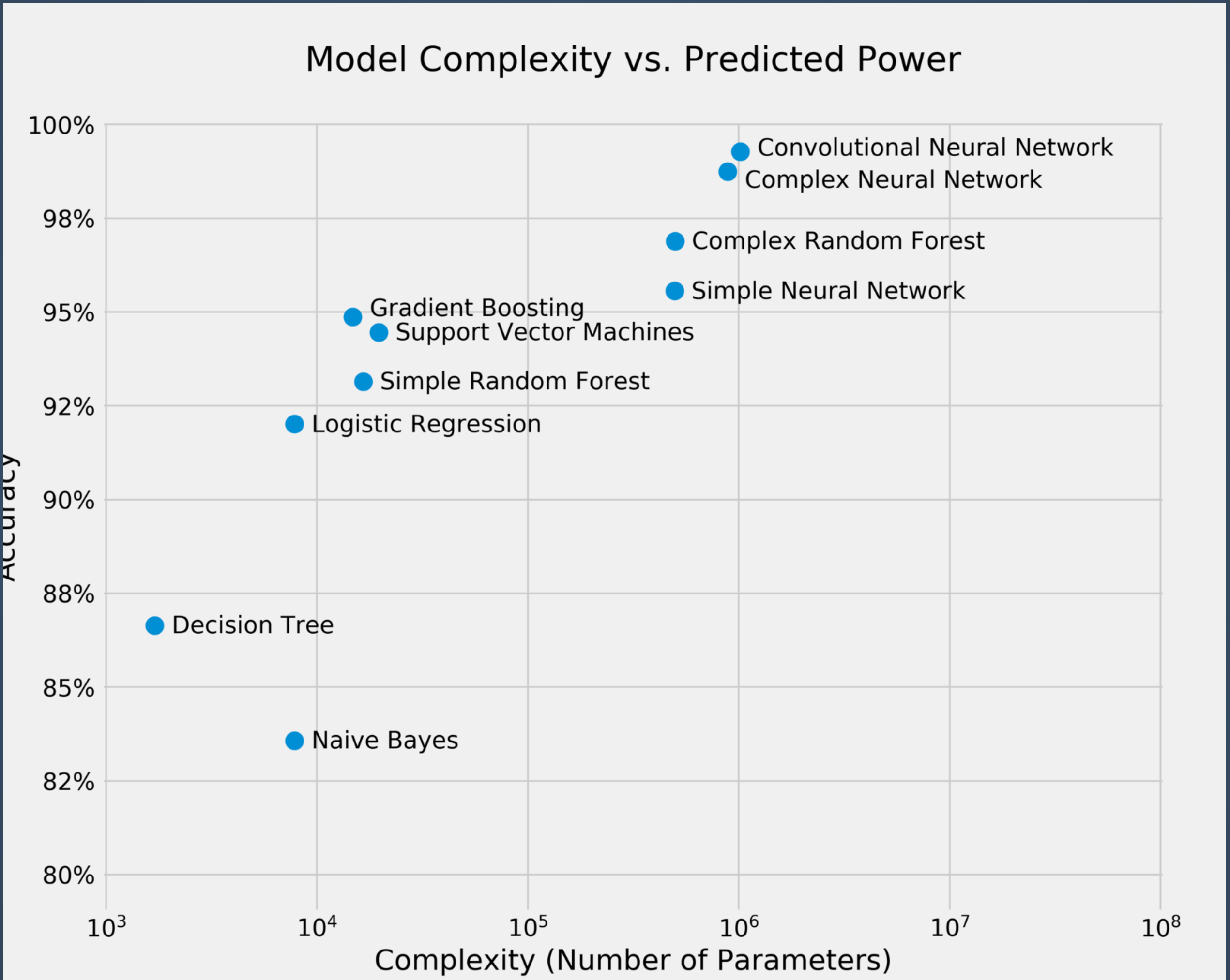
DEMYSTIFYING MACHINE LEARNING

Alejandro Correa Bahnsen

BLACK BOX MODELS

- Machine learning models are often dismissed on the grounds of lack of interpretability.
- When using advanced models it is nearly impossible to understand how a model is making a prediction.

