Does Model Choice Impact Classification Accuracy for Predicting Flu Vaccinations? Data from a 1-Million-Person Field Study

Jose Cervantez   Bethany Hsaio   Rob Kuan

Due: 11:59pm, May 5th, 2024

Contents

Executive Summary (1 page)

Introduction

Study Goal

Data Description

Methodology

Results

Detailed Analyses

Description of Data

Data variables

- `flu_vax_30_days`: whether the patient received a flu vaccination within 30 days of treatment
- `condition`: different text message content sent to the patient to encourage vaccination
- `day_of_text`: which day the text message was sent (1 of 3 days in September 2023)
- `SMS_twice`: whether the patient received a reminder message
- `flu_vax_previous_season`: whether the patient received a flu vaccination in the previous season
- `age`: the patient's age
- `male`: whether the patient is male
- `female`: whether the patient is female (indicator ommitted)
- `insurance`: the type of insurance that a patient has (e.g., Medicare, Medicaid, etc.)
- `prev_flu_vax_count`: the number of flu vaccinations the patient has received in the past 8 years
- `pharm_visits_last_yr`: the number of visits to the partner pharmacy in the last year where the patient made at least one pickup or transaction
- `last_vax_dow_30_min`: the day of week of the patient's last vaccination (rounded to the last 30 minutes)
- `last_vax_time_30_min`: the time of the patient's last vaccination (rounded to the last 30 minutes)
- `timezone`: the patient's timezone

Exploratory Data Analysis

Predictive Modeling

OLS w/ Classifier

- The OLS w/ Classifier model used an ordinary least squares (OLS) regression to predict the probability of a patient receiving a flu vaccination within 30 days of treatment. The model predictions were then converted to a binary classification (vaccinated or not) using a 50% probability threshold.

- When evaluated on the test set, the OLS w/ Classifier achieved an AUC of 0.763, indicating moderately strong predictive performance. The misclassification error was 0.118, meaning the model incorrectly predicted the vaccination status for about 11.8% of patients. Looking at the confusion table, the model correctly identified 180,113 patients who did not get vaccinated (true negatives) and 4 patients who did get vaccinated (true positives). However, it misclassified 24,031 vaccinated patients as not vaccinated (false negatives).

Logistic Regression

Next, we use logistic regression to predict whether an individual will get vaccinated given their covariates. Logistic regression maximizes the probability that the outcome of interest occurs, and we can interpret the output coefficients as probabilities that quantify the effect of each covariate on the log odds of vaccination. We use all available covariates to fit our model, and to make predictions, we use a threshold of 0.5. That

