Does Model Choice Impact Classification Accuracy for Predicting Flu Vaccinations? Analyzing a 1-Million-Person Vaccination Dataset

Jose Cervantez Bethany Hsaio Rob Kuan

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

Contents

CONFIDENTIAL - PLEASE DO NOT SHARE

This report uses data in partnership with a large pharmacy in the United States. Since the data is subject to HIPPA privacy requirements, all the data for the analysis was conducted on a separate secure server, in accordance with the contractual obligations with the pharmacy partner.

Since there are restrictions on how this data may be used, we ask that you do not share this report with anyone outside of grading purposes.

Executive Summary (1 page)

Introduction

Even though vaccines are a cost-effective method for decreasing mortality, morbidity, and healthcare expenses[1], promoting recommended annual vaccinations against dangerous diseases remains a major public health challenge. For example, for the 2022-23 flu season, 53.1% of the U.S. population choose not to receive the recommended influenza vaccine even though such vaccines can prevent subsequent illnesses, hospitalizations, and deaths[2].

In collaboration with a large U.S. pharmacy chain, we collected a 1-million-person dataset where text messages were sent to patients to encourage a flu vaccination. Using this data set, we compare the performance of different predictive models to test how accurately we can identify patients who are likely to be vaccinated for the flu 30 days after receiving a text message.

Study Goal

The primary objective of this study was to test if model choice impacts classification accuracy on patients' likelihood of receiving a flu vaccination within 30 days. We also aim to benchmark the predictive accuracy of the best model in predicting flu vaccination uptake, using information that is commonly available to pharmacies and providers. These results have implications for both policymakers, healthcare systems, and health providers who may want to predict vaccination uptake to inform planning and resource allocation.

Data Description

The dataset, obtained through our pharmacy partner, includes 1 million patients who received a text message flu vaccination reminder text message. The data captures patient demographics, vaccination history, text message reminder details, and pharmacy visit information. The primary outcome variable is flu vaccination within 30 days of receiving the reminder.

Methodology

We developed and evaluated six predictive models: OLS with Classifier, Logistic Regression, Relaxed Lasso with Logit, Relaxed Lasso with OLS, Random Forest, and Neural Network. Model performance was assessed using metrics such as area under the ROC curve (AUC), misclassification error, and confusion matrices.

Exploratory data analysis revealed associations between vaccination uptake and variables such as age, previous vaccination history, pharmacy visit frequency, and Medicare insurance status. These insights informed feature selection for the predictive models.

Results

The OLS with Classifier and Neural Network models achieved the highest AUC (0.763 and 0.764, respectively) and lowest misclassification error (0.118 for both) on the test set. However, the Neural Network predicted no vaccinations for all patients, indicating potential limitations.

The Logistic Regression, Relaxed Lasso with Logit, Relaxed Lasso with OLS, and Random Forest models achieved AUCs ranging from 0.740 to 0.762 and misclassification errors between 0.118 and 0.120.

While the OLS with Classifier and Neural Network had the best performance metrics, the OLS with Classifier is preferred due to its interpretability and lower false positive rate, which is crucial for informing vaccine stocking decisions.

[1] Leidner, A. J., Murthy, N., Chesson, H. W., Biggerstaff, M., Stoecker, C., Harris, A. M., Acosta, A., Dooling, K., & Bridges, C. B. (2019). Cost-effectiveness of adult vaccinations: A systematic review. Vaccine, 37(2), 226–234. https://doi.org/10.1016/j.vaccine.2018.11.056

[2] Flu Vaccination Coverage, United States, 2022–23 Influenza Season | FluVaxView | Seasonal Influenza (Flu) | CDC. (2023, October 10). https://www.cdc.gov/flu/fluvaxview/coverage-2223estimates.htm

Detailed Analyses

Description of Data

The dataset used in this study was obtained through a partnership with a large nationwide pharmacy chain, where text messages were sent to patients to encourage flu vaccinations. The text messages were sent in September 2023 during the start of the 2022-2023 flu vaccine season to patients who (1) were over the age of 18, (2) opted in to pharmacy communications, (3) received at least one vaccine at the pharmacy chain in the past. The primary outcome of interest being whether a patient received a flu vaccination within 30 days of receiving the reminder (`flu_vax_30_days`).

