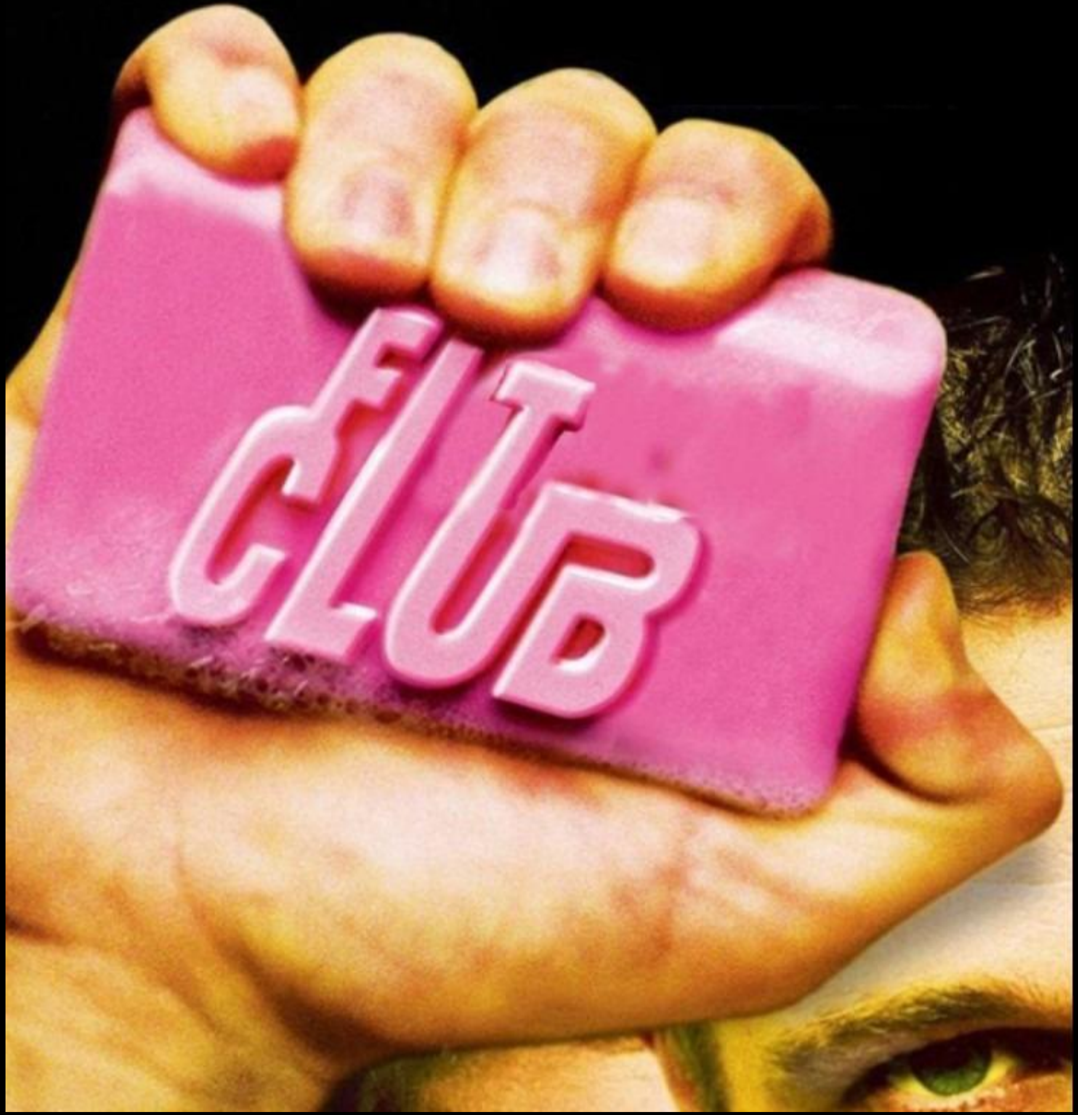


Phishing

Kaggle competition

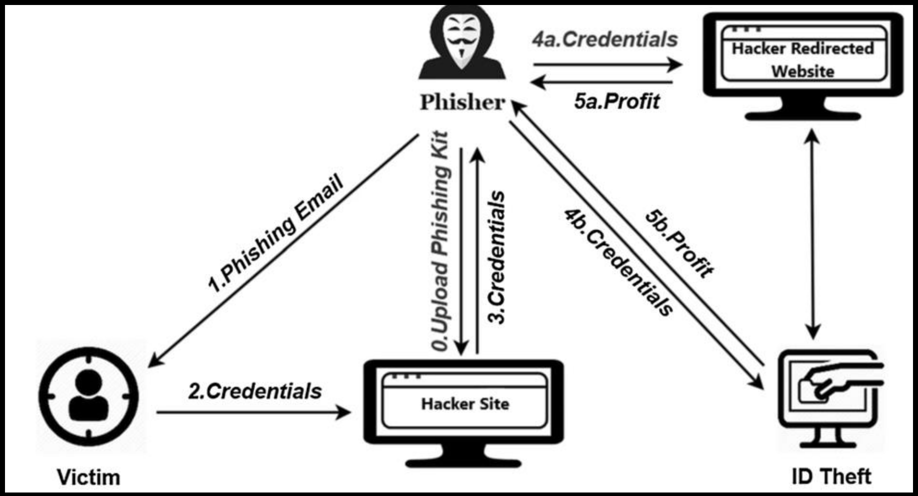
FIT CLUB

Ydata Jan 2021

THE PROBLEM

PHISHING ATTACKS HAVE BECOME ONE OF THE MOST PROMINENT ATTACKS FACED BY INTERNET USERS, GOVERNMENTS, AND SERVICE-PROVIDING ORGANIZATIONS. IN A

PHISHING ATTACKS ARE **AIMED AT COLLECTING SENSITIVE DATA** (LOGIN DETAILS, CREDIT CARD NUMBERS ETC.) BY USING SPOOFED EMAILS OR FAKE WEBSITES.



Phishing attack diagram [Forecast. (2017). Global fraud and cybercrime forecast.]

DATA OVERVIEW

- Phishing Emails in tabular form
- Long text – Email subject and content
- Short text – encoding
- Numerical – indexes
- Lables – binary {0,1}
- Imbalanced data – 90%/10%

