POLITECNICO MILANO 1863

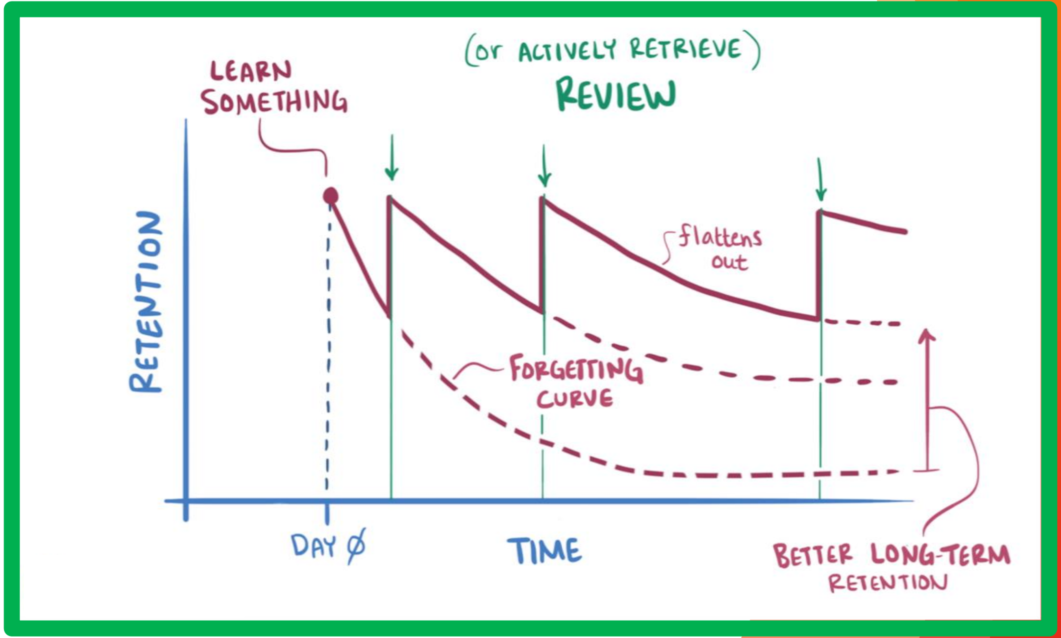
Nonparametric modelling for spaced repetition scheduling

Politecnico di Milano - 03/02/2021

Spaced Repetition

Spaced repetition is a method for **memorizing concepts**:

- No cramming, reviews spaced through time
- Increasing durations between reviews as one learns the item
- Software schedules each review



duolingo and Half-Life Regression

Duolingo is a language learning app:

- Relies on spaced repetition under the hood
- Half-Life Regression model to estimate the user’s probability of recalling an item at any point in time after the last review

Half-Life Regression paper + dataset:

- 2 weeks of real usage data
- 115’000 users
- 13 million word recall probabilities

duolingo Model

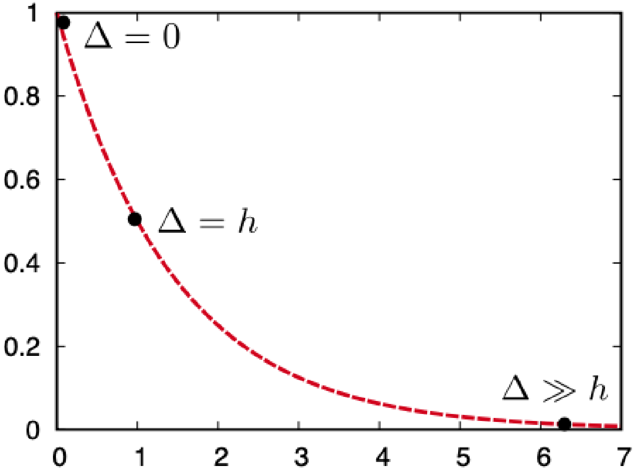
Forgetting Curve Model:
(Ebbinghaus, 1885)

lag time

$$p = 2^{-\Delta/h}$$

Half-life

p-recall



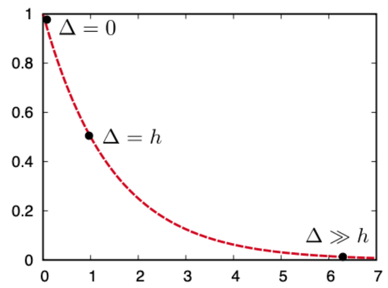
duolingo Model

lag time

Forgetting Curve Model: $$p = 2^{-\Delta/h}$$

(Ebbinghaus, 1885)

p-recall *Half-life*



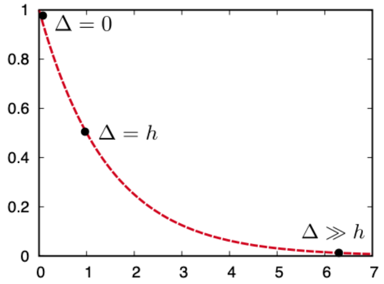
Half-Life Regression Model:

parameters

$$\hat{h}_\Theta = 2^{\Theta \cdot \mathbf{x}} \longrightarrow \hat{p}_\Theta = 2^{-\Delta/\hat{h}_\Theta}$$

feature variables

duolingo Model



Forgetting Curve Model:

