NamUs: Unidentified Persons

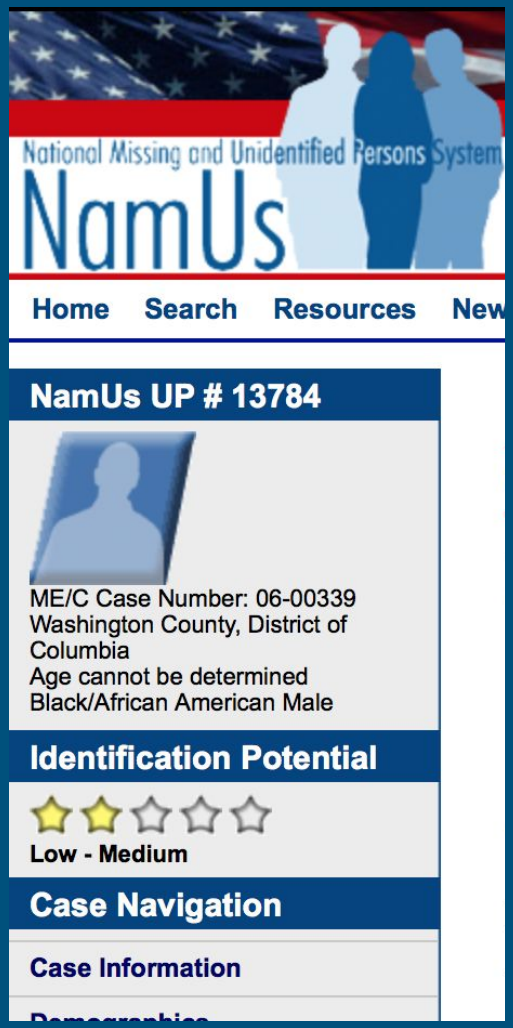
Using Machine Learning to Predict the “Identification Potential” 5-star rating

Dr. Jennifer A Stark
GitHub: JAStark/Namus_Project
Twitter: @_JAStark

Target Response: “Identification Potential”

- High ★★★★★ 5
- Medium-High ★★★★☆ 4
- Medium ★★★☆☆ 3
- Low-Medium ★★☆☆☆ 2
- Low ★☆☆☆☆ 1
- Extremely-Low ☆☆☆☆☆ 0

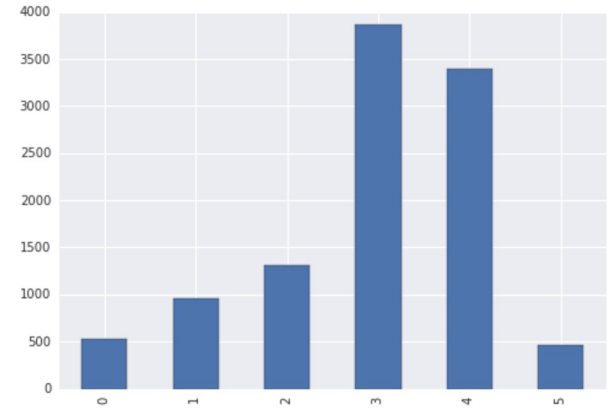
Web-scraped using `Requests` and `BeautifulSoup4`, and pickled



Some Data:

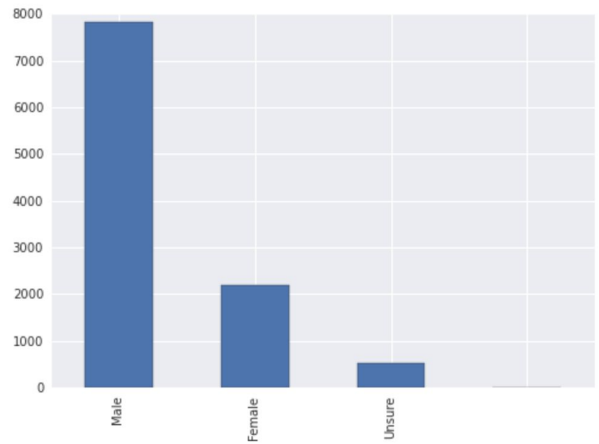
```
stars = [0,1,2,3,4,5]
namus.rating.value_counts().ix[list(stars)].plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x130f0a4e0>
```



