Prediction of H1N1 Vaccination Status

as a means to steer public health interventions

Summary

Goals:

Predict whether individuals chose to receive H1N1 vaccination.

Identify targets for future public health campaigns.

Analysis and predictive models are based upon survey results from the 2009 H1N1 Flu Survey.

Takeaways:

- Logistic Regression and Decision Tree models predict with ~85% accuracy
- Regression > Decision Tree models
- Future campaigns: focus on conveying risk, vaccine effectiveness, and doctor recommendations

Outline

- Scientific | Business Problem
- Data
- Methods
- Results
- Conclusions

Scientific Problem

Find a model that accurately predicts whether individuals received the H1N1 vaccine.

Confirm the model's performance in generalizing to unseen data, and explore alternative models and hyperparameters.

Business Problem

With Public Health Agencies as our stakeholders, identify key factors that we can focus on to increase the effectiveness of future vaccination campaigns.

Data

data set: 2009 H1N1 Flu Survey
[26,707 respondents to 35 survey questions]

survey topics:
- behavioral
- opinion
- background

entries: binary, ordinal, categorical

methods: pandas, scikit-learn

target variable: received H1N1 vaccine

Methods

Model Classes:

- Logistic Regression
- Decision Tree

Preprocessing steps:

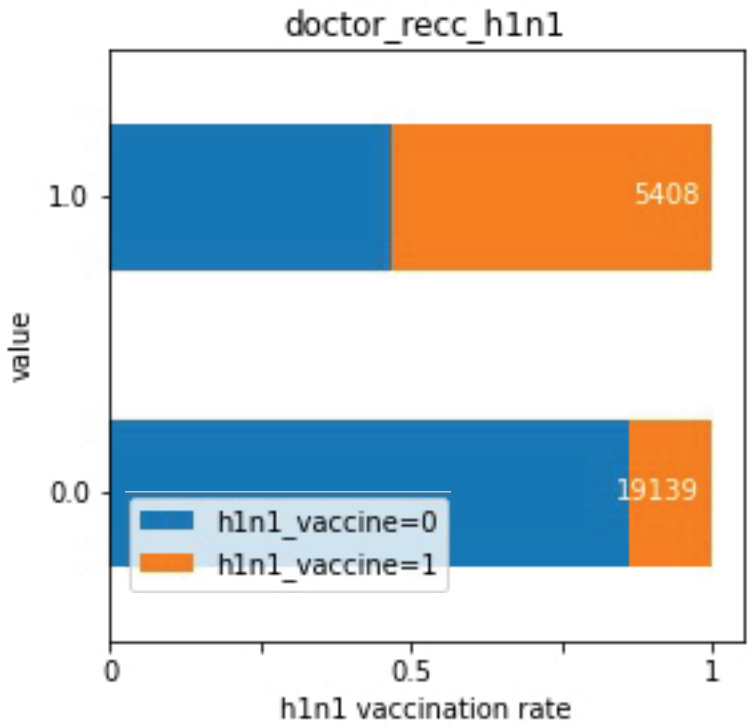
- -categorical variables -> one-hot encodings
- -scaling variables

Cross-validated evaluation of model fits (3-fold)