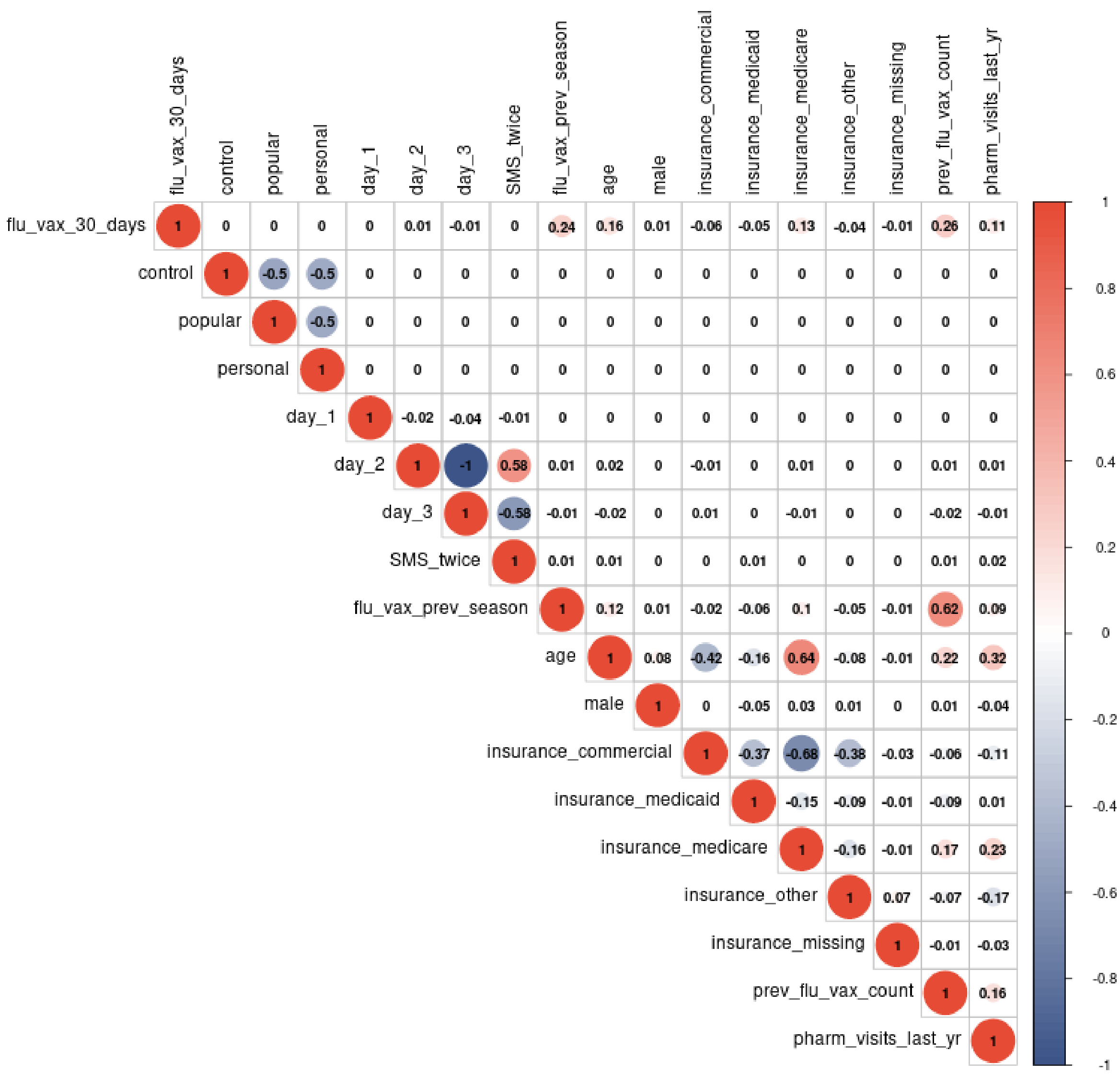

Figure 1: Spearman Correlation Plot of Key Variables

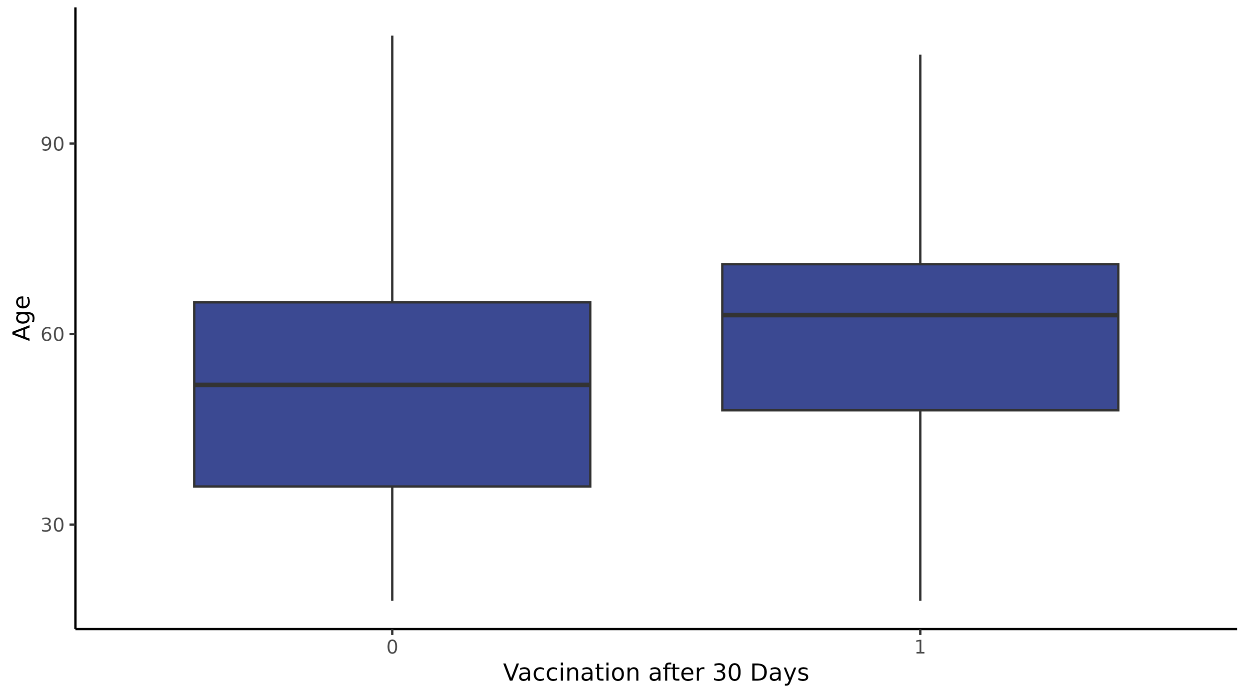

Figure 2: Boxplot of Vaccination (30 Days After Treatment) and Patient Age

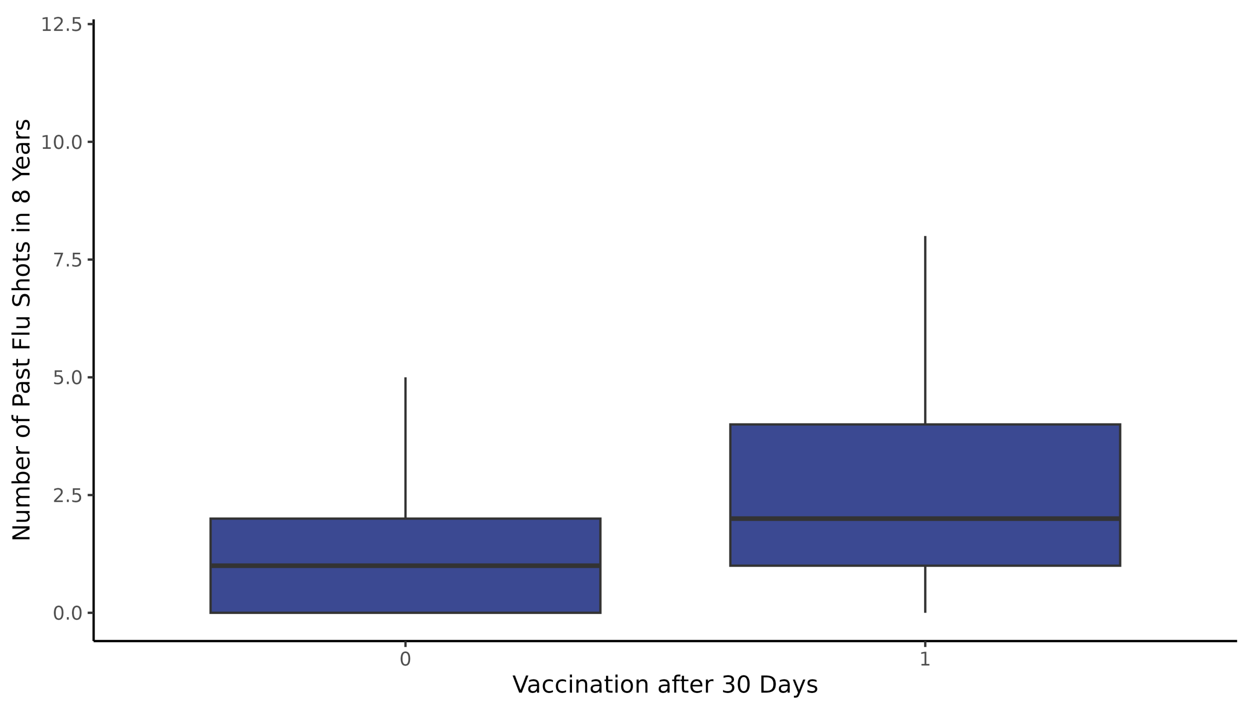

Figure 3: Boxplot of Vaccination (30 Days After Treatment) and Number of Past Flu Shots

