Predicting Seasonal Flu Vaccination

Tajesvi Bhat, Yasmine Hejazi, Emily Huang, Matt Lauritzen

Motivation & Data

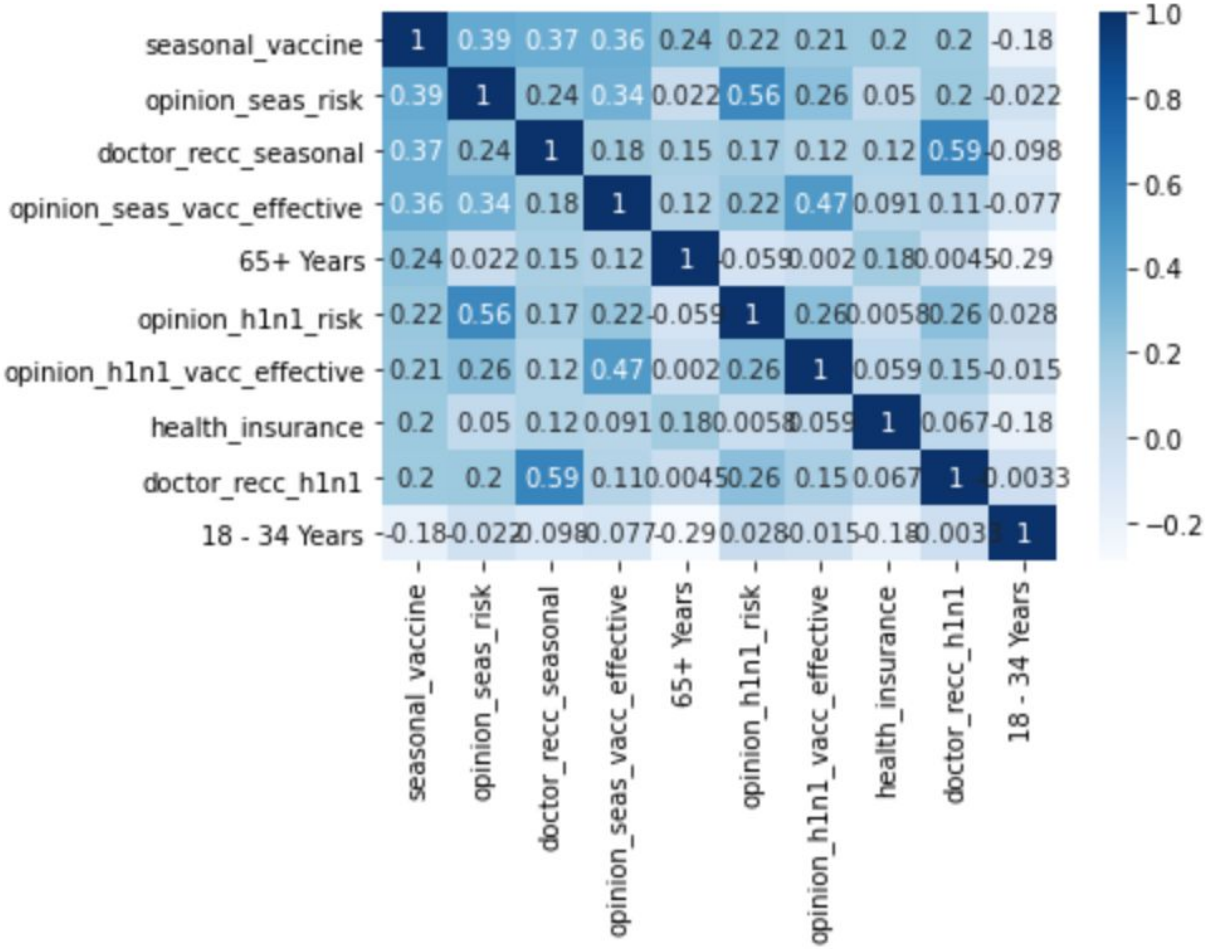
Motivation

- **Research Question: Can we predict whether people get the seasonal flu vaccine using information they shared about their backgrounds, opinions, and health behaviors?**
- Why?
  - Understanding characteristics vs personal vaccination patterns can provide guidance for future public health efforts. This space is becoming more popular given the Covid-19 pandemic and there are existing efforts to use machine learning in order to predict both vaccine efficacy and vaccination likelihood.
- Approach:
  - We chose to use a logistic regression model in addition to random forest classifier which would look into the top correlated features. The performance and accuracy of