The content of the text message reminder (`condition`) was varied to test the effectiveness of different messaging strategies. for example, one text message encouraged patients to get vaccinated at a *popular* time, while others encouraged patients to get vaccinated at a *personalized* time consistent with the time and day of week of thier last vaccination. To assess the potential impact of multiple reminders, some patients received a second message (`SMS_twice`). Lastly, the day the message was sent (`day_of_text`) was also recorded.

In addition to variables describing the text message, the dataset captures each patient's vaccination history. This includes whether they received a flu shot in the previous season (`flu_vax_previous_season`) and the total number of flu vaccinations they had received in the past 8 years (`prev_flu_vax_count`).

The dataset also includes demographic variables such as age (`age`), gender (`male`, `female`), and insurance type (`insurance`). These factors may influence a patient's likelihood of receiving a flu vaccination and are important to consider in the analysis.

To account for potential differences in patient behavior based on their level of engagement with the pharmacy, the dataset includes the number of visits that involved at least one prescription pickup or transaction in the past year (`pharm_visits_last_yr`).

Lastly, the study captures temporal patterns in vaccination behavior by recording the day of the week (`last_vax_dow_30_min`) and time (`last_vax_time_30_min`) of the patient's last vaccination. The patient's timezone (`timezone`) is also included as a general indicator of location (as any data that may identify a given participant is protected and cannot be shared).

This dataset combines intervention-specific variables with patient characteristics and historical behaviors. We aim to gain insights into the effectiveness of text message reminders in promoting flu vaccination uptake.

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

The exploratory data analysis section presents several visualizations that provide insights into the relationships between key variables and flu vaccination uptake within 30 days of receiving the text message reminder.

Figure 1 shows a Spearman correlation plot of the key variables. The plot reveals strong positive monotonic correlations between the outcome variable `flu_vax_30_days` and several predictors, including `prev_flu_vax_count`, `flu_vax_prev_season`, `age`, and `pharm_visits_last_yr` (Figure 1 - first row). This suggests that patients who have received more flu vaccinations in the past, got vaccinated in the previous season, are older, and visit the pharmacy more frequently are more likely to get vaccinated within 30 days of the reminder.