NLP Challenge

EDA

- BASIC FIDELITY –

  ALL SAMPLES HAVE LABELS

- DUPLICATES -

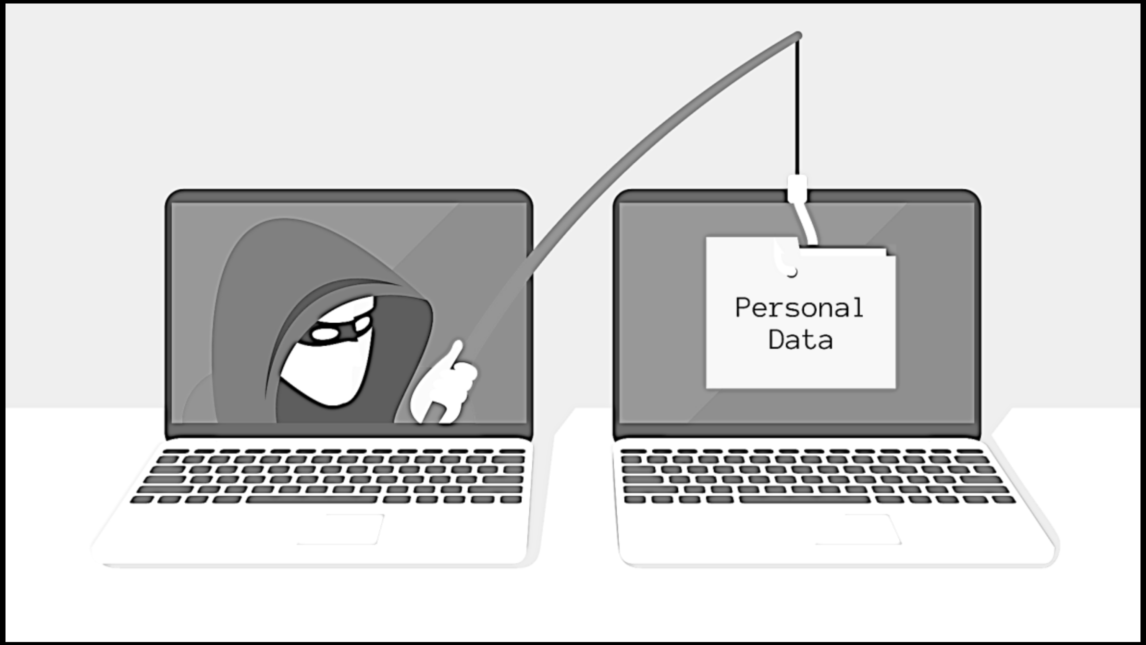
  DID NOT HANDLE DUPLICATES SINCE TEST DATA CONTAINS THEM AS WELL

- NAN HANDLING –

  NANS REPLACED BY EMPTY STRINGS

- SEARCHING FOR COMMON WORDS BY CLASS –

  PHISHING EMAILS' SUBJECT ARE CHARACTERIZED BY SPECIFIC WORDS THAT ARE RELATED TO THE PHISHING ACTION (E.G. ACCOUNT, BANK, PAYPAL, EBAY, UPDATE, PLEASE).



Fun fact – sorting labels by provided index divides the vector, first 2802 values are "1" and the rest are all 0. Using this information and manually attributing the labels gives a score of 0.96

REFERENCE ARTICLE

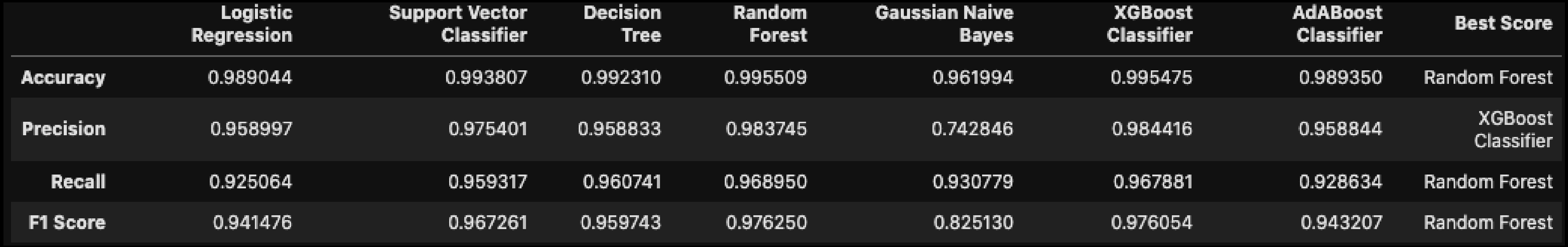
- BEFORE STARTING THE ANALYSIS WE CONDUCTED A LITERATURE SCAN SEARCHING FOR RELEVANT PAPERS THAT DEALT WIT SIMILAR CHALLENGES.
- WE WERE MOST EFFECTED BY THIS PAPER: AKINYELU, A. A., & ADEWUMI, A. O. (2014). CLASSIFICATION OF PHISHING EMAIL USING RANDOM FOREST MACHINE LEARNING TECHNIQUE. JOURNAL OF APPLIED MATHEMATICS, 2014.HTTPS://WWW.HINDAWI.COM/JOURNALS/JAM/2014/425731/.

  WE LEARNED THAT LINKS PRESENCE AND CHARACTERISTICS ARE KEY FEATURES FOR PHISHING EMAILS.
- IN THE NEXT SLIDE WE'LL SPECIFY THE FEATURES WE EXTRACTED. FEATURES THAT WERE CHOSEN BASED ON THIS ARTICLE WILL BE MARKED WITH # SIGN.