our results would be evaluated using Area under ROC and F1 scores. We fine tuned our hyperparameters in order to see which combinations on our model would yields the best results.

Data

- The data comes from the National 2009 H1N1 Flu Survey (NHFS)
- Size of dataset: 26,707 records with Each row = one survey respondent
- Top 10 Correlated Features:
    a. Opinion on risk without seasonal vaccine
    b. Seasonal vaccine was recommended by doctor
    c. Opinion on seasonal vaccine effectiveness
    d. Opinion on risk without H1N1 vaccine
    e. Opinion on H1N1 vaccine effectiveness
    f. Has health insurance
    g. H1N1 vaccine was recommended by doctor
    h. Has a chronic medical condition
    i. Level of concern about H1N1
    j. Is a healthcare worker
    k. Age group

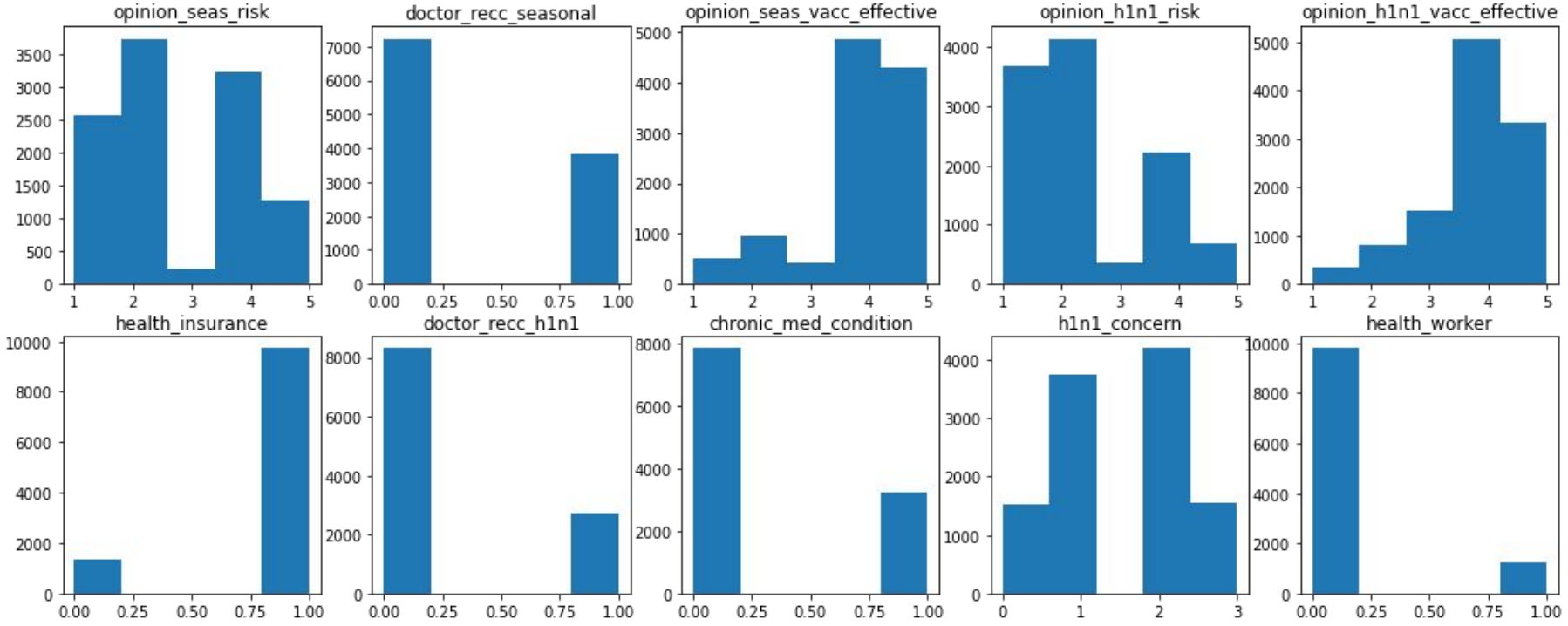
Data (continued)