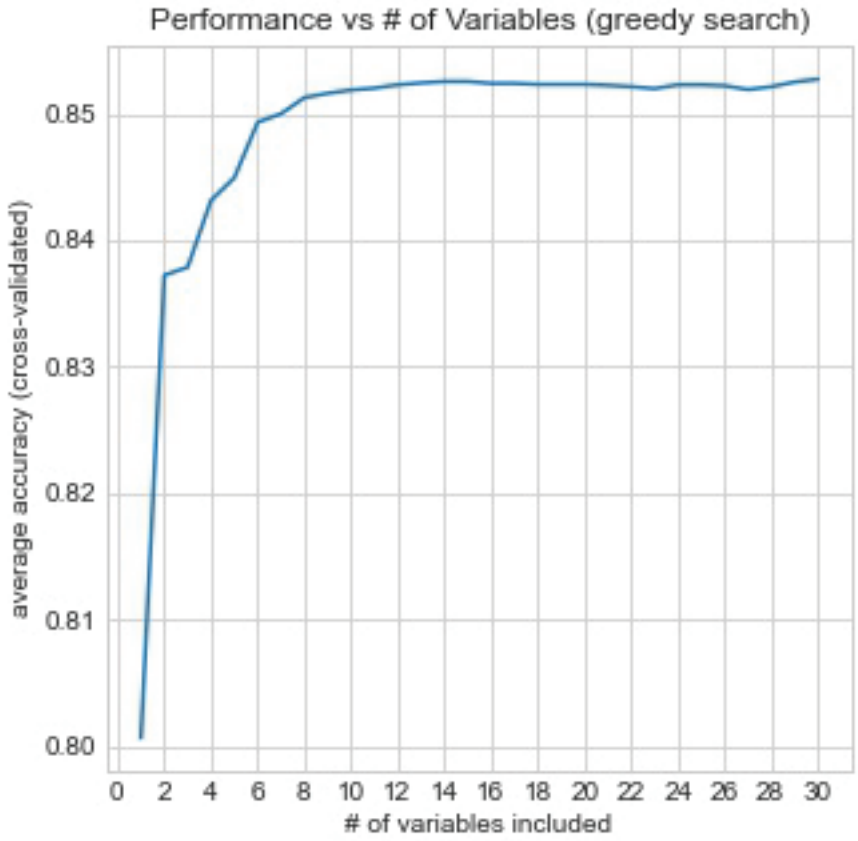
Hyperparameter Tuning

- -Regularization/Fitting (regression)
- -Depth, Features, etc (decision tree)

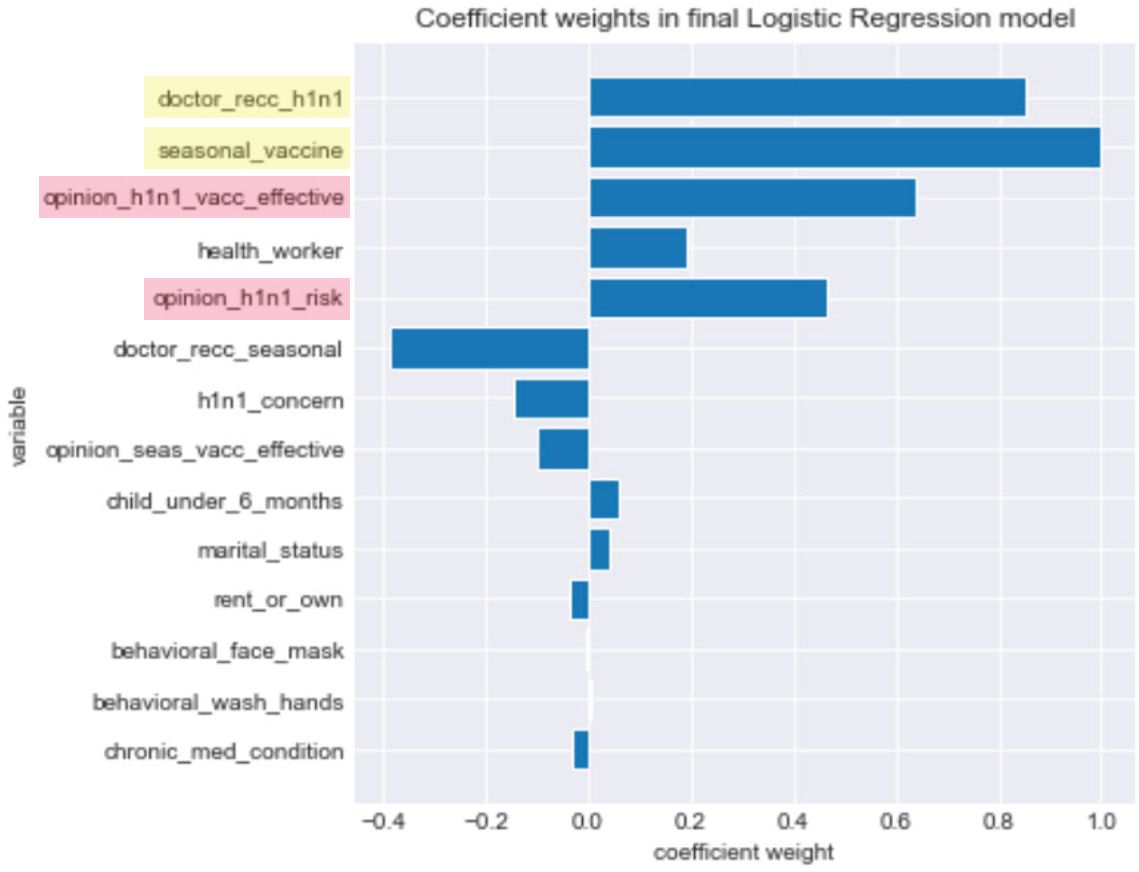
Doctor recommendation is predictive of vaccination



