

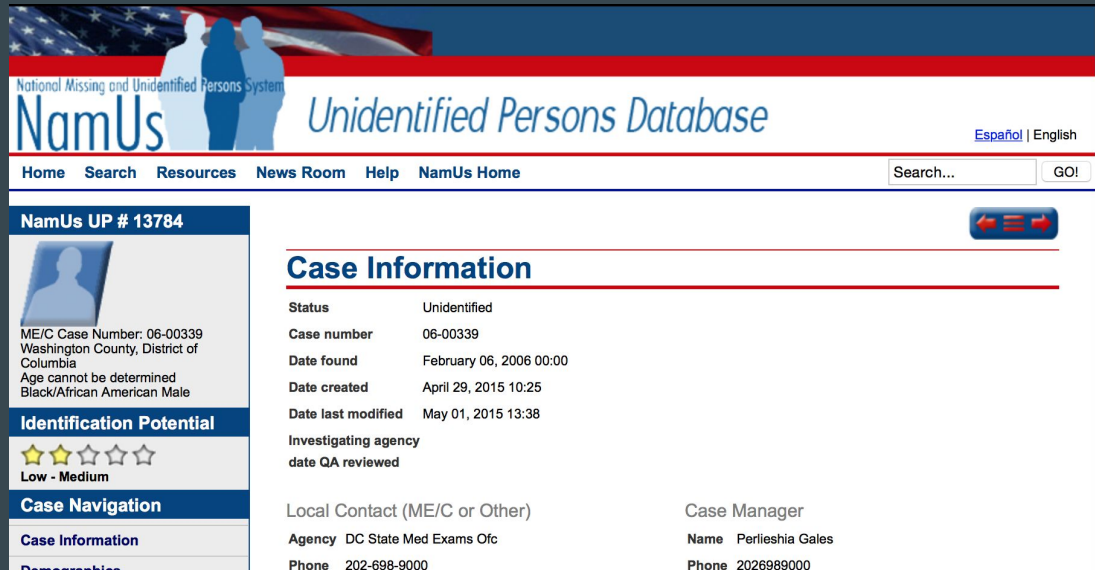
NamUs: Unidentified Persons



Using Machine Learning to Predict the “Identification Potential” 5-star rating

Target Response: “Identification Potential” ‘IP’

- High
- Medium-High
- Medium
- Low-Medium
- Low
- Extremely-Low ★



The screenshot displays the NamUs (National Missing and Unidentified Persons System) website. The header includes the NamUs logo and the text "Unidentified Persons Database". A navigation bar contains links for Home, Search, Resources, News Room, Help, and NamUs Home. A search bar is located on the right. The main content area shows details for "NamUs UP # 13784". On the left, there is a placeholder for a photo and case details: ME/C Case Number: 06-00339, Washington County, District of Columbia, Age cannot be determined, Black/African American Male. Below this is the "Identification Potential" section, which shows a star rating of 2 out of 5 stars and the text "Low - Medium". To the right of the case details is the "Case Information" section, which includes fields for Status (Unidentified), Case number (06-00339), Date found (February 06, 2006 00:00), Date created (April 29, 2015 10:25), Date last modified (May 01, 2015 13:38), Investigating agency, and date QA reviewed. At the bottom, there is a section for "Local Contact (ME/C or Other)" and "Case Manager", with details for Agency (DC State Med Exams Ofc), Name (Perlieshia Gales), and Phone (202-698-9000).

Scraped using Requests and BeautifulSoup, and put into Pandas

Recoded to 0, 1, 2, 3, 4, 5

Predicting Unidentified Case 5-Star Rating:

Star ratings are automatically calculated when data is entered into the system.

If I **can** predict the star rating:

1. Which features predict “Identification Potential”?
2. How do these compare with those identified by NamUs in their Help.PDF?

If I **cannot** predict the star rating:

1. This may indicate key information is not available to the public on NamUs.
2. Indicate a re-write of the Help.PDF is required.

Namus Help PDF

3 Stars

A fingerprint classification or fingerprint card has been entered or uploaded, AND/OR, information has been entered in at least one of the tooth boxes on the dental chart page, AND/OR the “Recognizable Face” option has been selected in the “Body Condition” section and a facial photo or artist’s rendering has been uploaded

4 Stars

5 Stars

The face is recognizable, a facial photo or rendering has been uploaded, fingerprint information has been entered or uploaded, a DNA profile has been established and specific tooth information has been entered.