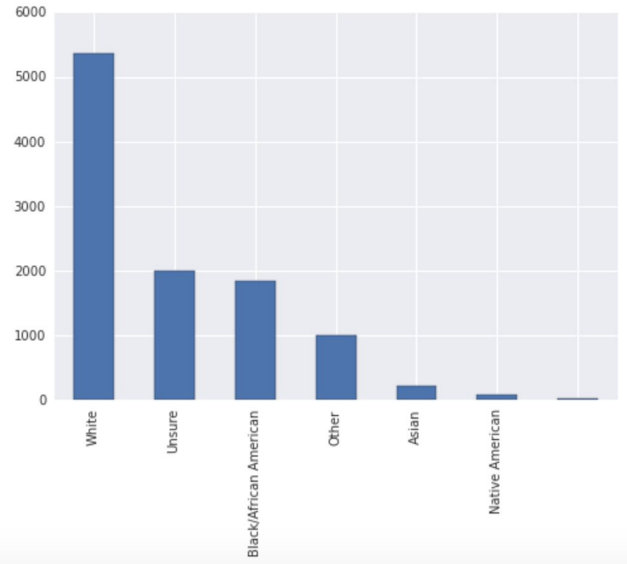

```
namus.sex.value_counts().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1366bb518>
```




```
namus.race.value_counts().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x115c04860>
```



Predicting Unidentified Case 5-Star Rating:

Stars 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?

Predicting Unidentified Case 5-Star Rating:

Stars are **automatically** calculated when data is entered into the system.

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.

New Data: Face Detection!

OpenCV3 compiled with Anaconda Python3 bindings

- Frontal face haar cascade
- Profile face lbp cascade (only works on right-facing faces)
  - lbp -> Local Binary Patterns. “Faster but less accurate than Haar”

`Scrape_images.py`

Code: GitHub.com/JAStark/WomenDataScientistsDC_MeetupNov2015

Datatypes -> Key Question -> My Assumption -> Binarize Everything!

Data types:

- Binary
- Continuous integers (height, weight, #images)
- Text (“options” list, like a dropdown menu)
- Long-form text (eg. descriptions)
- Images

Key Question:

- Or is it important what the data actually *says*?
- Is it the presence / absence of data that’s important? Known vs unknown

Binarize!

Examples of Binary Binarized and Continuous numeric Features:

Binary	**Binarized**	**Continuous Numeric**
jewelry	**_dna**	age_range (engineered)
footwear	**_dental**	images
clothing	**_fingerprints**	
tattoos	**_face**	
skeletal_findings	**face_images**	
	_sex	

Challenges in Binarization

```
namus['_dna'] = namus.dna.map({'Sample submitted - Tests complete':1,
                               'Samples submitted - Tests not complete':0,   #ch
                               'Complete - Insufficient DNA for profiling':0,
                               'No DNA information is currently available':0,
                               'Sample is currently not available':0,
                               'Sample available - Not yet submitted':0}) #chang
```

```
namus.hair_color.unique()
```

```
array(['', 'Brown', 'Unknown or Completely Bald', 'Gray or Partially Gray',
       'Black', 'Blond/Strawberry', 'Red/Auburn', 'White', 'Sandy',
       'Purple'], dtype=object)
```

```
namus['_haircolor'] = namus.hair_color.map({'Brown' : 1,
                                            'Unknown or Completely Bald': 0,  #would
                                            'Gray or Partially Gray':1,
                                            'Black':1,
                                            'Blond/Strawberry':1,
                                            'Red/Auburn':1,
                                            'White':1,
                                            'Sandy':1,
                                              '':0,
                                             'Purple':1})  #someone dyed their hair?
```