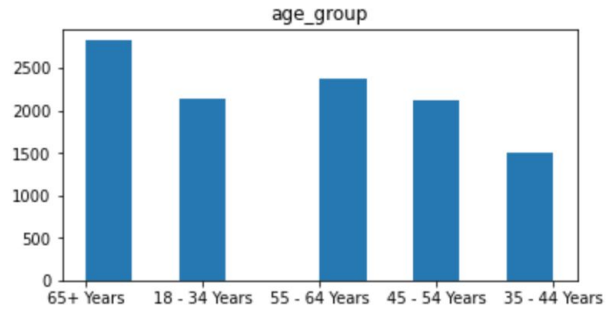



```
seasonal_vaccine               1.000000
opinion_seas_risk              0.390106
doctor_recc_seasonal           0.369190
opinion_seas_vacc_effective    0.361875
65+ Years                      0.244830
opinion_h1n1_risk              0.216625
opinion_h1n1_vacc_effective    0.205072
health_insurance               0.200858
doctor_recc_h1n1               0.198607
18 - 34 Years                 -0.178786
Name: seasonal_vaccine, dtype: float64
```

Summary Statistics

Feature Name	Type	Min	Max	Mean	# NaN
opinion_seas_risk	Ordinal	1	5	2.72	514
doctor_recc_seasonal	Binary	0	1	0.33	2160
opinion_seas_vacc_effective	Ordinal	1	5	4.03	462
65+ Years	Binary	0	1	0.26	0
opinion_h1n1_risk	Ordinal	1	5	2.34	388
opinion_h1n1_vacc_effective	Ordinal	1	5	3.85	391
health_insurance	Binary	0	1	0.88	12274
doctor_recc_h1n1	Binary	0	1	0.22	2160
chronic_med_condition	Binary	0	1	0.28	971
h1n1_concern	Ordinal	0	5	1.62	92
health_worker	Binary	0	1	0.11	804





Distribution of values for top 10 correlated features.

Summary of Results

Model	Train F1 Score	Test F1 Score	Train AUROC	Test AUROC
Logistic Regression 1	0.768	0.7691	0.7675	0.7685
Logistic Regression 2	0.7976	0.8069	0.7983	0.8045
Random Forest	0.9024	0.7993	0.9022	0.802

Approach & Experiments

Approach

Model 1: Baseline

- Baseline model using the distribution of values for the binary `seasonal_vaccine`: not vaccinated (0) and vaccinated (1). This gave us a baseline accuracy of 53.5% and 46.5% respectively.

Model 2: Logistic Regression 1

- First logistic regression with top 10 correlated numerical features, limited feature engineering

Model 3: Logistic Regression 2

- Logistic regression iteration with all numeric features. This iteration gave us the highest F1 score of 0.8311. This model converted each categorical feature into multiple one-hot encoded features and included all original numeric features.

