See on GitHub!

FAKE ACCOUNTS PREDICTOR

Machine learning capstone project
Diana Golman, PhD
2022

Future Technology

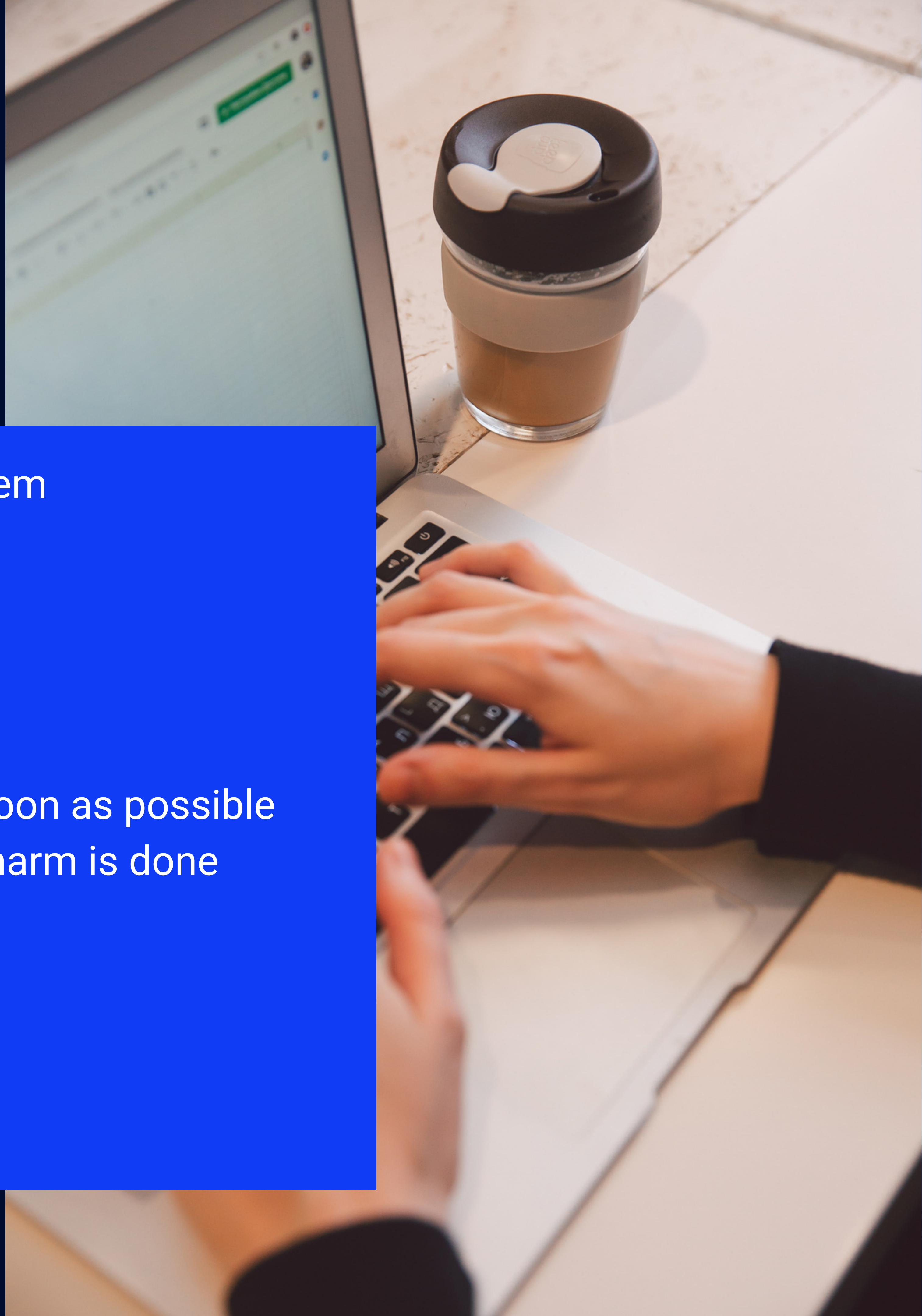
ARE FAKE ACCOUNTS A PROBLEM?