Examples of Binary Features:

- Capture what is *KNOWN* vs what is *UNKNOWN*

tattoos	piercings
scars_and_marks	skeletal_findings
n-hands_not_recovered	_dna
_dental	_fingerprints
_sex	_face
weight_bin	height_bin

`_face` (recoded from `Recognizable_face`):

```
In [48]: namusdb.recognizable_face.unique()
```

```
Out[48]:
```

```
array(['Not recognizable - Decomposing/putrefaction', 'Recognizable face',  
      'Not recognizable - Charred/burned',  
      'Not recognizable - Insect/animal activity',  
      'Not recognizable - Partial remains with soft tissues',  
      'Not recognizable - Partial skeletal parts only', '',  
      'Not recognizable - Near complete or complete skeleton',  
      'Not recognizable - Mummified',  
      'Not recognizable - Traumatic injuries'], dtype=object)
```

Examples of Text and Numeric Features:

-

tattoos_description	piercings_description
scars_and_marks_description	skeletal_findings_description
clothing_on_body	footwear
eyewear	jewelry
head_hair	facial_hair
body_hair	circumstances
images	age_range

Linear Regression Model

```
# Create features for linear regression: all features
linreg_features = ['all_parts_recovered',
                  'amputations',
                  'artificial_parts_aids',
                  'deformities',
                  'finger_toe_nails',
                  'foreign_objects',
                  'head_not_recovered',
                  'images',
                  'medical_implants',
                  'n-hands_not_recovered',
                  'n-limbs_not_recovered',
                  'organ_absent',
                  'other_distinctive_features',
                  'other_medical_information',
                  'piercings',
                  'prior_surgery',
                  'scars_and_marks',
                  'skeletal_findings',
                  'tattoos',
                  'torso_not_recovered',
                  '_sex',
                  '_dna',
                  '_dental',
                  '_fingerprints',
                  '_face',
                  '_l_eye',
                  '_r_eye',
                  'height_bin',
                  'weight_bin',
                  'age_range']
```

```
linreg = LinearRegression(
    normalize=True)
```

Null: RMSE = 1.19
split RMSE = 0.8159

Train-test-

A prediction of IP from 0-5 can be off by ~ 0.82 stars.

REGULARIZATION:

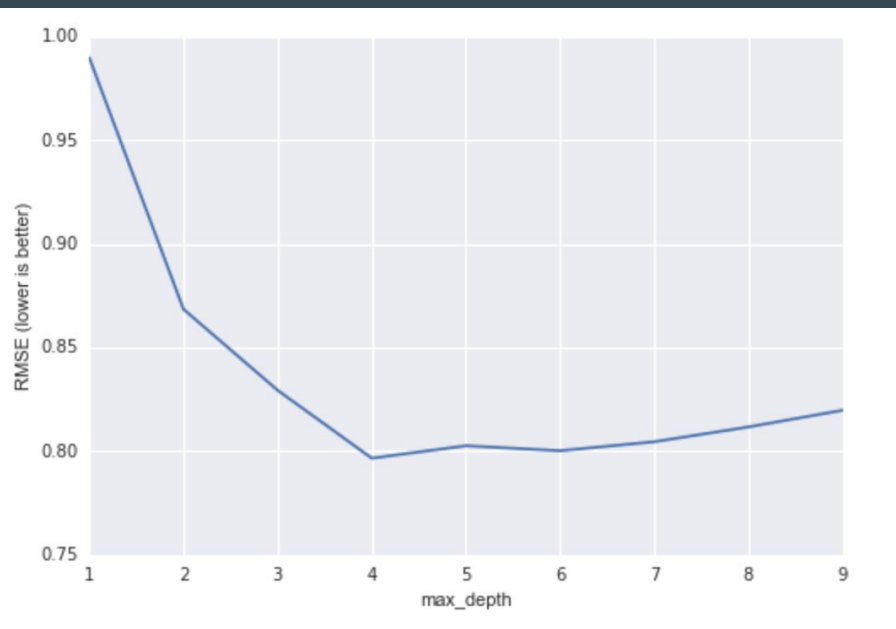
RidgeCV: RMSE = 0.8158

LassoCV: RMSE = 0.8154

Features with zero coefficients:

```
'artificial_parts_aids',  'deformities',
'foreign_objects',        'r_eye',
'medical_implants',      'torso_not_recovered'
```


Decision Tree Regression Model



```
treereg = DecisionTreeRegressor  
(max_depth=depth,  
random_state=1)
```

max_depth range 1-10, cv=100
Best: 4 splits, RMSE of 0.7966.

cv=1000, RMSE = 0.7316.

Improvement due to better sampling of
skewed data?

Regression Tree Splits!

FACE?

0? 5?

