Data Scientist I Assessment

Josh Jiayang Hu

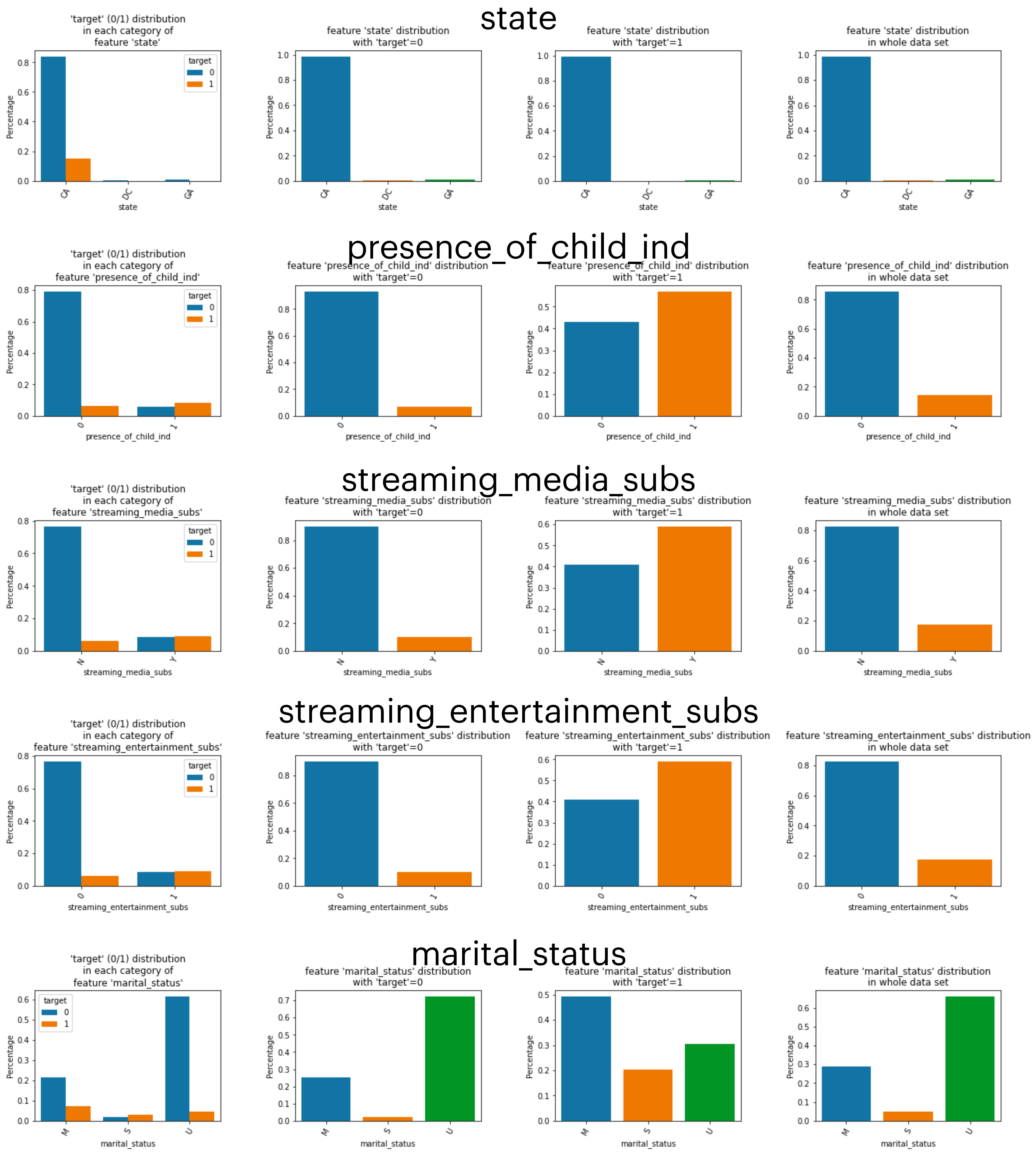
jiayang.hu@columbia.edu

917-657-3860

Table of Contents of the Notebook
As a Guideline

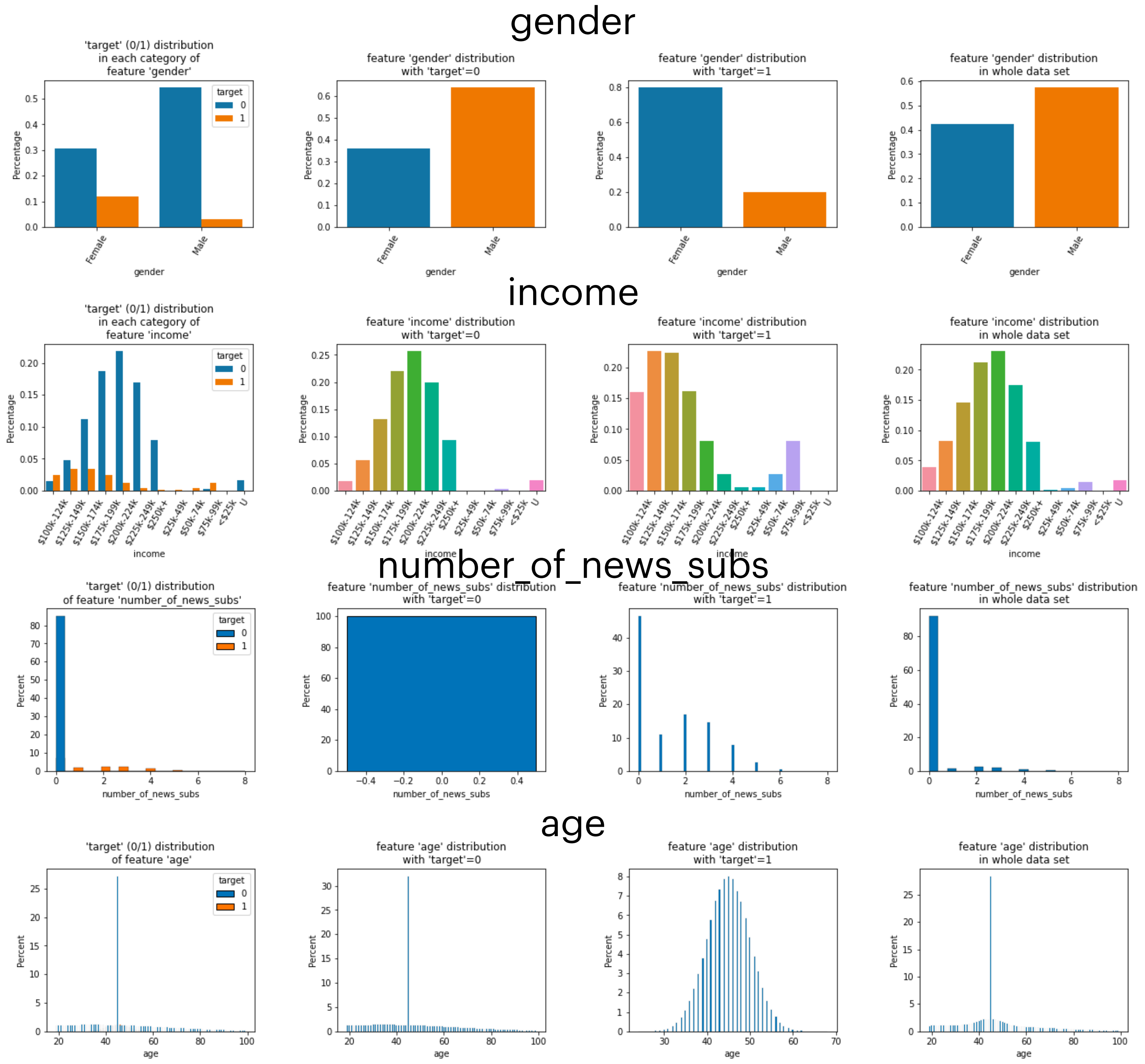
Visualize the data for each feature

1. Distribution of positive/negative cases
2. Distribution of values at negative case
3. Distribution of values at positive case
4. Distribution of values in the whole data set



The original is **imbalanced**.