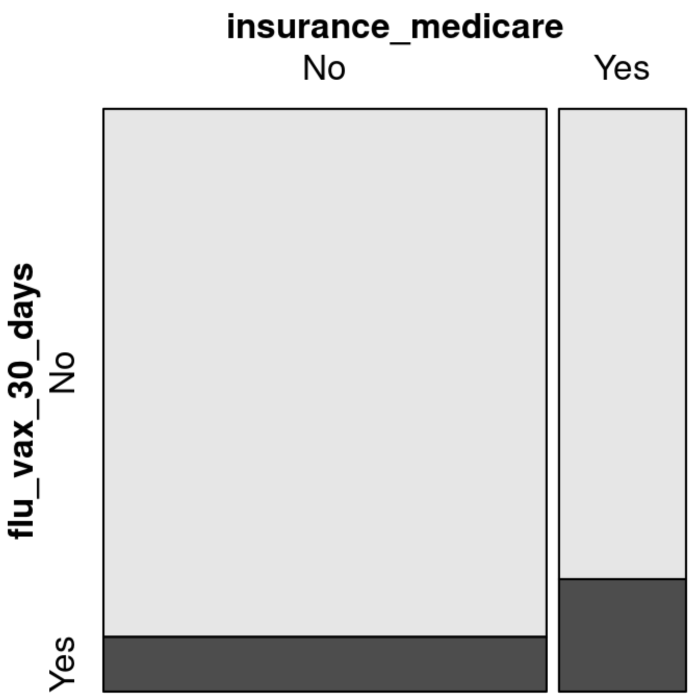

Figure 4: Mosaic Plot of Vaccination (30 Days After Treatment) and Medicare Insurance

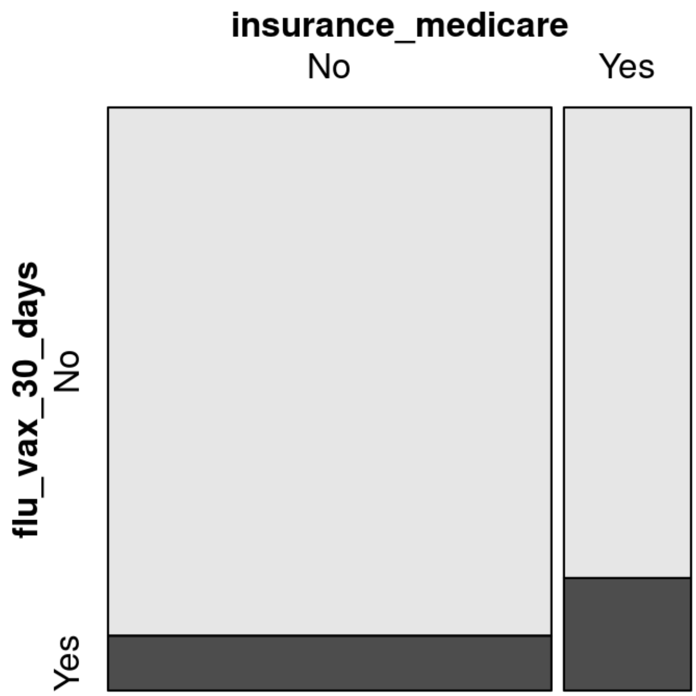

Figure 5: Mosaic Plot of Vaccination (30 Days After Treatment) and Medicare Insurance

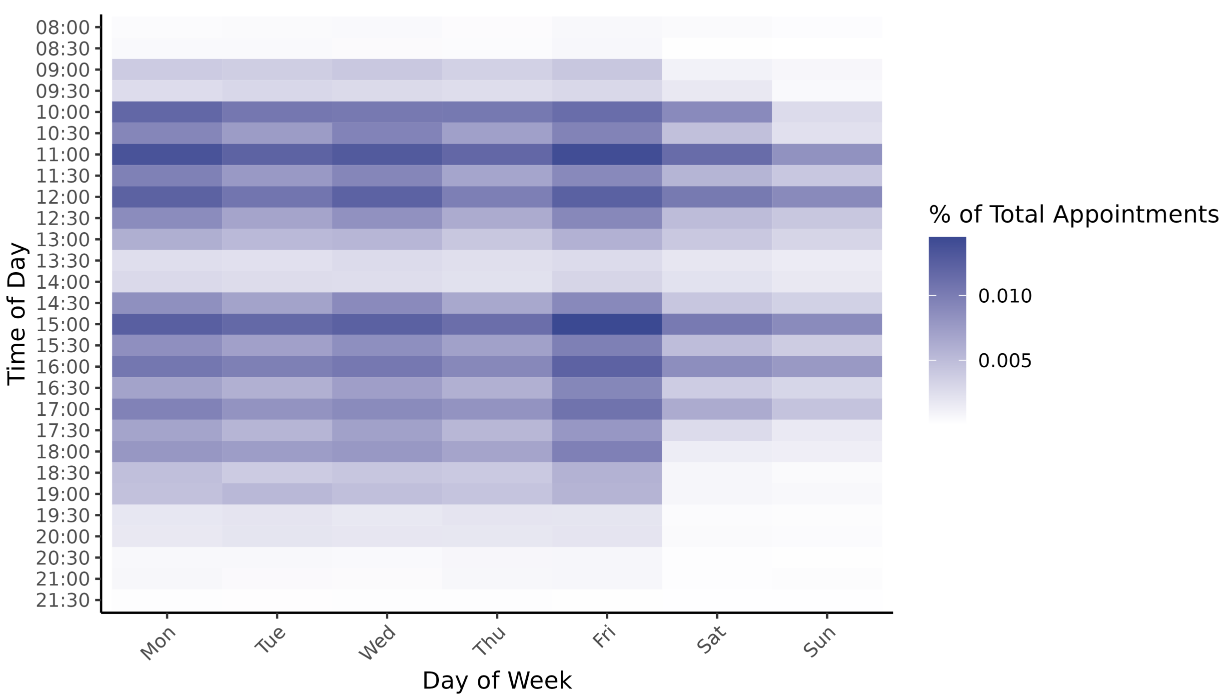

Figure 6: Heatmap of Last Vaccination Times

is, if the $\hat{y} \geq 0.5$, then we predict that the individual will get vaccinated. Using a threshold of 0.5 is more parsimonious and makes more intuitive sense than using a different threshold would.

