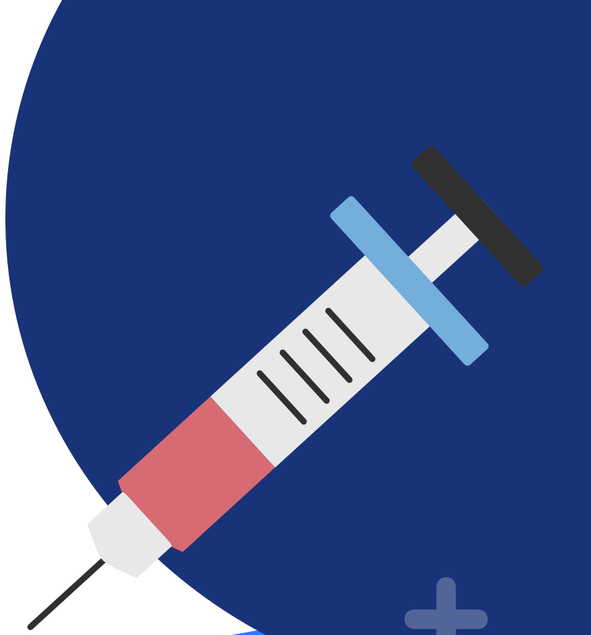
ReDI School of Digital Integration



Neuro Shot:
A Flu Shot Learning Project: Predict H1N1 and Seasonal Flu Vaccines

By:
Sergii Makhin
Diana Paola Americano Guerrero

Problem

- Data comes from the National 2009 H1N1 Flu Survey in the United States, so results should be considered within the context back then.
- However the importance of vaccination remains, especially in this post-pandemic world with anti-vaccination movements.
- Through this analysis we seek to know what factors (demographical, geographical, behavioral) influence vaccine uptake and can we use them to predict real world data.

ReDI School of Digital Integration

Training set:
26,707 rows, 36 features
Labels:
26,707 rows, 2 targets (`h1n1_vaccine`, `seasonal_vaccine`)
Test set:
26,708 rows, 36 features
Features include:
behavioral, opinion-based, demographic, and household information.
Target variables:
H1N1 vaccine: (21% vaccinated vs 79% not vaccinated).
Seasonal vaccine: (46% vaccinated vs 54% not vaccinated)
No meaningful outliers.

Source: DataDriven (2020) https://www.drivendata.org/competitions/66/flu-shot-learning.

Exploratory Data Analysis – 1

Demographics

- Age is fairly balanced but slightly skewed towards older groups (65+).
- Sex: More females (59%) than males (41%).
- Race: Majority White (79%), with smaller representation of other groups.
- Income: Most respondents above poverty threshold.
- Employment: Majority employed (54%), ~41% not in labor force, small unemployed group.

Potential Challenges

1. **Class imbalance** for H1N1 vaccine → may require resampling, class weights, or careful metric choice.
2. **High missingness** (~50%) in employment and insurance data → decisions needed: drop, impute, or encode "unknown".
3. **Mixed feature type**s (ordinal, categorical, binary, continuous) → preprocessing pipeline must handle appropriately.
4. **High-cardinality categories** (employment_industry, employment_occupation, hhs_geo_region) → need encoding strategies (target encoding, frequency encoding, etc.).
5. **Correlation among opinion variables** → risk of multicollinearity, might need dimensionality reduction or feature grouping.