Linear Regression Model

```
#  Create features for linear regression: all fea
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']
```

modified

\+
'hair_color'
'face_images'

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

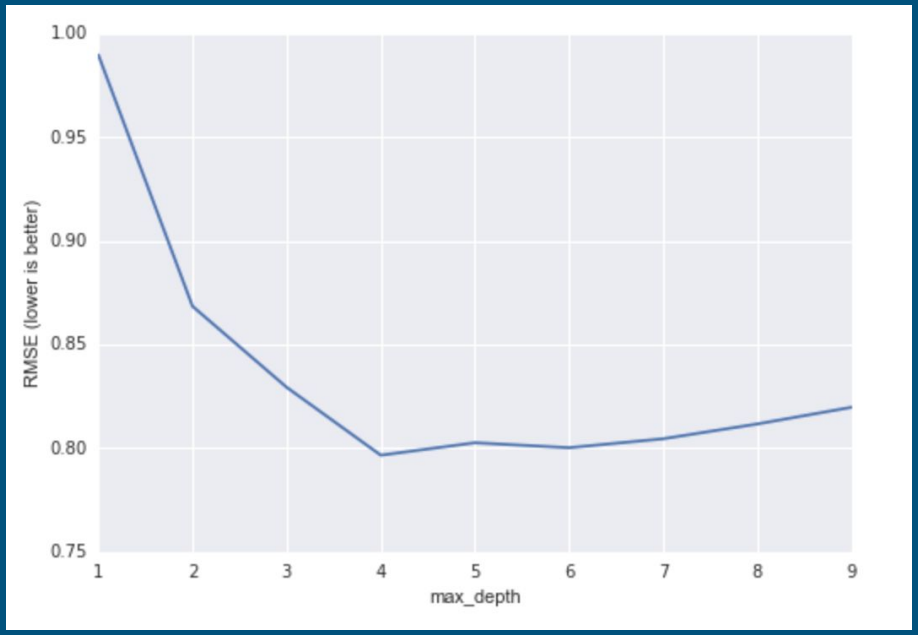
A prediction of IP from 0-5 can be off by ~ 0.82 stars. NEW 0.6972

REGULARIZATION:

RidgeCV: RMSE = 0.8158 NEW 0.6970
LassoCV: RMSE = 0.8154 NEW 0.6974

12	other_distinctive_features	0.164711
20	_sex	0.176840
24	_face	0.273228
30	face_images	0.313659
23	_fingerprints	0.425424
22	_dental	0.930528
21	_dna	1.344366