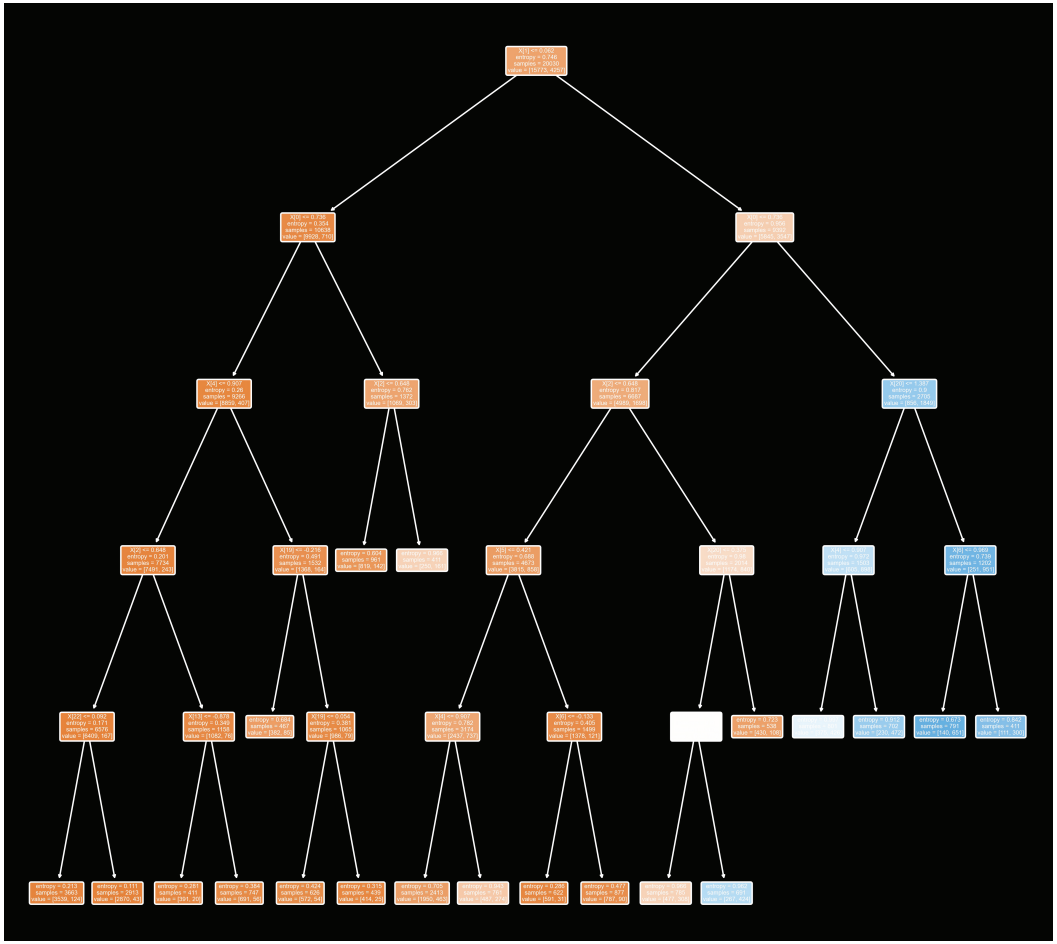
Using 10-14 variables is optimal for Logistic Regression



