HOW TO VALIDATE YOUR CLIENT CHURN MODEL

PyData London 2019 Elena Sharova

ABOUT ME

Senior Data Scientist at ITV since Jul-18

Previously worked for financial services

Have been using Python since 2013

Started building churn prediction models

at ITV for the HUB+ - ad-free VoD platform

TALK OUTLINE

Survival Analysis for Client Churn

Overview of Kaplan–Meier estimator and Proportional Hazards Model (CPH) Fitting CPH with lifelines

CPH Model Assumptions

Concordance Scores

Residuals

AUC

Other Methods

Conclusions

SURVIVAL ANALYSIS FOR CLIENT CHURN

SURVIVAL ANALYSIS FOR CLIENT CHURN

- All subscribers eventually churn…
- The TIME TO EVENT (*T*) is what we want to model. *T* is a random variable.
- How to predict the length of subscription time?
  - linear regression?
  - KNN or decision trees regression?

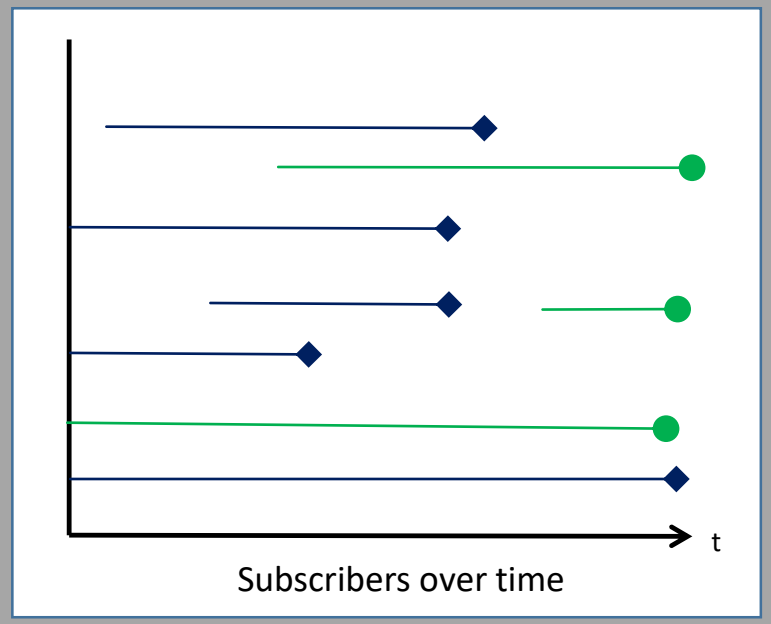
Survival Analysis used in medical statistics:

- guarantees $T \geq 0$
- provides a term-structure for tenures.

- This talk is about Cox Proportional Regression (non-time varying coefficients).

SURVIVAL ANALYSIS FOR CLIENT CHURN

- How do we measure time and when does the clock starts?
- Dealing with right-censored only.
- We assume that censorship time *C is non-informative*, and *T* and *C* are independent.
- In reality this may not be the case:
  - ITV schedule is seasonal
  - some viewers prefer one genre over others



SURVIVAL ANALYSIS FOR CLIENT CHURN

- Q – What data to calibrate the model on?

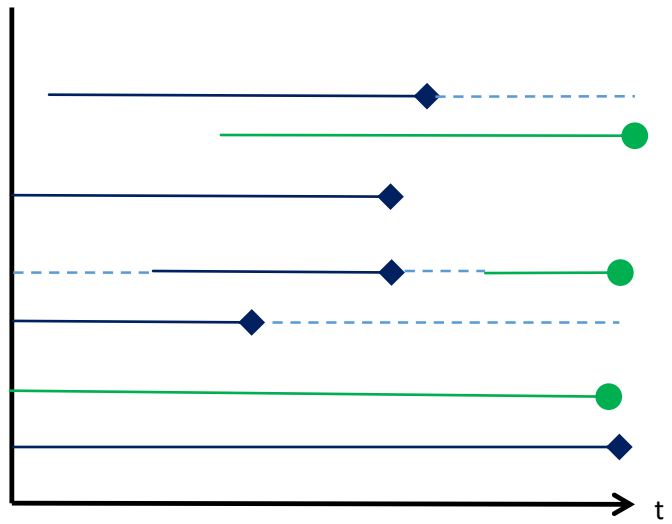
A – The data pertaining subscriptions only.

- Q – What to do with those who churn and re-subscribe?

A – Treat them separately. Depends on what the individual subscriptions mean to the business financially (e.g. can add tenures without discounting?).

- Q – What features should be used?

A – Features that describe the subscribers but do not 'leak' information about tenure (e.g. exclude total monthly paid, etc.)



Subscribers over time

Kaplan-Meier Estimator and Cox Proportional Hazards Model

Kaplan-Meier and Cox Proportional Hazards Model (CPH)

Time to event T is a positive random variable. It has a pdf:

$$f(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P(t \leq T < t + \Delta t)$$

A survival function *S(t)* is a probability that an individual survives longer than *t*:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - \int_0^t f(u)du$$

$$S(t) = 1 - F(t)$$

A hazard function is an instantaneous failure rate at time *t* given alive up to *t*:

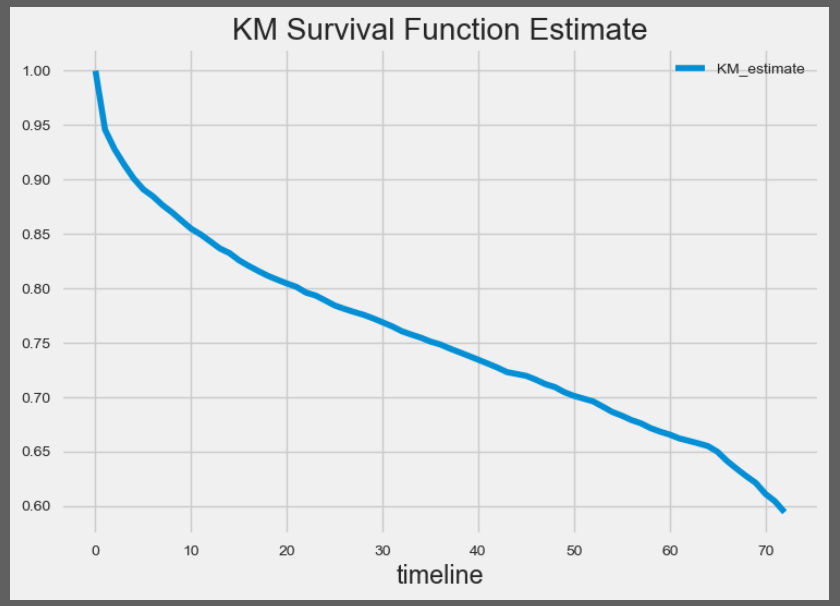
$$h(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) = \frac{f(t)}{S(t)} = \frac{-d\, ln[S(t)]}{dt}$$

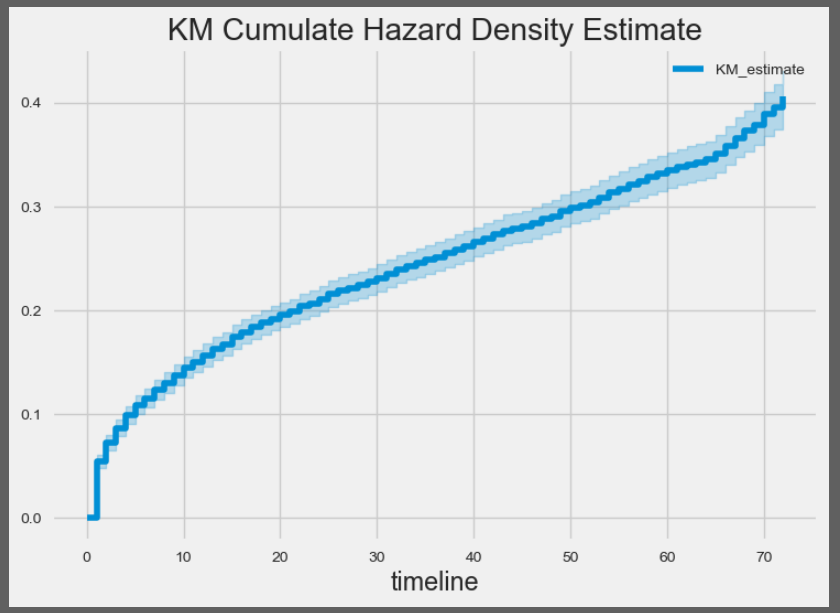
A cumulative hazard function at time *t*:

$$H(t) = \int_0^t h(u)du = -ln\,[S(t)]$$

Kaplan-Meier and Cox Proportional Hazards Model (CPH)

Kaplan-Meier and Cox Proportional Hazards





Kaplan-Meier and Cox Proportional Hazards Model (CPH)

How do we estimate the pdf? What do we model?

Kaplan and Meier (1958) proposed a nonparametric product-limit estimator for survival function $\boldsymbol{S(t)}$:

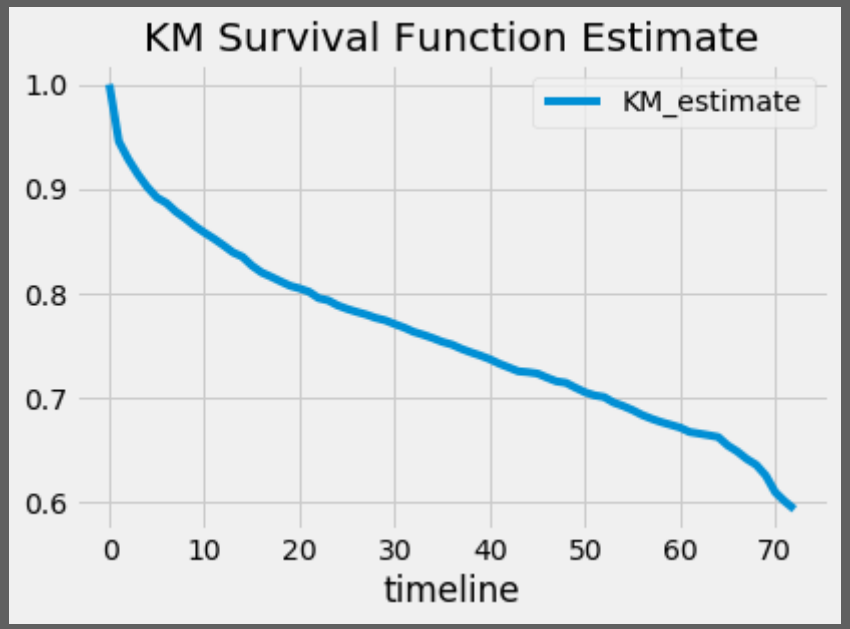
$$\widehat{\boldsymbol{S}}(\boldsymbol{t}) = \begin{cases} \mathbf{1}, & \boldsymbol{if\ t < t_1} \\ \prod_{\boldsymbol{t_i \le t}} \mathbf{1} - \frac{\boldsymbol{d_i}}{\boldsymbol{Y_i}}, & \boldsymbol{if\ t_1 \le t} \end{cases}$$

Kaplan-Meier estimator is conditional because an estimate at time t depends on the number of "deaths" and the number at risk of "death" at time $\boldsymbol{t_i}$.

