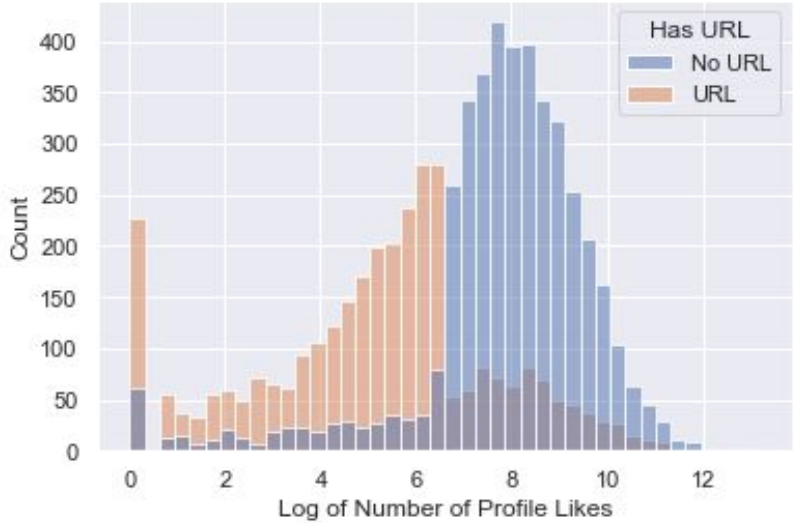
Social Media Prediction Kaggle Competition - L'Équipe

Martin Dallaire - 923013
Benjamin Sigman - 1060990
François Paugam - 20169017
François David - 20171906

IFT-6758 Data Science

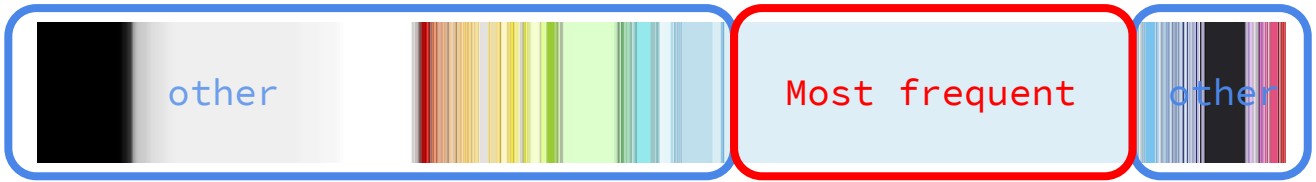
Data cleaning/feature engineering

- Uniformisation (lower case, identify Nan)
- Change URL to “Has Url” column
- Median imputation for Nan



Data cleaning/feature engineering

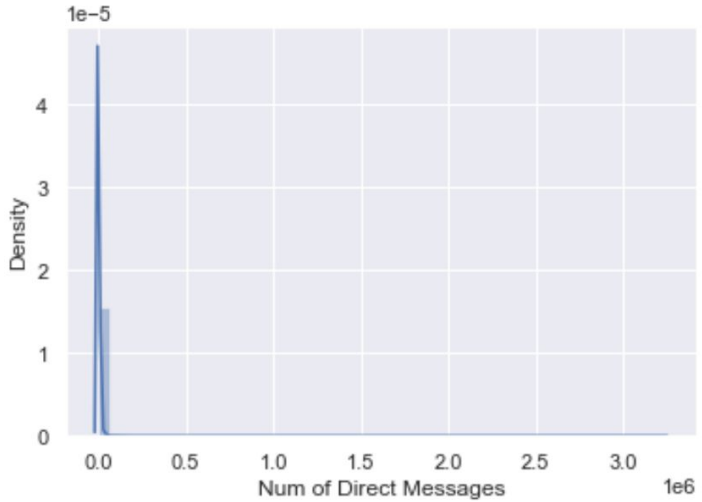
- Keep most frequent categories (that were correlated) and put remaining in “other”