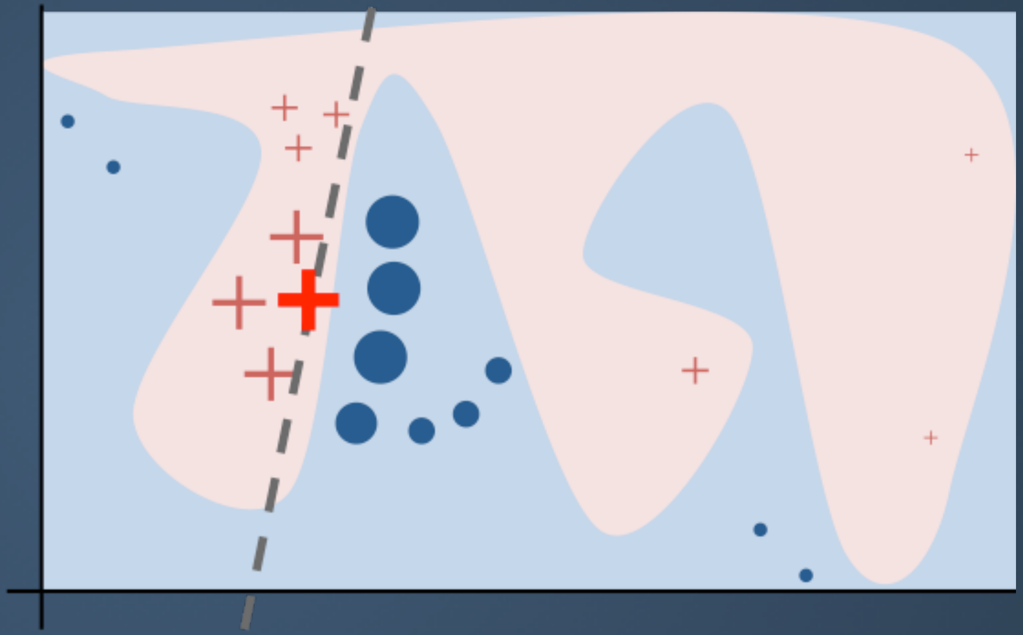
MNIST - ACCU VS # PARAMS



Notebook to create the plot

LIME

LIME stands for Local Interpretable Model-agnostic Explanations, and its objective is to explain the result from any classifier so that a human can understand individual predictions

