					
	coef	exp(coef)	se(coef)	z	Pr(> z)	
age	-0.00614	0.99388	0.001281	-4.79	1.66E-06	***
CumPage	0.009802	1.00985	0.000798	12.28	<2e-16	***
Page.Since.Last.Failure	-0.18366	0.83222	0.01597	-11.5	<2e-16	***
Page.30D	0.079857	1.083132	0.007174	11.13	<2e-16	***
Page.15D	0.081076	1.084454	0.008835	9.176	<2e-16	***
Page.7D	0.08296	1.0865	0.01274	6.51	7.54E-11	***
Page.2D	1.06239	2.89328	0.09942	10.69	<2e-16	***

Age: in month

Page Variables: in 100,000 pages (if the variable increases by 1 unit, that means the volume increase by 100,000 pages)

Coefficient interpretation

Obtained from univariate COX regression

Coef sign

Positive: Hazard rate (Failure possibility) increase when variable increase

Negative: Hazard rate (Failure possibility) decrease when variable increase

Exponential of coef

> 1: Hazard rate increase by $(\exp(\text{coef}) - 1)$ when variable increase 1 unit

< 1: Hazard rate decrease by $(1 - \exp(\text{coef}))$ when variable increase 1 unit

Negative coef of Age & Page.Since.Last.Failure:

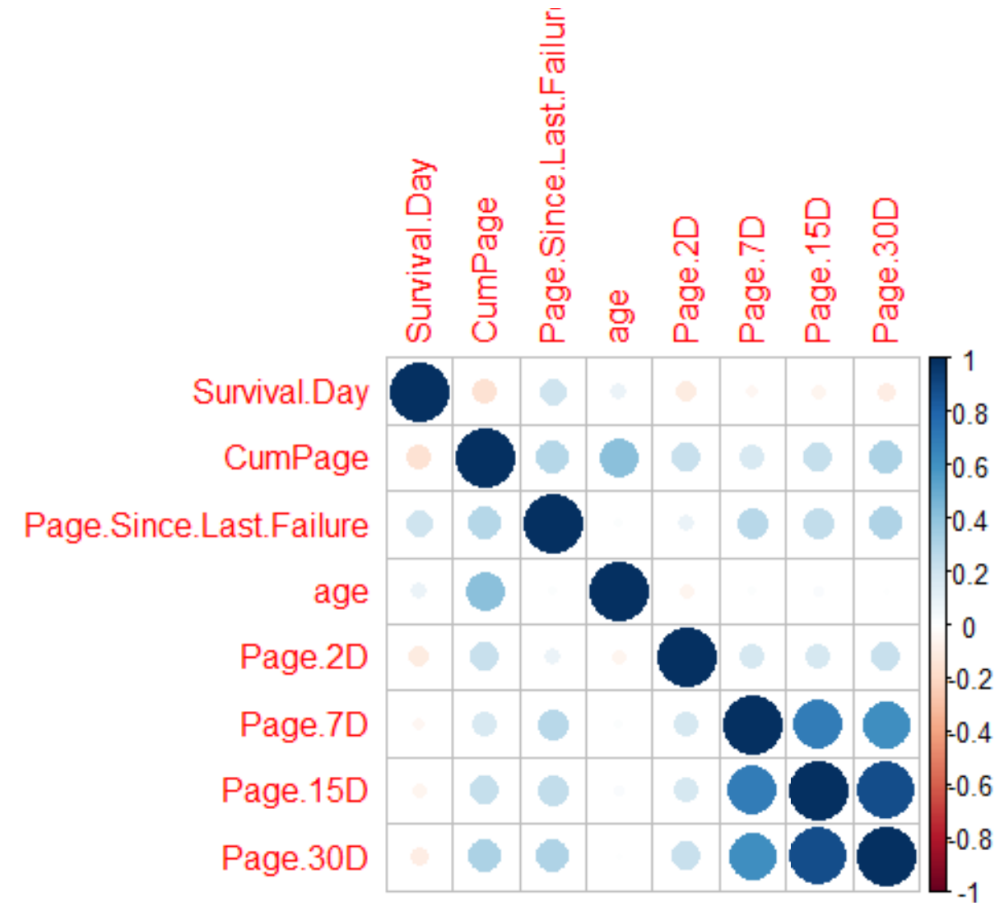
1. Intervention information
2. Repair data