Decision Tree Regression Model




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

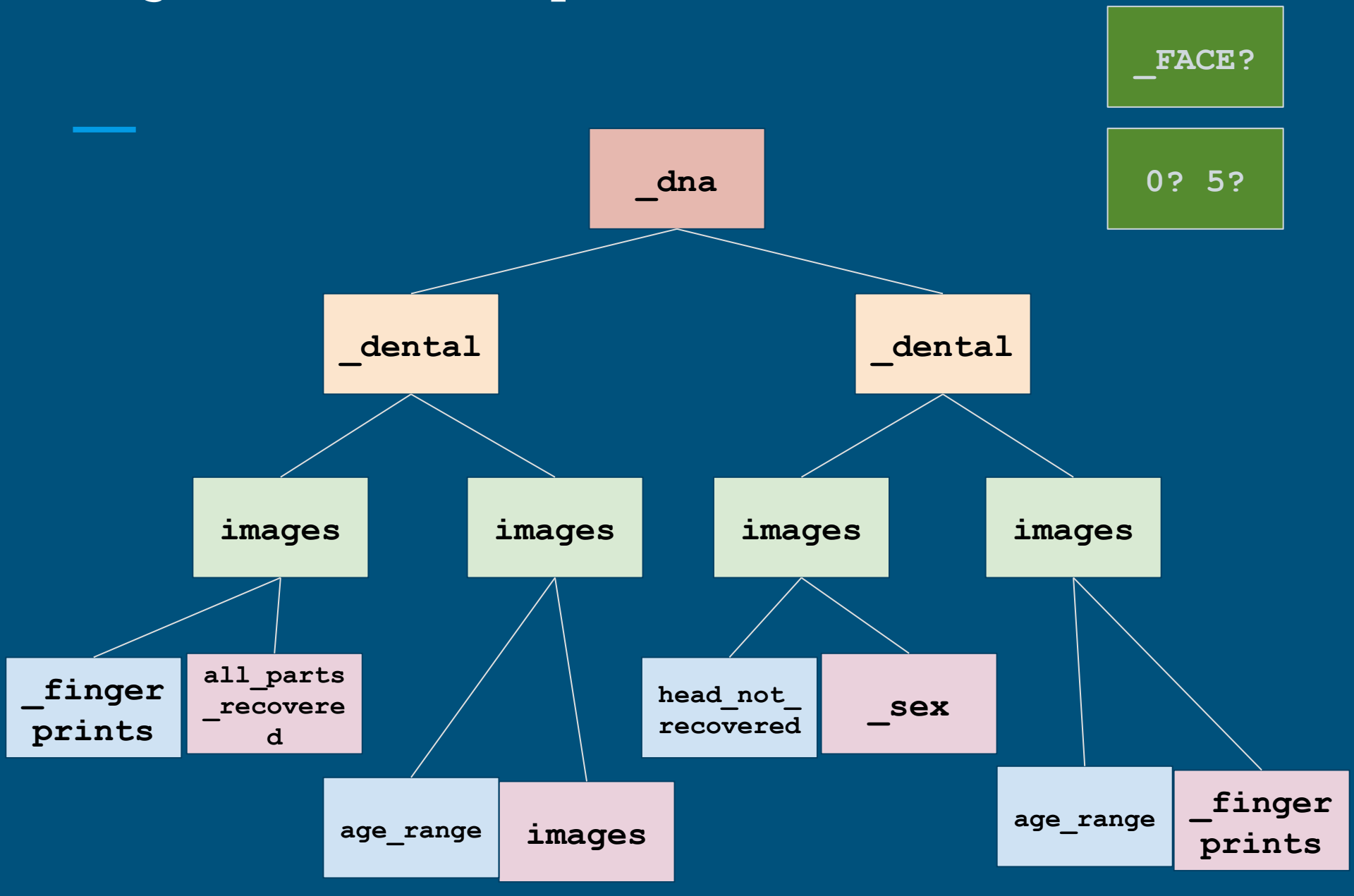
`max_depth range 1-10, cv=100`

Best: 4 splits, RMSE of 0.7966.

Best: 6 splits, RMSE 0f 0.6234

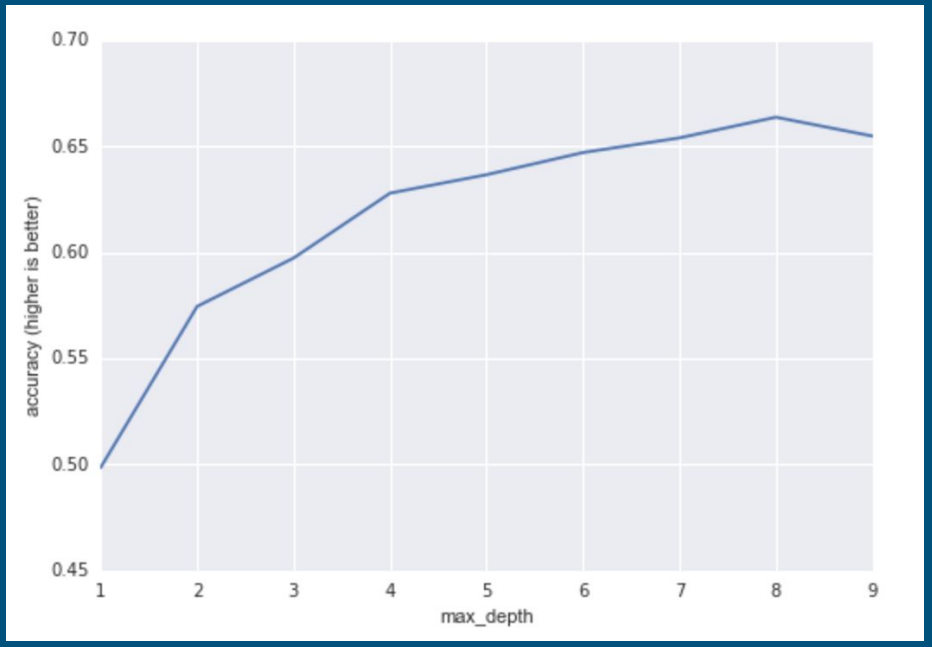
26	r_eye	0.010029
29	age_range	0.012340
24	_face	0.014764
23	_fingerprints	0.031734
7	images	0.074924
22	_dental	0.213598
21	_dna	0.613137

