ENPM 808W : Data Science

FINAL PROJECT PRESENTATION

Using fake news to predict party, demographics, and more!

Team Members: Garrett Hill, Denesh Narasimman, Prasanna Raghavan

Data Source

“Where” and “What”?

The data is extracted from a fake news survey that was conducted by Buzz Feed in the article “**Most Americans Who See Fake News Believe It, New Survey Says**”. The link for the article and the data is mentioned here:

https://github.com/BuzzFeedNews/2016-12-fake-news-survey

What does our data look like:

The data was in two CSVs file with some of the columns mentioned here:

1. Respondent ID
2. Headline ID (A-K)
3. Whether the headline is one of the five fake headlines or not
4. Order of viewing the headline by the respondent (1-6)
5. Whether the respondent recalls seeing or hearing about the headline before (Yes/No/Unsure)
6. Whether the respondent believed the headline to be accurate or not (very accurate/Somewhat accurate/Not very accurate/Not at all accurate).

To get a general idea of the data, we were also given some supplemental documents describing the variables.

Data Cleanup

What did we do to our data:

1. Initially, our data had a lot of variables which were beyond our understanding. We had to choose the variables that were relevant to us and remove the rest.
2. A new dataframe “clean_headlines” was created to save the data which was processed and cleaned.
3. Here, with the help of some pandas functions like .replace(), .drop(), .get_dummies() we were able to process the raw data to a version which was suitable for us. We also used functions like .apply() and .merge() to merge the raw data and the headline data.
4. In this dataframe, we managed to extract as much information possible of each participant combined, with their observations when compared to the original data which had separate observation for each participant.

What the output of our data cleanup stage is:

1. The output of our data cleanup had only the features that we wanted that were extracted using Principal Component Analysis to highlight the various features of the large raw data. This gave us a correlation matrix between the features and we were able to choose the features with the highest correlation for the project.

Most important predictive components for if a participant will accurately classify fake news

```
array(['DP_GENAGE', 'DP_INCOME', 'DP_USHHI2_der', 'DWD10', 'DWD4',
       'GRID_DWD11_10_DWD11', 'GRID_DWD11_1_DWD11', 'GRID_DWD11_2_DWD11',
       'GRID_DWD11_3_DWD11', 'GRID_DWD11_6_DWD11', 'GRID_DWD11_7_DWD11',
       'GRID_DWD11_8_DWD11', 'GRID_DWD9_1_DWD9', 'GRID_DWD9_3_DWD9',
       'GRID_DWD9_4_DWD9', 'GRID_DWD9_5_DWD9', 'GRID_DWD9_6_DWD9',
       'HADD_ZipCode_US', 'HCAL_AGGLO_CODE_US', 'HCAL_REGION1_US',
       'INDHH1030', 'INDHH1032', 'MARK_START_TIME_DWD',
       'MRK_ORDER_DWD11_5_MRK_ORDER_DWD11', 'MRK_SMPSRC', 'QMktSize_2_1',
       'QMktSize_83_1', 'QUOTA_DWD', 'QUOTA_HEADSLINES_ATOE2', 'STATE',
       'US09KAB16', 'US09KAB19', 'US09KAB_AG_merged01',
       'US09KAB_AG_merged04', 'US09KAB_AG_merged06',
       'US09KAB_AG_merged12', 'US09KAB_AG_merged17', 'US09KAG02',
       'US09KAG05', 'US09KAG06', 'US09KAG09', 'USMAR2', 'Weightvar_y',
       'resp_age_long', 'usmar2_der'], dtype='<U33')
```

```
{'A': ['GRID_DWD11_6_DWD11',
  'GRID_DWD11_8_DWD11',
  'GRID_DWD11_2_DWD11',
  'GRID_DWD11_7_DWD11',
  'GRID_DWD11_3_DWD11'],
 'B': ['MRK_ORDER_DWD11_5_MRK_ORDER_DWD11',
  'GRID_DWD11_6_DWD11',
  'GRID_DWD11_3_DWD11',
  'GRID_DWD11_8_DWD11',
  'GRID_DWD11_10_DWD11'],
 'C': ['GRID_DWD11_10_DWD11',
  'GRID_DWD9_3_DWD9',
  'GRID_DWD9_1_DWD9',
  'GRID_DWD9_5_DWD9',
  'MRK_SMPSRC'],
 'D': ['INDHH1030',
  'MARK_START_TIME_DWD',
  'INDHH1032',
  'DP_USHHI2_der',
  'QMktSize_83_1'],
 'E': ['GRID_DWD11_6_DWD11',
  'GRID_DWD11_1_DWD11',
  'GRID_DWD11_3_DWD11',
  'GRID_DWD9_5_DWD9',
  'QMktSize_2_1'],
 'F': ['US09KAB_AG_merged06',
  'US09KAB_AG_merged12',
  'HADD_ZipCode_US',
  'US09KAB_AG_merged04',
  'QUOTA_DWD'],
 'G': ['GRID_DWD9_6_DWD9',
  'US09KAB_AG_merged01',
  'US09KAG09',
  'usmar2_der',
  'US09KAG06'],
 'H': ['DWD10', 'US09KAB16', 'US09KAB19', 'DP_INCOME', 'US09KAG02'],
 'I': ['GRID_DWD9_4_DWD9',
  'DP_GENAGE',
  'US09KAG05',
  'US09KAG06',
  'GRID_DWD9_1_DWD9'],
 'J': ['QUOTA_HEADSLINES_ATOE2',
  'HCAL_REGION1_US',
  'resp_age_long',
  'DWD4',
  'Weightvar_y'],
 'K': ['US09KAB_AG_merged06',
  'US09KAB_AG_merged17',
  'HCAL_AGGLO_CODE_US',
  'STATE',
  'USMAR2']}
```

Exploratory Data Analysis

What interesting trends did we notice:

- With some quick library based Primary Component Analysis (PCA), we can see that most components have little to no predictive power between each other (correlation)
- Some of the main predictors for “important components” (mostly headline accuracy) can be seen on the following slide

Main Focus

How did our bespoke features compare to the defaults?

- Not very well, our model doesn't do any better than random chance at predicting accuracy based on a given participant's background information
- We predict that this is due to a misunderstanding of certain data columns and what their values actually represent in real world terms

How does party impact belief in fake news?

- So far, it appears that party is not as strong as an indicator for fake news as having voted for Trump is, but party is a strong indication of who voted for Trump
- Notably, both parties believed more than 50% of fake news presented to them

Visualized existing/new features and impacts

- When looking at age, it appears to have a strong correlation with the number of fake news stories believed, with younger people believing fewer fake news stories on average
- Notably state of resident did not significantly impact recall accuracy (gullible people live everywhere…)

Conclusions

- As it turns out, our bespoke k-best feature picking does not produce a well trained model, or at least not nearly as accurate as when we get a library to handle the entire training pipeline
- The library based model gets ~95% accuracy with <3 features when predicting if a respondent will accurately judge a headline as fake or not (more than 3 features will overfit to data)
- Our custom model gets ~50% accuracy with >3 features when predicting if a respondent will accurately judge a headline as fake or not