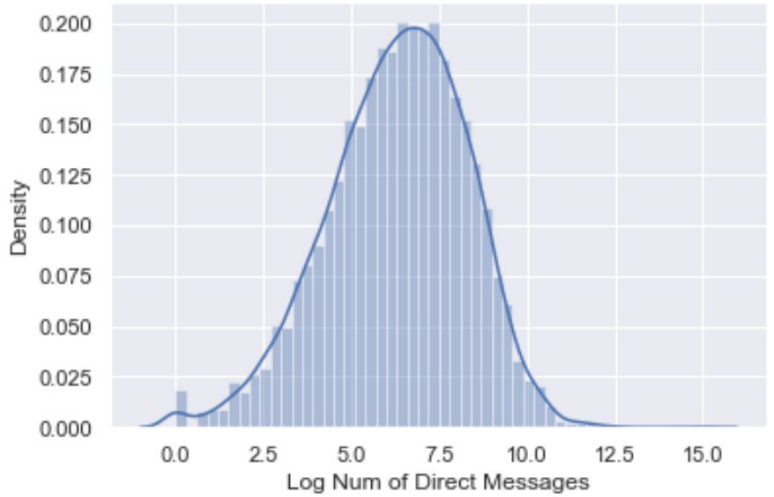


Profile Page color

Data cleaning/Feature engineering

- Log transform of numerical features





Data cleaning/Feature engineering

- Estimated clicks since creation
- Bag-of-words for Location
- Target mean for categorical data
- Failed creative methods: Sin/Cos on UTC

Keep the features that are correlated to the target

Model Selection (First steps)

- Some(log of)features decently linearly correlated with **log of target values (r>0.3)**
  - Linear models (Linear regression, SVM)
- Many categorical features: Try decision tree-based models.
- Ensemble models involving decision trees: Random Forests

Model Selection (Ensembling)

-Try more tree-based ensembling models:

  -LightGBM, great right out of the box. Scores 1.71 by itself on average.

  -XGBoost, needed fine tuning. Performed poorly out of the box. After hyperparameter tuning, it performed almost as well as LightGBM.

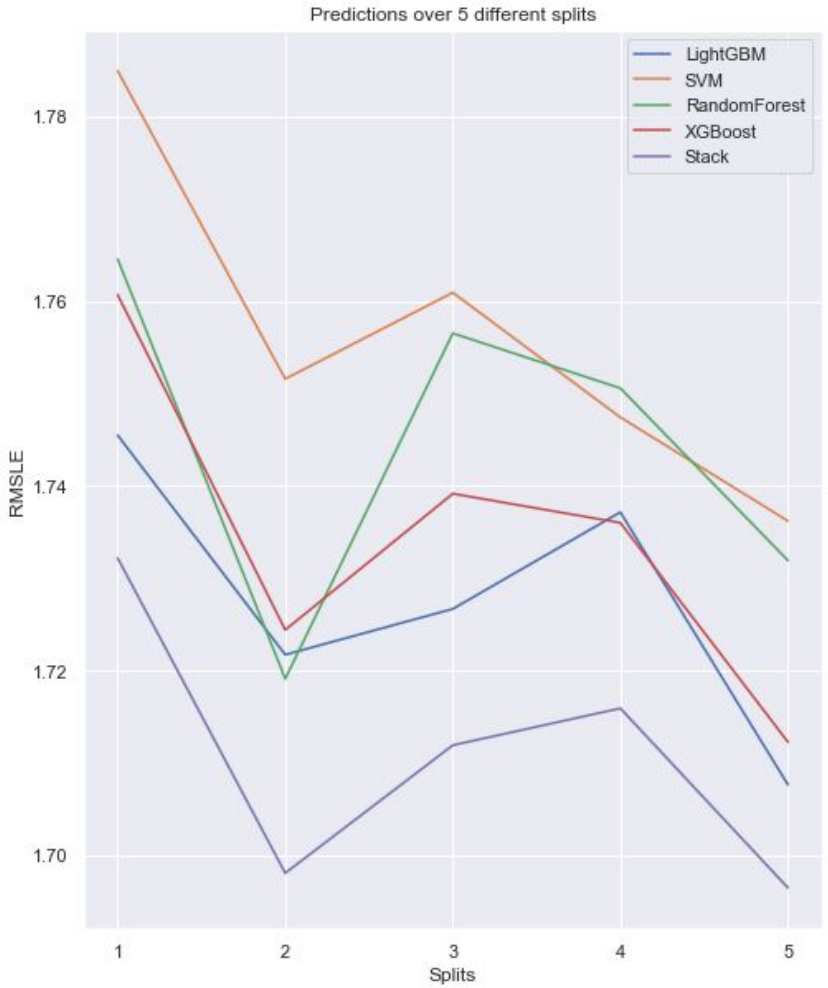
  -Many models performing between 1.71 and 1.85.
  Conclusion: Stacking regressor!