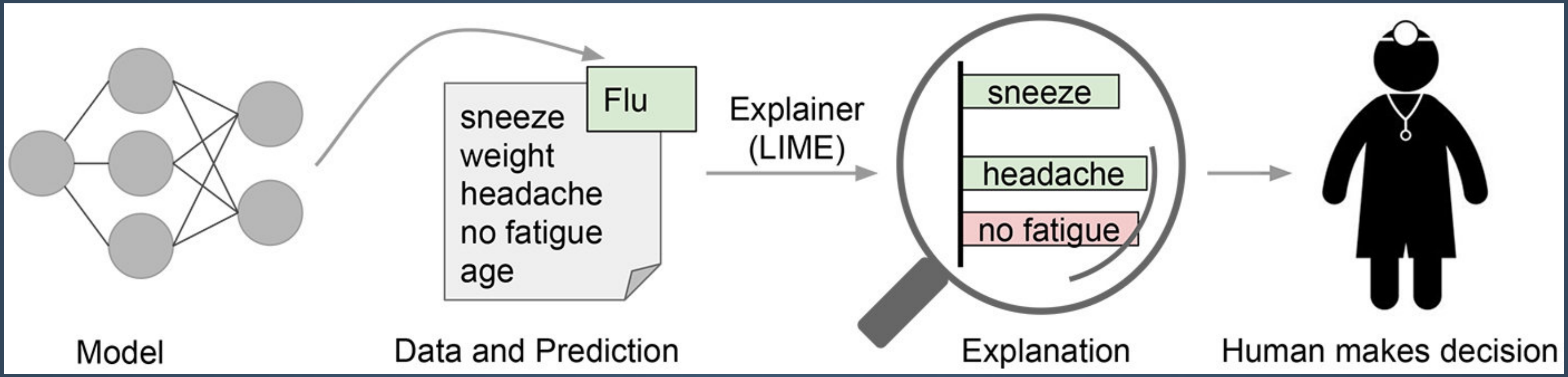


LIME

- An *interpretable representation* is a point in a space whose dimensions can be interpreted by a human.
- LIME frames the search for an interpretable explanation as an optimization problem. Given a set **G** of potentially interpretable models, we need a measure **L(f,g,x)** of how poorly the interpretable model **g∈G** approximates the original model **f** for point **x** this is the loss function. We also need some measure **Ω(g)** of the complexity of the model (e.g. the depth of a decision tree). We then pick a model which minimizes both of these

$$\xi(x) = \operatorname{argmin}\ g \in G\ L(f,g,x)+\Omega(g)$$

LIME

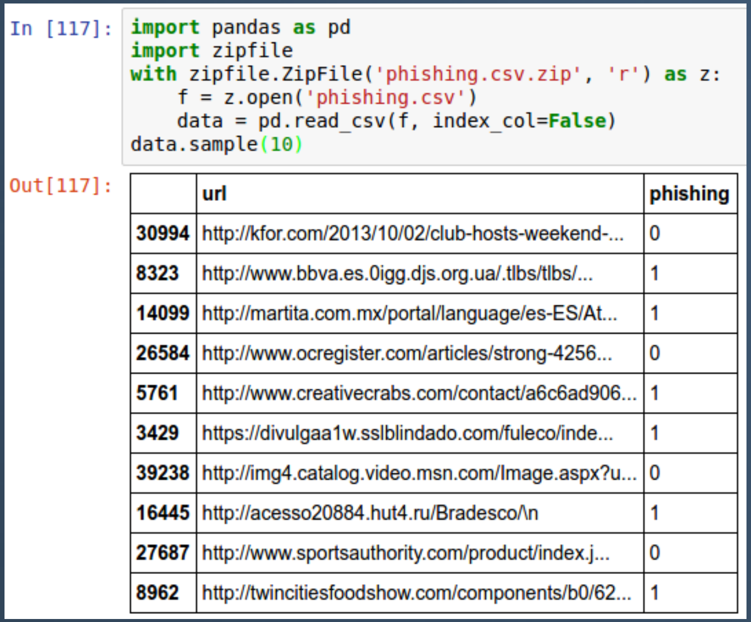

Flu
sneeze
weight
headache
no fatigue
age
Explainer
(LIME)
sneeze
headache
no fatigue
Model
Data and Prediction
Explanation
Human makes decision

LIME EXAMPLE

URL PHISHING DETECTION

URL PHISHING CLASSIFIER

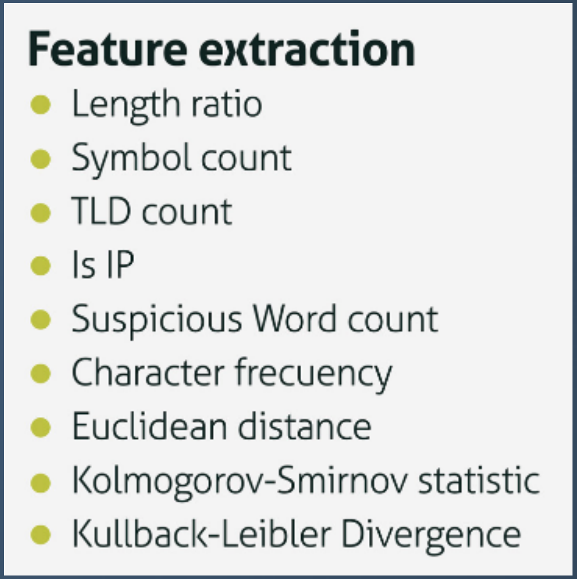
Objective: Evaluate phishing probability using only the web site URL



In [117]:
```
import pandas as pd
import zipfile
with zipfile.ZipFile('phishing.csv.zip', 'r') as z:
    f = z.open('phishing.csv')
    data = pd.read_csv(f, index_col=False)
data.sample(10)
```