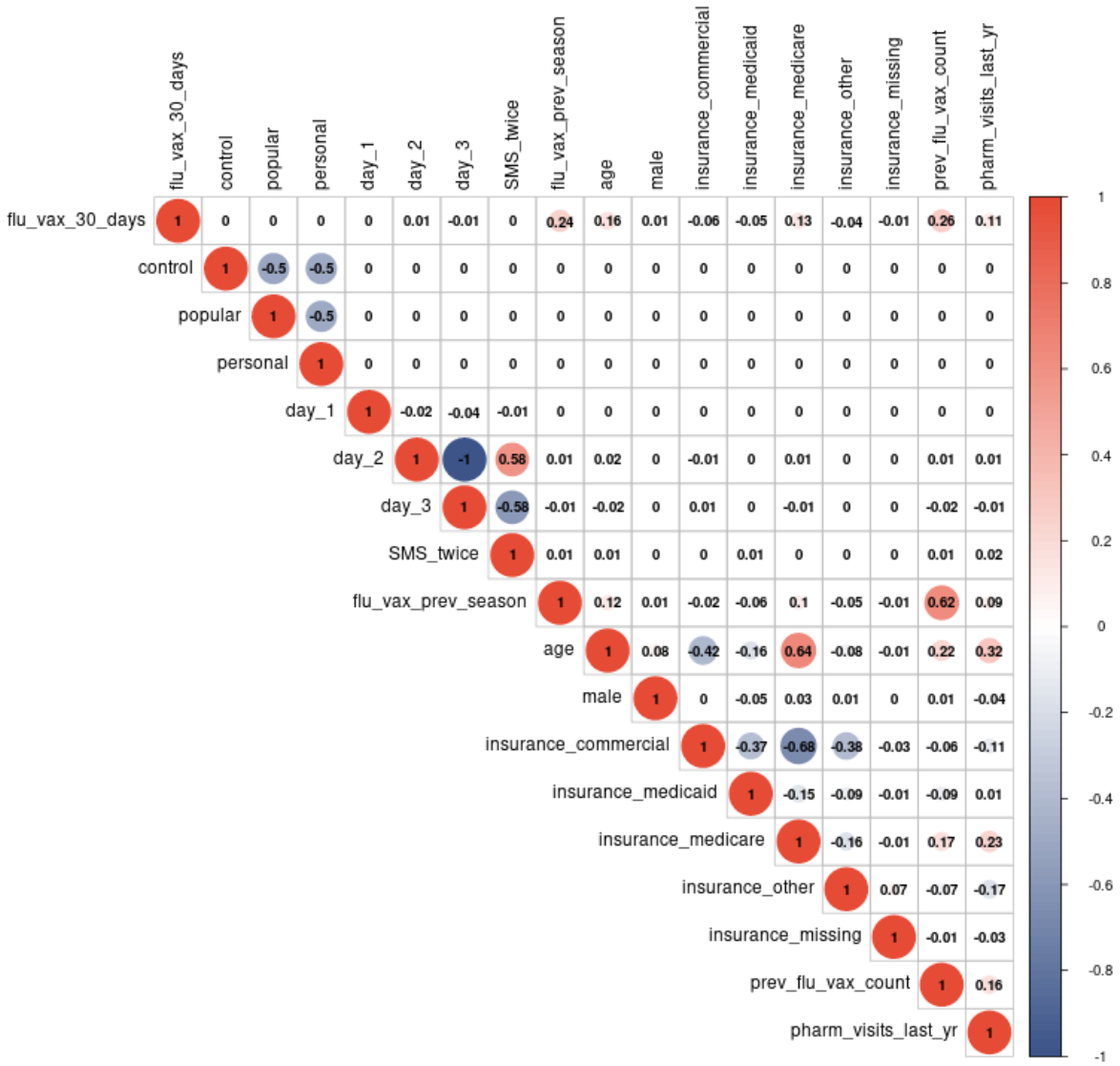

Figure 1: Spearman Correlation Plot of Key Variables

Figures 2 and 3 present boxplots comparing the distribution of age and the number of past flu shots between patients who did and did not get vaccinated within 30 days. In both cases, the mean values for vaccinated patients are higher, indicating that vaccinated patients tend to be older and have received more flu shots in the past compared to unvaccinated patients.

