Classify URL Reputation

Randy Grant (DSML)

Introduction

Issue: The client, CyberCents, needs a way to able to identify URLs that have a likelihood of not being benign prior to a DNS request being sent to a DNS sinkhole

Solution: With a list of URLs that have a known reputation of benign or not benign, train a classification model to identify URLs as being "benign" or "not benign" to be deployed as a preliminary check of URLs before being sent to the DNS server

Goal: Create a model that correctly identifies URLs more than 90% of the time

Methods

Data Used:

- Kaggle.com sample datasets

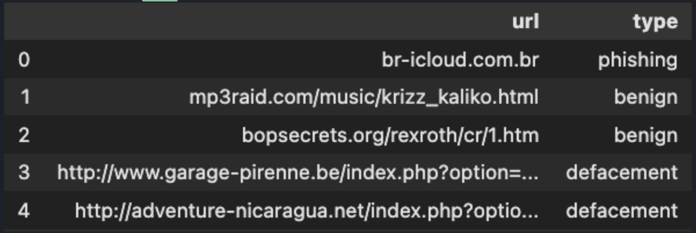
Tools Used:

- Python for model development
- Excel

Metrics Created:

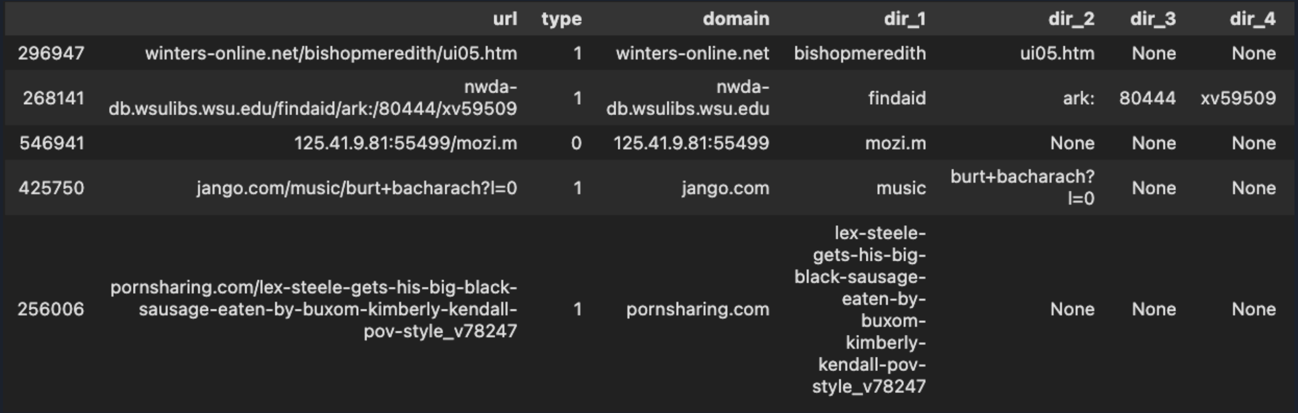
- Scores of Baseline Models
- Scores of Best Hyperparameter Tuned Models
- Rate of False and True Positives
- Final Ensemble Model Score and Confusion Matrix

Sample Data



	url	type
0	br-icloud.com.br	phishing
1	mp3raid.com/music/krizz_kaliko.html	benign
2	bopsecrets.org/rexroth/cr/1.htm	benign
3	http://www.garage-pirenne.be/index.php?option=...	defacement
4	http://adventure-nicaragua.net/index.php?optio...	defacement