(Ebbinghaus, 1885)

lag time

$$p = 2^{-\Delta/h}$$

p-recall *Half-life*

Half-Life Regression Model:

parameters

$$\hat{h}_\Theta = 2^{\Theta \cdot \mathbf{x}} \rightarrow \hat{p}_\Theta = 2^{-\Delta/\hat{h}_\Theta}$$

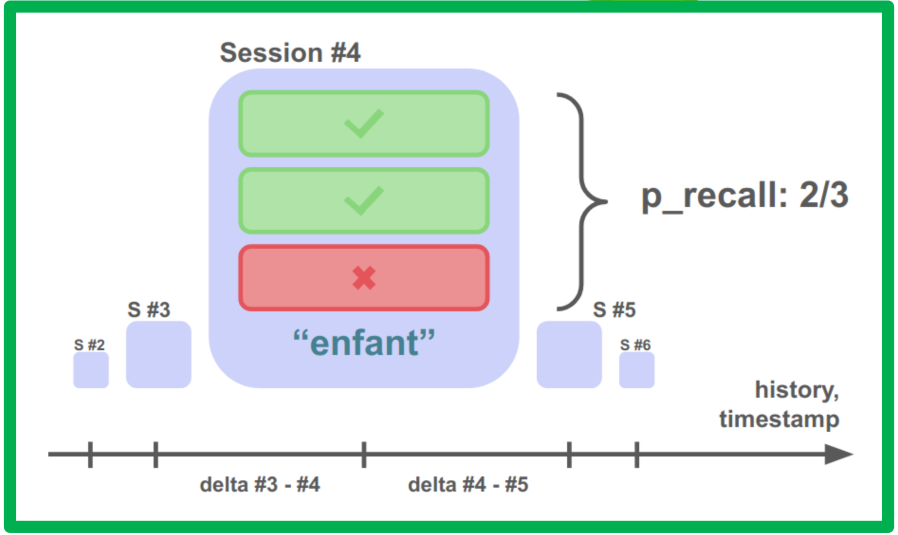
feature variables

GOALS:

develop the best model in order to improve the users’ experience

DATA EXPLORATION

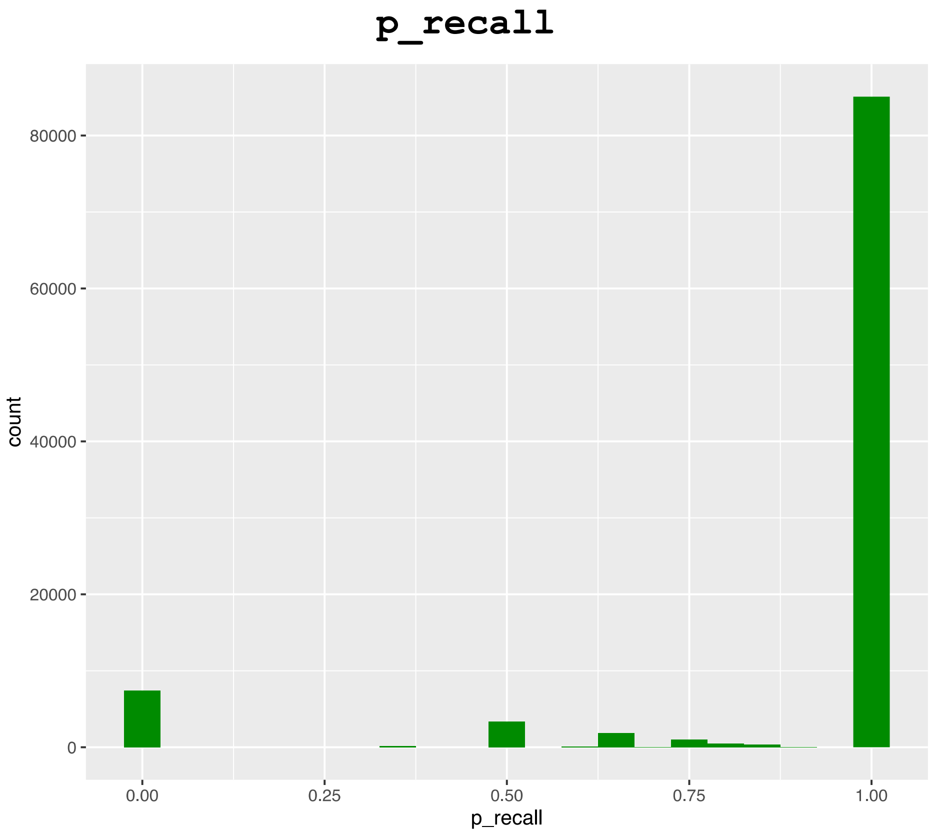
Our Data



- More than **12 million** "sessions"
- More than **115,000** users
- More than **19,000** words in 5 languages