Kaplan-Meier and Cox Proportional Hazards Model (CPH)

Kaplan-Meier product-limit estimator example (telco dataset):

	removed	observed	censored	entrance	at risk	survival	
event_at		d_i	c_i		Y_i		
0	4	0	4	4718	4718	=(1-0/4718)	1
1	417	257	160	0	4714	=1*(1-257/4714)	0.94548154
2	150	78	72	0	4297	=0.945*(1-78/4297)	0.92831897
3	136	64	72	0	4147	=0.928319*(1-64/4147)	0.9139924
4	120	54	66	0	4011	=0.91399*(1-54/4011)	0.90168497



Kaplan-Meier and Cox Proportional Hazards Model (CPH)

$\widehat{S}(t)$ is a product-limit of data-derived ratios. Additional information about subjects (e.g. age, gender, etc.) is not taken into account.

Cox Proportional Hazards Model (1972) is a semi-parametric regression model for $h(t)$.

Let's say we have information (vector of *k* covariates X_i) for $i = 1, \dots, N$ individuals. We also know their time-to-event and the outcome (death or censorship) - (T_i, δ_i, X_i). Then:

$$h_i(t) = h_0(t) \cdot exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})$$

Baseline hazard function for time *t*

Linear regression form

Kaplan-Meier and Cox Proportional Hazards Model (CPH)

CPH $h_i(t)$ consists of:

-time-varying baseline hazard (same for all subjects, but varies with time)

-time-invariant partial hazard that depends on subject's values for covariates.

hazard functions of two subjects $h_i(t)$ and $h_j(t)$ relate via a constant ratio of their partial hazards. This is the reason for 'proportional' in CPH.

LIFELINES

Python package can be used to fit CPH to data (original developer Cameron Davidson-Pilon).

Other Python packages:

sebp / scikit-survival

(original developer Sebastian Pölsterl).

tick.survival

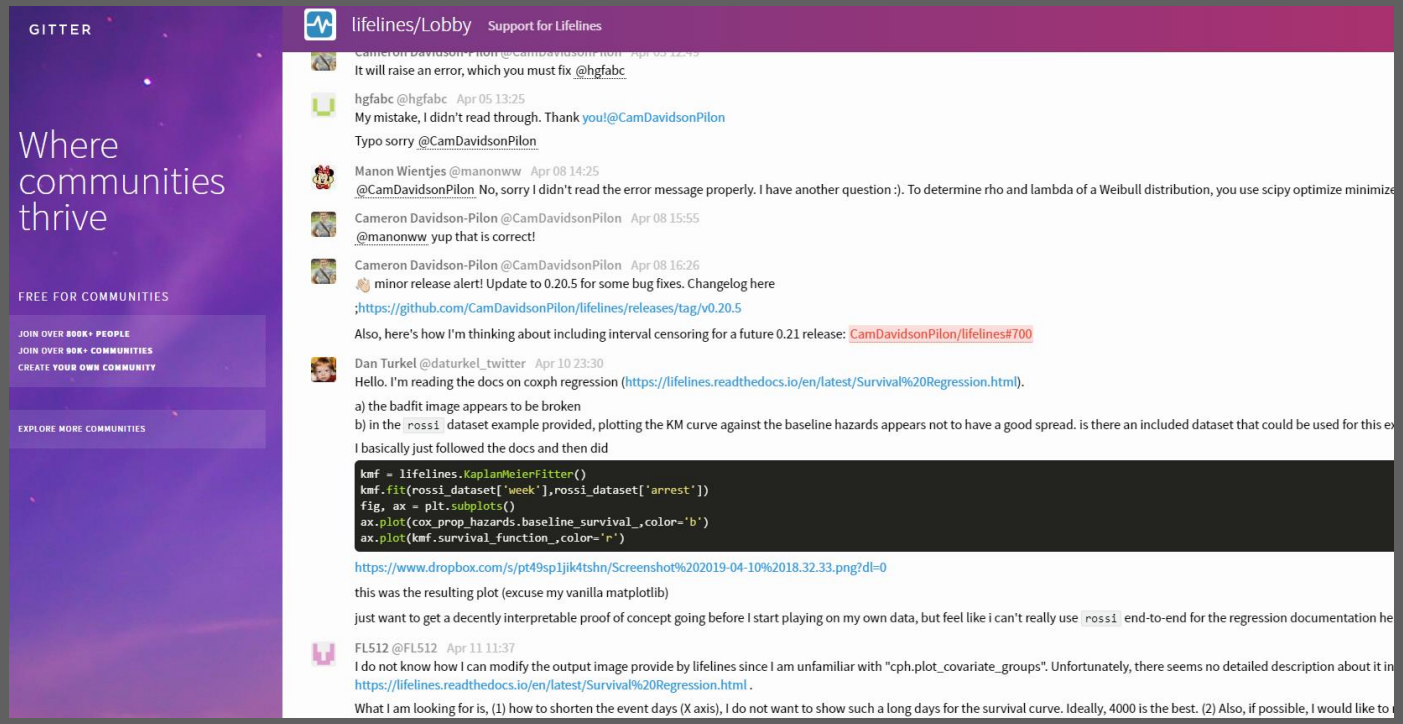
Python package for modelling point Hawkes processes.

PySurvival

Implements CPH and Survival Forest.

Kaplan-Meier and Cox Proportional Hazards Model (CPH)

lifelines has a support channel: https://gitter.im/python-lifelines/Lobby



Kaplan-Meier and Cox Proportional Hazards Model (CPH)

Derivation in lifelines package:

$$\underbrace{h(t|x)}_{\text{hazard}} = \overbrace{b_0(t)}^{\text{baseline hazard}} \underbrace{\exp\overbrace{\left(\sum_{i=1}^{n} b_i(x_i - \overline{x_i})\right)}^{\text{log-partial hazard}}}_{\text{partial hazard}}$$

Standardised covariates

Regression coefficients are found using MPLE and Newton-Raphson iterative search method. Ties in *T*'s are resolved with Efron's method.

