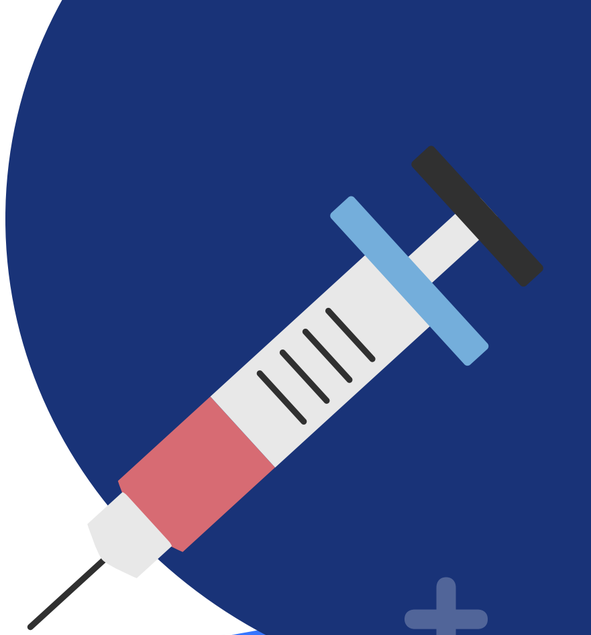
ReDI School of Digital Integration



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

By:
Sergii Makhin
Diana Paola Americano Guerrero

Berlin, Germany
09/10/2025

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.

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

Data Overview

- Train features: 26,707 × 36 | Test features: 26,708 × 36
- Labels: 26,707 × 2 → h1n1_vaccine, seasonal_vaccine
- Feature mix: behavioral & opinions (ordinal/binary), demographics & household, geography, employment

Targets (class balance)

- H1N1: 21.25% yes / 78.75% no
- Seasonal: 46.56% yes / 53.44% no

Missingness (key patterns)

- High: employment_industry (~50%), employment_occupation (~50%), health_insurance (~46%)
- Mid: income_poverty (~16%), education, marital_status, employment_status, rent_or_own (5–10%)
- Paired gaps: doctor_recc_h1n1 & doctor_recc_seasonal (~8%)
- Takeaway: missingness is informative (non-random), especially across socio-economic items

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.

Main Insights / Results

Data quality & distributions

- Ordinal scales and binaries sit in expected discrete buckets; no meaningful outliers
- Household counts capped at 0–3 (3 = “3+”); no extremes

Encodings & artifacts

- Created **aligned** train_encoded / test_encoded (same columns)
- High-card features (employment_industry, employment_occupation) **target-encoded per target** (K-Fold, no leakage)
- Saved correlation heatmap with hierarchical clustering → reports/figures/correlation_heatmap_clustered.png

Correlation & redundancy

- No pairs ≥ 0.9; at 0.8 only expected correlations:
  - One-hot siblings (e.g., rent_or_own_*, health_insurance_*)
  - Missingness indicators co-occurring (marital_status_Missing ↔ employment_status_Missing)
  - Target encodings across the two targets (*_te_h1n1 ↔ *_te_seasonal)
- Decision: keep all features; none are truly redundant in meaning

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.

ReDI School of Digital Integration

Key Steps & Decisions (WHAT and WHY)

Imputation (WHAT)

- Categorical gaps → "Missing" label
- doctor_recc_* (binary) → 0
- Ordinal/numeric (incl. household) → median
- **WHY**: preserve signal from non-response; keep scales sensible; avoid bias

Leakage control (WHAT)

- Medians fit on train then applied to test
- Target encoding via K-Fold means (per target)
- **WHY**: prevent look-ahead; generalize honestly

Encoding (WHAT)

- Ordinal mapping: age_group, education, income_poverty
- One-hot: small/medium nominal (sex, marital_status, rent_or_own, health_insurance, race, employment_status, census_msa, hhs_geo_region)
- Target encoding: high-card (employment_industry, employment_occupation) → *_te_h1n1_vaccine, *_te_seasonal_vaccine
- **WHY**: model-agnostic, low dimensionality, retain target signal

Redundancy policy (WHAT)

- Review at $|r| \geq 0.8$; keep expected pairs (one-hot siblings, missingness, cross-target TE)
- **WHY**: correlated ≠ redundant; pairs encode different semantics

ReDI School of Digital Integration

Feature Ranking Analysis

H1N1 Vaccine

- Doctor recommendation for H1N1 Vaccine → strongest predictor ($\chi^2 \approx 3308$, Cramér's $V \approx 0.39$).
- Perceived risk of H1N1 → high ($\chi^2 \approx 1914$, $V \approx 0.33$).
- Perceived risk of seasonal flu → moderate ($\chi^2 \approx 1222$, $V \approx 0.26$).
- Doctor recommendation for seasonal flu → $\chi^2 \approx 892$, $V \approx 0.22$.
- Health insurance status → moderate ($\chi^2 \approx 840–729$, $V \approx 0.22–0.24$).
- Beliefs in H1N1 vaccine effectiveness ($\chi^2 \approx 496$, $V \approx 0.30$) & seasonal vaccine effectiveness ($\chi^2 \approx 243$, $V \approx 0.19$).
- Being a health worker → moderate ($\chi^2 \approx 672$, $V \approx 0.17$).
- Personal concern about H1N1 → weak ($\chi^2 \approx 201$, $V \approx 0.12$).