This model obtains an AUC of 0.7624 and a misspecification error of 0.119. Looking into the breakdown of errors, this model correctly predicted that 179,014 individuals would not get vaccinated and that 843 individuals would get vaccinated but incorrectly predicted that 1,099 individuals got vaccinated (false positives) and that 23,192 did not get vaccinated (false negatives).

Comparing our logistic regression model with our OLS model, we see very similar results of the AUC and misspecification error. However, the OLS regression model outperforms the logistic regression model in both false positives, while the logistic regression model outputs fewer false negatives. Given the close performance of these two models, we may want to consider the interpretability of the models as well as the kinds of mistakes they make to evaluate which model we would prefer. If we want to ensure that we will not be overly optimistic, then we will prefer the model with a lower false positive rate, which in this case is the OLS regression model. Should these vaccination predictions be used to inform vaccine stocking decisions, an inflated estimate could lead to wasted vaccines. On the other hand, if we want to be conservative and not over-predict the number of individuals who would like to get vaccinated, then we would prefer the model with a lower false negative rate, which in this case is the logistic regression model.

Relaxed LASSO with Logit

In order to more effective compare the OLS and logistic regression models, we also run a relaxed LASSO with logit model. As with the above models, we use a threshold of 0.5 to determine if an individual is predicted to have gotten vaccinated or not. By creating a model that only incorporates the most important variables, we can compare which variables are selected for the OLS versus logistic regression models, which can then inform our evaluation of which model is more suitable for this task. [TODO: selected covariates]

This model achieves an AUC of 0.7404 and a misspecification error of 0.120. It correctly predicts that 178,570 individuals did not get vaccinated and that 1,048 individuals did get vaccinated but incorrectly predicts that 1,543 individuals got vaccinated even though they did not (false positives) and that 22,987 individuals did not get vaccinated when they actually did (false negatives). Using LASSO slightly decreases accuracy metrics in terms of both the AUC and misspecification error when compared to the logistic regression model. As with the regular OLS and logistic regression models, the relaxed LASSO with OLS model outperforms the relaxed LASSO with logit model.

Relaxed LASSO with OLS

- The Relaxed Lasso with OLS model first used a relaxed version of the Lasso (least absolute shrinkage and selection operator) regression to select important features, and then fit an OLS regression using those selected variables to predict flu vaccination probabilities. A 50% threshold was applied to classify patients as vaccinated or not based on the predicted probabilities.
- The Relaxed Lasso with OLS achieved an AUC of 0.744 on the test set. The misclassification error was 11.78%. Examining the confusion table, this model correctly predicted 180,040 non-vaccinated patients (true negatives) and 65 vaccinated patients (true positives), while incorrectly classifying 23,970 patients as not vaccinated when they actually were (false negatives) and 73 as vaccinated when they were not (false positives).

Random Forest

The above models have high false negative rates, which may indicate that linear and logistic regressions cannot fully capture the relationship between the covariates and vaccination propensities. As such, we run a random forest model, which can allow for more flexibility. Our random forest model has mtry = 4, meaning that we split on four randomly selected predictors at each split, and ntree = 500, meaning that our forest consists of 500 trees. We arrived at these hyperparameters via manual tuning as the required R packages for finding the optimal hyperparameters were not available on the secure server.

The random forest model achieves an AUC of 0.7489 and a misspecification error of 0.119. Despite the additional flexibility offered by the random forest model, the AUC is lower and there is no meaningful difference in the misspecification error when compared to those of the OLS and logistic regression models. Considering the confusion matrix, the random forest model correctly predicted that 179,382 individuals were not vaccinated (true negatives) and 494 were vaccinated (true positives) but incorrectly predicted tat 731 individuals were vaccinated when they were not (false positives) and 23,541 individuals were not vaccinated when they were (false negatives). Again, there does not appear to be improvement in the random forest model when compared to the OLS and logistic regression models.

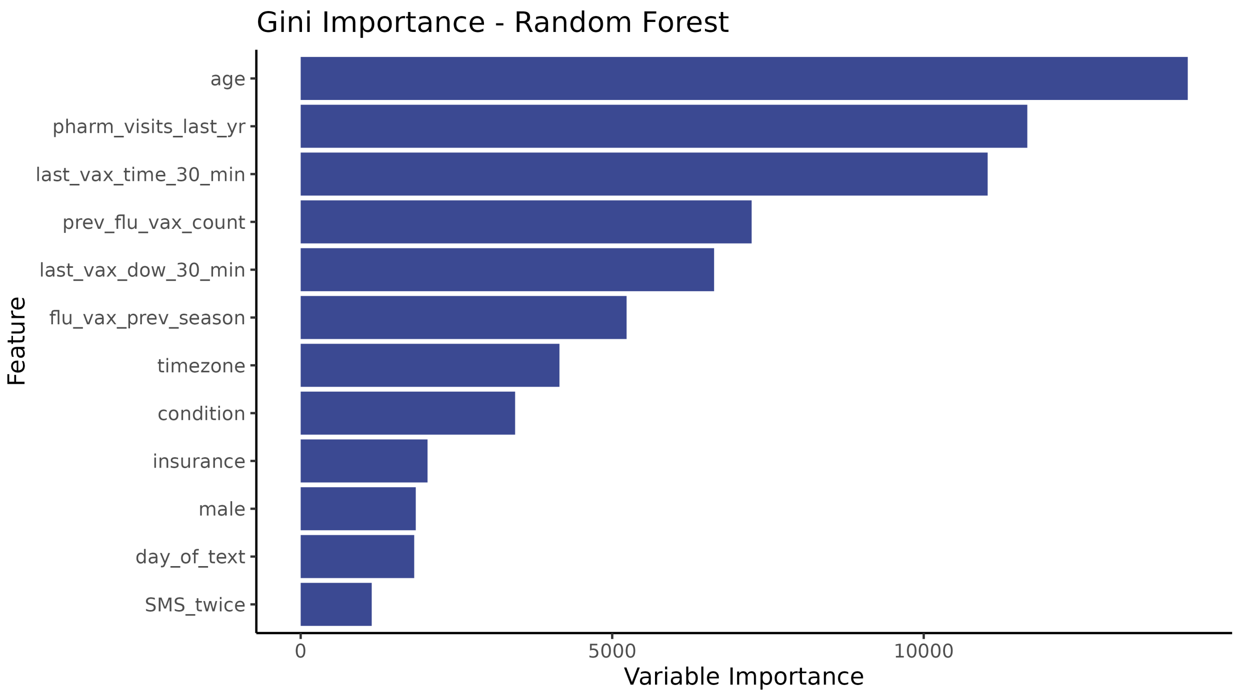

Figure 7: Random Forest Variable Importance - Gini

Neural Network

The final model that we consider is a neural network with 1 hidden layer of 10 nodes with a logistic activation function and an output layer that uses a sigmoid activation. We can interpret the outputs as the probability that an input individual and their associated covariates has been vaccinated. We implemented our neural network using the nnet package in R, and to train our neural network, we run it over 100 epochs. Due to computational resource constraints, we were not able to tune our hyperparameters or run the neural network for more epochs. Notably, our neural network predicts 0 for every input, suggesting that hyperparameter tuning or a different architecture may be needed to yield more informative results.

