Prediction of
H1N1 Vaccination Status

Summary

Accurately predict whether individuals chose to receive the vaccination for the H1N1 virus.

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

Takeaways:

- Both Logistic Regression and Decision Tree models yielded accurate predictions (~85% accuracy)
- Both generalized well from training to test data
- With optimized hyperparameters, Regression models had slightly better performance than Decision Tree models

Outline

- Scientific 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.

Data

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

survey topics:
- behavioral: (exposure | infection prevention | medication)
- opinion: (infection risk | vaccine effectiveness)
- background: (sex, race, education, housing situation)

entries: binary (yes/no), ordinal (rating 1-5), categorical (label)

methods: pandas, scikit-learn

target variable: received H1N1 vaccine (yes/no)

Methods

Model Classes:

- Logistic Regression
- Decision Tree

Preprocessing steps:

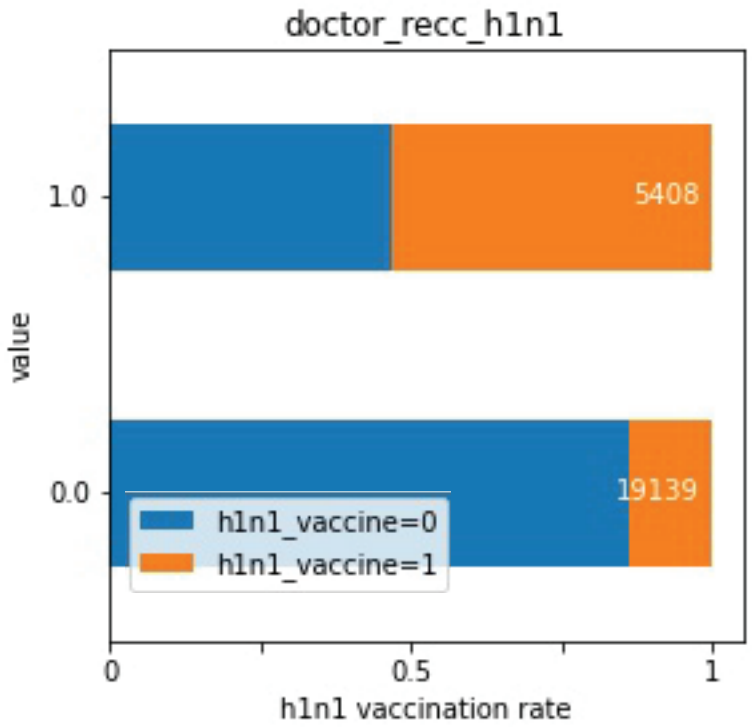
- -convert categorical variables to one-hot numerical encodings
- -scale each variable to have unit variance