Regression Tree Splits!



Decision Tree Classification Model

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



`max_depth range 1-10, cv=100`

Best: 8 splits!, Accuracy 64%

Best: 8 splits! Accuracy 77%

6-splits: 0-Star confused with 1-Star

5-Stars dominate *some* leaves!

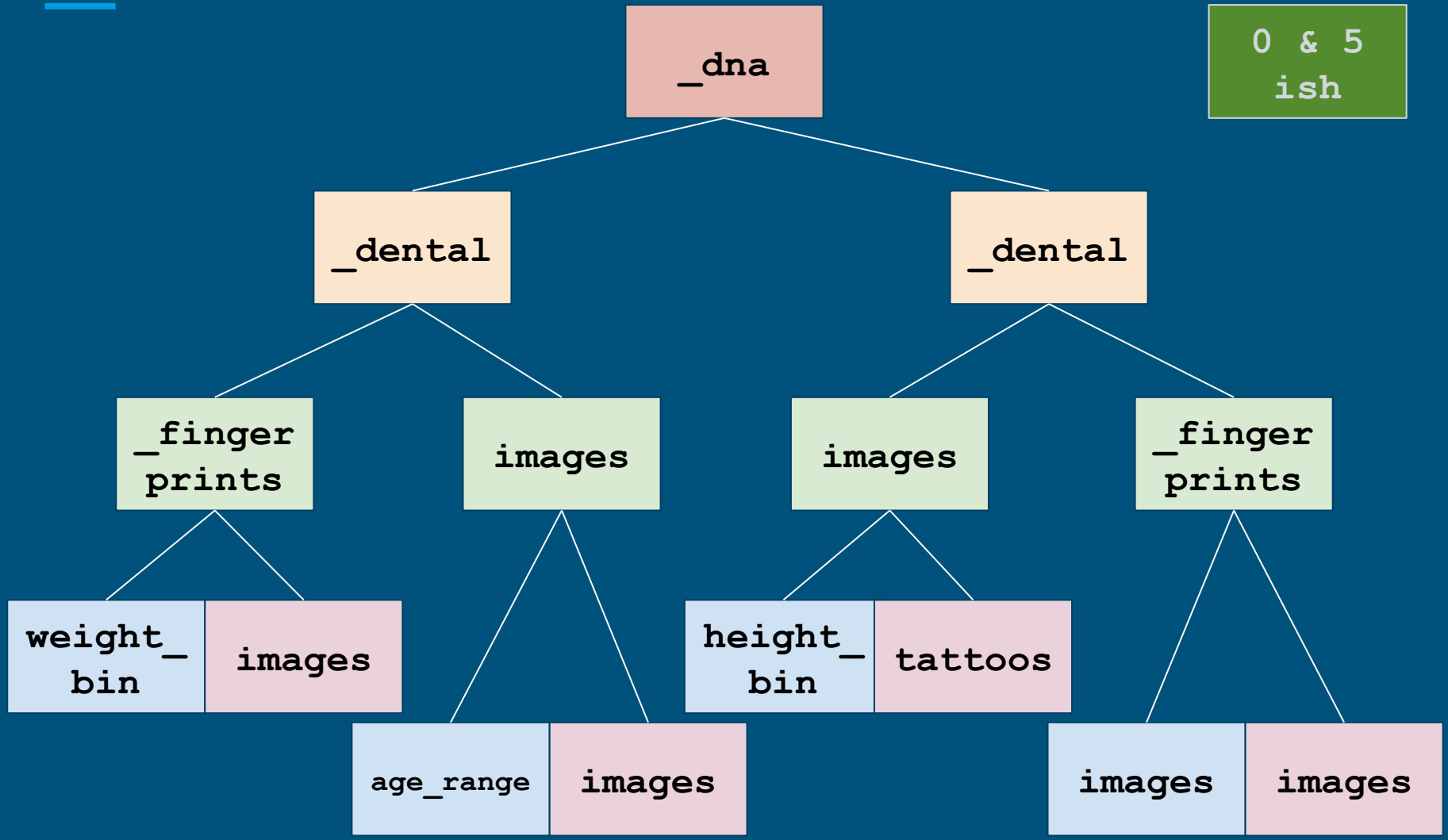
27	height_bin	0.019130
29	age_range	0.033670
31	face_images	0.044708
23	_fingerprints	0.070071
7	images	0.129285
22	_dental	0.249570
21	_dna	0.361732