The baseline hazard function is estimated from cumulative partial hazard estimates, grouped by tenure.


Kaplan-Meier and CPH Demo

CPH MODEL ASSUMPTIONS

CPH Model Validation – Checking Assumptions

CPH Model Assumptions

CPH Model is a regression model with *at least* the following underlying assumptions:

1) Structure of the model is assumed to be correct:
   - Linearity and multiplicative covariate effect on the hazard rate
2) Proportionality of the hazards of any two subjects(when covariates are time invariant).
3) No significantly influential data points.
4) The model can accurately predict churn for different customer risk groups.

CPH Model Assumptions

ASSUMPTIONS

1) Structure of the model is assumed to be correct:
   - Linearity and multiplicative covariate effect on the hazard.
2) Proportionality of the hazards of any two subjects.
3) No significantly influential data points.
4) The model can accurately predict churn for different customer risk groups.

HOW TO CHECK

1) Martingale Residuals*, C-Index and Uno Concordance
2) Schoenfeld Residuals
3) Deviance* and score Residuals
4) AUC, Brier Score

Deviance residuals can also be used to validate (1)

Concordance Scores and Residuals

Concordance Scores and Residuals – C-Score

Concordance Index Definition:

proportion of pairs that are concordant in duration *T* vs. estimated hazard *H*:

$H_1 > H_2$ $T_1 < T_2$	$H_1 < H_2$ $T_1 < T_2$
$H_1 > H_2$ $T_1 > T_2$	$H_1 < H_2$ $T_1 > T_2$

Concordance Scores and Residuals – C-Score

lifelines package provides Concordance Index calculation.

Definition: proportion of pairs that are concordant in duration vs. estimated hazard.

observation	age	observed	duration	p_hazard
0	10	1	6	0.775902
1	12	1	5	0.829296
2	15	1	4	0.916357
3	34	1	3	1.724427
4	9	1	3	0.750508
5	11	1	4	0.802155
6	22	1	7	1.156713
7	28	1	2	1.412327

```
*Observation 7 = (2, 1.41).
*Observation 0 = (6, 0.77).
7 vs. 0 is concordant (6 > 2 and 0.77 < 1.41).
*Observation 1 = (5, 0.83).
7 vs. 1 is concordant (5 > 4 and 0.83 < 1.41).
*Observation 2 = (4, 0.92).
7 vs. 2 is concordant (4 > 2 and 0.92 < 1.41).
*Observation 3 = (3, 1.72)
7 vs. 3 is discordant (3 > 2 but 1.72 > 1.41).
...
```

=15/26 (58%)

In the absence of ties, C-score will make $\frac{N!}{n!\cdot(N-n)!}$ comparisons. Tied observations are excluded from comparison (e.g. observation 3 cannot be compared to observation 4).

Concordance Scores and Residuals – C-Score

However, censored observations are compared even in the presence of ties.

- Censored observation may result in the upward bias in C-score.

H. Uno et.al [2] proposed an estimator free of censoring distribution.

Definition:

proportion of pairs that are concordant in duration vs. estimated hazard,

weighted by censoring probability derived with Kaplan-Meier estimator

The observed difference between C-score and C-score IPCW depends on the distribution of censorship and which records are censored.

*IPCW – inverse probability of censoring weights

Martingale Residuals

Martingale residuals compare *observed* to *expected at time T*

$$\boldsymbol{res_{mar} = event\ [1\ or\ 0]_i\ - H_{i,T}}$$

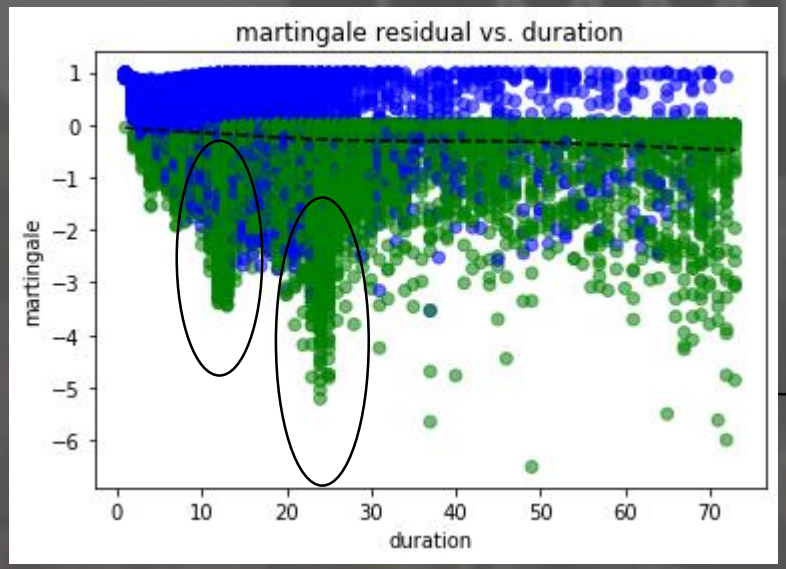
$\boldsymbol{res_{mar}}$ in [1,-inf) for uncensored and in [0, -inf) for censored observations.