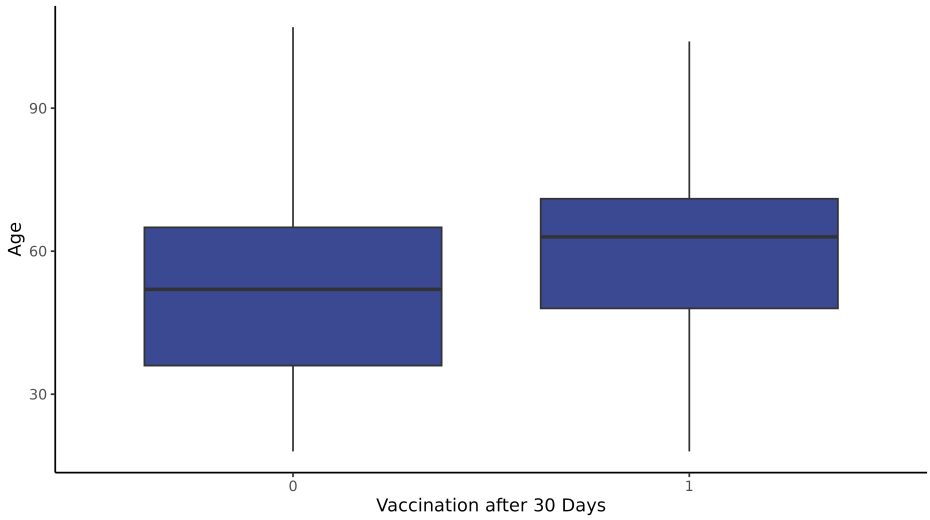

Figure 2: Boxplot of Vaccination (30 Days After Treatment) and Patient Age

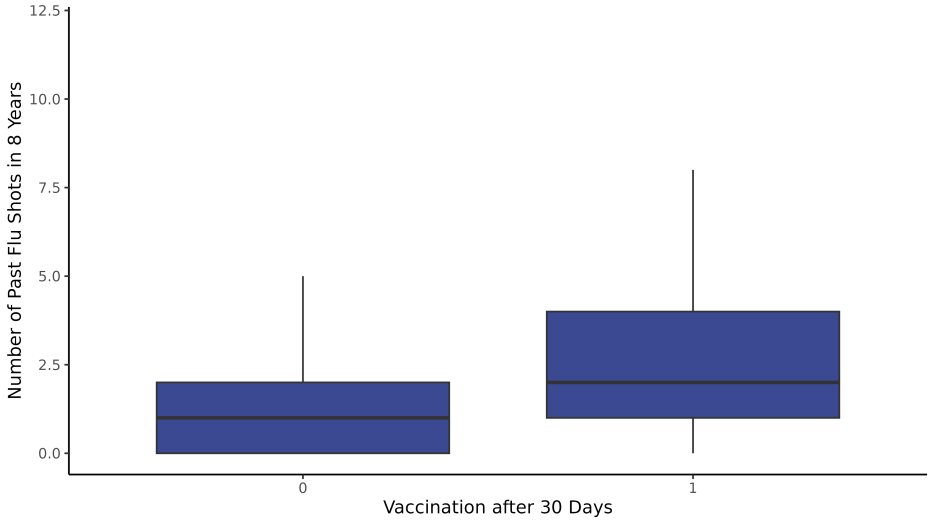

Figure 3: Boxplot of Vaccination (30 Days After Treatment) and Number of Past Flu Shots

Figures 4 and 5 display mosaic plots examining the relationship between Medicare insurance and flu vaccination within 30 days. The plots show that a higher proportion of patients with Medicare insurance get vaccinated compared to those without Medicare. This suggests that insurance type, specifically Medicare coverage, may influence a patient's likelihood of getting a flu shot after receiving the reminder.

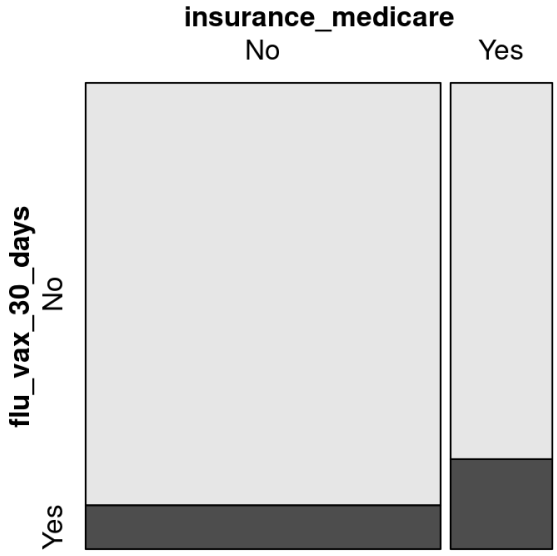

Figure 4: Mosaic Plot of Vaccination (30 Days After Treatment) and Medicare Insurance