Out[117]:

	url	phishing
30994	http://kfor.com/2013/10/02/club-hosts-weekend-...	0
8323	http://www.bbva.es.0igg.djs.org.ua/.tlbs/tlbs/...	1
14099	http://martita.com.mx/portal/language/es-ES/At...	1
26584	http://www.ocregister.com/articles/strong-4256...	0
5761	http://www.creativecrabs.com/contact/a6c6ad906...	1
3429	https://divulgaa1w.sslblindado.com/fuleco/inde...	1
39238	http://img4.catalog.video.msn.com/Image.aspx?u...	0
16445	http://acesso20884.hut4.ru/Bradesco/\n	1
27687	http://www.sportsauthority.com/product/index.j...	0
8962	http://twincitiesfoodshow.com/components/b0/62...	1

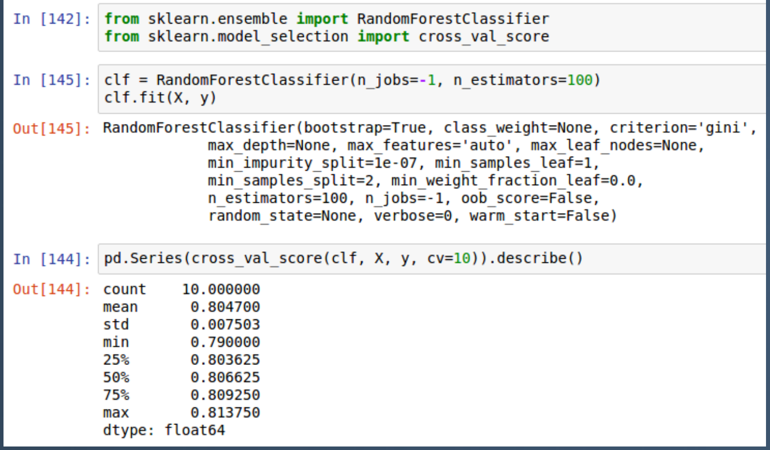


Feature extraction

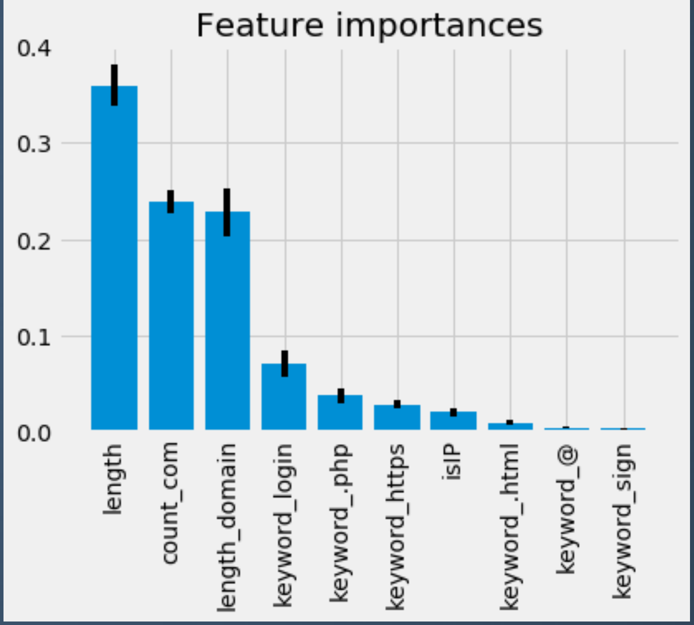
- Length ratio
- Symbol count
- TLD count
- Is IP
- Suspicious Word count
- Character frecuency
- Euclidean distance
- Kolmogorov-Smirnov statistic
- Kullback-Leibler Divergence