- Geography, household composition, and race → near zero Cramér's V, very low effect.

Seasonal Flu Vaccine

- Perceived risk of seasonal flu → strongest predictor ($\chi^2 \approx 2795$, $V \approx 0.39$)
- Doctor recommendation for seasonal flu → very strong ($\chi^2 \approx 2422$, $V \approx 0.36$).
- Age group → older adults more likely to get vaccinated ($\chi^2 \approx 1370$, $V \approx 0.29$).
- Belief in seasonal vaccine effectiveness) → very strong ($\chi^2 \approx 991$, $V \approx 0.40$).
- Opinion of H1N1 risk → moderate ($\chi^2 \approx 866$, $V \approx 0.22$).
- Doctor recommendation for H1N1 → somewhat relevant ($\chi^2 \approx 840$, $V \approx 0.20$).
- Chronic medical conditions → moderate ($\chi^2 \approx 558$, $V \approx 0.17$).
- Being a health worker → weaker influence than for H1N1 ($\chi^2 \approx 384$, $V \approx 0.13$).

- Geography, child under 6 months, antiviral medication use → very low χ^2 and Cramér's V.

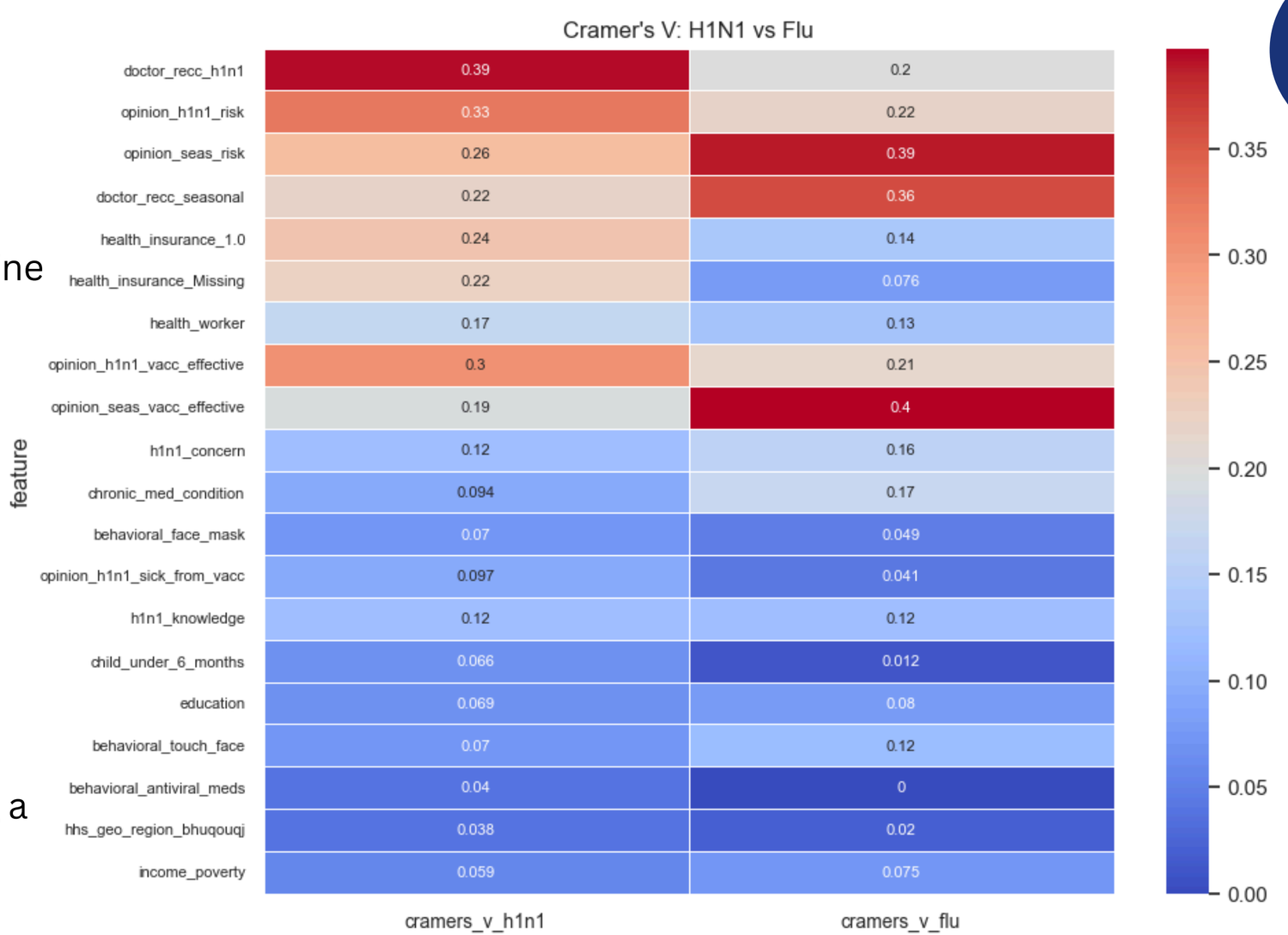
ReDI School of Digital Integration

Feature Ranking Analysis

Conclusions

- 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 seem to influence uptake of the other (such as doctor recommendation and risk).

→ Next step for feature analysis is to try using a more complex feature selection model.



ReDI School of Digital Integration

ReDI School of Digital Integration

Thank You.

For Your Attention

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