Cross Validation: 80% to train the model, 20% to test the error rate

n= 3202, number of events= 3202

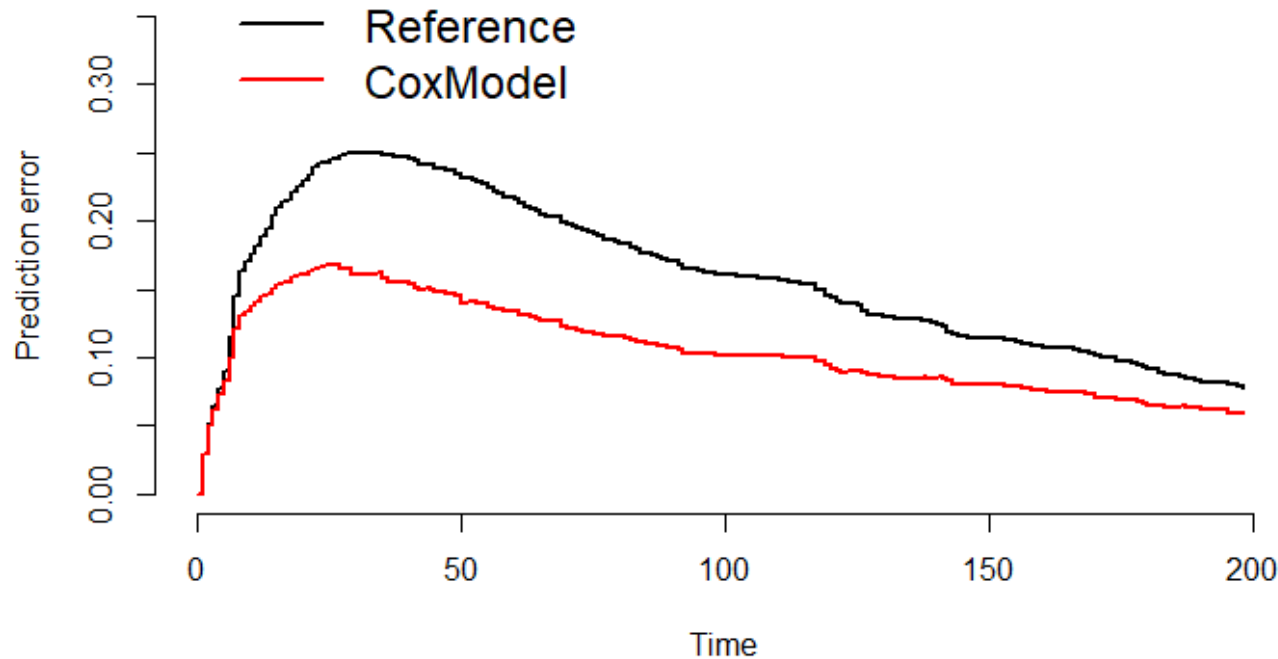
	coef	exp(coef)	se(coef)	z	Pr(> z)	
age	-0.012156	0.987918	0.001491	-8.154	3.33e-16	***
CumPage	0.014601	1.014708	0.001243	11.744	< 2e-16	***
factor(Label)1	0.518590	1.679658	0.058406	8.879	< 2e-16	***
factor(Label)2	-0.473851	0.622600	0.057986	-8.172	3.33e-16	***
Page.Since.Last.Failure	-0.666448	0.513530	0.023903	-27.882	< 2e-16	***
Page.30D	0.635647	1.888243	0.027499	23.115	< 2e-16	***
Page.15D	-0.591557	0.553465	0.030644	-19.304	< 2e-16	***
Page.7D	0.649968	1.915479	0.032183	20.196	< 2e-16	***
Page.2D	0.452661	1.572491	0.150890	3.000	0.0027	**

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1



Model 1 - Performance

- ❖ **Basic Survival Model:** Day 33, Error rate 0.2499809
- ❖ **Our Cox Model:** Day 26, Error rate 0.1689168



B-Score for Error Rate

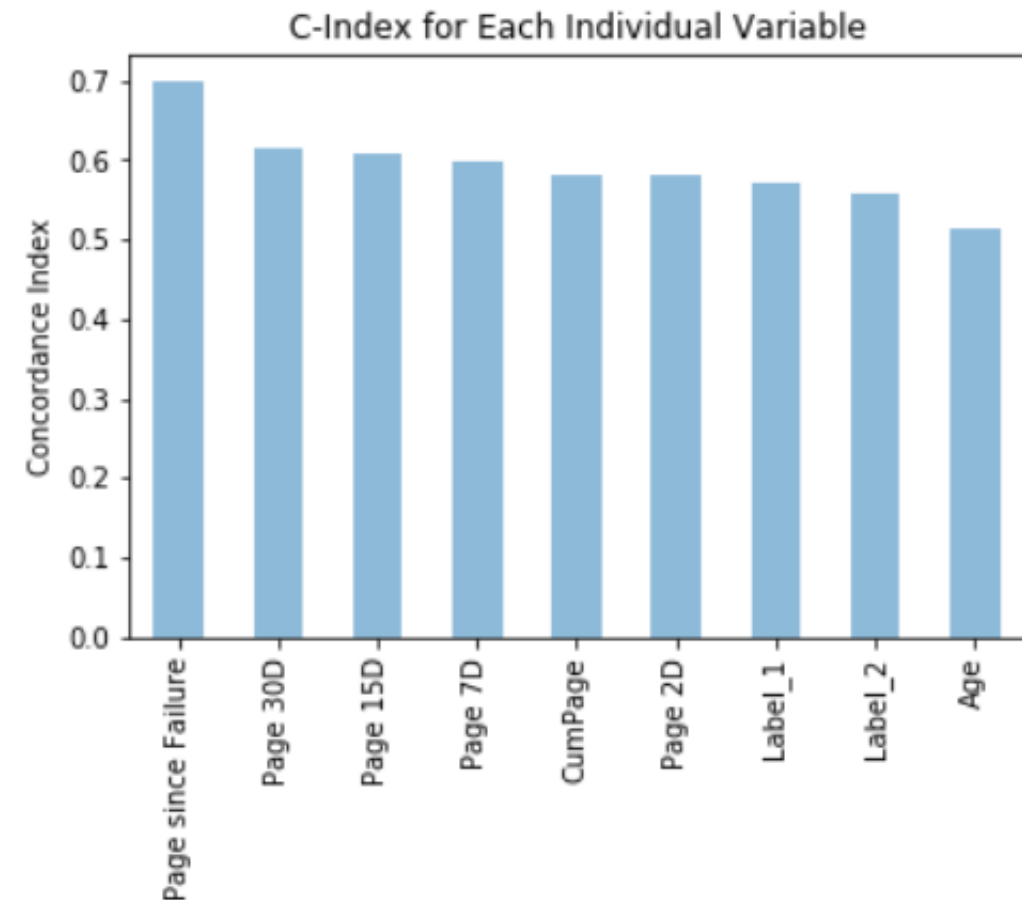
Brier score (B-Score) is a weighted average of the squared distances between the observed survival status and the predicted survival probability of a model.