Classification Tree Splits!

(depth8 too big to render png…
I have depth 6 in repo)

_FACE?

0 & 5 ish



Random Forest Regression

After Tuning:

```
rfreg = RandomForestRegressor(n_estimators=270 290,
max_features=5 7, oob_score=True, random_state=1)    rfreg.
fit(X, y)                                             R² Out-
of-Bag Score: 0.56262711423348488                    R² Out-
of-Bag Score: 0.69827779742411078
```

0	all_parts_recovered	0.015528
24	_face	0.022618
23	_fingerprints	0.039977
7	images	0.107092
29	age_range	0.180272
22	_dental	0.197411
21	_dna	0.276950

24	_face	0.019658
31	face_images	0.029949
23	_fingerprints	0.033139
7	images	0.075582
29	age_range	0.119363
22	_dental	0.182083
21	_dna	0.391180

Random Forest Classification

After Tuning:

```
rfclf_test = RandomForestClassifier(n_estimators=65 90,
max_features=10 10, random_state=1, oob_score=True)
rfclf_test.fit(X, y)
```

R^2 Out-of-Bag Score: **0.65** Accuracy: **60%**
R^2 Out-of-Bag Score: **0.77** Accuracy: **74%**

0	all_parts_recovered	0.025640
24	_face	0.027298
23	_fingerprints	0.046634
22	_dental	0.099698
7	images	0.112156
21	_dna	0.183071
29	age_range	0.240513

24	_face	0.022283
30	face_images	0.027830
23	_fingerprints	0.042146
7	images	0.082232
22	_dental	0.121611
29	age_range	0.159470
21	_dna	0.308432

Reverse-Engineering is Hard Because:

Humans!

