Name-based Prediction

418 Final Project
Yuki Urata

Dataset Used

- FBI Wanted API
  - Public API provided by the Federal Bureau of Investigation (FBI), gives access to the current list of wanted fugitives and persons of interest.
  - Created to Increase Public Awareness, enable developers to create tools, and to support transparency.
  - API returns information on
    - Title (name)
    - Sex
    - Race
    - Nationality
    - Dates of birth
    - Subjects (description of the case)
  - >1000 rows
  - Why use this dataset?
    - Dataset with real-life full names are uncommon

Model

- Random Forest
  - Name features used
    - Name length
    - Number of Words
    - Vowel ratio
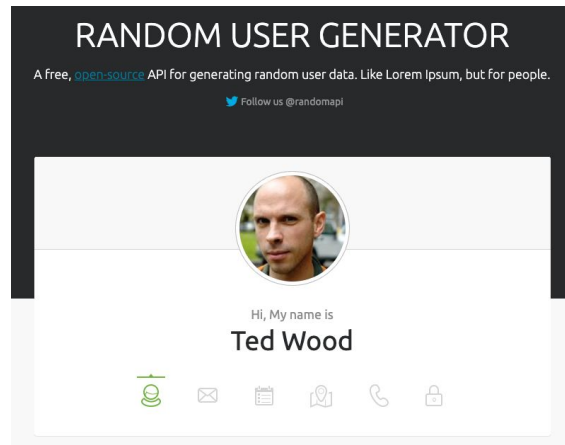   - Consonant ratio
    - First Letter
  - Model accessed via **Flask API**
- K-Nearest Neighbour
  - Outputs a similar name to the input
  - Used **randomuser.me API** to create dataset
    - Generates a fake, but realistic user data
    - Dataset created by creating 5000 fake users

  - Predicts:
    - Birth Month
    - Gender
    - Current Age
    - Subject

RANDOM USER GENERATOR

A free, open-source API for generating random user data. Like Lorem Ipsum, but for people.

Follow us @randomapi

Hi, My name is
Ted Wood

Demo Shiny-app

- Red line in Histogram shows where input name locates
- Random Name Generator included (randomuser api)

Generate Random Name

Name: Manuella Moens
Gender: Female
DOB: 1948-10-14
Age: 76

Generate

Name-Based Predictor

Enter a name:

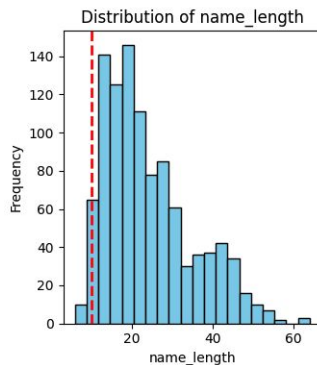Yuki Urata

Select mode:

birth_month

Predict!

Reset

Prediction

7.0

Select variable to plot:

name_length

Similar Name: ZHOU SHUAI

Distribution of name_length

Improvements

- Update Original Dataset + Model
  - No need to use API for list data if not up to date
- Use Better Dataset
  - FBI dataset is highly biased, use other non-biased name-included datasets (babynames…)
  - Could have started off using randomuser.me for dataset
- Better UI
  - Change layout, word size, colours to increase readability
- Receive Feedback
  - Create a Feedback window to improve model prediction and see accuracy
  - Use LLM to compare results