PREPROCESSING AND FEATURE EXTRACTION

- CONTENT/SUBJECT MISSING (BOOL)
- DATA TYPE IS HTML (BOOL) #
- NUMBER OF UPPERCASE/LOWERCASE CHARS (NUM)
- CONTENT/SUBJECT LENGTH (NUM)
- CONTENT/SUBJECT IS IN ASCII (BOOL)
- NUMBER OF LINKS IN CONTENT/SUBJECT (NUM) #
- LINKS PRESENCE IN CONTENT/SUBJECT (BOOL)
- SUSPICIOUS WORDS IN THE LINKS TEXT (BOOL) #
- PHISHING RELATED WORDS IN SUBJECT (BOOL)
- PHISHING RELATED WORDS IN CONTENT (BOOL) #

QUICK'N'DIRTY MODEL



	Logistic Regression	Support Vector Classifier	Decision Tree	Random Forest	Gaussian Naive Bayes	XGBoost Classifier	AdABoost Classifier	Best Score
Accuracy	0.989044	0.993807	0.992310	0.995509	0.961994	0.995475	0.989350	Random Forest
Precision	0.958997	0.975401	0.958833	0.983745	0.742846	0.984416	0.958844	XGBoost Classifier
Recall	0.925064	0.959317	0.960741	0.968950	0.930779	0.967881	0.928634	Random Forest
F1 Score	0.941476	0.967261	0.959743	0.976250	0.825130	0.976054	0.943207	Random Forest

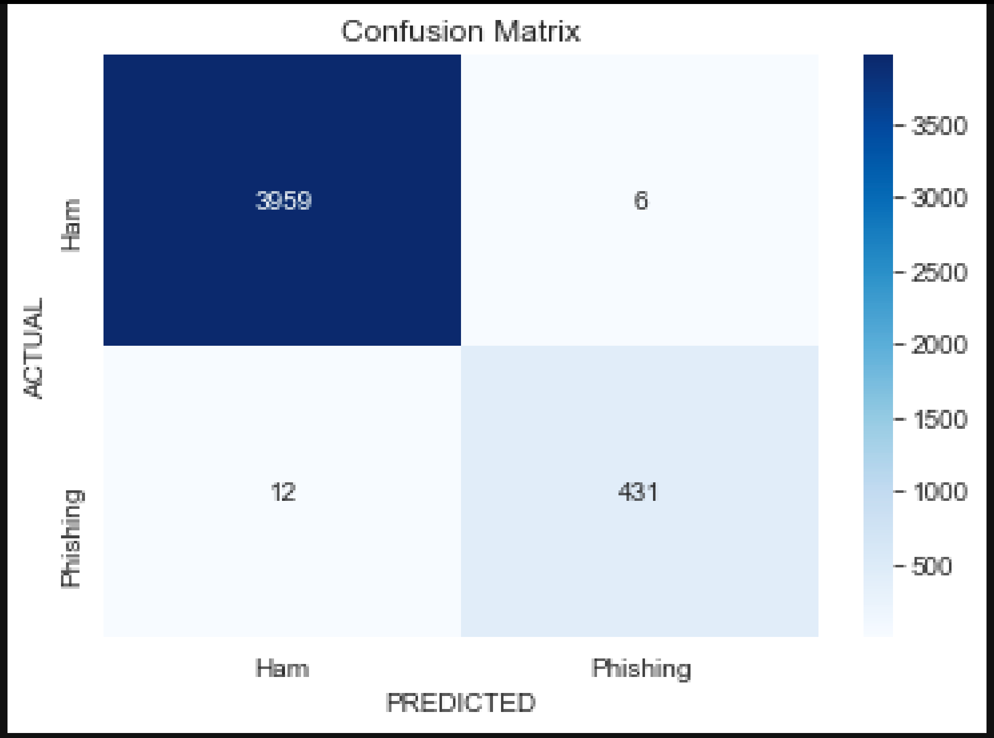
- Although the data is imbalanced, the models (besides NB) perform well, probably due to the domain specific feature extraction. Therefore we will not perform resampling or apply cost sensitive approach.

- Most models demonstrate close to perfect performance. Therefor it will be hard to optimize it.

- We'll persue with Random Forest which gives overall (slightly) better performance than the rest of the classifiers.