Despite predicting all 0s, the neural network obtains an AUC of 0.7644, which is the highest of all of the models, and a misspecification error of 0.118. Of these predictions, 180,113 are true negatives and 24,035 are false negatives. While the neural network slightly outperforms all of the other models in terms of the AUC, the comparable misspecification error shows that machine learning may not always lead to improvements in performance. Additionally, given the significantly higher resource requirements for running a neural network and the uninformative results, simpler models are likely better suited for this problem.

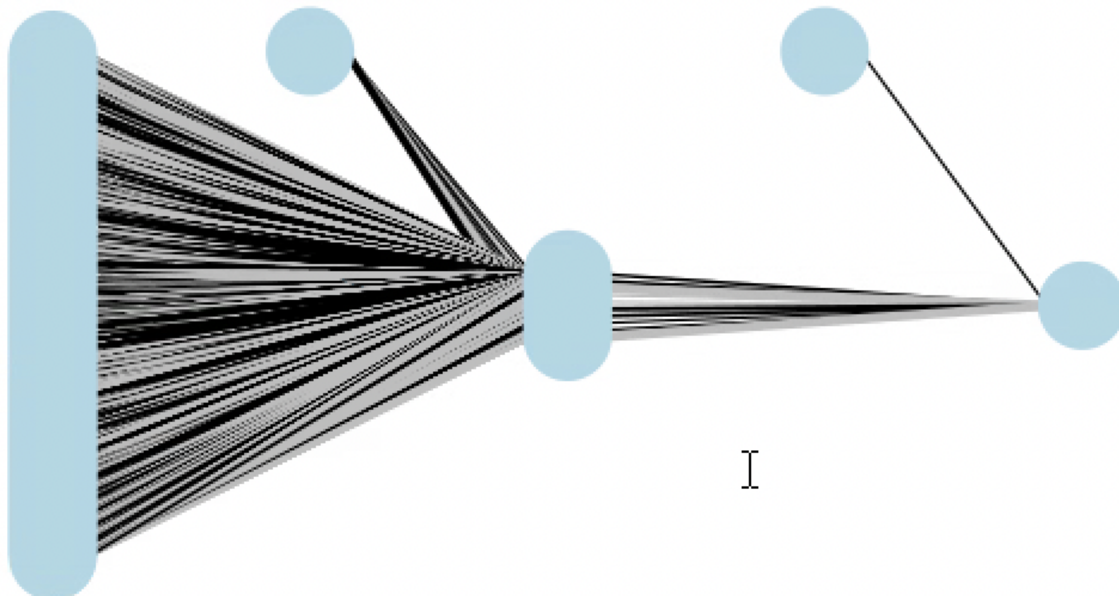

Figure 8: Neural Network Architecture

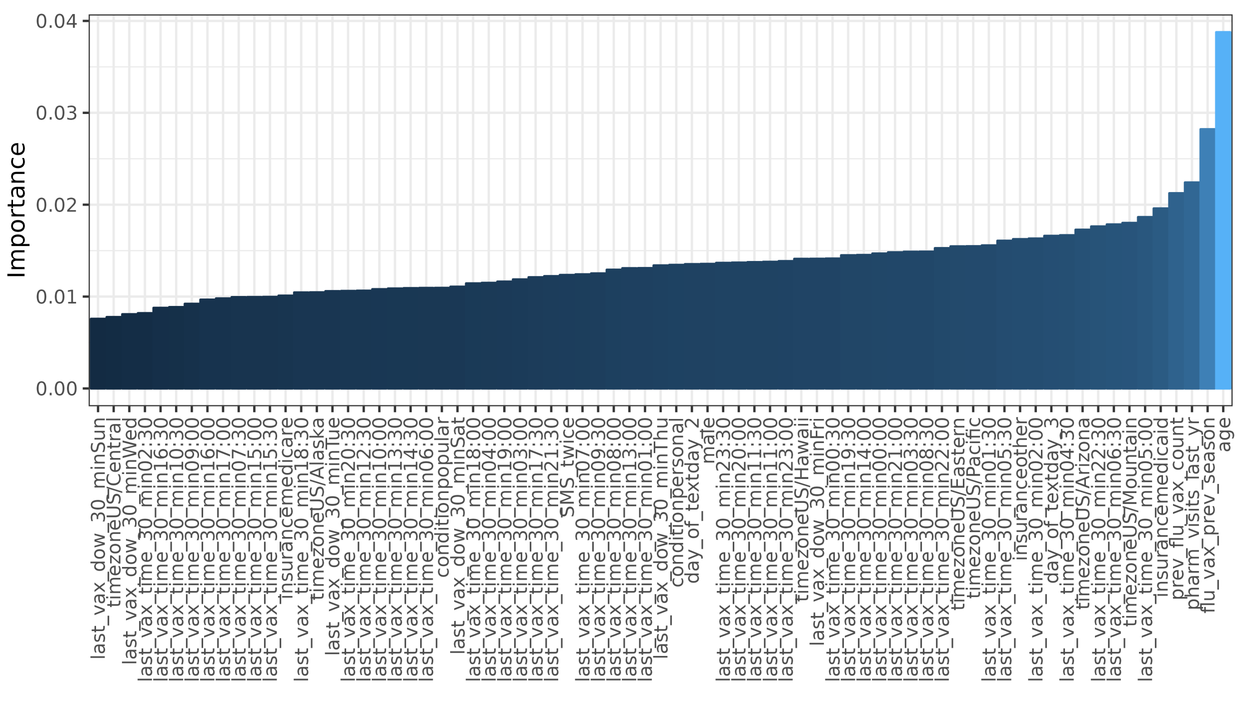

Figure 9: Neural Network Variable Importance - Garson

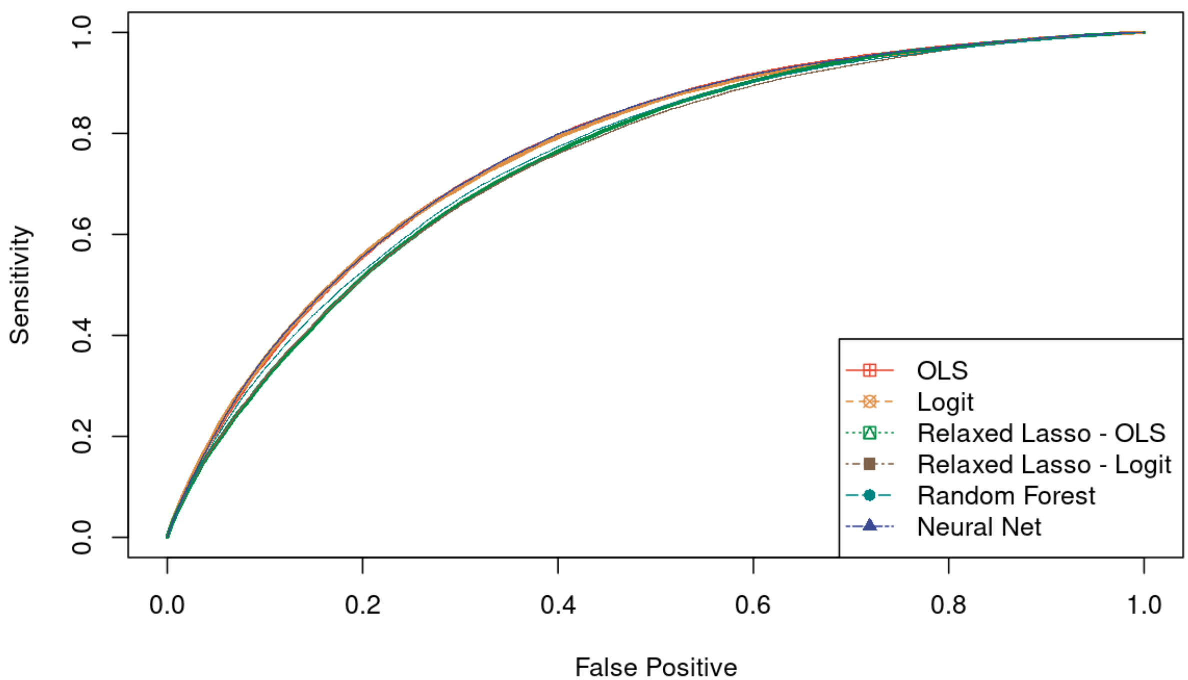

Figure 10: ROC Curve Comparison of Different Models

Conclusions

Appendix

OLS Regression

Table 1: OLS Regression Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	-0.08393	0.01914	-4.385	1.158e-05
conditionpopular	0.0024	0.0009504	2.525	0.01158
conditionpersonal	0.002722	0.0009509	2.862	0.004205
day_of_textday_2	0.0181	0.01397	1.296	0.195
day_of_textday_3	0.01461	0.01395	1.047	0.2949
SMS_twice	-0.008328	0.001436	-5.8	6.646e-09
flu_vax_prev_season	0.09021	0.0009431	95.65	0
age	0.001115	2.871e-05	38.84	0
male	0.0016	0.0007949	2.012	0.0442
insurancemedicaid	-0.01738	0.001518	-11.45	2.292e-30
insurancemedicare	0.02429	0.001217	19.97	1.131e-88
insuranceother	-0.00739	0.001442	-5.125	2.981e-07
prev_flu_vax_count	0.02981	0.0002589	115.1	0
pharm_visits_last_yr	0.001377	4.286e-05	32.13	2.891e-226
last_vax_dow_30_minTue	-0.004296	0.001375	-3.125	0.001781