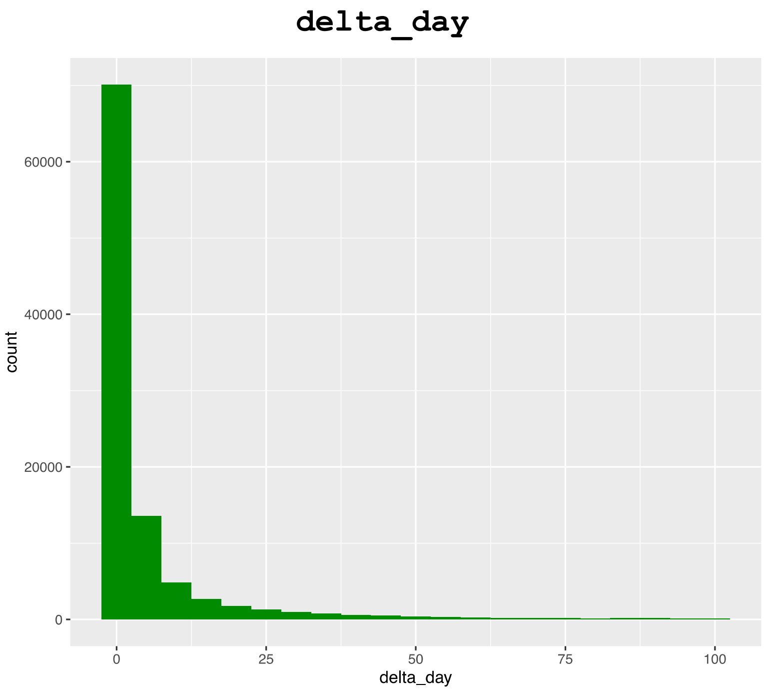
- `p_recall` - proportion of exercises from this lesson/practice where the word/lexeme was correctly recalled
- `timestamp` - UNIX timestamp of the current lesson/practice
- `delta` - time (in seconds) since the last lesson/practice that included this word/lexeme
- `user_id` - student user ID who did the lesson/practice (anonymized)
- `learning_language` - language being learned
- `ui_language` - user interface language (presumably native to the student)
- `lexeme_id` - system ID for the lexeme tag (i.e., word)
- `lexeme_string` - lexeme tag (see below)
- `history_seen` - total times user has seen the word/lexeme prior to this lesson/practice
- `history_correct` - total times user has been correct for the word/lexeme prior to this lesson/practice
- `session_seen` - times the user saw the word/lexeme during this lesson/practice
- `session_correct` - times the user got the word/lexeme correct during this lesson/practice