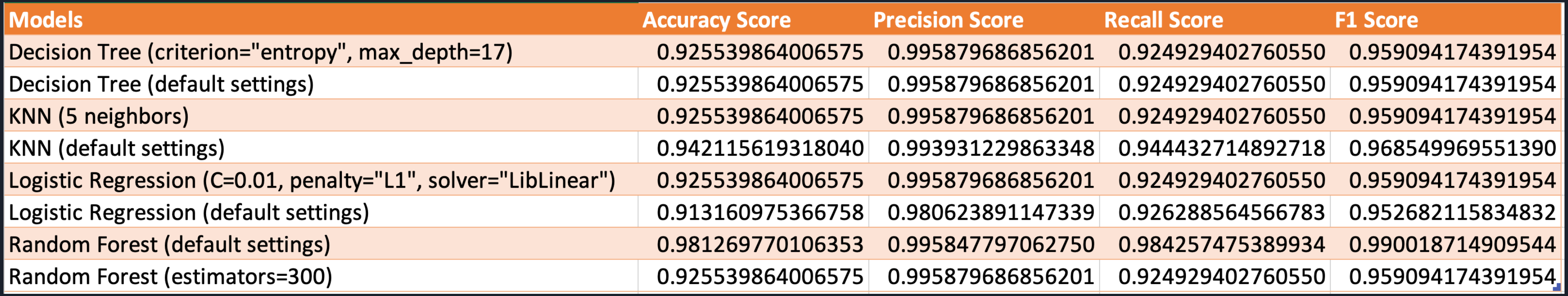
Raw



	url	type	domain	dir_1	dir_2	dir_3	dir_4
296947	winters-online.net/bishopmeredith/ui05.htm	1	winters-online.net	bishopmeredith	ui05.htm	None	None
268141	nwda-db.wsulibs.wsu.edu/findaid/ark:/80444/xv59509	1	nwda-db.wsulibs.wsu.edu	findaid	ark:	80444	xv59509
546941	125.41.9.81:55499/mozi.m	0	125.41.9.81:55499	mozi.m	None	None	None
425750	jango.com/music/burt+bacharach?l=0	1	jango.com	music	burt+bacharach?l=0	None	None
256006	pornsharing.com/lex-steele-gets-his-big-black-sausage-eaten-by-buxom-kimberly-kendall-pov-style_v78247	1	pornsharing.com	lex-steele-gets-his-big-black-sausage-eaten-by-buxom-kimberly-kendall-pov-style_v78247	None	None	None