Catfishing and fake accounts are a problem for media companies as they lose users' trust and loyalty. This is especially relevant to dating apps as people plan to meet in person

Safety problem

Users loyalty

Identify as soon as possible before any harm is done



FAKE ACCOUNTS PREDICTOR

PROJECT'S GOAL

Create a model that will predict whether an account is fake or not.

Fake accounts will be identified as soon as possible

METHOD

Dataset

A publicly available dataset (on Kaggle) was used. It has 17 features and 65325, user accounts identified as fake or not fake. https://www.kaggle.com/krpurba/fakeauthentic-user-instagram?select=user_fake_authentic_4class.csv

METHOD

Data wrangling

PROBLEMS WITH THE DATA	SOLUTIONS
UNINFORMATIVE AND INCOMPLETE DATA	THREE FEATURES THAT WERE DROPPED
MISSING VALUES IN FORM OF "-1"	THESE ROWS DROPPED
DUPLICATED ROWS	THESE ROWS DROPPED
FEATURES WITH A SKEWED DISTRIBUTION	TRIM THE DATA AT 99%
OUTLIERS	CHOSEN OVER INTERQUARTILE RANGE

Features used in Modeling

- pic: Picture availability | Value 0 if the user has no profile picture, or 1 if has
- link: Link availability | Value 0 if the user has no external URL, or 1 if has
- cap_zero_per: Percentage of captions that has almost zero (<=3) length
- no_image_per: Percentage of non-image media i.e. image, video, and carousel.
- loc_tag: Percentage of posts tagged with location
- class: 2-class User classes: r (real/authentic user), f (fake user)
- posts_a: Number of total posts that the user has ever posted-trimmed
- flw_a: Number of followers-trimmed
- flg_a: Number of following-trimmed
- likes_a: Engagement rate (ER) is commonly defined as (num likes) divide by (num media) divided by (num followers)- trimmed
- hash_a: Average number of hashtags used in a post- trimmed
- cap_avg_a: The average number of characters of captions in media- trimmed
- comment_r_a: Similar to ER like, but it is for comments- trimmed
- post_interval_a: Number of total posts that the user has ever posted- trimmed.