last_vax_dow_30_minWed	-0.003579	0.001353	-2.646	0.008146
last_vax_dow_30_minThu	-0.003933	0.001395	-2.82	0.004803
last_vax_dow_30_minFri	0.002652	0.001326	2	0.04546
last_vax_dow_30_minSat	-0.0006752	0.001538	-0.439	0.6607
last_vax_dow_30_minSun	-0.003585	0.001664	-2.155	0.03114
last_vax_time_30_min12:30	-0.009462	0.002305	-4.106	4.028e-05
last_vax_time_30_min13:00	-0.01574	0.002341	-6.724	1.772e-11
last_vax_time_30_min13:30	-0.01802	0.00313	-5.758	8.513e-09
last_vax_time_30_min14:00	-0.01718	0.003006	-5.715	1.096e-08
last_vax_time_30_min14:30	-0.008429	0.002276	-3.704	0.0002127
last_vax_time_30_min15:00	0.00461	0.002063	2.235	0.02541
last_vax_time_30_min15:30	-0.006773	0.002286	-2.962	0.003054
last_vax_time_30_min16:00	-0.002343	0.002097	-1.117	0.2639
last_vax_time_30_min16:30	-0.01107	0.00233	-4.753	2.006e-06
last_vax_time_30_min17:00	-0.003952	0.002165	-1.826	0.06791
last_vax_time_30_min17:30	-0.002274	0.002439	-0.9324	0.3511
last_vax_time_30_min18:00	-0.002978	0.002308	-1.29	0.1971
last_vax_time_30_min18:30	-0.006947	0.002745	-2.531	0.01136
last_vax_time_30_min19:00	-0.006806	0.002666	-2.553	0.01069
last_vax_time_30_min19:30	-0.008503	0.003739	-2.274	0.02297
last_vax_time_30_min20:00	-0.01111	0.003764	-2.951	0.003167
last_vax_time_30_min20:30	-0.0127	0.005914	-2.148	0.03175
last_vax_time_30_min21:00	-0.002086	0.00635	-0.3286	0.7425
last_vax_time_30_min21:30	-0.009138	0.009964	-0.9171	0.3591
last_vax_time_30_min22:00	-0.016	0.0101	-1.585	0.1131
last_vax_time_30_min22:30	-0.04286	0.0174	-2.463	0.01378
last_vax_time_30_min23:00	-0.008279	0.02009	-0.4121	0.6802
last_vax_time_30_min23:30	0.01823	0.01321	1.381	0.1673
last_vax_time_30_min00:00	-0.0209	0.04063	-0.5145	0.6069