Model 4: Random Forest

- Random Forest with top 28 important features (feature importance > 0.01). Hyperparameter tuning for numTrees and maxDepth

Evaluation Metrics

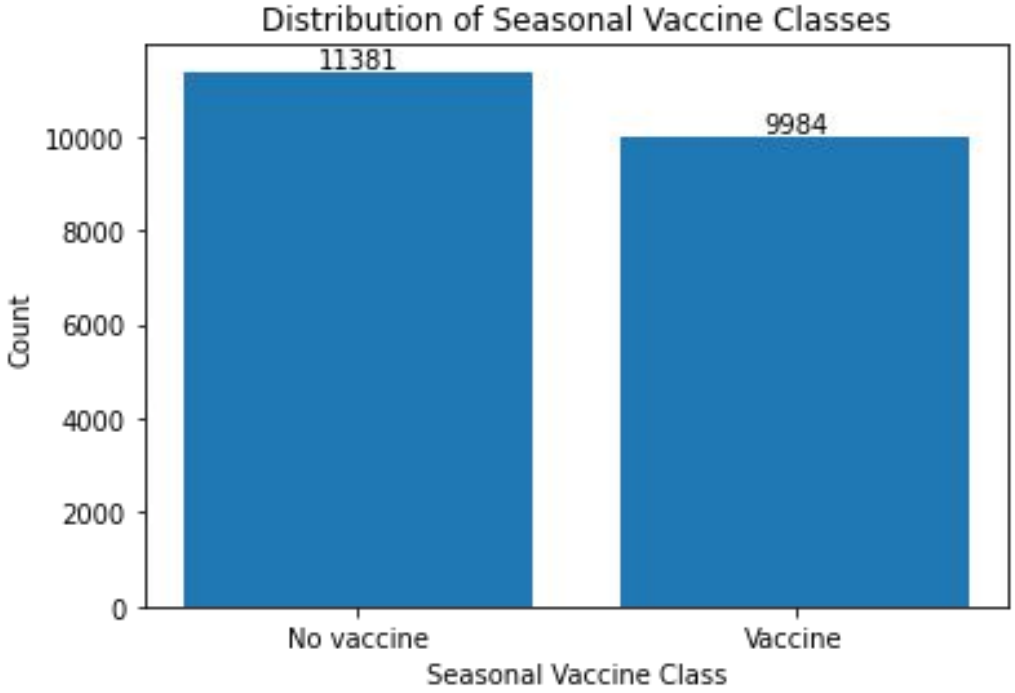
- ROC AUC
  - Competition metric. Higher value = stronger performance
- F1 Score
  - A balanced score of precision and recall; good for measuring imbalanced datasets. Higher score = stronger performance

Model 1: Baseline

Predict majority class for all examples.

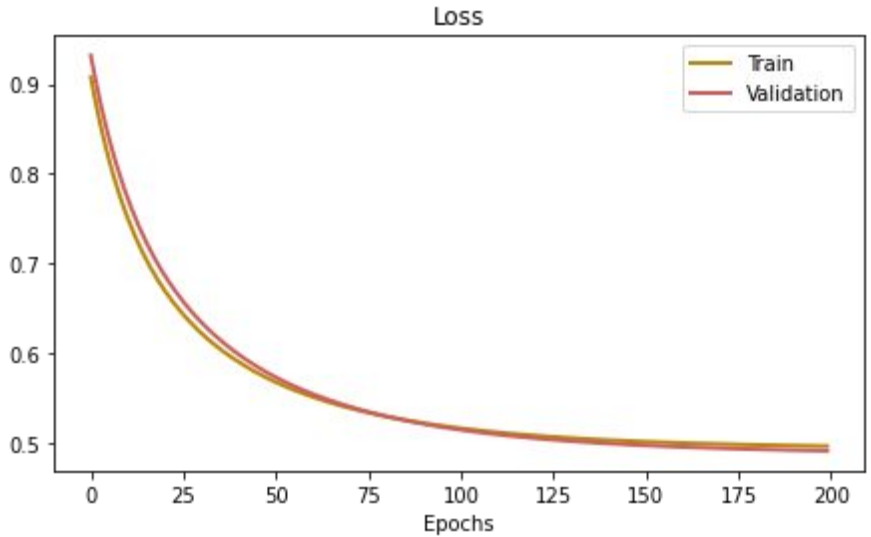
Majority Class = No vaccine (0)