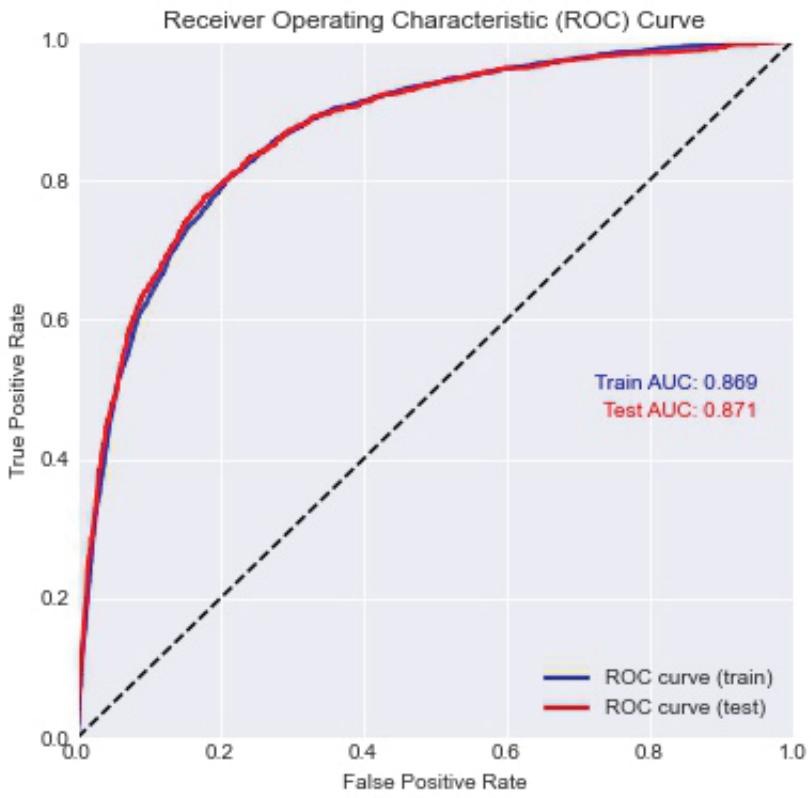
Opinions, Doctor Recommendation, & Seasonal Vaccine are key predictive factors

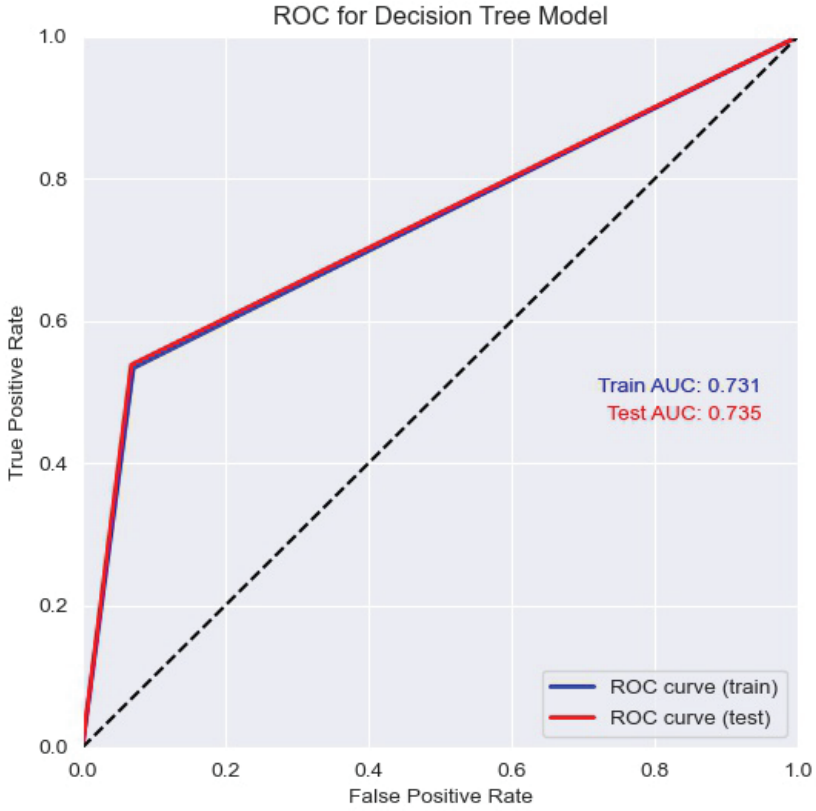


First Decision Tree Nodes Rely on Same Variables



Logistic Regression slightly outperforms Decision Tree





Conclusions

We could predict individual's H1N1 vaccination status with > 85% accuracy.

- Both Logistic Regression and Decision Tree models did well

Future public health campaigns should:

- Focus on getting **doctors to recommend** vaccination
- Encourage individuals to get **seasonal flu vaccines**
- Strengthen communication on **risk** and **vaccine effectiveness**

Thanks for your attention.