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

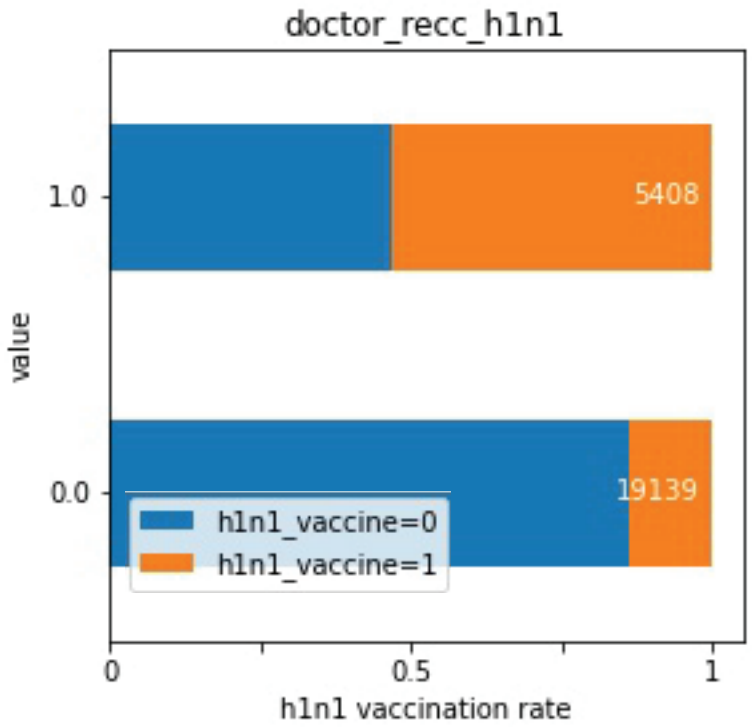
Hyperparameter Tuning

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

Doctor recommendation is predictive of vaccination

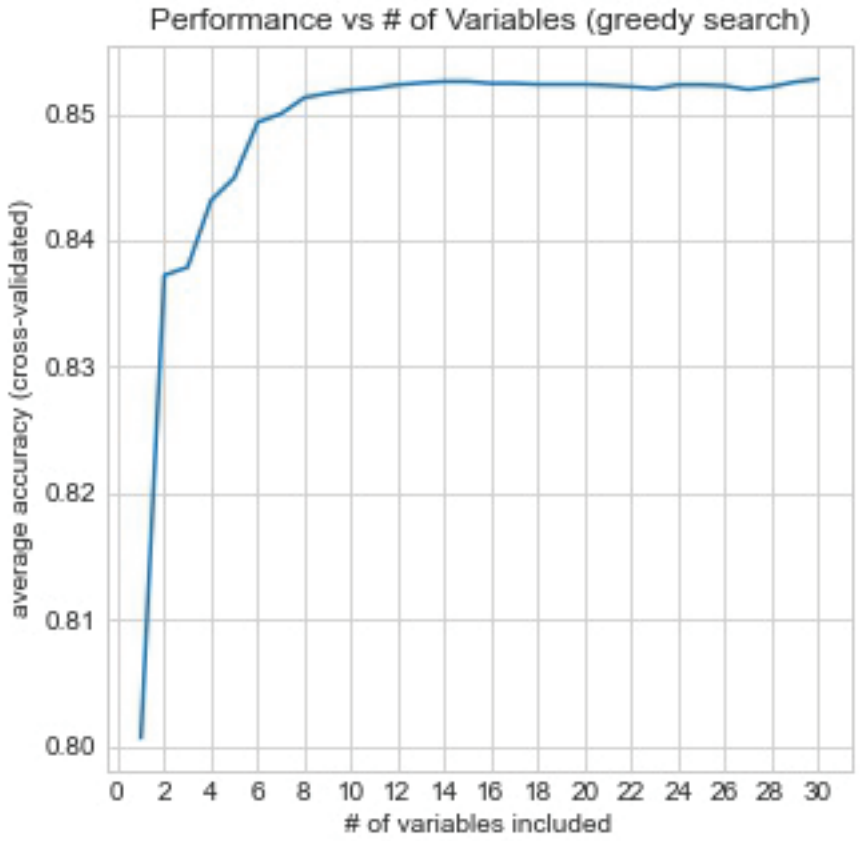


Doctor recommendation is predictive of vaccination

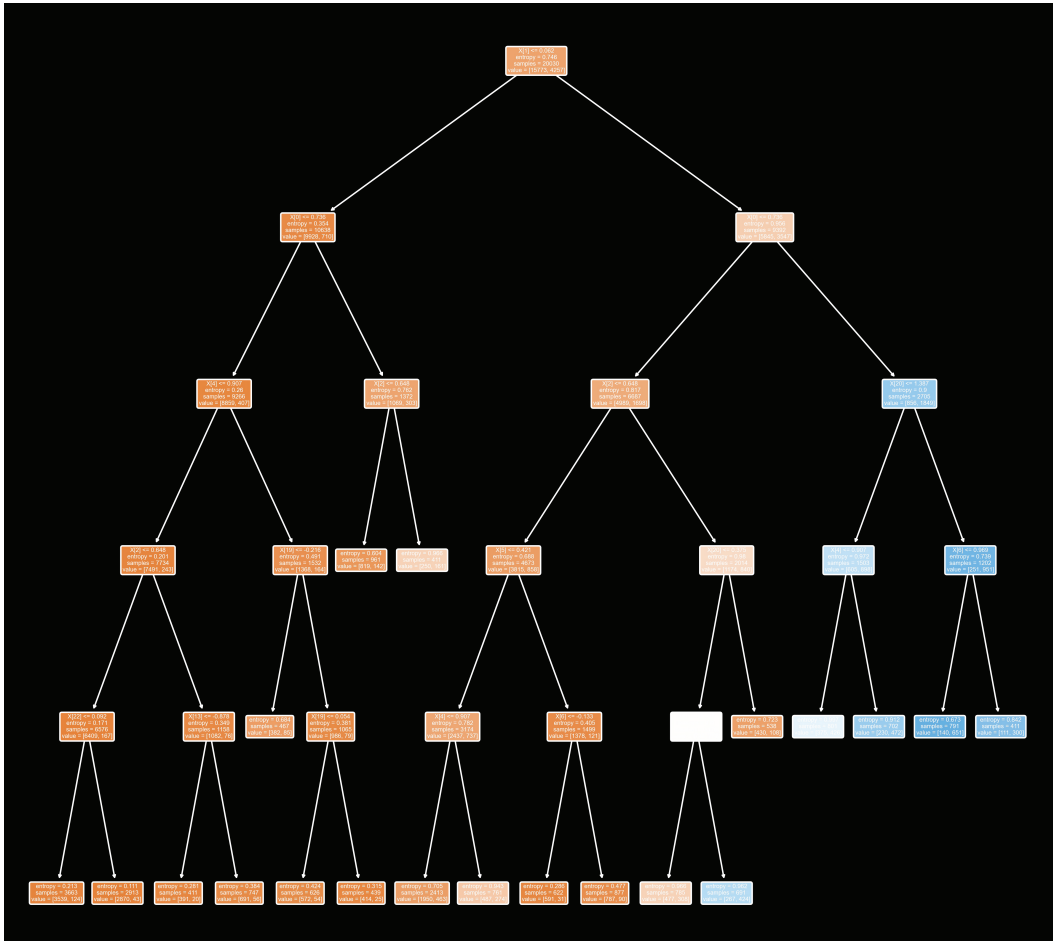