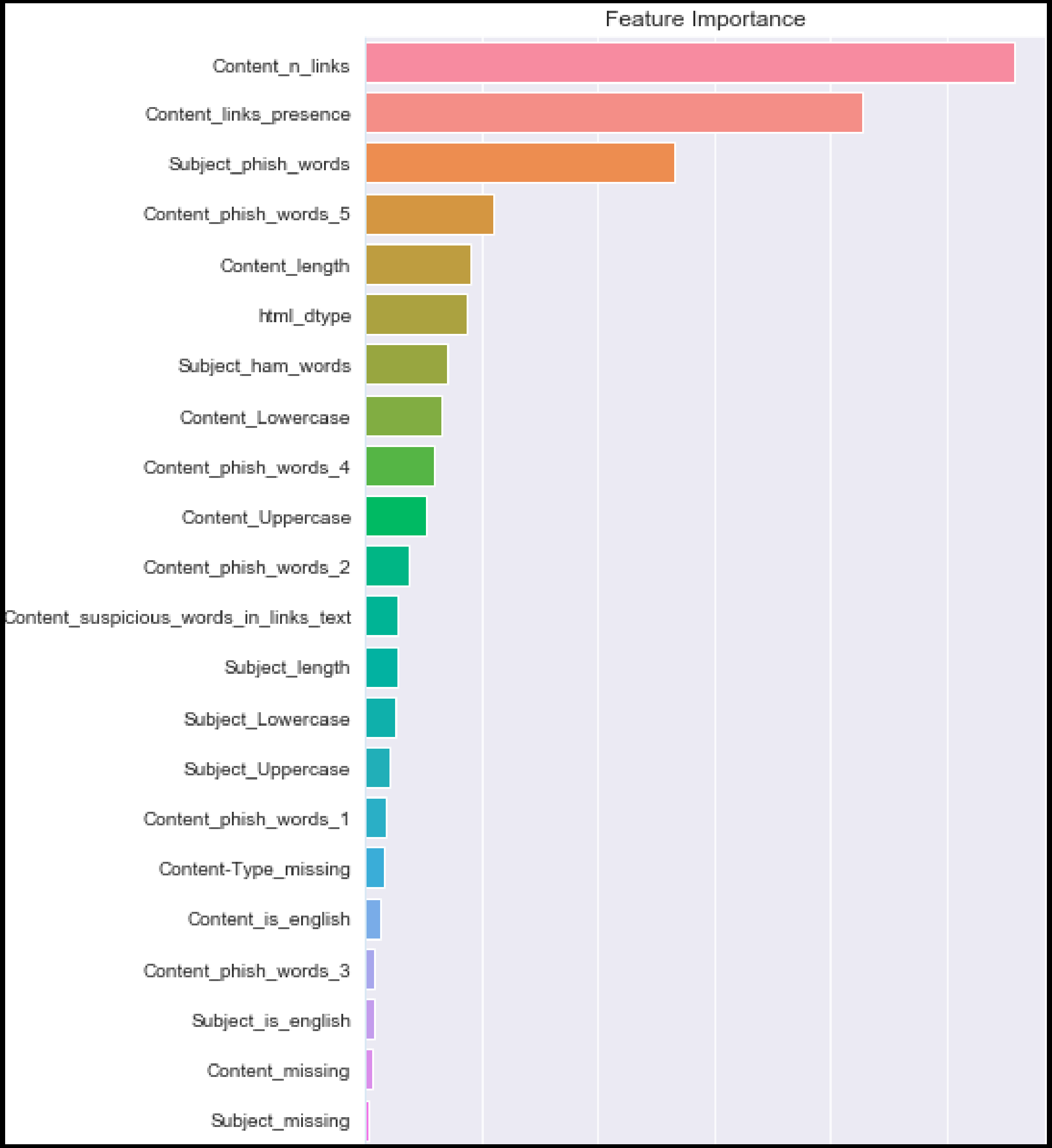
RESULTS FOR RANDOM FOREST WITH DEFAULT SETTINGS

- TRAIN-TEST-VAL SPLIT:
  - TRAIN 75%
  - VALIDATION 15%
  - TEST 10%
- MODEL TESTING TIME: 0.082 SECONDS
- VALIDATION F1 SCORE: 0.98