term	estimate	std.error	statistic	p.value
last_vax_time_30_min00:30	0.05162	0.04635	1.113	0.2655
last_vax_time_30_min01:00	-0.03088	0.03687	-0.8376	0.4023
last_vax_time_30_min01:30	-0.05397	0.05373	-1.005	0.3151
last_vax_time_30_min02:00	-0.05891	0.04216	-1.397	0.1623
last_vax_time_30_min02:30	-0.03681	0.05291	-0.6957	0.4866
last_vax_time_30_min03:00	0.03413	0.04531	0.7533	0.4513
last_vax_time_30_min03:30	-0.03745	0.07162	-0.5229	0.6011
last_vax_time_30_min04:00	-0.02771	0.03488	-0.7944	0.427
last_vax_time_30_min04:30	-0.0618	0.06203	-0.9963	0.3191
last_vax_time_30_min05:00	0.00366	0.0217	0.1687	0.866
last_vax_time_30_min05:30	0.05216	0.03358	1.553	0.1204
last_vax_time_30_min06:00	0.006928	0.01544	0.4488	0.6536
last_vax_time_30_min06:30	-0.04728	0.02399	-1.971	0.04874
last_vax_time_30_min07:00	0.01704	0.01124	1.517	0.1294
last_vax_time_30_min07:30	0.00171	0.01479	0.1156	0.908
last_vax_time_30_min08:00	0.01274	0.007746	1.644	0.1001
last_vax_time_30_min08:30	-0.01675	0.00743	-2.254	0.02417
last_vax_time_30_min09:00	0.00641	0.003163	2.026	0.04272
last_vax_time_30_min09:30	-0.007261	0.003459	-2.099	0.03578
last_vax_time_30_min10:00	0.01127	0.002229	5.058	4.236e-07
last_vax_time_30_min10:30	-0.001207	0.00236	-0.5114	0.609
last_vax_time_30_min11:00	0.006527	0.00207	3.153	0.001617
last_vax_time_30_min11:30	-0.007871	0.002296	-3.428	0.0006087
timezoneUS/Alaska	0.02635	0.03956	0.6661	0.5053
timezoneUS/Arizona	0.04055	0.01322	3.068	0.002152
timezoneUS/Central	0.02629	0.01293	2.034	0.042
timezoneUS/Eastern	0.03144	0.0129	2.437	0.0148
timezoneUS/Hawaii	0.05917	0.01344	4.402	1.07e-05
timezoneUS/Mountain	0.02892	0.01332	2.172	0.02986
timezoneUS/Pacific	0.03816	0.01293	2.951	0.00317

Table 2: OLS Regression Type II Anova

term	sumsq	df	statistic	p.value
condition	0.9039	2	4.897	0.007472
day_of_text	1.174	2	6.361	0.001729
SMS_twice	3.105	1	33.64	6.646e-09
flu_vax_prev_season	844.5	1	9150	0
age	139.2	1	1509	0
male	0.3737	1	4.049	0.0442
insurance	57.83	3	208.9	2.062e-135
prev_flu_vax_count	1224	1	13258	0
pharm_visits_last_yr	95.26	1	1032	2.891e-226
last_vax_dow_30_min	4.056	6	7.325	7.577e-08
last_vax_time_30_min	34.39	47	7.927	3.294e-52
timezone	14.1	7	21.83	1.048e-29
Residuals	56519	612371	NA	NA

Logitic Regression

Table 3: Logistic Regression Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	-4.514	0.2191	-20.6	2.673e-94
conditionpopular	0.02622	0.01037	2.527	0.0115
conditionpersonal	0.03116	0.01038	3.002	0.002679
day_of_textday_2	0.1781	0.154	1.156	0.2475
day_of_textday_3	0.1433	0.1538	0.932	0.3513
SMS_twice	-0.08375	0.01564	-5.354	8.608e-08
flu_vax_prev_season	1.056	0.009948	106.1	0
age	0.01528	0.0003344	45.7	0
male	0.009871	0.008613	1.146	0.2517
insurancemedicaid	-0.358	0.02105	-17.01	7.407e-65
insurancemedicare	0.09614	0.01205	7.98	1.46e-15
insuranceother	-0.1975	0.01843	-10.72	8.581e-27
prev_flu_vax_count	0.2111	0.002242	94.17	0
pharm_visits_last_yr	0.01303	0.0004149	31.41	1.578e-216
last_vax_dow_30_minTue	-0.0504	0.01496	-3.369	0.0007537
last_vax_dow_30_minWed	-0.03745	0.01459	-2.567	0.01026
last_vax_dow_30_minThu	-0.04518	0.01517	-2.978	0.002904
last_vax_dow_30_minFri	0.03195	0.01422	2.247	0.02466
last_vax_dow_30_minSat	0.0033	0.01672	0.1974	0.8435
last_vax_dow_30_minSun	-0.03668	0.01856	-1.977	0.04808
last_vax_time_30_min12:30	-0.09639	0.02469	-3.904	9.443e-05
last_vax_time_30_min13:00	-0.2139	0.02726	-7.846	4.287e-15
last_vax_time_30_min13:30	-0.2037	0.03643	-5.593	2.233e-08
last_vax_time_30_min14:00	-0.2031	0.03535	-5.746	9.124e-09
last_vax_time_30_min14:30	-0.1009	0.02473	-4.08	4.496e-05
last_vax_time_30_min15:00	0.04379	0.02142	2.045	0.0409
last_vax_time_30_min15:30	-0.0673	0.02443	-2.754	0.00588
last_vax_time_30_min16:00	-0.01807	0.02226	-0.8115	0.4171
last_vax_time_30_min16:30	-0.1098	0.0255	-4.307	1.658e-05
last_vax_time_30_min17:00	-0.03405	0.02344	-1.453	0.1463
last_vax_time_30_min17:30	-0.003898	0.02656	-0.1468	0.8833
last_vax_time_30_min18:00	-0.02198	0.02545	-0.8636	0.3878
last_vax_time_30_min18:30	-0.0591	0.03084	-1.916	0.05532
last_vax_time_30_min19:00	-0.06304	0.03017	-2.089	0.03668
last_vax_time_30_min19:30	-0.08575	0.04386	-1.955	0.05059
last_vax_time_30_min20:00	-0.1145	0.0438	-2.613	0.00897
last_vax_time_30_min20:30	-0.1513	0.07116	-2.127	0.03344
last_vax_time_30_min21:00	0.005858	0.07333	0.07988	0.9363
last_vax_time_30_min21:30	-0.1134	0.1244	-0.9119	0.3618
last_vax_time_30_min22:00	-0.204	0.1273	-1.602	0.1091
last_vax_time_30_min22:30	-0.7028	0.2656	-2.646	0.008155
last_vax_time_30_min23:00	-0.116	0.2638	-0.4397	0.6602
last_vax_time_30_min23:30	0.2407	0.1396	1.724	0.08469
last_vax_time_30_min00:00	-1.051	1.017	-1.033	0.3016
last_vax_time_30_min00:30	0.6212	0.4626	1.343	0.1793
last_vax_time_30_min01:00	-0.9216	0.7381	-1.249	0.2118
last_vax_time_30_min01:30	-1.001	1.03	-0.9726	0.3308
last_vax_time_30_min02:00	-8.731	26.46	-0.33	0.7414
last_vax_time_30_min02:30	-0.9261	1.035	-0.8949	0.3708
last_vax_time_30_min03:00	0.4946	0.5456	0.9064	0.3647