variable	accuracy	accuracyBASE
doctor_recc_h1n1	0.800652	0.787546
opinion_h1n1_risk	0.788969	0.787546
h1n1_concern	0.787546	0.787546
health_worker	0.787546	0.787546
household_children	0.787546	0.787546
household_adults	0.787546	0.787546

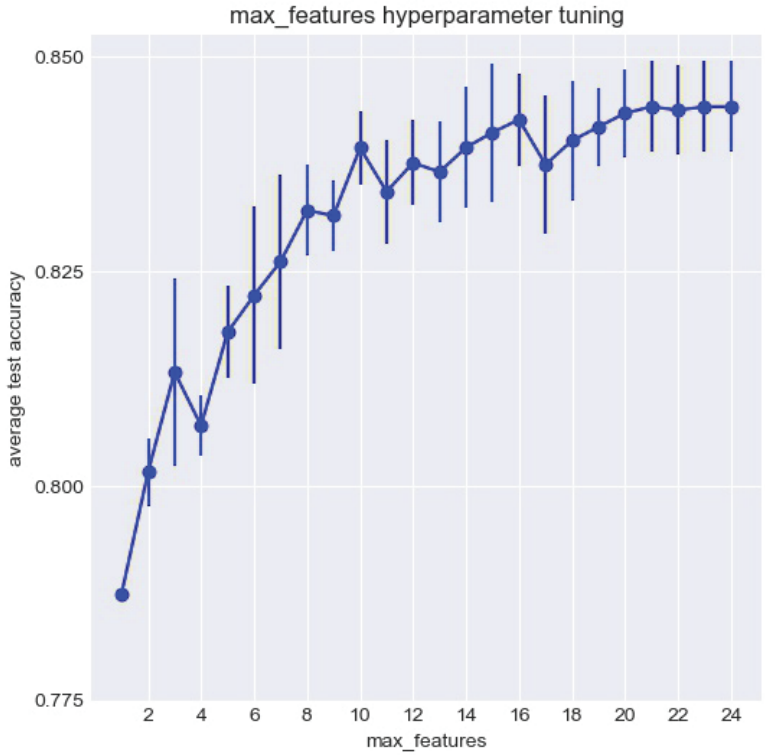
Logistic Regression identifies 10-14 optimal variables