1. Most data (~99%) collected in CA, which potentially introduces other biases, e.g. race, education, occupation, political leaning, etc.
2. Obvious singularity at age 45 (~28% of total), while other ages are ~1% - 2%
3. 'gender' feature is sensitive, special caution needed to use this feature at decision making. In a more formal business, probably should not include this feature.



For detailed analysis, please see the **Google Colab** Notebook (anyone has this link has access):
https://colab.research.google.com/drive/1ZzO2LIPaI5dpvSAoA9wQN-c1SycUc3Py?usp=sharing

Two major models tested:

- Logistic Regression
- Random Forest Classifier

Model choice:

- Both models good
- Both give ~95% of accuracies for both training data and test data
- For business purposes, choose **logistic regression** due to ease of interpretation.

Metrics for feature importance:

- Coefficients (weight) of logistic regression
- Feature importances (mean reduction of impurity (gini here)) of random forest classifier
- Harder to interpret the meaning behind the reduction of impurity (e.g. is it a more important feature to avoid or include), but good to use as a reference

Data choice:

- Turn off 'state'
- Turn off 'age'
- Keep 'gender' for test purposes, drop in a real formal model

Conclusions:

- The **most important** feature : *'number_of_news_subs'*. Strongly suggests people with news subscriptions are very likely to become targets (listening to podcasts). This can suggest things like *more news content in the future PodNN, or more PodNN commercials on other news media*.
- From the '*income*' feature, people with income lower than $150k are more likely to become non-targets, especially those with income between $100k - $150k. While PodNN may want to target on people with income more than $150k (e.g. *more luxury products commercials*).
- '*marital_status*' has a mild effect. Married people has no effect on the decision. While PodNN may be *interested in single people, and want to avoid those whose marital status unknown*.
- People with children will likely to be the targets. Those with no children will likely not to be the targets. So, PodNN may want to target (marketing-wise or content-wise) on people with children.

Random Forest Classifier
Feature Importance

	feature_category	feature_importance
0	presence_of_child_ind_0	0.091204
1	presence_of_child_ind_1	0.100242
2	streaming_media_subs_N	0.069279
3	streaming_media_subs_Y	0.049826
4	streaming_entertainment_subs_0	0.043851
5	streaming_entertainment_subs_1	0.055638
6	marital_status_M	0.017887
7	marital_status_S	0.033610
8	marital_status_U	0.047532
9	gender_Female	0.044514
10	gender_Male	0.045836
11	income_$100k-124k	0.006444
12	income_$125k-149k	0.007597
13	income_$150k-174k	0.018180
14	income_$175k-199k	0.006529
15	income_$200k-224k	0.004288
16	income_$225k-249k	0.009238
17	income_$250k+	0.006853
18	income_$25k-49k	0.000271
19	income_$50k-74k	0.001194
20	income_$75k-99k	0.004390
21	income_<$25k	0.000018
22	income_U	0.001416
23	number_of_news_subs	0.334164

Logistic Regression
Feature Importance

	feature_category	feature_importance
0	presence_of_child_ind_0	-3.734562
1	presence_of_child_ind_1	3.043066
2	streaming_media_subs_N	-0.985703
3	streaming_media_subs_Y	0.296803
4	streaming_entertainment_subs_0	-0.985703
5	streaming_entertainment_subs_1	0.296803
6	marital_status_M	0.000000
7	marital_status_S	1.536612
8	marital_status_U	-1.537005
9	gender_Female	0.630450
10	gender_Male	-1.321701
11	income_$100k-124k	-3.577537
12	income_$125k-149k	-4.215358
13	income_$150k-174k	3.994005
14	income_$175k-199k	3.538958
15	income_$200k-224k	2.709969
16	income_$225k-249k	1.842327
17	income_$250k+	0.933652
18	income_$25k-49k	-0.762545
19	income_$50k-74k	-1.753740
20	income_$75k-99k	-2.619760
21	income_<$25k	-1.425580
22	income_U	0.481277
23	number_of_news_subs	84.199667

For detailed analysis, please see the **Google Colab** Notebook (anyone has this link has access): https://colab.research.google.com/drive/1ZzO2LIPaI5dpvSAoA9wQN-c1SycUc3Py?usp=sharing