EDA
Exploratory Data Analysis

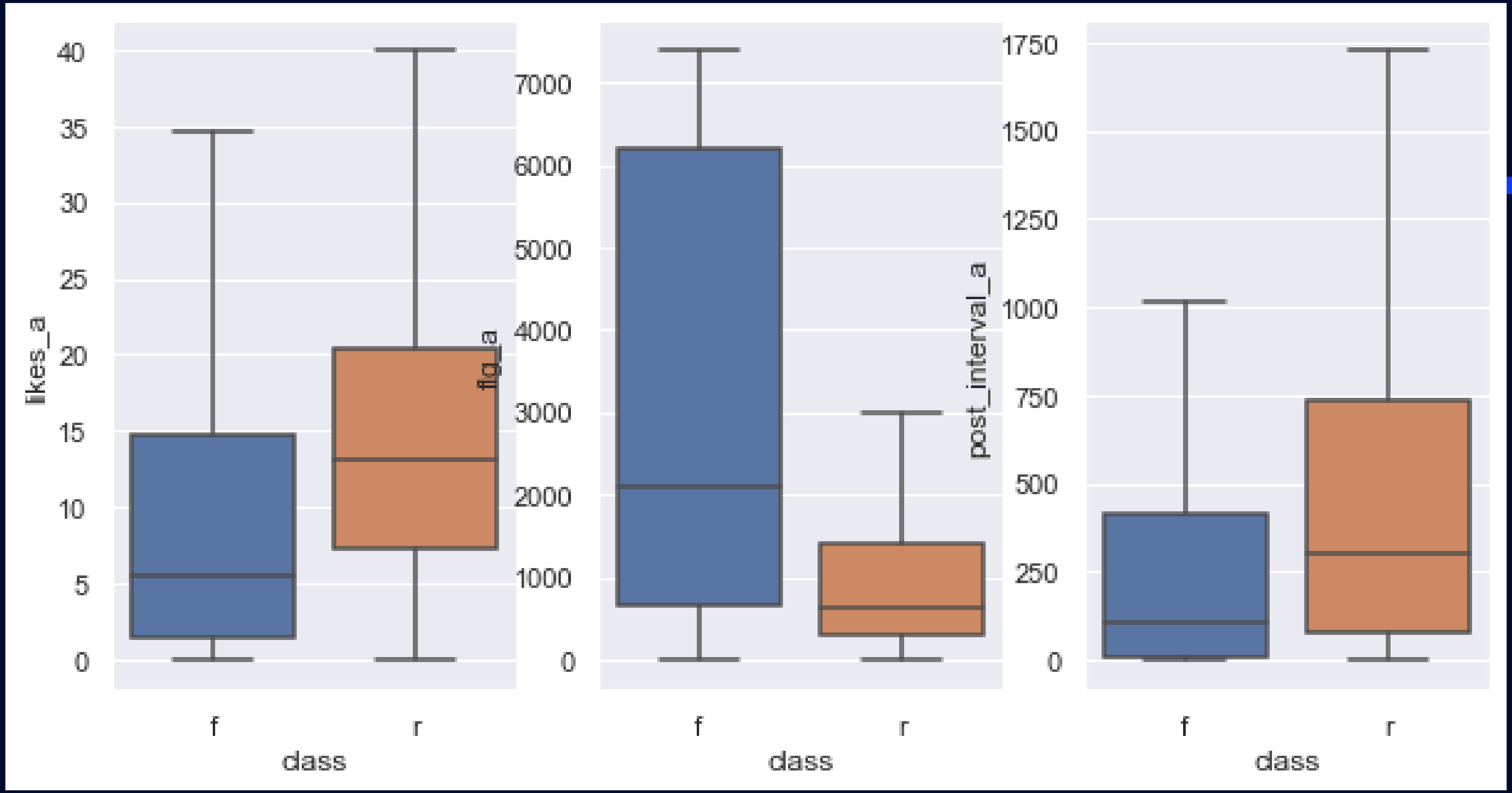
distributions, after trimming

correlations between all features

features are significantly different from each other in the real compared to fake accounts: t-tests