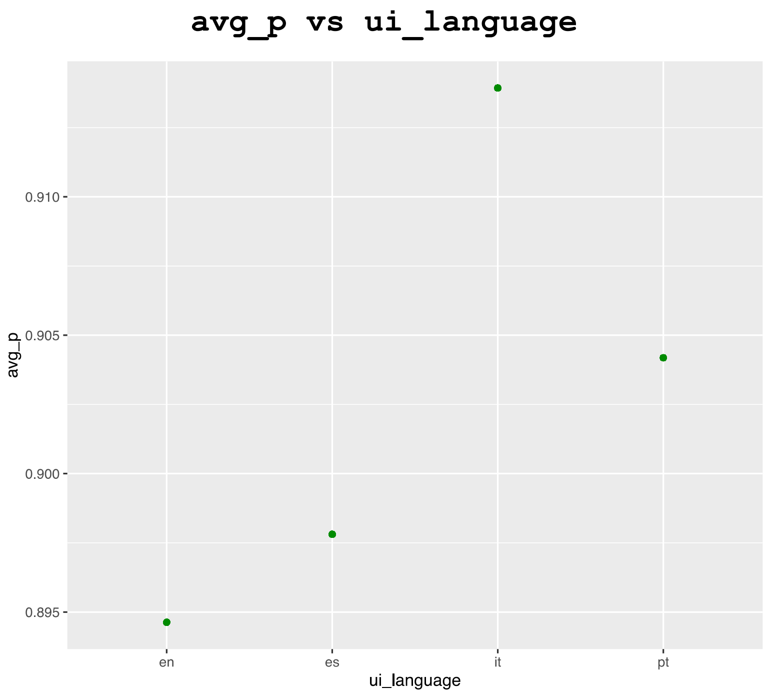
Our Data



p_recall
80000
60000
40000
20000
0
count
0.00
0.25
0.50
0.75
1.00
p_recall

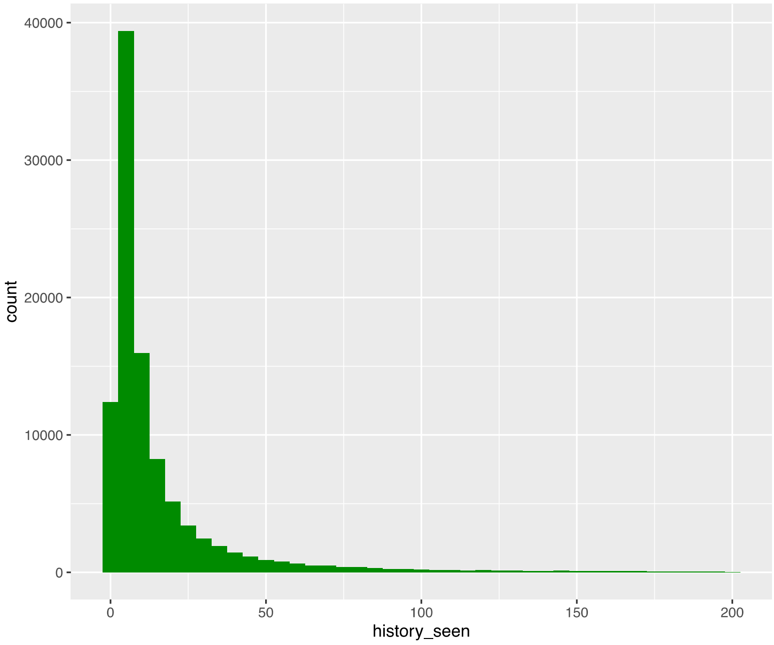
Our Data


delta_day
count
60000
40000
20000
0
0
25
50
75
100
delta_day

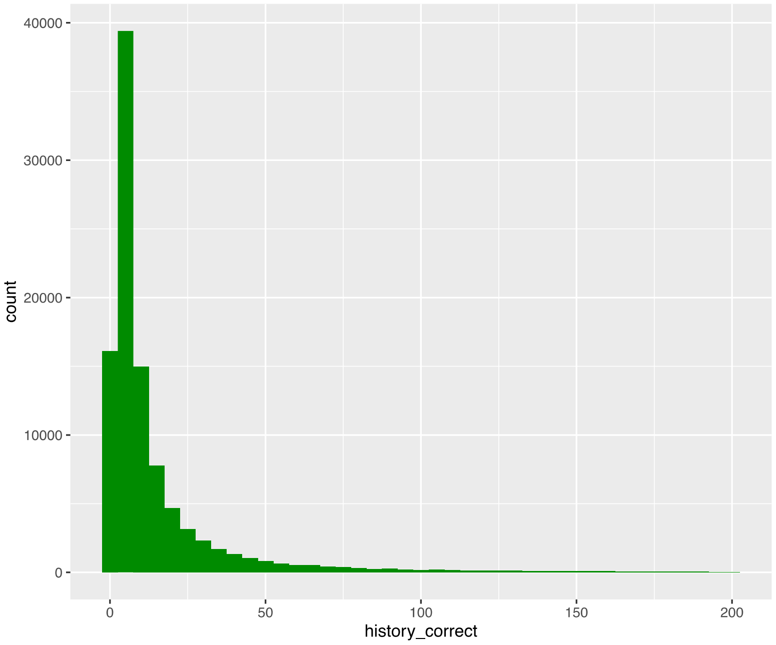

avg_p vs ui_language
avg_p
0.910
0.905
0.900
0.895
en
es
it
pt
ui_language

Our Data

history_seen



history_correct

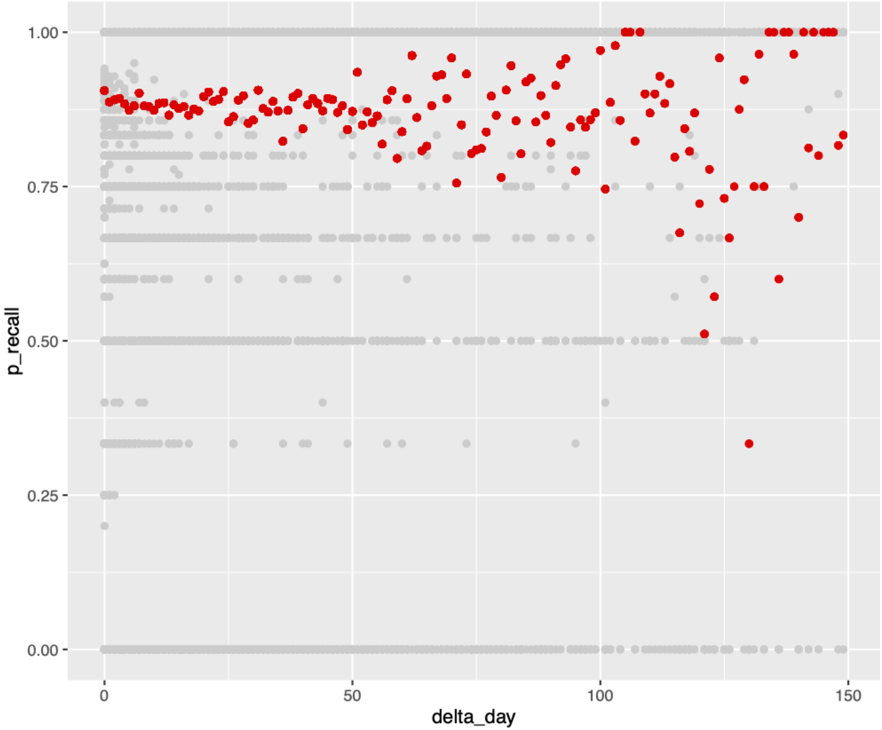


Our Data

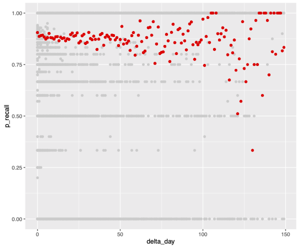
We created some **new variables**:

- **`delta_day`** = `delta / (24*3600)`
- **`word_length`** = length of the word
- **`p_history`** = `history_correct` / `history_seen`
- **`learning_language_tag`** **=** we manually grouped the word by `learning_language`; we came up with 3 groups; learning_language_tagis 0 if the user is learning French, 1 if the user is learning Italian or Portuguese and 2 otherwise.
- **`ui_language_tag`** **=** we manually grouped the word by `ui_language`; we came up with 2 groups: `ui_language_tagis` is 1 if the user is learning Italian
- **`avg_user_p`** = mean value of the `p_recall` scores of every user (if the user is new, we consider the mean value of the whole dataset)

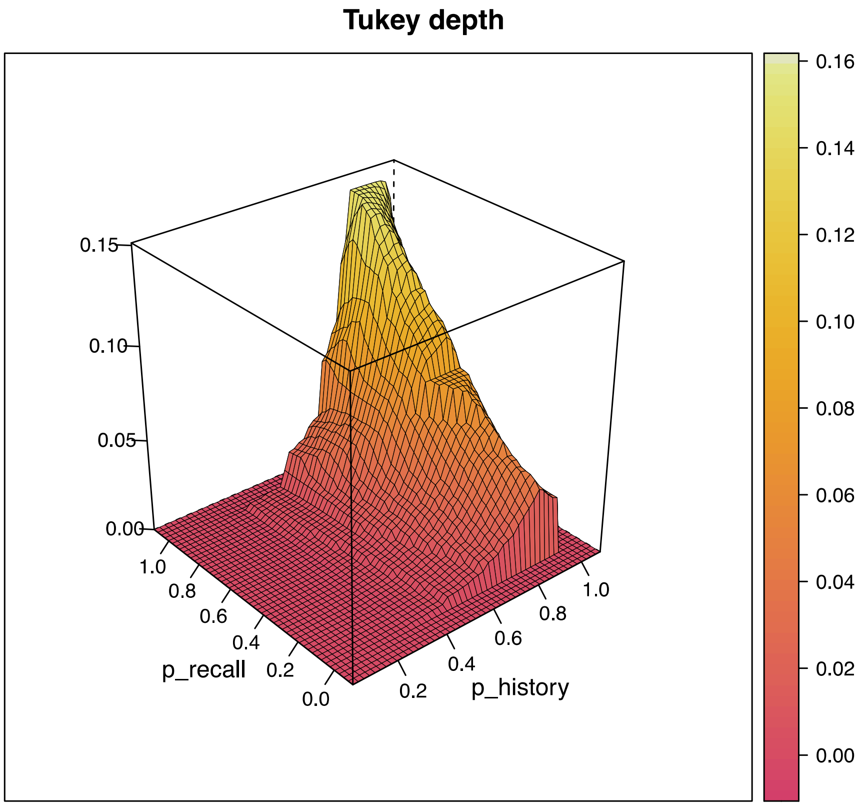
p_recall ~ delta_day



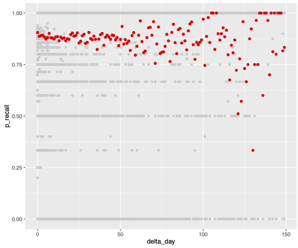
p_recall ~ delta_day



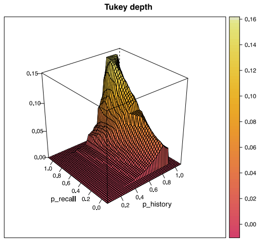
p_recall ~ p_history



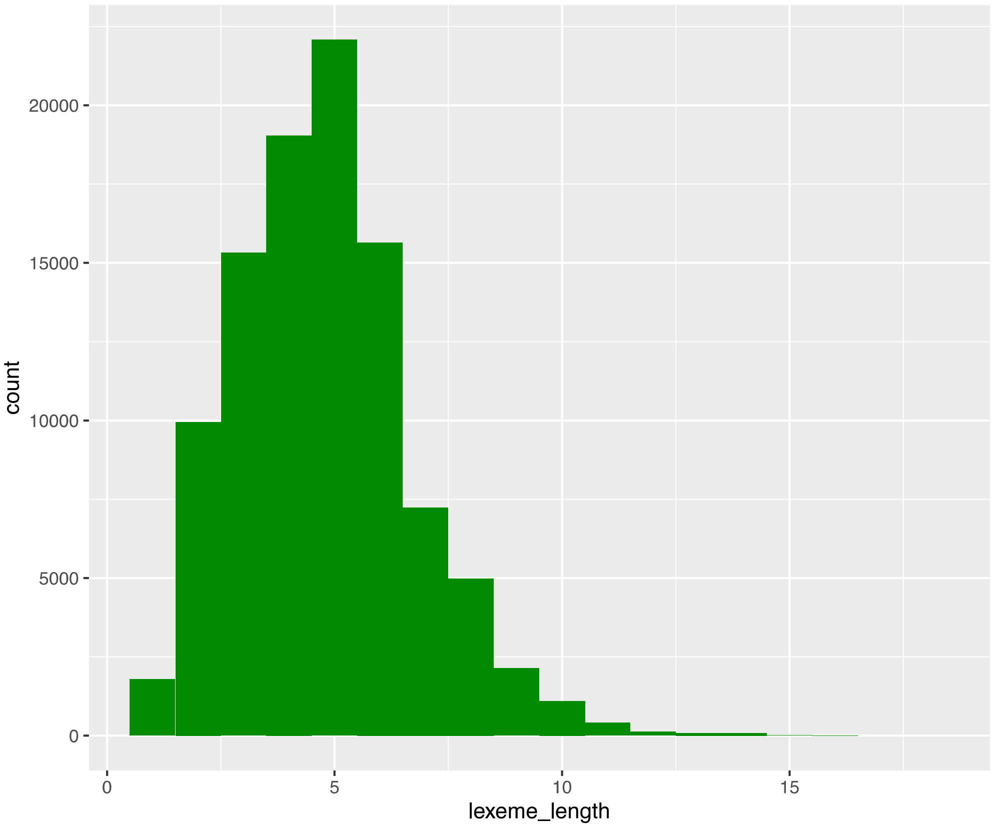
p_recall ~ delta_day