$+\ \boldsymbol{res_{mar}}$ → at T $\boldsymbol{H_{i,T}}$ was underestimated

$-\ \boldsymbol{res_{mar}}$ → at T $\boldsymbol{H_{i,T}}$ was overestimated or censored observation

Martingale Residuals

Martingale residuals can be plotted against durations



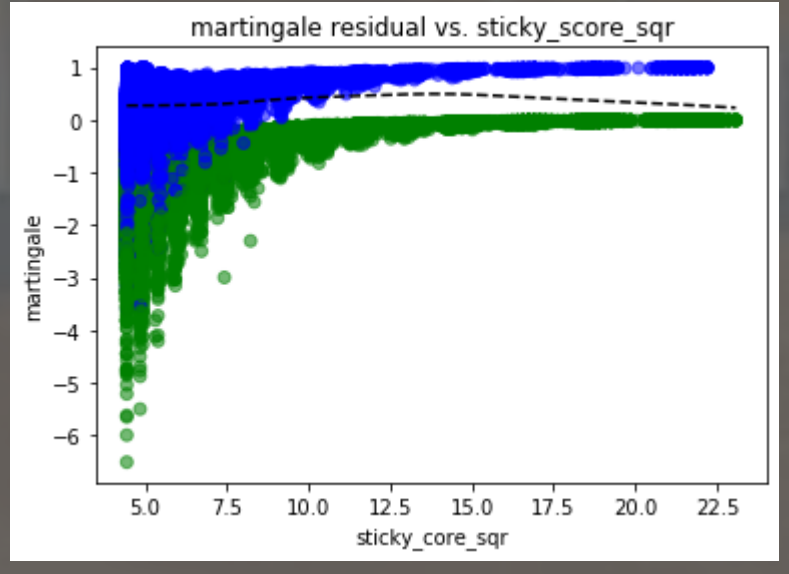
Censored observations with 12 and 24 months durations

***skew makes it hard to spot outliers*

Martingale Residuals

Martingale residuals can also be plotted against model covariates to examine the relationship between model error and covariate values.

To spot non-linearity, add LOWESS lines [8] (scatterplot smoothing lines).



Sticky score = genre diversity and monthly stickiness-loyalty

Deviance Residuals

Deviance residual are a transformation on martingale residuals to achieve symmetry:

$$\boldsymbol{res_{dev}} = \boldsymbol{sign(res_{mar})} \cdot \sqrt{-\boldsymbol{2} \cdot [\boldsymbol{res_{mar}} + \boldsymbol{I_i} \cdot \log(I_i - res_{mar})]}$$

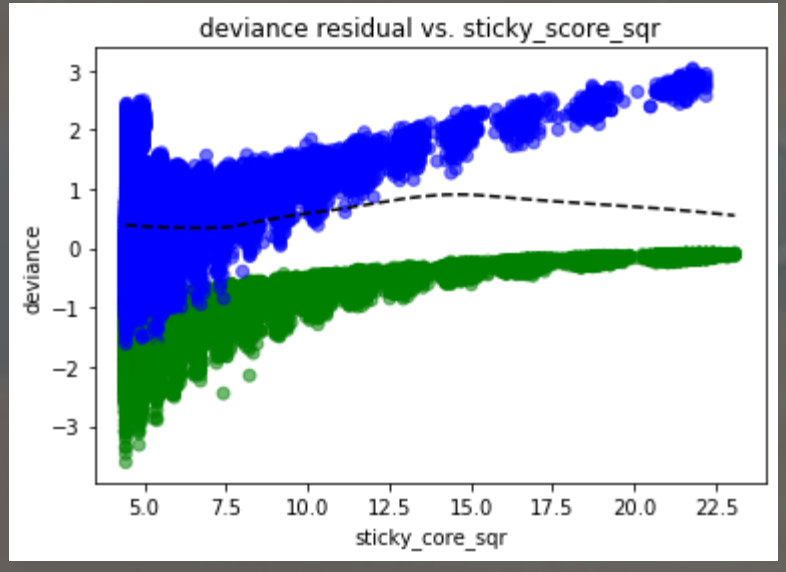
where $\boldsymbol{I_i}$ is the event indicator for i's observation.

Deviance residuals have the same sign as martingale residuals.

Can be used to examine the functional model fit and identification of highly influential data points.

Deviance Residuals

Deviance residuals plotted against model covariates to examine the relationship between model error and covariate values.



Sticky score = genre diversity and monthly stickiness-loyalty

Large sticky scores are generally predictive of longer subscription duration.

Schoenfeld Residuals

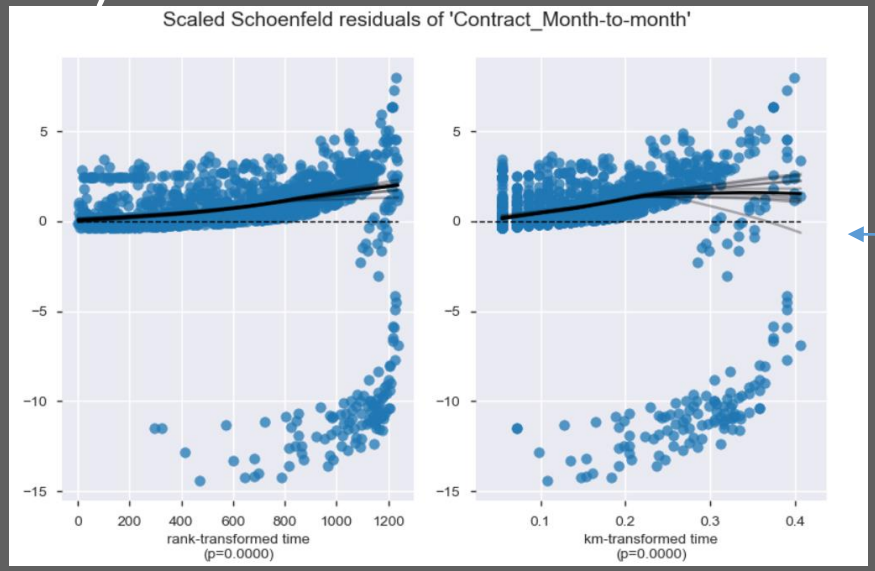
Schoenfeld residuals are used to test the proportional hazards assumption.