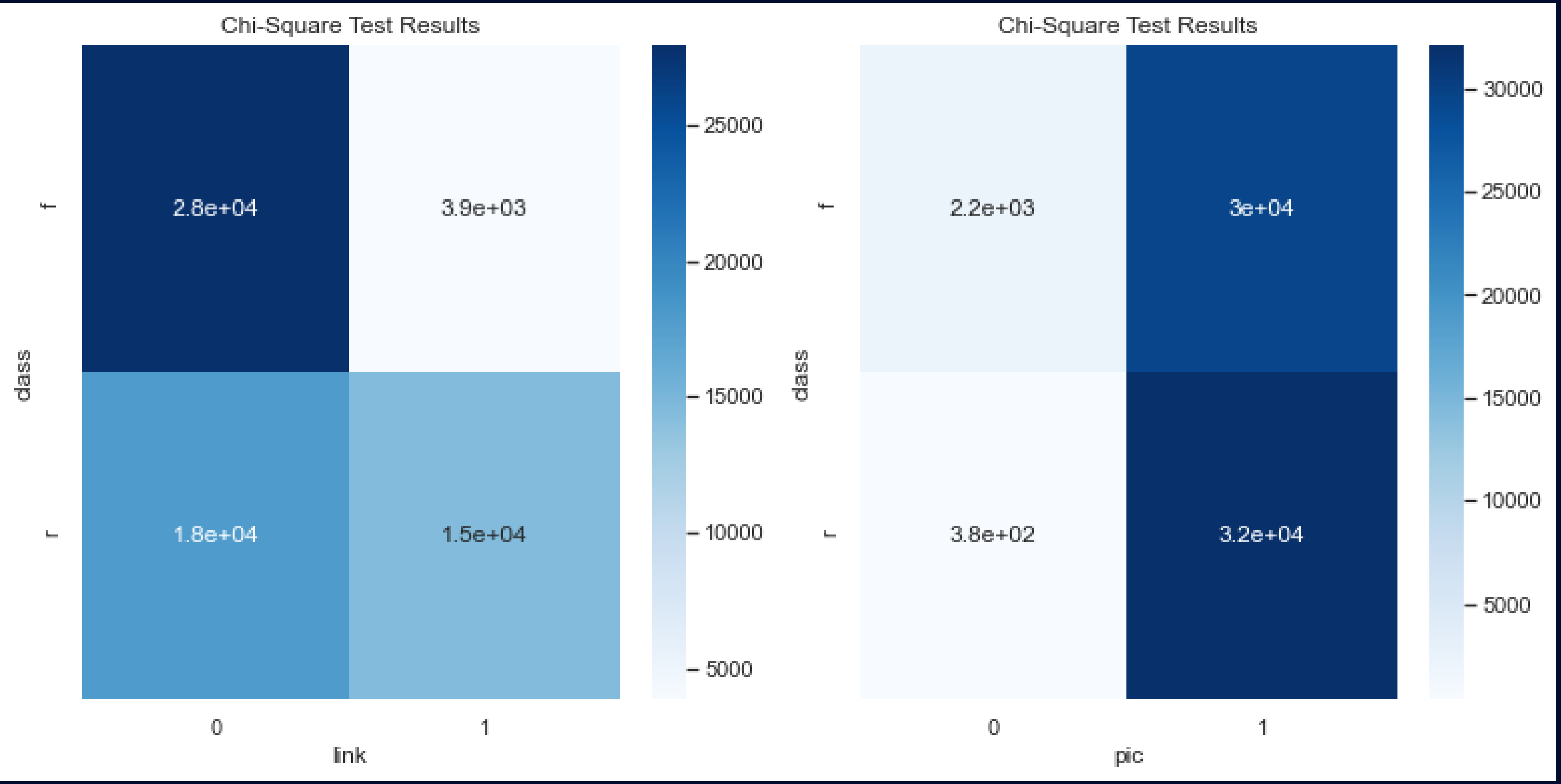
chi-square nonparametric test

LIKES, FOLLOWING OTHERS, AND INTERVALS BETWEEN POSTINGS



All features were found significantly different between real and fake accounts

LINKS TO OTHER WESITES, AND A PROFILE PICTURE



chi square nonparametric test

All features were found significantly different between real and fake accounts

PREPROCESSING

Since the features were on different scales, all data was scaled before modeling.

The categorical variables were dummy encoded.
The target classes were: fake-1 and real-0