URL PHISHING CLASSIFIER

Train a random forest

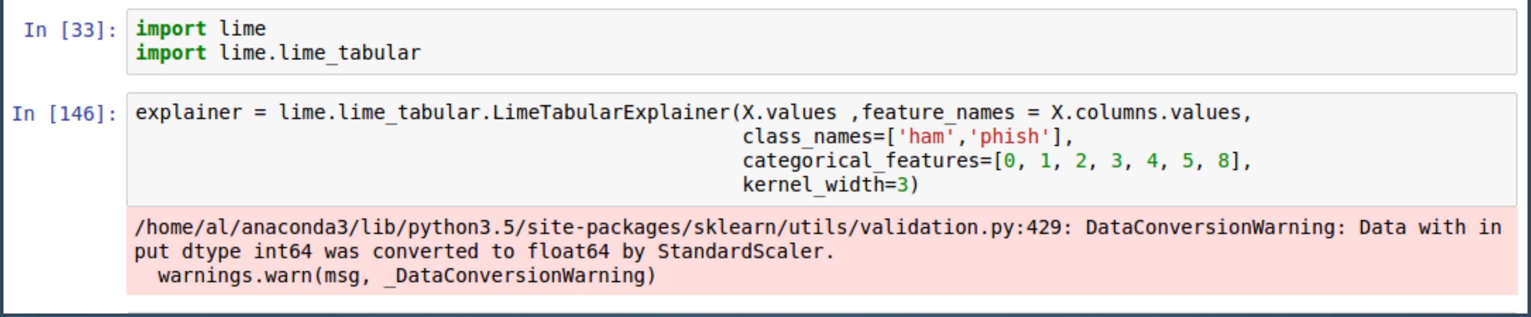


```
In [142]: from sklearn.ensemble import RandomForestClassifier
          from sklearn.model_selection import cross_val_score

In [145]: clf = RandomForestClassifier(n_jobs=-1, n_estimators=100)
          clf.fit(X, y)

Out[145]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      min_impurity_split=1e-07, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=100, n_jobs=-1, oob_score=False,
                      random_state=None, verbose=0, warm_start=False)

In [144]: pd.Series(cross_val_score(clf, X, y, cv=10)).describe()

Out[144]: count    10.000000
          mean      0.804700
          std       0.007503
          min       0.790000
          25%       0.803625
          50%       0.806625
          75%       0.809250
          max       0.813750
          dtype: float64
```



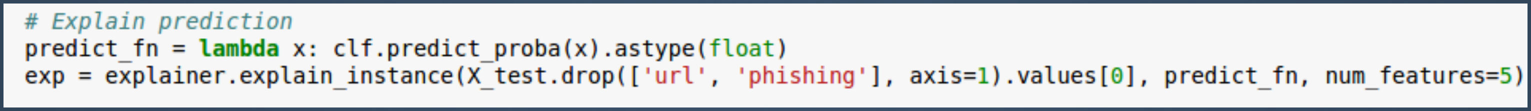
LIME EXAMPLE

Fit lime explainer



```
In [33]:  import lime
           import lime.lime_tabular

In [146]:  explainer = lime.lime_tabular.LimeTabularExplainer(X.values ,feature_names = X.columns.values,
                                                              class_names=['ham','phish'],
                                                              categorical_features=[0, 1, 2, 3, 4, 5, 8],
                                                              kernel_width=3)

           /home/al/anaconda3/lib/python3.5/site-packages/sklearn/utils/validation.py:429: DataConversionWarning: Data with in
           put dtype int64 was converted to float64 by StandardScaler.
             warnings.warn(msg, _DataConversionWarning)
```