FEATURE SELECTION

- MOST IMPORTANT FEATURES ARE: NUMBER OF LINKS IN CONTENT, WHETHER THERE ARE LINKS, PRESENCE OF SUSPICIOUS PHISH WORDS, CONTENT LENGTH AND WHETHER THE TYPE OF EMAIL IS HTML.
- IMPORTANTLY, GIVEN THE KNOWN PROBLEMS WITH THE IMPURITY-BASED FEATURE IMPORTANCE , WE COMPUTED A PERMUTATION BASED APPROACH AND COMPARED TO OUR RESULTS. THE ORDER OF IMPORTANCE WAS PRESERVED (SEE NOTEBOOK).
- WE DECIDED TO DROP THE FEATURES THAT DON'T HELP THE CLASSIFIER.
- EVENTUALLY WE WERE LEFT WITH 22 FEATURES.

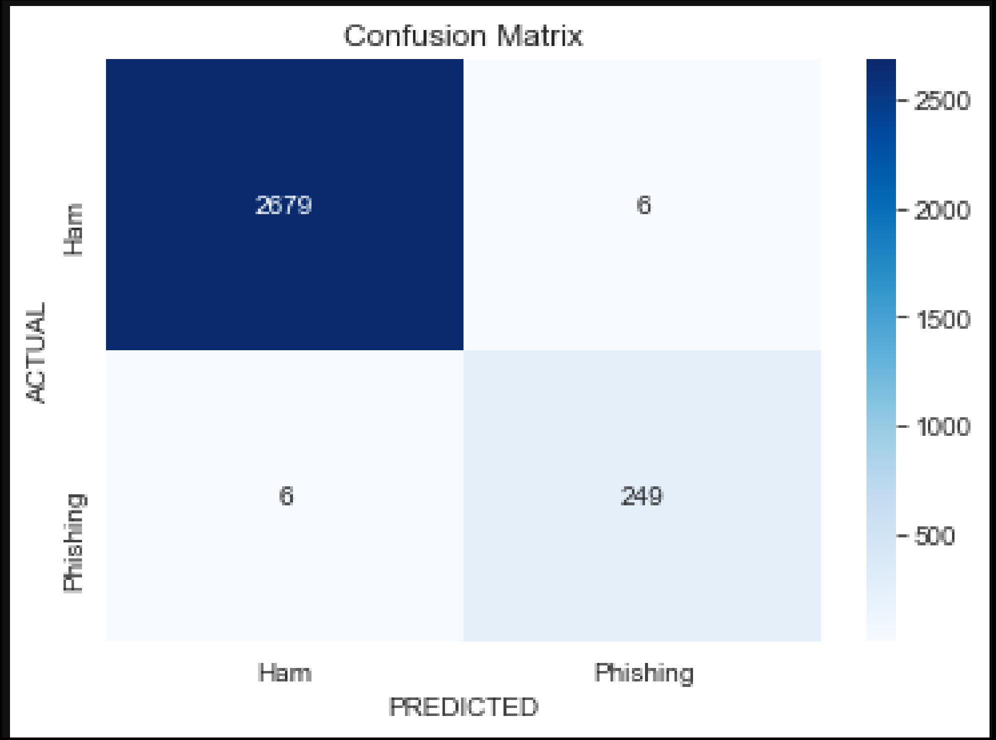


ERROR ANALYSIS

- We trained the model on the 22-features training set and predicted the labels for the validation set. We then inspected the samples that were misclassified.
- Example for a normal email subject that was misclassified as phishing:
  'Yahoo! Bill Pay - New E-Bill'
- Example for a phishing email subject that was misclassified as normal:
  'In Reference to Your February Account Summary
- The misclassified samples reassured that a more thorough feature extraction should be applied.