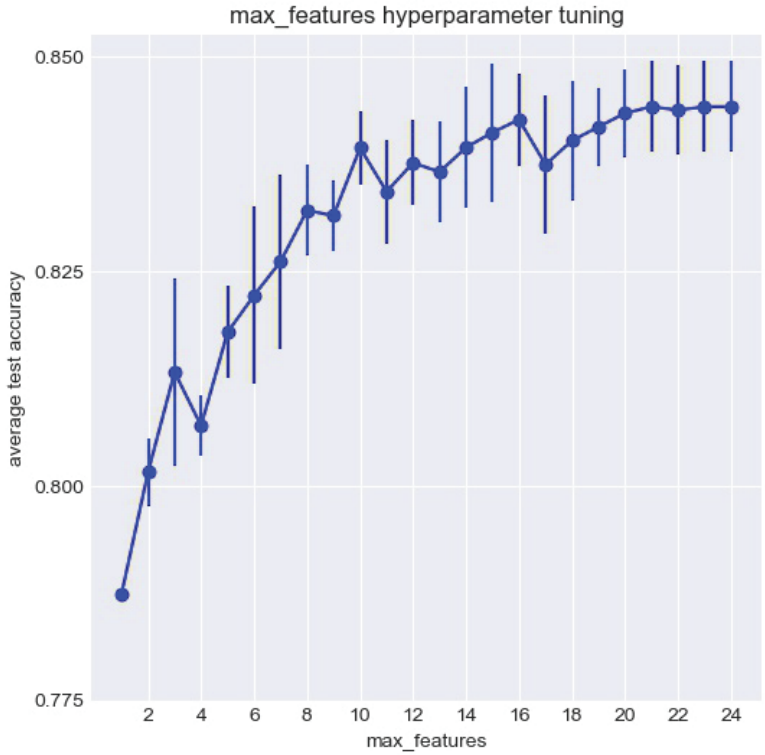
Christopher Henry

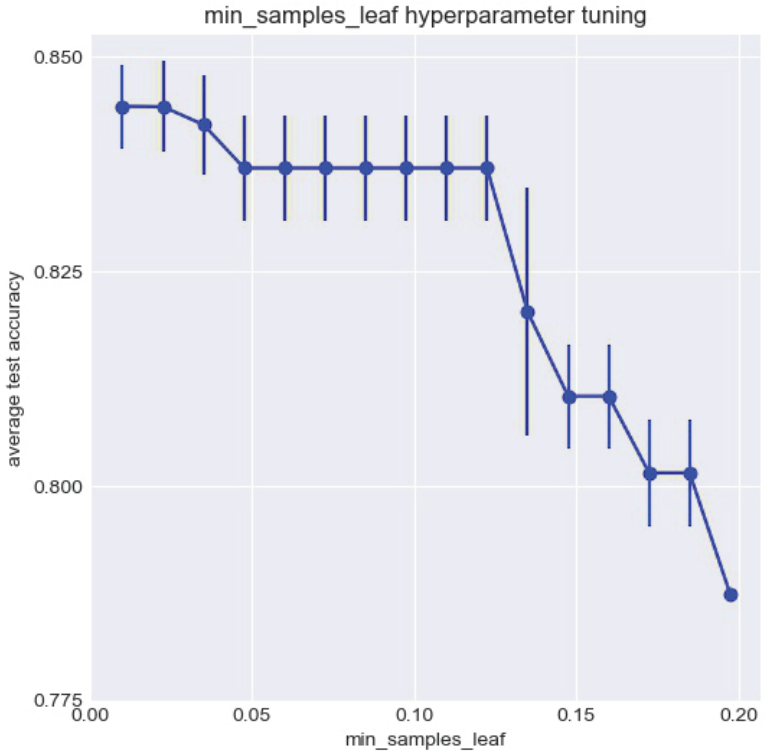
Email: chenrynyc@gmail.com

GitHub: @christopheraaronhenry

LinkedIn: https://www.linkedin.com/in/christopher-a-henry/

Optimal decision tree has many features, small min_samples_leaf





Additional Considerations

- Inclusion of nonlinear transformations of the data (pairwise products) did not further increase accuracy
- Other nonlinear transformations (squaring, higher-order-products) could possibly prove informative
- Stable performance (across models, regularization, and training/test sets) suggests that we may be in a near optimal regime for classifying these data