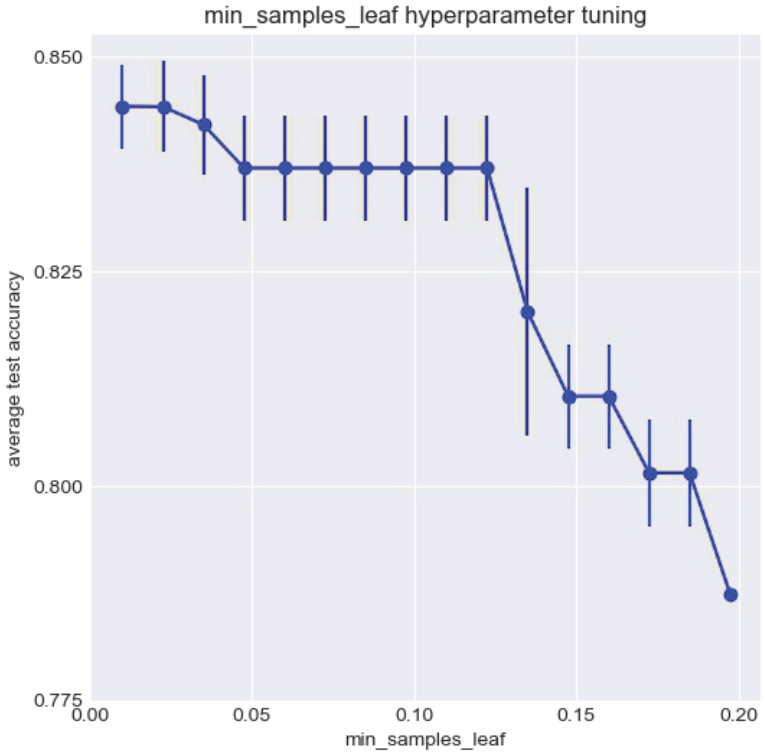


Decision tree optimal model architecture

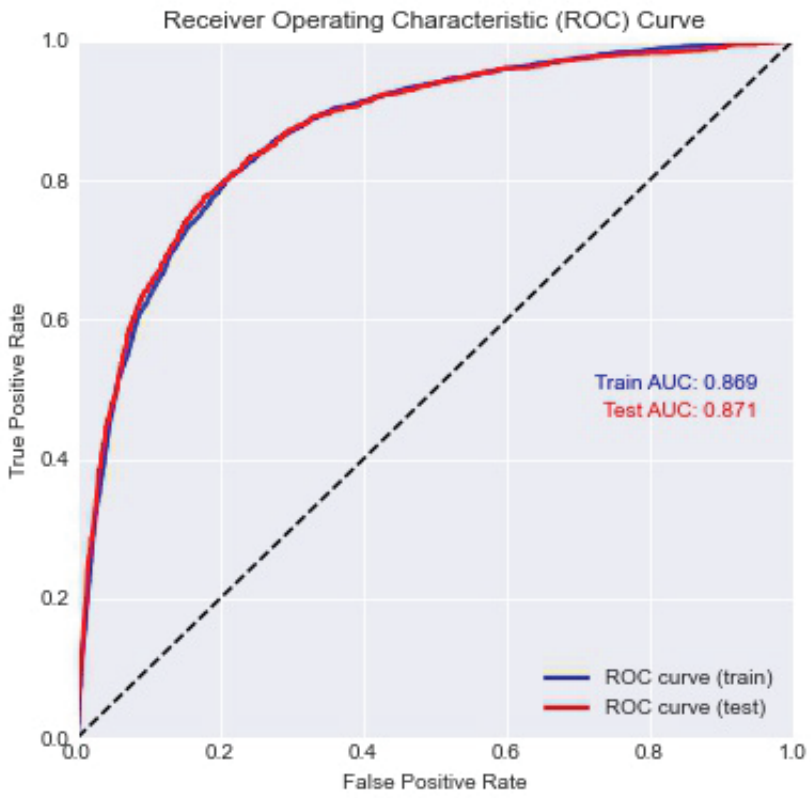


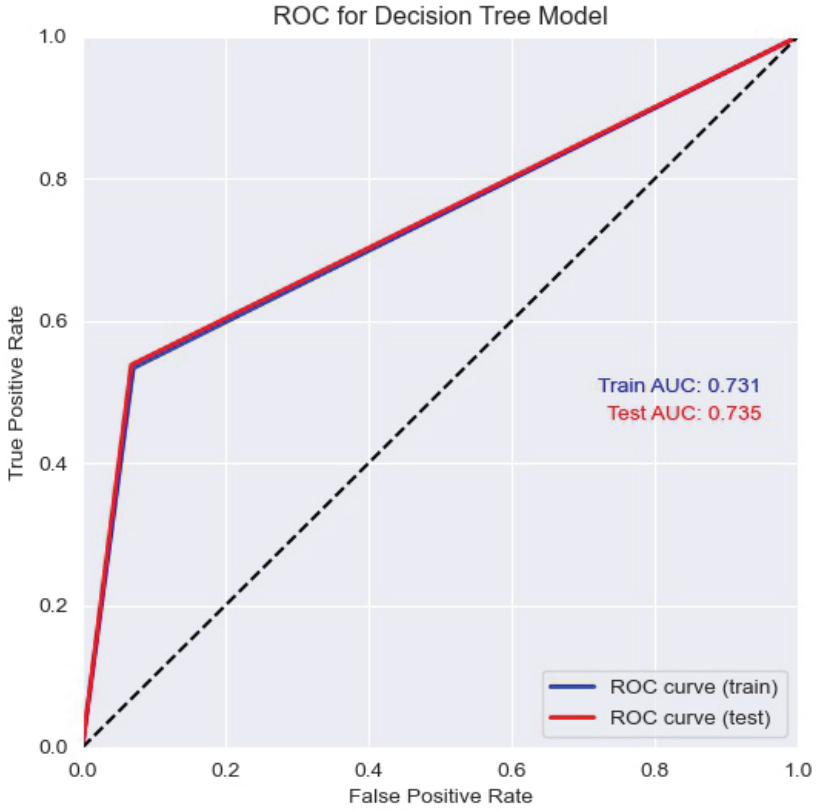
Optimal decision tree has many features, small min_samples_leaf





Logistic Regression slightly outperforms Decision Tree





Conclusions

We were able to predict individual's H1N1 vaccination status with > 85% accuracy (and significantly above the null model).

- Both Logistic Regression and Decision Tree models did well
- Both relied on a few key variables (opinion of H1N1 risk, doctor vaccine recommendation) and performance grew with adding 10-15 variables.
- Hyperparameter tuning led to models that generalized equally well between training/test data, indicating a good fit

Thanks for your attention.

Christopher Henry

Email: chenrynyc@gmail.com

GitHub: @christopheraaronhenry

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

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