- Humans who wrote the Algorithm / Documentation

1 Star

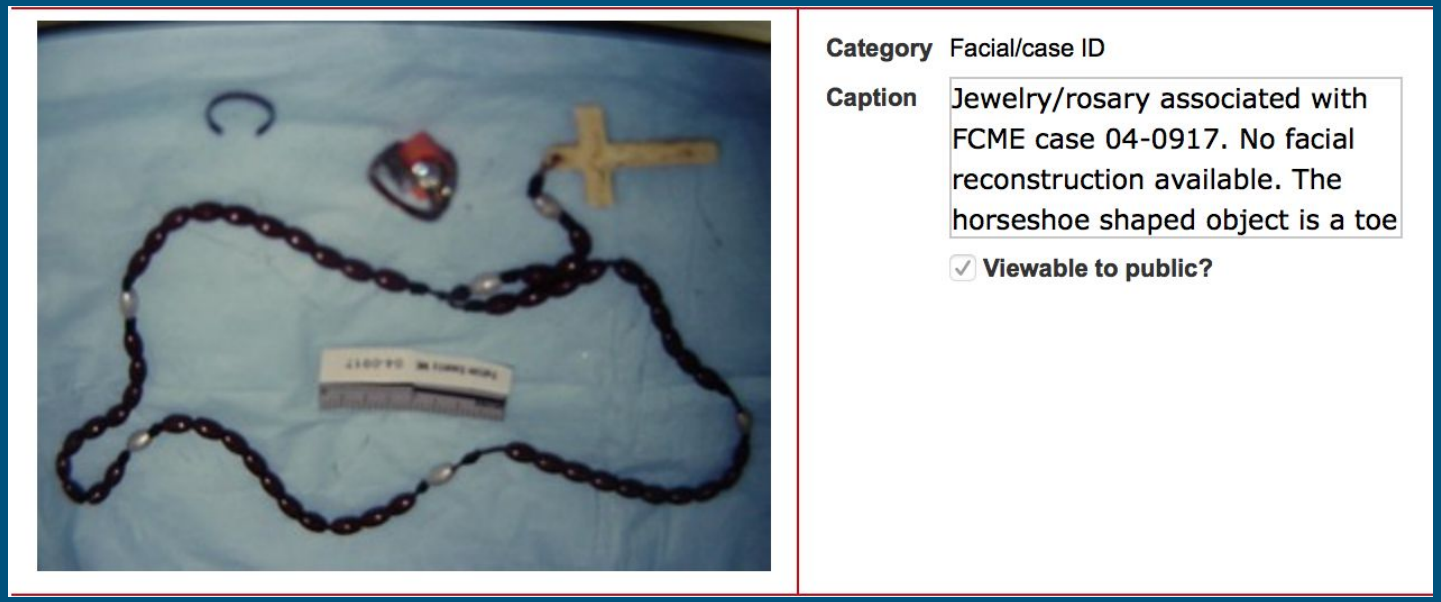
The case information includes a case number, the date the body or body part was found, the county and state where the body or body part was found and the condition of the body. Entries are also required for the estimated age group, race, sex, weight and height, but these entries may be listed as “unsure” or “cannot estimate.”

case_number	date_found	county	state	recognizable_face	race	sex	weight	height	est_age
A98-3798	September 12, 1998 00:00	Richmond	Georgia	Recognizable face	Black/African American	Male	161,	68,\n\t\t\t\t\tEstimated	Adult

Reverse-Engineering is Hard Because:

Humans!

- Humans who wrote the Algorithm / Documentation
- Humans who input the data - 5-star requires facial *images* (+DNA, fingerprints, recognizable face and dental)



Reverse-Engineering is Hard Because:

Humans!

- Humans who wrote the Algorithm / Documentation
- Humans who input the data
- Foolish Human who tries to reverse engineer the Algorithm (me!)
  - Incorrect assumptions
  - Not a pro at machine learning *OR* coding in general - tons more for me to learn!
    - Face detection was not perfect! Low-res, contrast, head-models...
  - Don’t have full access to full documentation
  - Don’t have full access to case details

Improvements?

- Regression better than classification? Difficult to compare performance between %age accuracy and RMSE and R^2...
- Better exploration / visualization of model performance (eg confusion matrices, AUC - better than accuracy for unbalanced data, BUT need to fiddle with it to make it work with multinomial data)
- Ensembling (0, 1, 5 stars have distinct requirements)
- Ordinal Logistic Regression? Made for multiple, ordered classes.
  - Not yet available on scikit-learn
  - Available from GitHub Account fabiano/mord
  - “Collection of Ordinal Regression algorithms in Python, following a scikit-learn compatible API

New Questions and Future Additions:

- Is Binarization the way to go!?

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.

New Questions and Future Additions:

- Is Binarization the way to go!?
- Submit a FOIA (Freedom of Information Act) request?
  - 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...

Thank you! Questions?