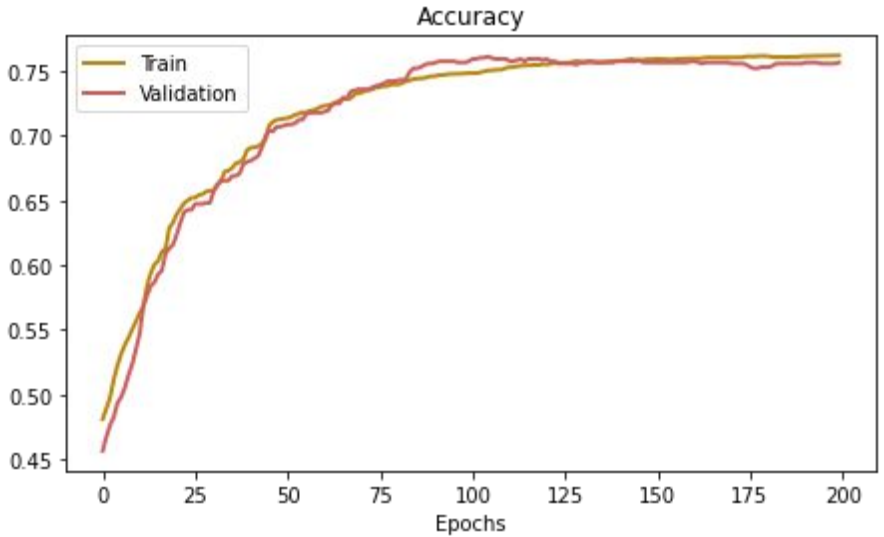
- Accuracy: 0.5412
- AUROC: 0.5
- F1 Score: 0



Model 2: Logistic Regression #1 (10 Features)

Baseline logistic regression model with top 10 correlated numerical features





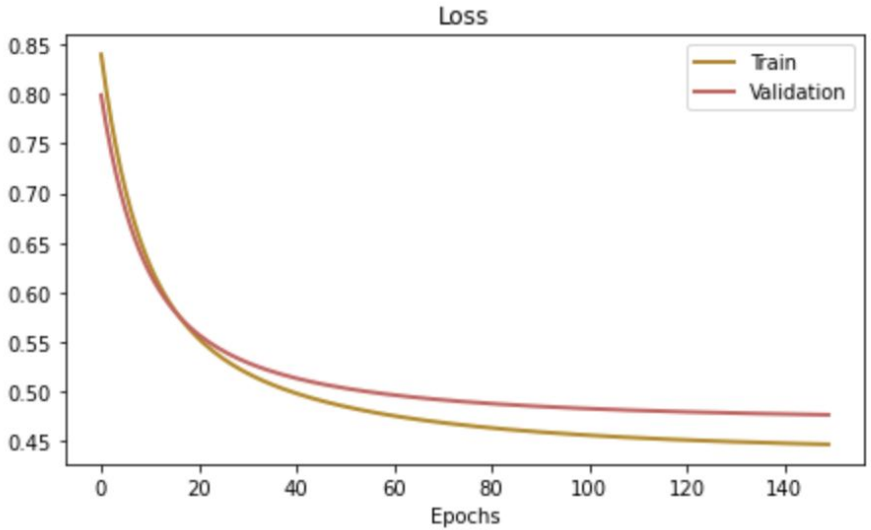
Model 2: Logistic Regression #1 (10 Features)