It is defined for each covariate, as the difference between the observed and the average in the risk-set at each duration.

- It is not defined for censored observation

Schoenfeld Residuals

lifelines package implements scaled Schoenfeld residuals (scaled by the variance-covariance matrix of coefficients)

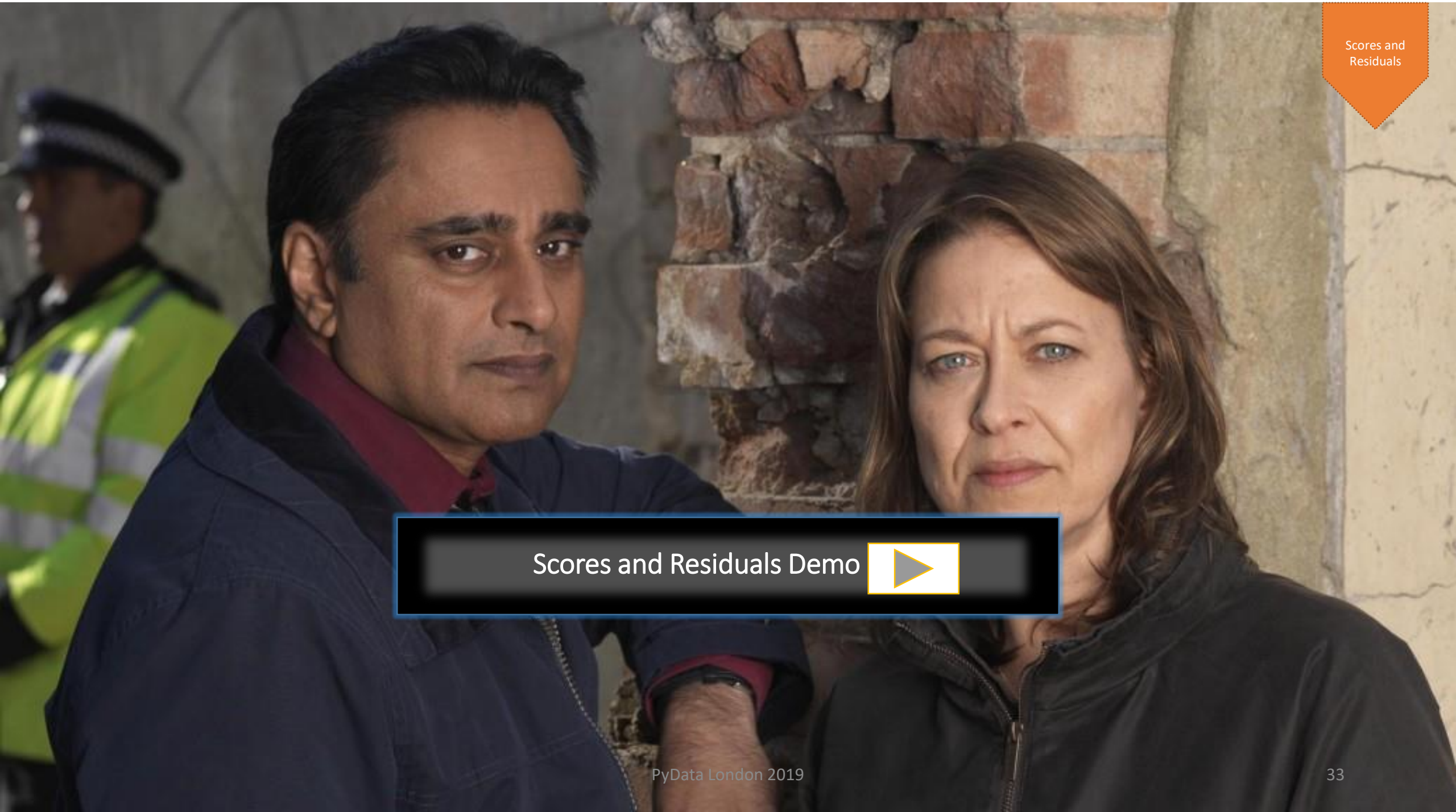


LOWESS curves - locally estimated scatterplot smoothing curves. Deviations from a constant line are violations of the PH assumption

Scores and Residuals

HOW TO FIX IT?

- Small concordance scores
  - introduce new features (are you over or under-estimating the hazards?)
- Large martingale residuals with non-random patterns (non-constant lowess lines)
  - introduce interactive terms or consider squared, cubed, etc. covariates.
- Proportionality assumption does not hold
  - stratify by the covariate that breaks it or move to time-varying model. lifelines library allows to perform non-interactive stratification, meaning that the baseline hazard function varies by the strata. There is also methods for an interaction-model stratification where the regression coefficients vary by the strata [6].



Scores and Residuals Demo

Area Under the ROC Curve - AUC

AUC For Survival Analysis

AUC

AUC is a measure of model's sensitivity vs. specificity. Taking a set of covariates as *M*,