MODELS

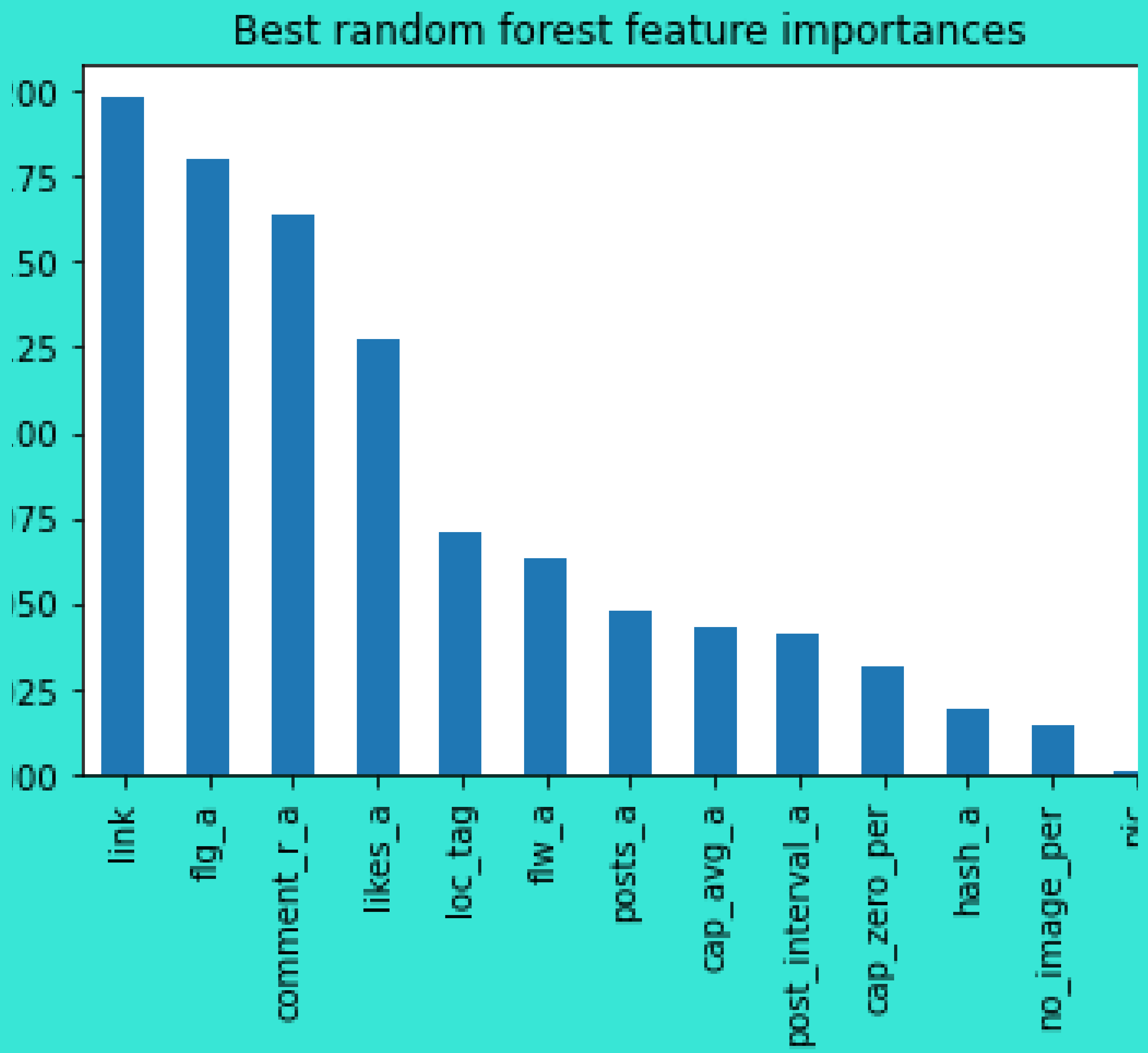
Several classifier models were examined , starting with simpler options because of the small number of features

Model	precision	recall	f1-score	ROC-AUC	Crossvalidated Train score mean	Crossvalidated Test score mean
Logistic regression(best c)	0.81	0.81	0.81	0.8091	0.8751	0.8750
Ridge Classifier	0.81	0.81	0.81	0.8069	0.87419	0.87414
Support Vector Classifier: with parameters tuning	0.83	0.84	0.84	0.8340	0.8336	0.8340

MODELS

More robust models were examined with parameters tuning and threshold adjustment

Model	precision	recall	f1-score	ROC-AUC	Crossvalidated Train score mean	Crossvalidated Test score mean
Random Forest with parameters tuning	0.96	0.82	0.88	0.8921	0.9545	0.9522
Random Forest with threshold adjustment	0.91	0.87	0.89	0.8911	0.9544	0.9522
HistGradientBoostingClassifier:with parameters tuning	0.94	0.83	0.89	0.8921		
HistGradientBoostingClassifier:with threshold adjustment	0.91	0.87	0.89	0.8907	0.9579	0.9564