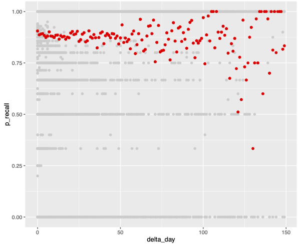
p_recall ~ p_history



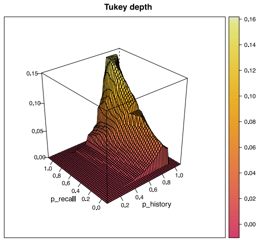
lexeme_length



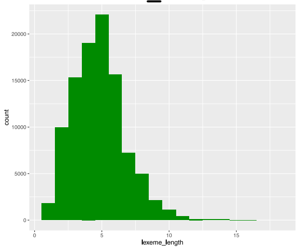
p_recall ~ delta_day



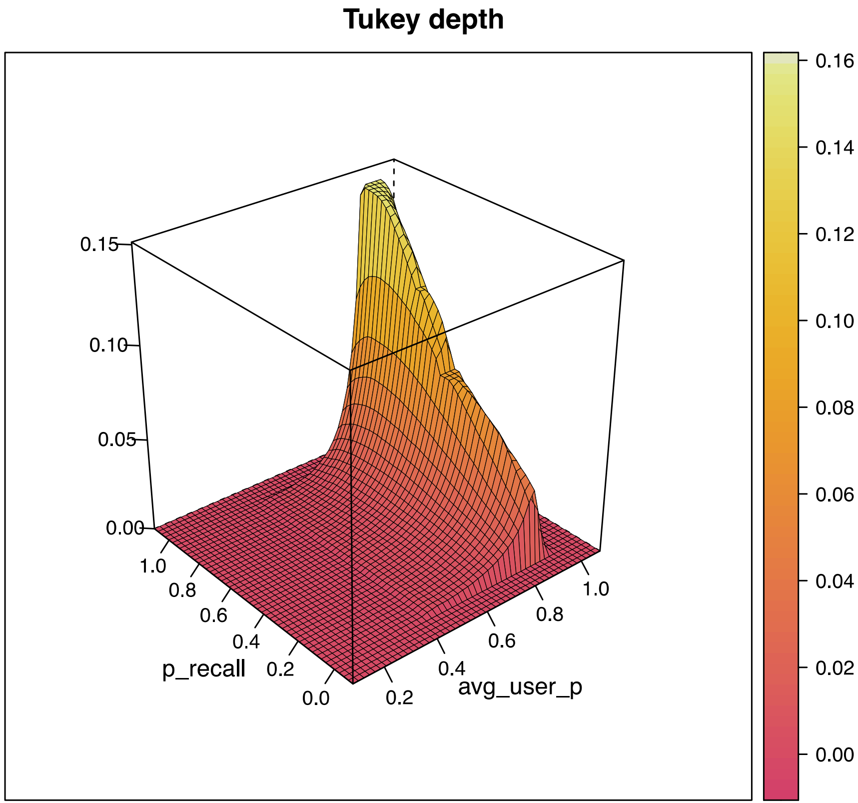
p_recall ~ p_history



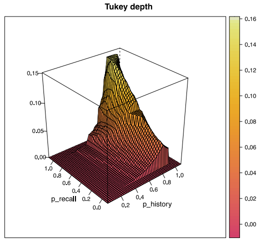
lexeme_length



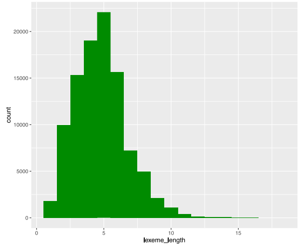
p_recall ~ avg_user_p



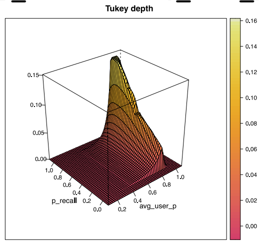
p_recall ~ p_history



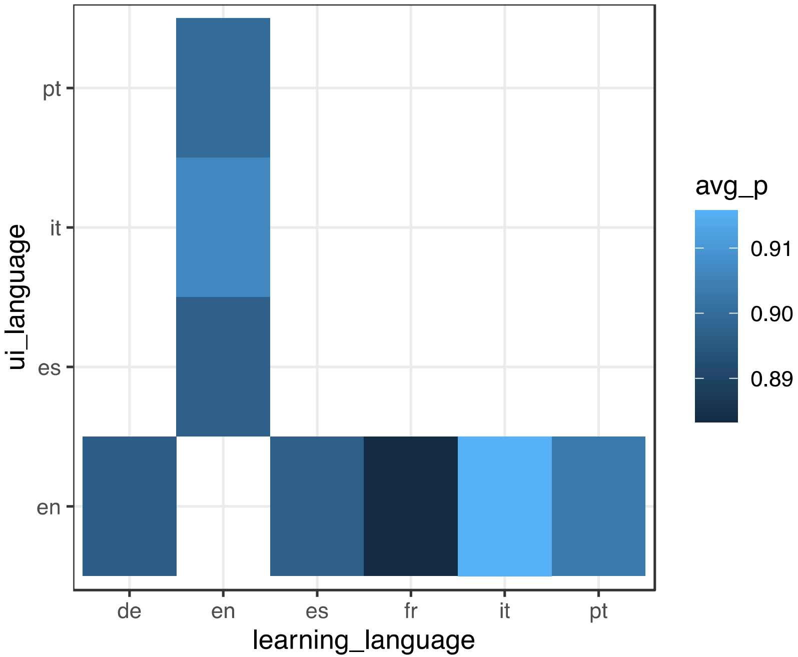
lexeme_length



p_recall ~ avg_user_p



heatmap



learning_language_tag

ui_language_tag

OUR GOALS

Initial goal

- Help Duolingo improve their user experience by making better decision about which word to test the user on

- Fit a model which makes better predictions about **`p_recall`**, in order to better prioritize between words

OUR MODELS

Model 1

p_recall

delta_day (d)

p_history (h)

word_length (w)

learning_language_tag (l_l)