term	estimate	std.error	statistic	p.value
last_vax_time_30_min03:30	-8.441	45.71	-0.1847	0.8535
last_vax_time_30_min04:00	-0.8311	0.7343	-1.132	0.2577
last_vax_time_30_min04:30	-1.074	1.046	-1.027	0.3046
last_vax_time_30_min05:00	0.01183	0.2566	0.04609	0.9632
last_vax_time_30_min05:30	0.5352	0.3216	1.664	0.09608
last_vax_time_30_min06:00	0.09147	0.1641	0.5575	0.5772
last_vax_time_30_min06:30	-0.6955	0.3543	-1.963	0.04966
last_vax_time_30_min07:00	0.1842	0.117	1.574	0.1156
last_vax_time_30_min07:30	0.03315	0.1734	0.1911	0.8484
last_vax_time_30_min08:00	0.1586	0.079	2.008	0.04464
last_vax_time_30_min08:30	-0.159	0.08292	-1.918	0.0551
last_vax_time_30_min09:00	0.07111	0.03314	2.146	0.03187
last_vax_time_30_min09:30	-0.07083	0.03706	-1.911	0.05601
last_vax_time_30_min10:00	0.103	0.02281	4.516	6.311e-06
last_vax_time_30_min10:30	-0.01397	0.02459	-0.568	0.5701
last_vax_time_30_min11:00	0.05595	0.02131	2.626	0.008639
last_vax_time_30_min11:30	-0.08211	0.02426	-3.385	0.0007117
timezoneUS/Alaska	0.3704	0.4183	0.8856	0.3758
timezoneUS/Arizona	0.4706	0.1566	3.005	0.002653
timezoneUS/Central	0.3215	0.154	2.088	0.0368
timezoneUS/Eastern	0.3811	0.1537	2.48	0.01315
timezoneUS/Hawaii	0.636	0.1582	4.02	5.825e-05
timezoneUS/Mountain	0.3876	0.1578	2.456	0.01404
timezoneUS/Pacific	0.4548	0.154	2.953	0.003144

Table 4: Logistic Regression Type II Anova

term	statistic	df	p.value
condition	10.41	2	0.005478
day_of_text	10.77	2	0.004575
SMS_twice	28.72	1	8.368e-08
flu_vax_prev_season	11803	1	0
age	2122	1	0
male	1.313	1	0.2518
insurance	543.7	3	1.611e-117
prev_flu_vax_count	8553	1	0
pharm_visits_last_yr	956.9	1	4.093e-210
last_vax_dow_30_min	51.16	6	2.752e-09
last_vax_time_30_min	381	47	7.633e-54
timezone	146.4	7	2.293e-28

Relaxed Lasso - OLS Regression

Table 5: Relaxed Lasso (OLS) Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0255	0.001376	-18.53	1.235e-76
age	0.001108	2.87e-05	38.61	0
insurancemedicaid	-0.01823	0.001505	-12.11	8.976e-34

term	estimate	std.error	statistic	p.value
insurancemedicare	0.02521	0.001205	20.92	4.164e-97
prev_flu_vax_count	0.04274	0.0002231	191.6	0
pharm_visits_last_yr	0.001362	4.289e-05	31.76	3.53e-221

Table 6: Relaxed Lasso (OLS) Type II Anova

term	sumsq	df	statistic	p.value
age	139.9	1	1491	0
insurancemedicaid	13.77	1	146.8	8.976e-34
insurancemedicare	41.04	1	437.5	4.164e-97
prev_flu_vax_count	3442	1	36692	0
pharm_visits_last_yr	94.63	1	1009	3.53e-221
Residuals	57456	612439	NA	NA

Relaxed Lasso - Logistic Regression

Table 7: Relaxed Lasso (Logit) Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0255	0.001376	-18.53	1.235e-76
age	0.001108	2.87e-05	38.61	0
insurancemedicaid	-0.01823	0.001505	-12.11	8.976e-34
insurancemedicare	0.02521	0.001205	20.92	4.164e-97
prev_flu_vax_count	0.04274	0.0002231	191.6	0
pharm_visits_last_yr	0.001362	4.289e-05	31.76	3.53e-221

Table 8: Relaxed Lasso (Logit) Type II Anova

term	sumsq	df	statistic	p.value
age	139.9	1	1491	0
insurancemedicaid	13.77	1	146.8	8.976e-34
insurancemedicare	41.04	1	437.5	4.164e-97
prev_flu_vax_count	3442	1	36692	0
pharm_visits_last_yr	94.63	1	1009	3.53e-221
Residuals	57456	612439	NA	NA