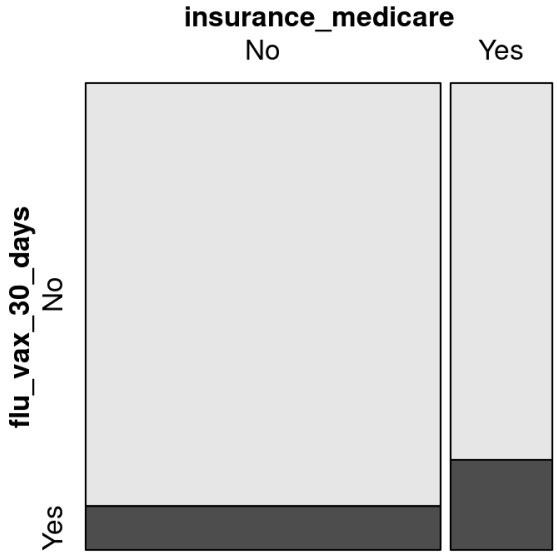

Figure 5: Mosaic Plot of Vaccination (30 Days After Treatment) and Medicare Insurance

Finally, Figure 6 presents a heatmap of the last vaccination times for patients. The heatmap reveals patterns in the timing of past vaccinations, with higher vaccination rates on Monday, Wednesday, and Friday compared to other weekdays and weekends. Vaccination times also clustered in the late morning (11am) and early afternoon hour (3pm). If there is a pattern or routine to how and when patients get vaccinated, this information could be useful to predict the likelihood of subsequent vaccination thirty days after receiving a text message.

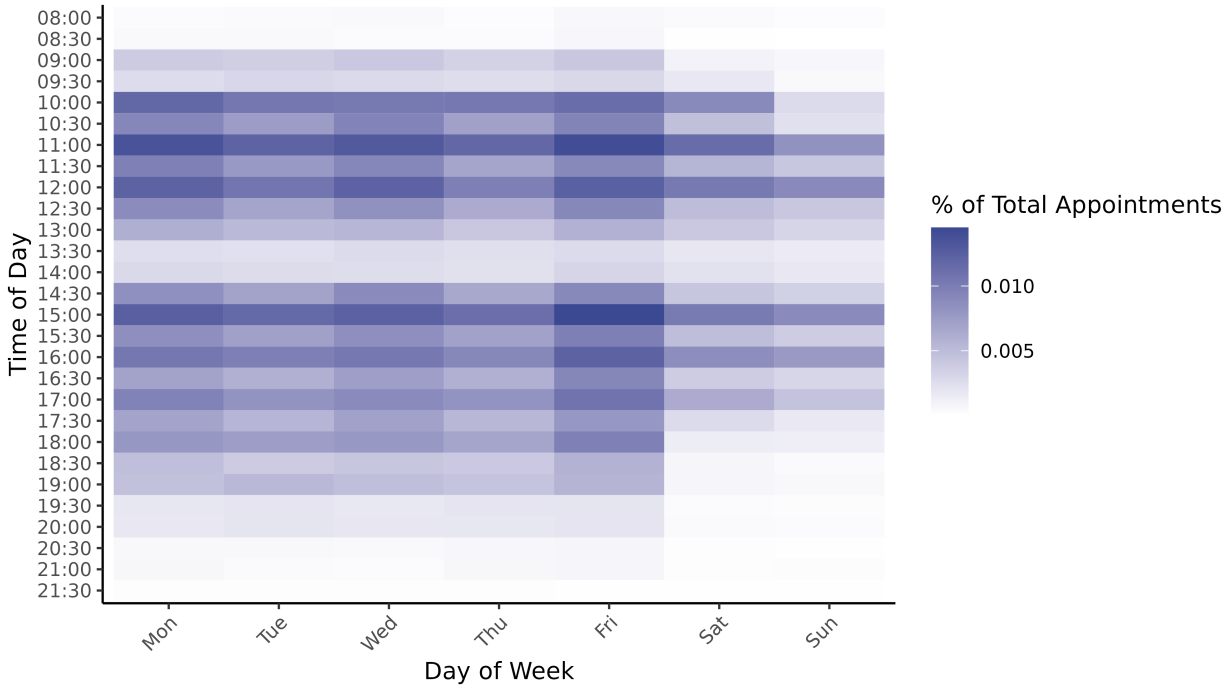

Figure 6: Heatmap of Last Vaccination Times

Predictive Modeling

To predict whether a patient will receive a flu vaccination within 30 days of the text message reminder, we developed and evaluated six different predictive models: OLS with Classifier, Logistic Regression, Relaxed Lasso with Logit, Relaxed Lasso with OLS, Random Forest, and Neural Network.