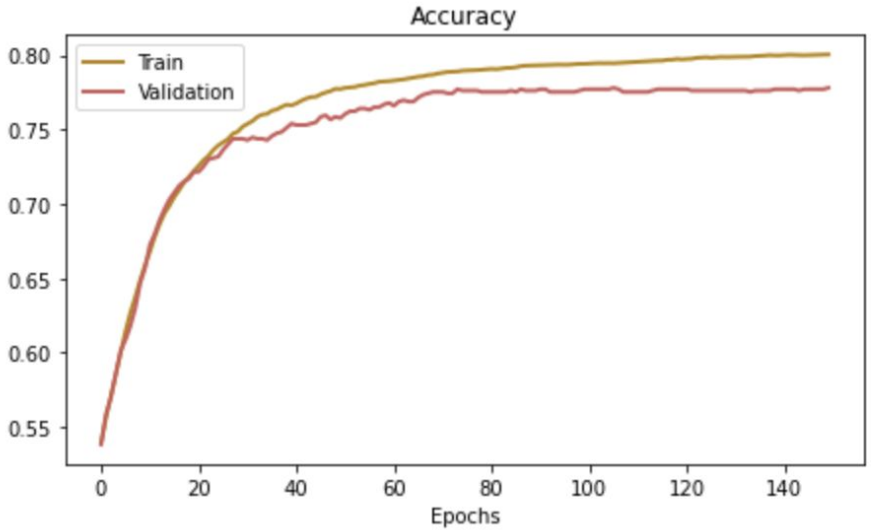
Baseline logistic regression model with top 10 correlated numerical features

F1 Score	AUROC	Learning Rate	# Epochs	Optimizer
0.7587	**0.7554**	0.1	100	SGD
0.7661	**0.7656**	0.1	150	SGD
0.7691	**0.7685**	0.1	200	SGD
0.769	**0.7681**	0.1	300	SGD
0.7686	**0.7677**	0.1	200	Adam

Model 3: Logistic Regression #2 (all features)

Baseline logistic regression model with all features, including one-hot encoding





Model 3: Logistic Regression #2 (all features)

Baseline logistic regression model with all features, including one-hot encoding

F1 Score	AUROC	Learning Rate	# Epochs	Optimizer
0.7958	**0.7934**	0.1	100	SGD
0.8069	**0.8045**	0.1	150	SGD
0.8048	**0.8027**	0.1	200	SGD
0.8057	**0.8030**	0.1	300	SGD
0.8022	**0.8001**	0.01	50	Adam

Model 3: Logistic Regression #2 (all features)

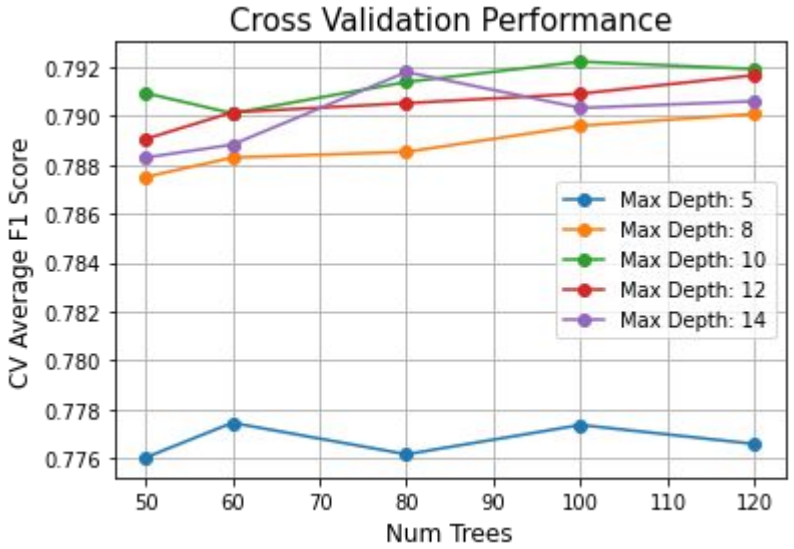
Baseline logistic regression model with all features, including one-hot encoding

Feature	Weight
opinion_seas_risk	0.6180
opinion_seas_vacc_effective	0.5920
doctor_recc_seasonal	0.5775
65+ Years	0.4087
18 - 34 Years	-0.2434
…	…
Not Married	-0.0037