Explain an instance



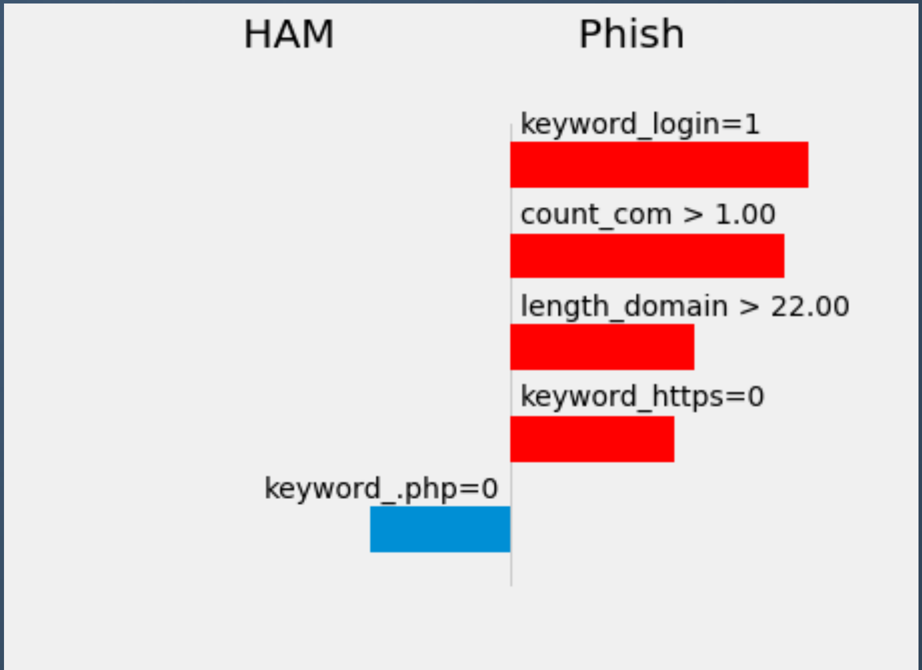
```
# Explain prediction
predict_fn = lambda x: clf.predict_proba(x).astype(float)
exp = explainer.explain_instance(X_test.drop(['url', 'phishing'], axis=1).values[0], predict_fn, num_features=5)
```

LIME EXAMPLE

Example Phishing URL

Url = http://login.paypal.com.convexcentral.com/Update/ab770f624342b07b71e56c1bae5d9bcb/

Phishing probability
1.0

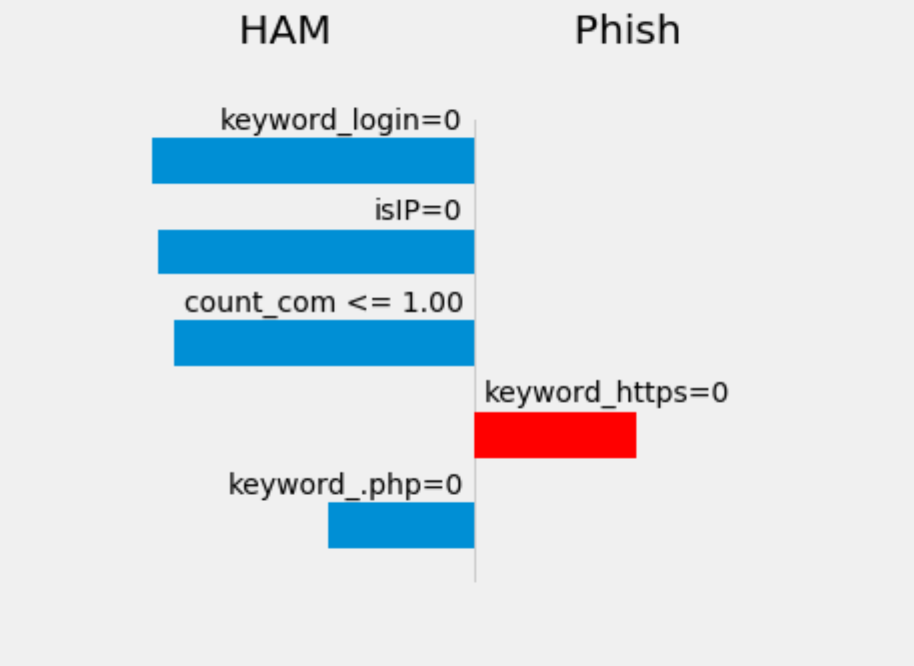


LIME EXAMPLE

Example Phishing URL

Url = `http://www.redeyechicago.com/entertainment/tv/redeye-banshee-ivana-mili`...

Phishing probability
0.0283



THANK YOU

FULL NOTEBOOK IN

HTTPS://GITHUB.COM/ALBAHNSEN/TALK_DEMYSTIFYING_MACHINE_LEARNING