_dna

_dental

_dental

images

images

images

images

`_fingerprints`

```
all_parts
_recover
d
```

images

age_range

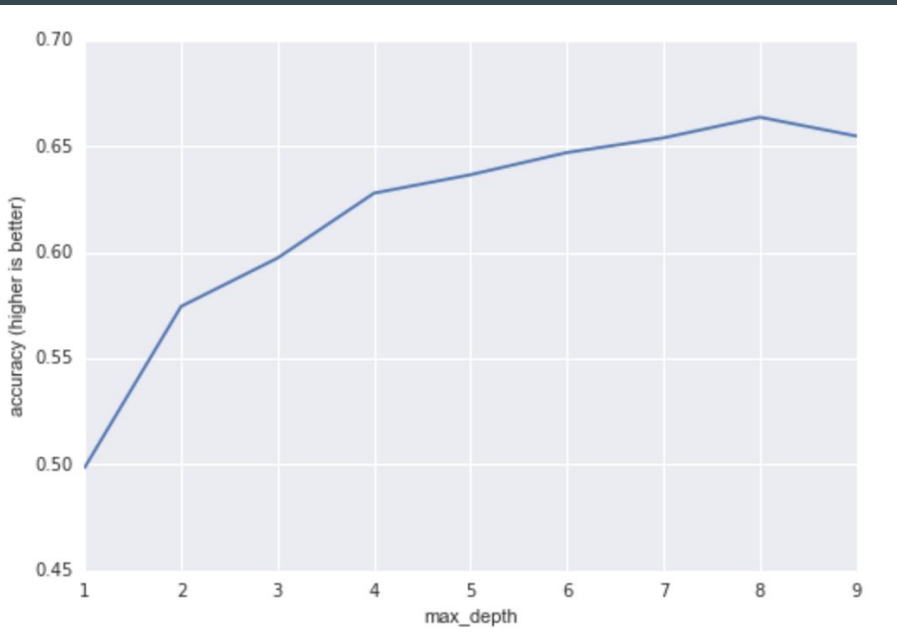
```
head_not_
recovered
```

sex

_finger prints

age_range

Decision Tree Classification Model



```
treeclf = DecisionTreeClassifier  
(max_depth=depth,  
random_state=1)
```

max_depth range 1-10, cv=100
Best: 8 splits! , Accuracy 64%

4 Splits: All 5-Star cases in one Leaf (gini 0.65)

Classification Tree Splits!

FACE?

0 & 5
ish

_dna

_dental

_dental

```
_finger
prints
```

images

images

_finger prints

```
weight_
bin
```

images

```
height_
bin
```

tattoos

age_range

images

images



images

Random Forest Regression

After Tuning:

```
rfreg = RandomForestRegressor(n_estimators=270,  
max_features=5, oob_score=True, random_state=1)  
  
rfreg.fit(X, y)
```

R^2 Out-of-Bag Score: **0.56262711423348488**



24	 _face 	0.023
23	_fingerprints	0.039
7	images	0.107
29	age_range	0.180
22	_dental	0.197
21	_dna	0.277

Random Forest Classification

After Tuning:

```
rfclf_test = RandomForestClassifier(n_estimators=65,  
max_features=10, random_state=1, oob_score=True) rfclf_test.fit  
(X, y) scores =  
cross_val_score(rfclf, X, y, cv=10, scoring='accuracy') predicted  
= cross_val_predict(rfclf, X, y, cv=10)
```

R² Out-of-Bag Score: 0.65

24	 _face 	0.03
23	_fingerprints	0.05
7	_dental	0.09
29	images	0.11
22	_dna	0.18
21	age_range	0.24

Discussion / Conclusions

- Regression better than classification? Difficult to compare performance between %age accuracy and RMSE and R^2 ...
- Why is `'_face'` so hard to get? Yet it's critical in the NamUs model. Why?
 - This may be due to smaller numbers of 5-Star cases
 - This may also be the reason 5-Star ratings are hard to predict.
 - Same goes for the 0-Stars
- Add binary `hair_color` !!?
- Images are a **count** and not separated by actual **FACE** images and **NON-FACE** images
- Improve Decision Tree performance by merging some features together?
- One Versus Rest better able to model the 0-Star and 5-Star?
- Ensembling
- Ordinal Logistic Regression? Made for multiple, ordered classes.
- Continue working on the Text data...current NB accuracy 48% (wowzers)
- Better exploration / visualization of model performance (eg AUC)

New Questions and Future Additions:

- Submit a FOIA?
 - Request data on solved cases - what features lead to the resolution?
 - Do they update their algorithm based on evidence from solved cases?
- Create an interactive map for the find-locations and case summary
- Create other visuals