ui_language_tag (l_{ui})

We resort to a **logistic GAM** with some penalized cubic regression spline terms:

$$\boxed{g(\mathbb{E}[p|\Delta = d, H = h, L_l = l_l, L_{ui} = l_{ui}, W = w]) = \beta_0 + \beta_1 l_l + \beta_2 l_{ui} + f_1(d) + f_2(h) + f_3(w)}$$

Model 2

p_recall

- ~~delta_day~~ (d)
- ~~p_history~~ (h)
- ~~word_length~~ (w)
- ~~learning_language_tag~~ (l_l)
- ~~ui_language_tag~~ (l_{ui})
- avg_user_p(u)

We resort to a **logistic regression model:**

$$\boxed{g(\mathbb{E}[P|U=u]) = \beta_0 + \beta_1 u}$$

- `avg_user_p` **seems to be the most important variable…**

- **… here we started questioning the validity of Duolingo's data…**

- **… are we sure that exists any time dependence ?**

???

GLM

- We reintroduce the dependence to **`delta_day`** to study its significance

`p_recall`

- **`delta_day (d)`**
- **`avg_user_p (u)`**

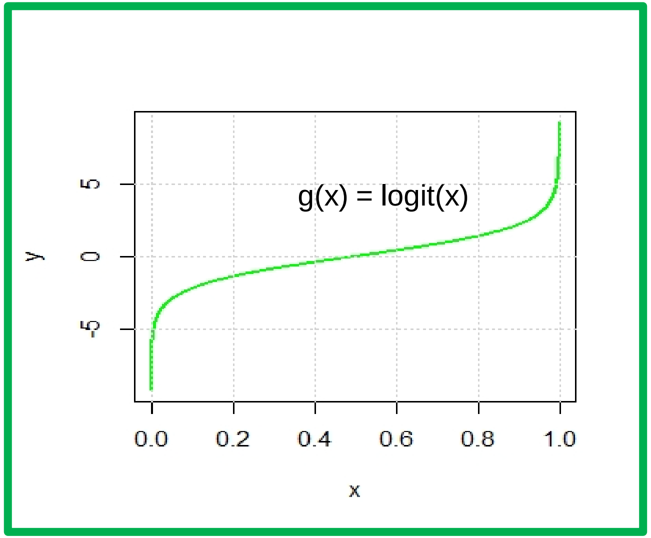
$$f(d) = \sum_{i=1}^{K} \beta_{1+i} b_i(d)$$