In survival analysis:

$$\boldsymbol{sensitivity = P(M > c | T \leq t)}$$

$$\frac{TP}{TP + FN}$$

$$\boldsymbol{specificity = P(M \leq c | T > t)}$$

$$\frac{TN}{TN + FP}$$

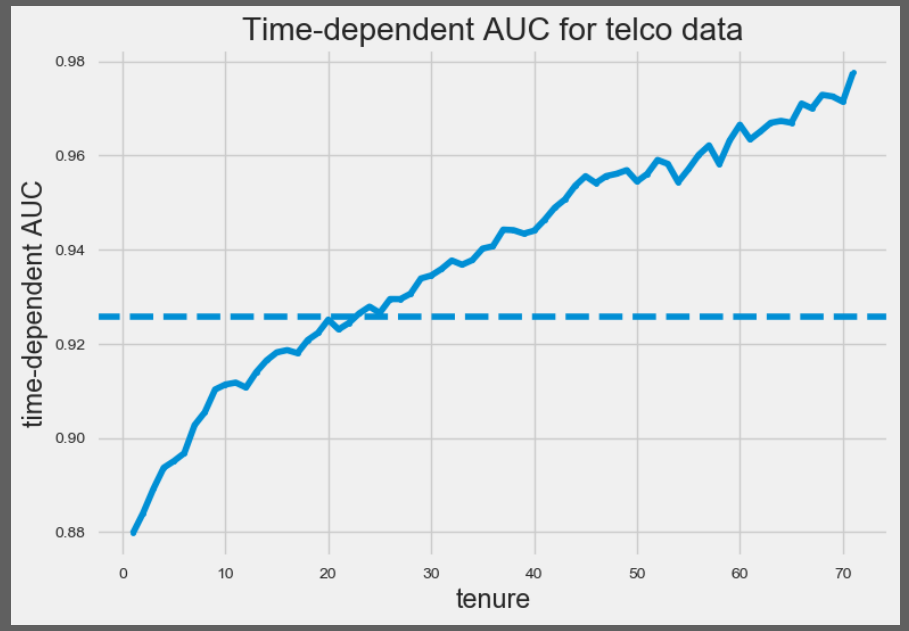
AUC is effectively *covariates – to- outcome concordance score*, and is calculated for each tenure.

*scikit-survival implements AUC calculation for survival analysis. But there are many different definitions in literature (see [3])

AUC For Survival Analysis

AUC

We can calculate AUC for each covariate, tenure and compute an average AUC.



Here, mean AUC is 92.5%.

AUC For Survival Analysis

AUC

A word of warning on AUC…

"Sensitivity and specificity only tell you something obliquely about prediction. They tell you something about the observed error proportions for specific tests or algorithms, but not about uncertainties for future observations or events and directly about the quality of the prognostication."

Drew Griffin Levy on https://www.fharrell.com/post/mlconfusion/

Other Methods

Brier Score

Other Methods

Brier Score* is a score function for accuracy of probabilistic prediction for some time T:

$$BrierScore_T = \frac{\sum_{i=1}^{N}\left(f_{i,t} - I_{i,t}\right)^2}{N}$$

The smaller the score the better the model is calibrated.

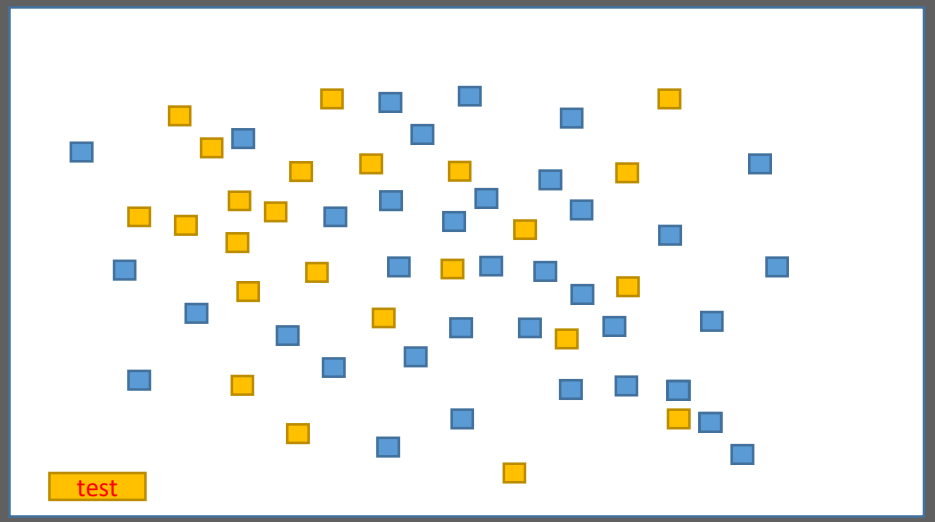
Output score should be compared with the score obtained using the 'null model' e.g. Kaplan-Meier.

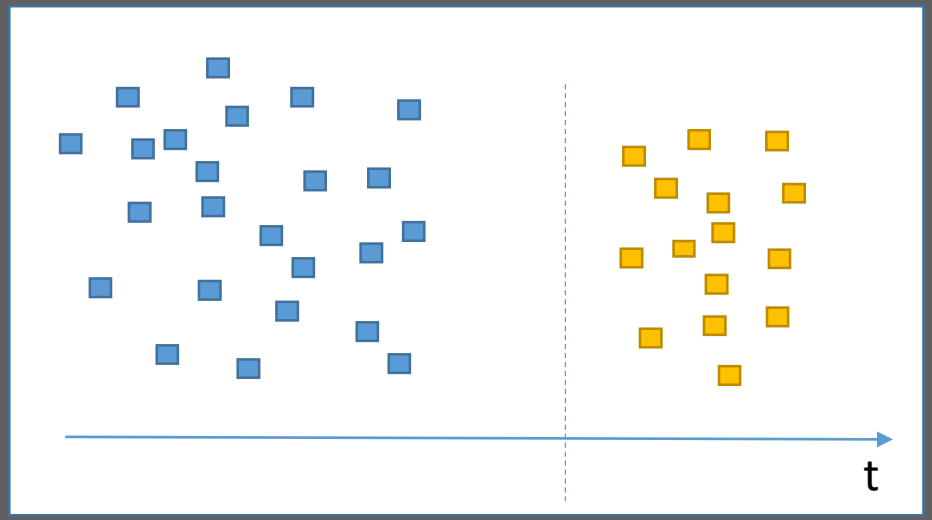
* Originally proposed by Glenn W. Brier in 1950

scikit-learn implements Brier score loss: sklearn.metrics.brier_score_loss

Scores and AUC Tests

Concordance scores, AUC and Brier Score Loss can be estimated on out-of-sample as well as out-of-time-sample data [7].