MODEL OPTIMIZATION WITH OPTUNA

- WE USED THE VALIDATION SET FOR OPTIMIZATION.
- THE HYPER-PARAMETERS WE OPTIMIZED WERE: MAX_DEPTH, N_ESTIMATORS, MAX_FEATURES, MIN_SAMPLES_SPLIT, MIN_SAMPLES_LEAF, MAX_SAMPLES.
- THAN WE USED THESE PARAMETERS TO TRAIN THE MODEL ON THE TRAINING SET AND PREDICTING FOR THE TEST SET. THESE ARE OUR FINAL RESULTS:



CONCLUSIONS

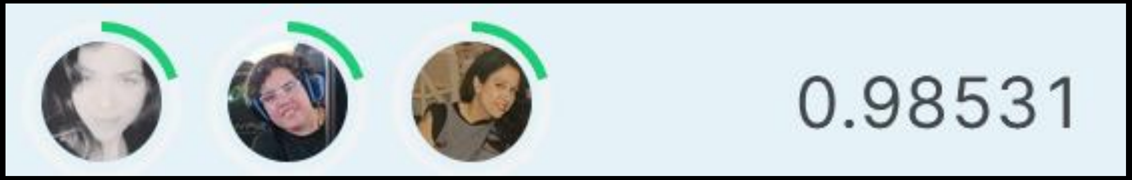
- THE DATASET FEATURES (E.G. MANY DUPLICATES) AND THE RELATIVELY EASY TO CHARACTERIZE DOMAIN FEATURES (I.E. SPECIFIC METHODS FOR PHISHING) MADE THE PERFORMANCE OF THE BASELINE MODELS VERY CLOSE TO OPTIMAL, THEREFOR LEAVING SMALL ROOM FOR IMPROVEMENT AND OPTIMIZATION.
- ALL MODELS GAVE COMPARABLE PERFORMANCE THEREFOR THERE IS NOT MUCH NEED TO UNDERSTAND WHY RANDOM FOREST GAVE THE BEST OF THEM. POSSIBLE REASON IT GAVE BETTER BASELINE PERFORMANCE THAN BOOSTING ALGORITHMS IS THAT ITS HYPER PARAMETERS DONT REQUIRE AS MUCH TUNING AS THOSE OF BOOSTING.
- MOST IMPORTANT FEATURES FOR PREDICTION WERE, AS SUSPECTED, RELATED TO THE CHARACTERISTICS OF PHISHING EMAILS - PRESENCE OF LINKS AND THEIR AMOUNT, HTML DATA TYPE OF CONTENT AND PHISHING RELATED WORDS IN THE EMAILS SUBJECT.

WAYS TO FURTHER IMPROVE PREDICTIONS

- perform error analysis more thoroughly and consider adding / changing features.
- Using more features we extracted. Thus far they seeded redundant as the used features gave such good results from the get go. We could try and incorporate them, assessing which contribute more and leave just them.
- Adding feature describing if email contains *spoof* URLs (current feature is just having URLs). Phishing websites are common entry points, copying the behavior of legitimate websites to harvest user details.
- Using domain knowledge in the form of known phishing terms.
- try estimating performance of tuned boosting algorithms.

THE RESULTS SUBMITTED TO KAGGLE

- THE PREDICTIONS THAT GAVE US THE HIGHEST SCORE WERE DERIVED WITH A DIFFERENT MODEL THAT IS NOT DESCRIBED IN THE MAIN NOTEBOOK/THIS PRESENTATION.
- THE MODEL:
  - PREPROCESSING: LOWERING, REMOVING NON-ALPHANUMERIC CHARACTERS, AND TOKENIZING
  - FEATURES: 300-LENGTH VECTOR GENERATED USING WORD2VEC WITH N-GRAMS ON A COLUMN COMBINING SUBJECT AND CONTENT
  - MODEL: ADABOOST WITH LEARNING RATE OF 0.1 AND 4000 ESTIMATORS AS STOPPING CONDITION
  - PARAMETER OPTIMIZATION – MANUAL.
- WE CHOSE TO PRESENT THE OTHER NOTEBOOK SINCE THE FEATURES WERE MORE MEANINGFUL AND THEREFOR INTERPRETABLE.
- FOR CODE SEE NOTEBOOK: "FITCLUB_ADABOOST_FOR_KAGGLE"



0.98531