$$\boxed{g(\mathbb{E}[P|U = u, \Delta = d]) = \beta_0 + \beta_1 u + f(d)}$$

$$g(\mu) = log\left(\frac{\mu}{1-\mu}\right)$$

`p_recall` transformation

- `p_recall` ∈ [0,1]
- We want to do some permutation tests for the significance of **`delta_day`**…
- … but they are quite problematic in a logistic regression setting
- We decided to abandon this setting and go back to linear models, by **transforming our response:**



g(`p_recall`) ∈ (-∞,+∞)

LM-g (full)

- We are making inference on our transformed data

`p_recall`

`delta_day (d)`

`avg_user_p (u)`

$$\boxed{\mathbb{E}[g(P)|U = u, D = d] = \beta_0 + \beta_1 u + f(d)}$$

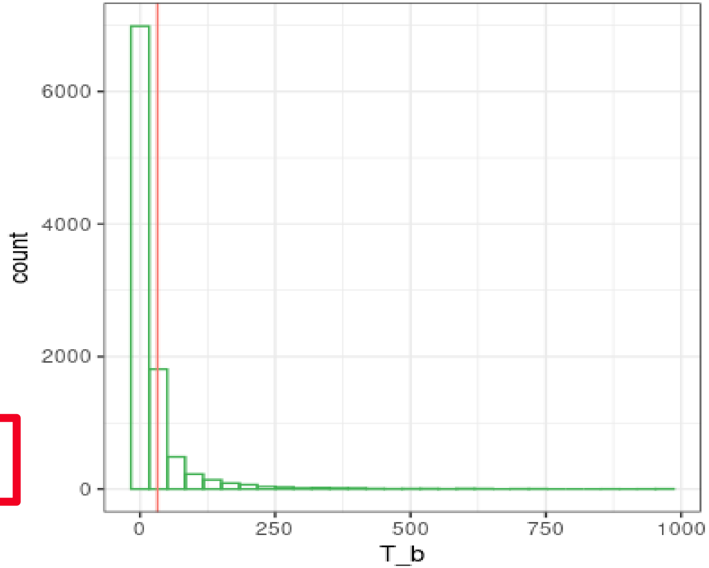
Permutation Test

Given our GAM model: $g(P)|U = u, D = d \sim \beta_0 + \beta_1 u + f(d) + \varepsilon$

we study the following test: $H_0 : f(d) = 0 \quad vs \quad H_1 : f(d) \neq 0$

adopting the *Freedman and Lane* scheme and choosing as **Test Statistic:**

$$T = \|f(d)\|_{L^2(d_1,d_2)}$$



p-value = 0.18

We can ignore delta_day!!!

LM-g (reduced)

- We ignore the time dependence:

`p_recall` — `avg_user_p (u)`

$$\boxed{g(P)|U = u \sim \beta_0 + \beta_1 u + \varepsilon}$$

- Permutation Test for the significance of avg_user_p:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$$

p-value = 0

RESULTS

Model	MAE↓	AUC↑
HLR	0.128	0.538
GLM	0.146	**0.626**
LM-g full	0.109	0.602
LM-g reduced	0.109	0.599
constant 1	**0.104**	n/a
constant $\overline{p} = 0.859$	0.175	n/a

- No evidence of any relationship between **`p_recall`** and **`delta_day`** (i.e. between the probability of recalling a word and the lag time of the word review)

- According to our analysis and our statistical (non-parametric) tests it seems like **the only important feature** for the inference of **`p_recall`** is the users' previous performances (**`avg_user_p`**), i.e. their ability to learn new words.

- A **constant *p* = *1* model** reaches a smaller MAE

THANK YOU !!!

POLITECNICO
MILANO 1863

References

- David Freedman and David Lane. “A nonstochastic interpretation of reported significancelevels”. In:Journal of Business & Economic Statistics1.4 (1983), pp. 292–298.
- John Gerring and Daniel Pemstein. “A Political Science Peer Review and Publication Con-sortium”. In:PS: Political Science and Politics(2020)
- Douglas M Potter. “A permutation test for inference in logistic regression with small-andmoderate-sized data sets”. In:Statistics in medicine24.5 (2005), pp. 693–708
- B. Settles and B. Meeder. “A Trainable Spaced Repetition Model for Language Learn-ing”. In:Proceedings of the Association for Computational Linguistics (ACL). ACL, 2016,pp. 1848–1858.doi:10.18653/v1/P16-1174.url:http://www.aclweb.org/anthology/P16-1174
- Burr Settles.Replication Data for: A Trainable Spaced Repetition Model for LanguageLearning. Version V1. 2017.doi:10.7910/DVN/N8XJME.url:https://doi.org/10.7910/DVN/N8XJME