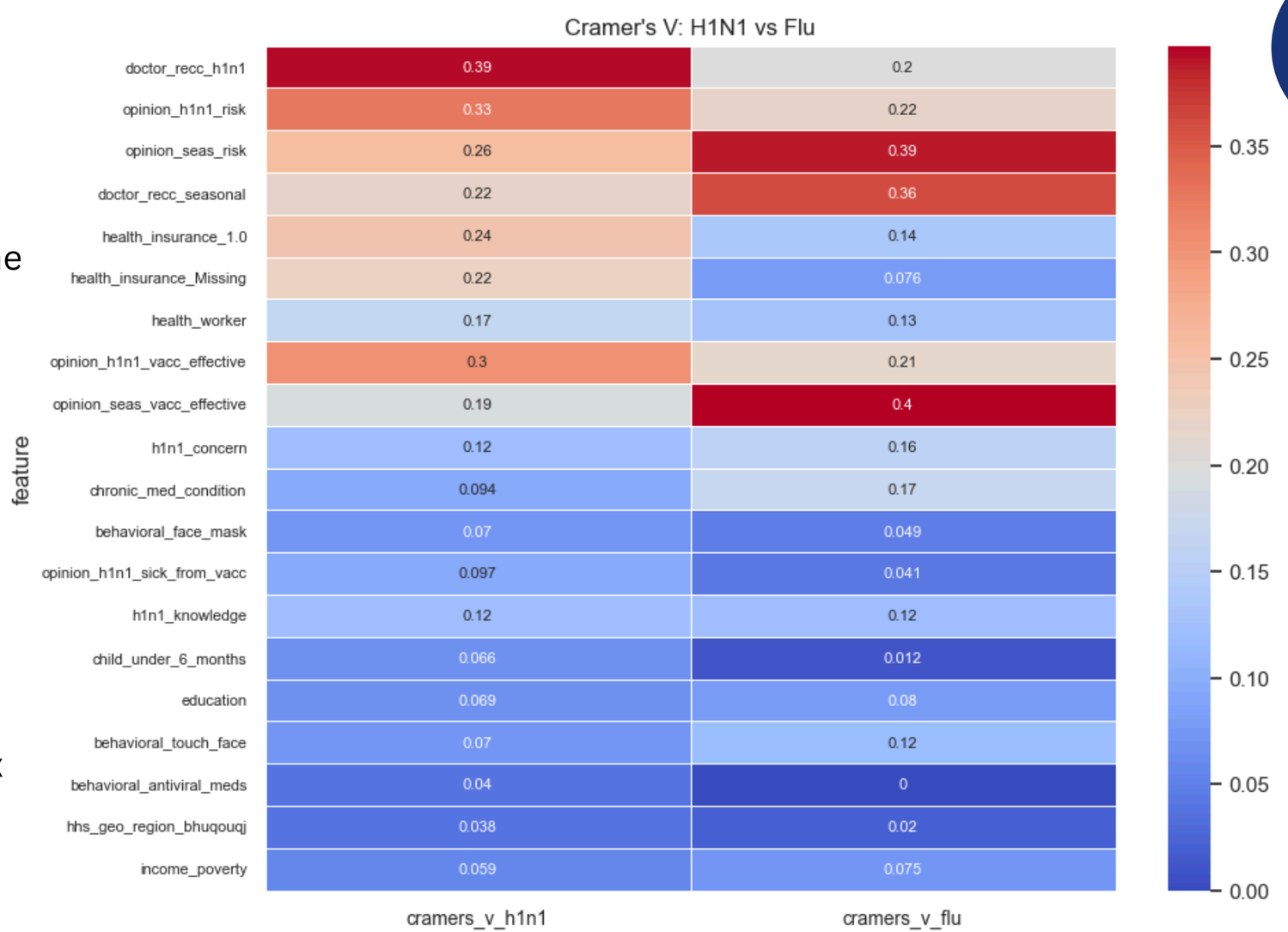
ReDI School of Digital Integration

Exploratory Data Analysis – 2

Feature Analysis

- Doctor recommendation is a shared important feature for both vaccines.
- Perceived personal risk and belief in vaccine effectiveness are also highly associated, health insurance also plays a moderate association.
- Demographics and geography have low association compared to opinion and behavior.
- Perceptions about one vaccine might influence uptake of the other.

→ Next step for feature selection is a model-based feature selection utilizing more complex models (ex. gradient boosting).



ReDI School of Digital Integration

ReDI School of Digital Integration

Thank You.

For Your Attention

https://github.com/americano-diana/neuro_flushot/