Conclusions

lifelines and scikit-survival

* excellent tools to perform survival analysis
* come with good model validation utilities.

Survival Model Validation is extensively covered in biostatistics research and is becoming more accessible through Python libraries to the data scientists.

It is important to use more than one validation tool to achieve better understanding of the built model and greater confidence in its usability.

References

[1] F.Harrell's Blog on Statistically Efficient Ways to Quantify Added Predictive Value of New Measurements https://www.fharrell.com/post/addvalue/

[2] H.Uno, et.al. "On the C-Statistics For Evaluating Overall Adequacy of Risk Prediction Procedures With Censored Survival Data." Harvard University Biostatistics Working Paper Series (101). 2009.

[3] P.J.Heagerty. Course slides: https://faculty.washington.edu/heagerty/Courses/Montpellier/montpelier-3.pdf

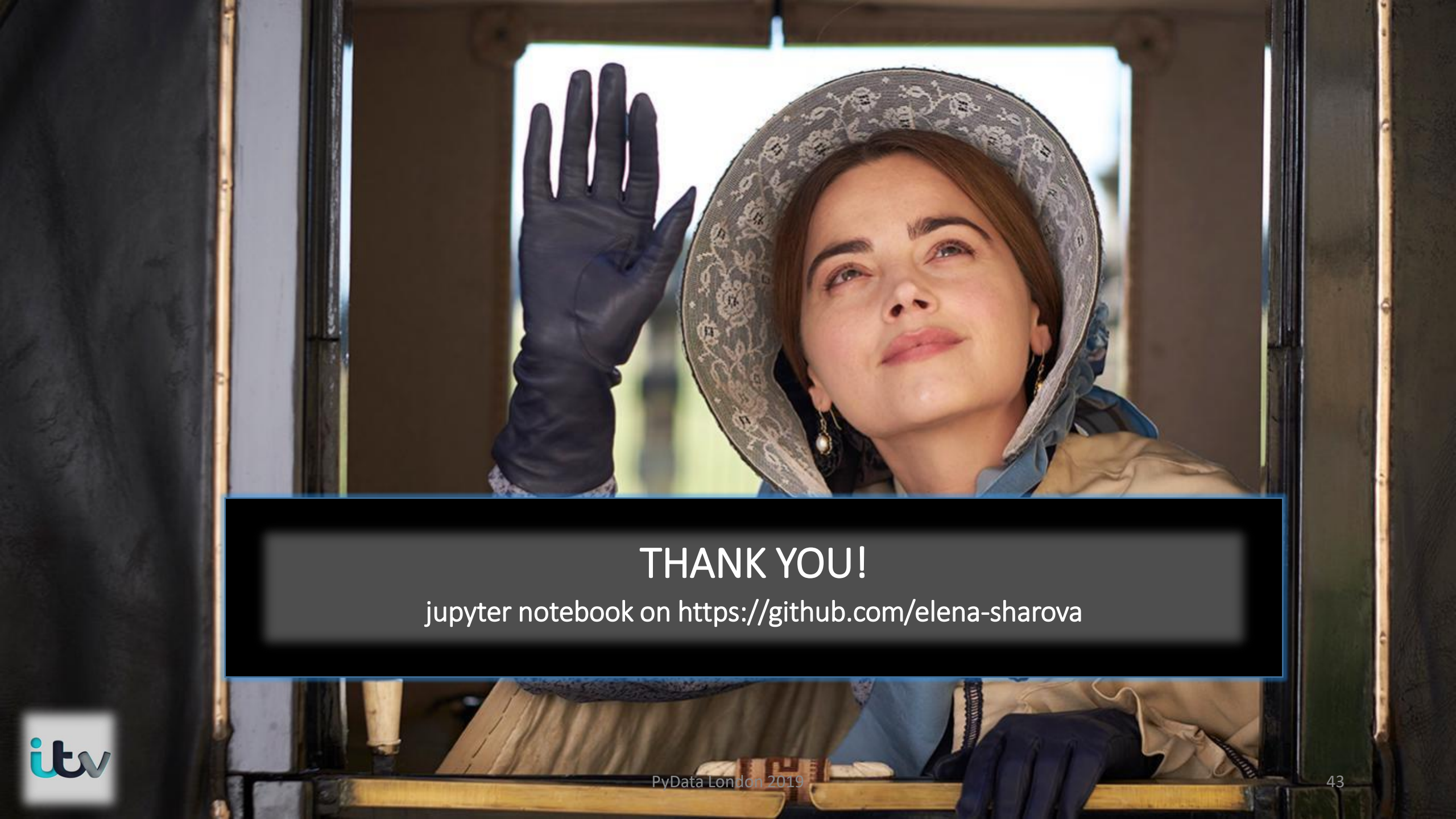
[4] lifelines package (citing): https://zenodo.org/record/3267531

[5] scikit-survival package (citing on bottom of the page): https://github.com/sebp/scikit-survival

[6] M.Abdelaal, S.Zakria. "Modelling Survival Data by Using Cox Regression Model". American Journal of Theoretical and Applied Statistics. 2015. 4(6). 504-512.

[7] E.Lima, C.Mues, B.Baesens. "Monitoring and backtesting churn models", Expert Systems with Applications, 38, 2011 975-982.

[8] F.Harrell, K.Lee, D. Mark. "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors." Statistics in Medicine, vol 15, 361-387, 1996.



THANK YOU!

jupyter notebook on https://github.com/elena-sharova