Model 4: Random Forest

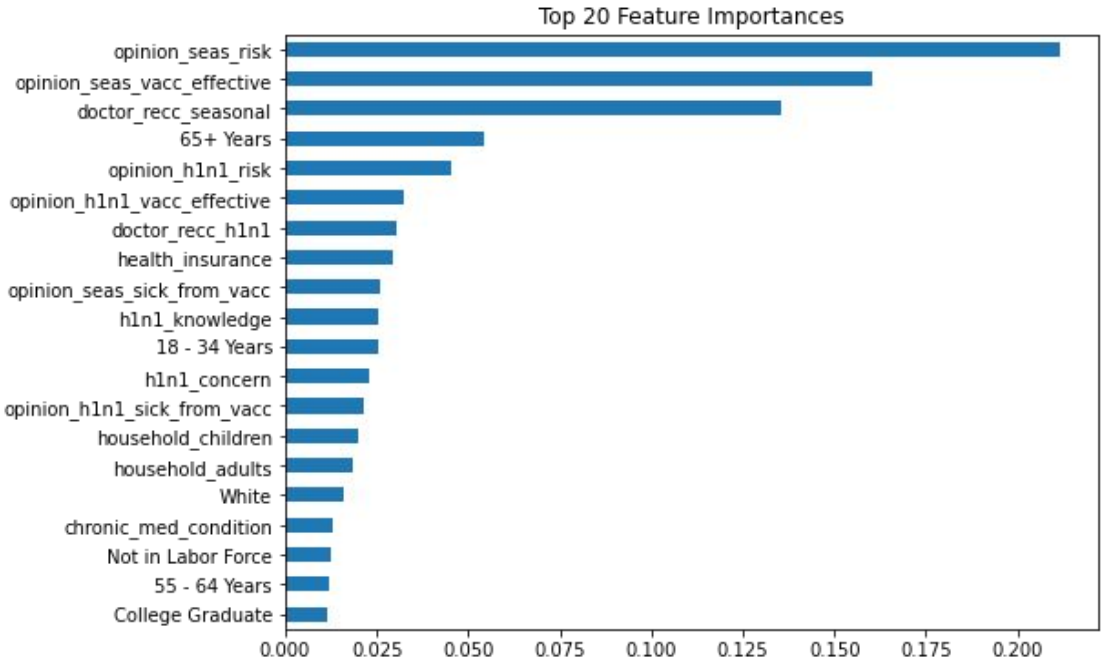
RF model with top 28 features

F1 Score	AUROC	NumTrees	MaxDepth
0.7969	**0.7994**	80	10
0.7958	**0.7987**	80	12
0.7933	**0.7958**	80	14
0.7993	**0.802**	100	10
0.795	**0.7973**	100	12
0.7923	**0.7954**	100	14



Model 4: Random Forest

RF model with top 28 features



Conclusion

Key Results:

- The highest model F1 score and AUROC were obtained from our Logistic Regression Model that took into account all numeric features of the dataset:

Model	Train F1 Score	Test F1 Score	Train AUROC	Test AUROC
Logistic Regression 2	0.7976	0.8069	0.7983	0.8045

What we Learned:

- Predicting user behavior is incredibly complex
- Data quality is as important as algorithm quality!
- Importance of proper documentation for our work!

Avenues for Future Work

- Predict H1N1 vaccination likelihoods
- Increase the number of respondents and records
- Collect data from 2009 till now and perform time series analysis, or compare this model's performance on 2009 data with data from 2022.
- How to handle nulls effectively
- How can this be misused? With the increased relevance and growth of the Anti-Vacc movement, can such models be used against Anti-Vaccers or vice versa?

Thank you for listening!

Any questions?

Contributions and Github Link

Link: [UC-Berkeley-I-School/w207-sec03-bhat-hejazi-huang-lauritzen (github.com)]

	Emily	Matt	Tajesvi	Yasmine
Data Research	X	X	X	X
Data Cleaning	X	X	X	X
Model Setup		X		X
Hyperparameter Tuning		X		X
Presentation Slides	X		X	