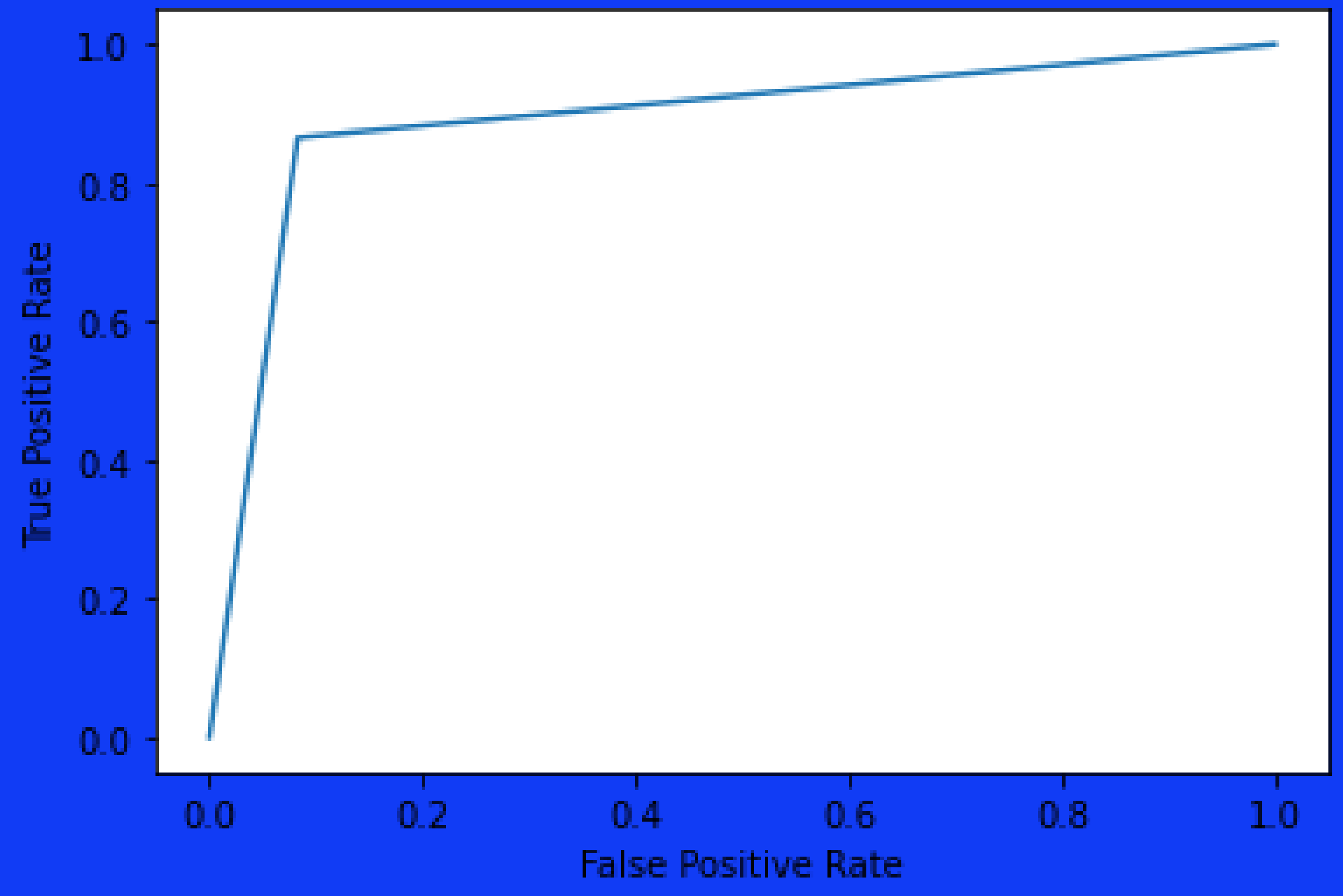
RANDOM FOREST CLASSIFIER

A closer look

Mean= **0.9544**

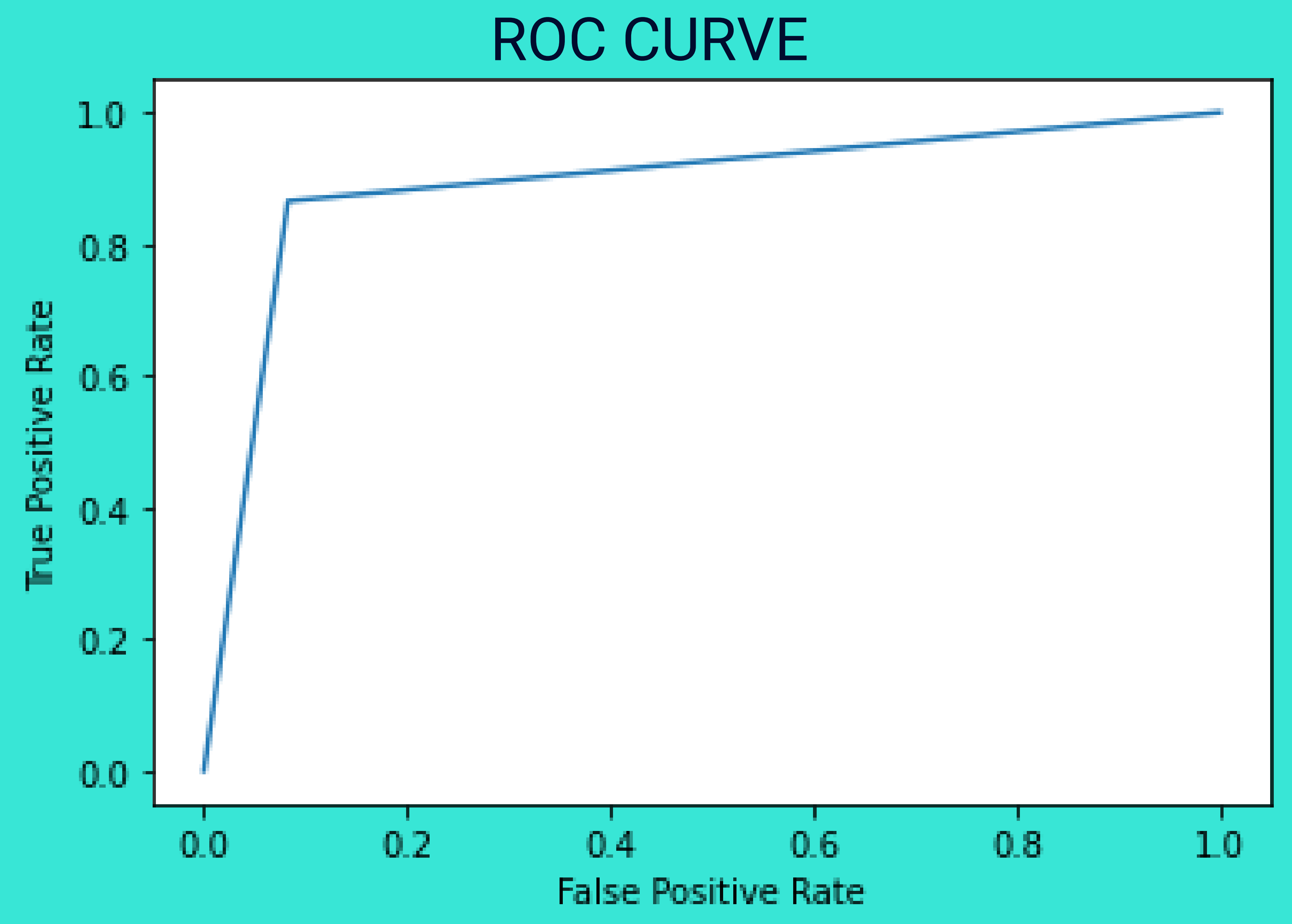
Standard deviation = **0.0058**

ROC CURVE



MODEL SELECTED

HistGradientBoosting Classifier with threshold adjustment was selected



Mean= **0.9564**

Standard deviation = **0.0039**

Conclusion

The HistGradient Boosting Classifier from the experimental module was faster than standard Boosting models and had very similar results to Random Forest. This model also had a lower cross-validated standard deviation than the random Forest(**Hist 0.0039** vs **RF 0.0058)**, meaning we have more stability in the predicted mean of our model. Since the models are similar the more stable model was chosen.

Using this model in a social network, social media, dating app security analysis can help predict fake accounts and use some protective measures before any harm is done.


www.reallygreatsite.com
THANK YOU
Future Technology