Model 2 - Variable Predictive Performance

Concordance Index (C-Index)

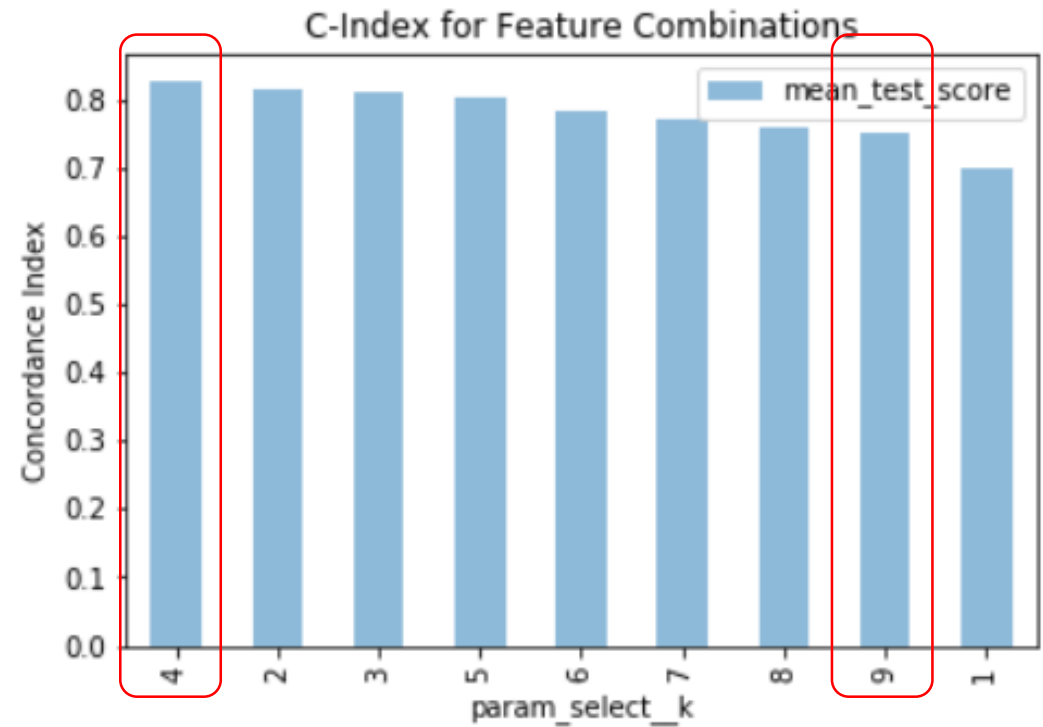
Generalization of the area under ROC curve, called Harrell's Concordance Index.

- Value of 0.5 denotes a random model
- Value of 1.0 denotes a perfect model
- Value of 0.0 denotes a perfectly wrong model



Model 2 - Feature Selection

- Use Grid Search with Cross Validation = 5
- 4 - Feature Combination achieved highest C-Index, and provided the best model
 - Page Since Failure
 - 30D Page Volume
 - 15D Page Volume
 - 7D Page Volume
- Fit best model with whole dataset
- C - Index before feature selection: 0.75
- C - Index after feature selection: 0.83



Demonstration of Model

Kindly see demonstration in R Studio

Recommendations

1. Explore deeper in survival analysis
2. Collect more data to train the model
3. Include maintenance data in the model
4. Update K-Means cluster labels when adding new asset features
5. Set auto-alarm reminders

Q & A

Thank you