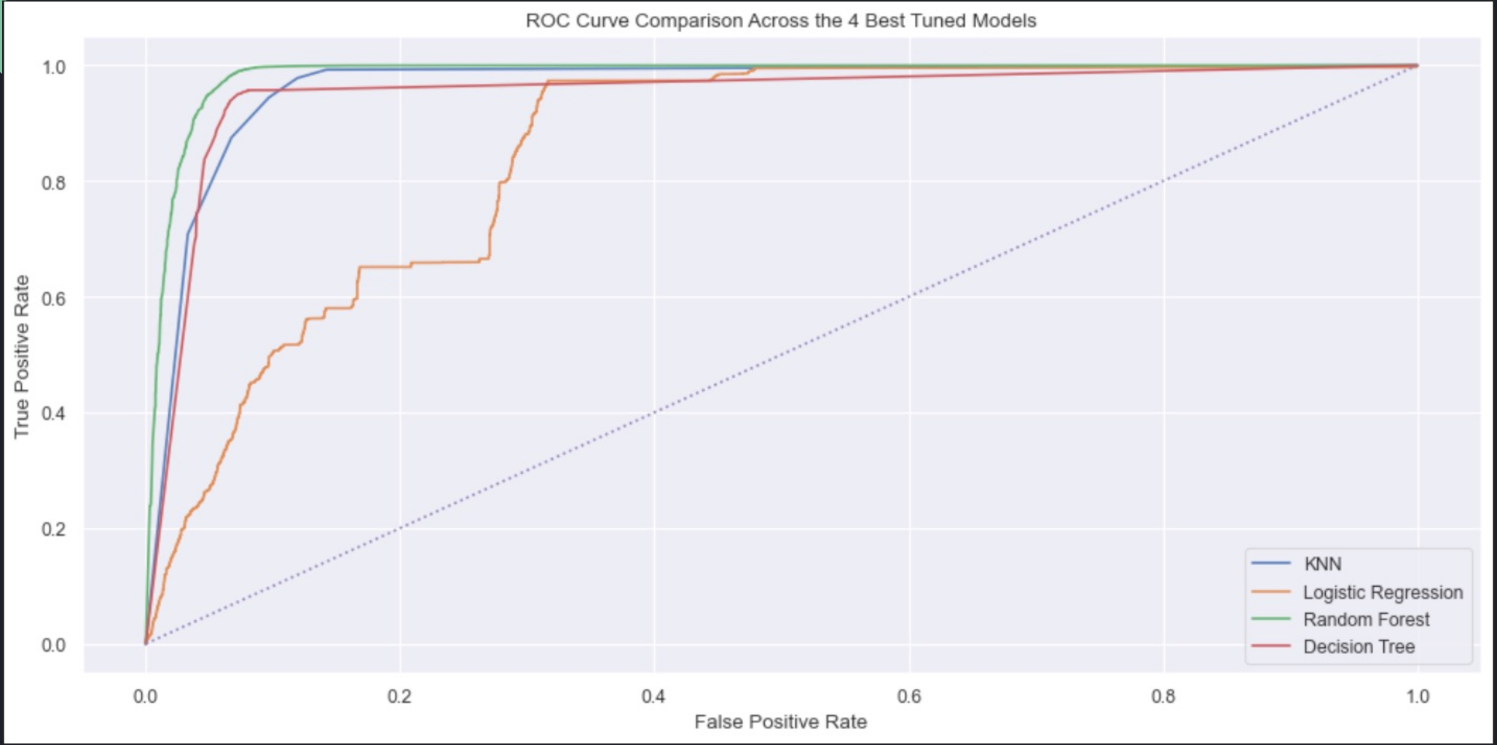
Training Data Snippet

Raw Scores

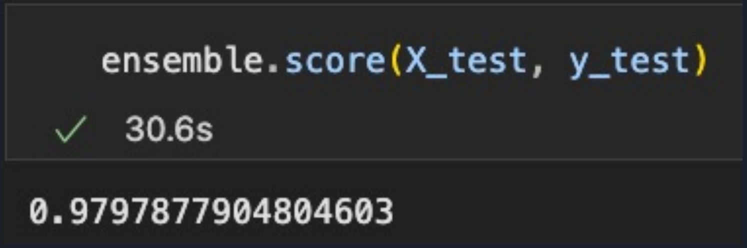


Models	Accuracy Score	Precision Score	Recall Score	F1 Score
Decision Tree (criterion="entropy", max_depth=17)	0.925539864006575	0.995879686856201	0.924929402760550	0.959094174391954
Decision Tree (default settings)	0.925539864006575	0.995879686856201	0.924929402760550	0.959094174391954
KNN (5 neighbors)	0.925539864006575	0.995879686856201	0.924929402760550	0.959094174391954
KNN (default settings)	0.942115619318040	0.993931229863348	0.944432714892718	0.968549969551390
Logistic Regression (C=0.01, penalty="L1", solver="LibLinear")	0.925539864006575	0.995879686856201	0.924929402760550	0.959094174391954
Logistic Regression (default settings)	0.913160975366758	0.980623891147339	0.926288564566783	0.952682115834832
Random Forest (default settings)	0.981269770106353	0.995847797062750	0.984257475389934	0.990018714909544
Random Forest (estimators=300)	0.925539864006575	0.995879686856201	0.924929402760550	0.959094174391954

Rate of False and True Positives with Tuned Models



Ensemble Model - Chosen as Deployable Model

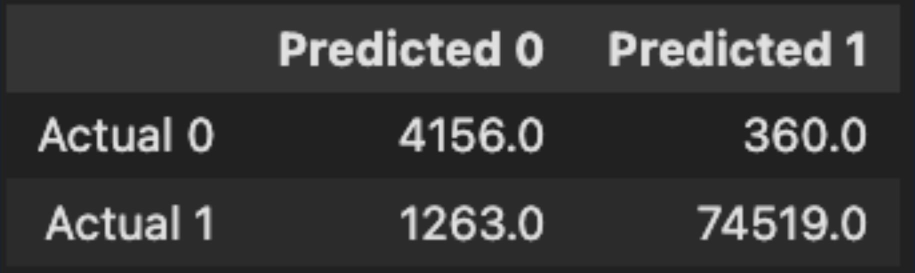


```
ensemble.score(X_test, y_test)
✓ 30.6s

0.9797877904804603
```

Score

Confusion Matrix



	Predicted 0	Predicted 1
Actual 0	4156.0	360.0
Actual 1	1263.0	74519.0

References

365 Data Science, ML in Python course, section 4.9

https://splunkbase.splunk.com/app/2734/

https://web.archive.org/web/20161130185550/http://fsecurify.com/using-machine-learning-detect-malicious-urls/

https://stackoverflow.com/a/50084009

https://datascienceparichay.com/article/pandas-split-column-by-delimiter/

https://stackoverflow.com/questions/37335598/how-to-get-the-length-of-a-cell-value-in-pandas-dataframe

https://stackoverflow.com/questions/49234374/how-to-count-vowels-and-consonants-in-pandas-dataframe-both-uppercase-and-lower

https://stackoverflow.com/questions/13174468/how-do-you-join-all-items-in-a-list

https://machinelearningmastery.com/modeling-pipeline-optimization-with-scikit-learn/

https://machinelearningmastery.com/modeling-pipeline-optimization-with-scikit-learn/

https://towardsdatascience.com/ensemble-learning-in-sklearn-587f21246e8d

https://stackoverflow.com/questions/332289/how-do-i-change-the-size-of-figures-drawn-with-matplotlib

https://towardsdatascience.com/ensemble-learning-in-sklearn-587f21246e8d

Thank you!

Questions?