Model performance

Similar performances.

LightGBM performs better of al solo models (on average)

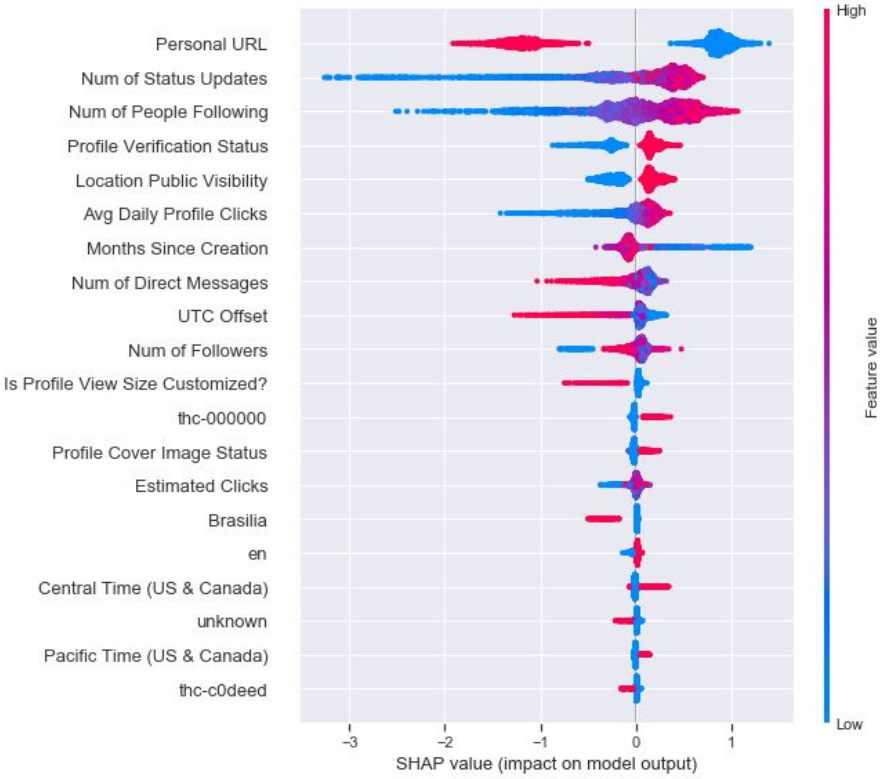
Stack performs best almost every time.



Validation

- 50-fold cross validation
- Hyperparameters tuning (gridSearchCv)
- Reduce the bias because more data points are considered (compared to val-split)
- Reduces Variances as more data points are used in the validation set.
- Reduces overfitting

Shapley Values - Interpretability/Feature Selection



- On the log transformation of the Number of Profile Likes.
- Red means upper range of feature value, blue means the lower range.
- Points on the right contributes to predicting higher values and the ones on the left contributes to predicting lower values.

Thanks!


Uppsala universitet 1852–1916
Det gamla universitetet. Festskrift till Anders Hallberg
Studenten. Staden och Samingen
UPPSALA UNIVERSITE