The OLS with Classifier and Logistic Regression models serve as our baseline approaches, using linear and logistic regression techniques, respectively. The Relaxed Lasso models (with Logit and OLS) extend these approaches by incorporating feature selection to identify the most informative predictors.

We also explore two machine learning models: Random Forest and Neural Network. The Random Forest model allows for a more flexible, non-linear relationship between the predictors and the outcome, while the Neural Network model has the potential to capture complex patterns in the data.

For each model, we evaluate its performance using metrics such as the area under the ROC curve (AUC) and misclassification error, as well as examining the confusion matrix to assess the model's ability to correctly predict vaccinated and unvaccinated patients. Our dataset was divided into three partitions: a 60% training set, a 20% testing set, and a 20% holdout validation set. The testing set was used to select the best-performing model and the validation set was used to evaluate the final model out-of-sample performance.

By comparing the performance of these six models, we aim to identify the approach that best predicts flu vaccination uptake within 30 days of the text message reminder.

OLS w/ Classifier

The OLS with Classifier model utilized ordinary least squares (OLS) regression to estimate the probability of a patient getting vaccinated for influenza within 30 days of receiving the treatment message. The predicted probabilities were then converted into binary classifications of vaccinated or not vaccinated using a threshold of 50%. The covariate with the highest magnitude as well as statistical significance is whether an individual received a vaccine in the previous season (see Appendix for full model summary and ANOVA).

When applied to the test set, the OLS with Classifier model achieved an area under the ROC curve (AUC) of 0.763, suggesting it has moderately good discriminatory power in identifying patients who will and will not get vaccinated. The overall misclassification error was 0.118, meaning the model's predictions were incorrect for 11.8% of patients.

Table 1: Confusion Matrix for OLS Regression with Classifier

	Actual - 0	Actual - 1
Predicted - 0	180113	24031
Predicted - 1	0	4

Examining the confusion matrix provides additional insight into the model's performance. It correctly identified a large number of patients who did not get vaccinated (180,113 true negatives) and a small number who did get vaccinated (4 true positives). However, the model struggled more with false negatives, incorrectly predicting 24,031 patients would not get vaccinated when they actually did. This suggests the model may have a hard time discriminating between patients who do and do not get vaccinated.

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 is, if the $P(vaccination) \geq 0.5$, then we predict that the individual will get vaccinated.

In this model, as with the OLS model, the variable with the largest impact (i.e., the largest magnitude) is whether an individual got a flu vaccine in the previous season. The coefficient is 1.056, meaning that an individual getting vaccinated in the previous season increases their log odds of being vaccinated this season by 105.6%. Other significant covariates include what condition the individual was in, whether they received two reminder text messages, their age, their insurance, and the number of previous vaccines they received. These results show that history is a powerful predictor of what someone will do. Although the magnitudes of the coefficients are smaller on the treatments, the coefficients are still significant and positive, indicating that an individual being in a treatment condition and not a control condition increases their log odds of getting vaccinated (see Appendix for full model summary and ANOVA).

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).

Table 2: Confusion Matrix for Logistic Regression with Classifier

	0	1
0	179014	23192
1	1099	843

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 effectively compare the OLS and logistic regression models, we also run a relaxed LASSO with logit model. We do not force in any covariates. 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.

LASSO selects the following variables to include in the logistic regression: age, insurance (Medicaid or Medicare), the number of previous flu shots received, and the number of pharmacy visits in the last year. All of these covariates were stastically significant at a <0.001 alpha level in the original logistic regression, and all of these covariates are significant in this model, indicating that there is no need for backwards selection. Similar to the original logistic regression, history is a powerful predictor; here, the previous flu vaccination count is the most informative covariate in terms of magnitude. However, the original model found whether an individual had been vaccinated in the past year to be more informative, and it is interesting that LASSO did not select that covariate to be included.

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.

Table 3: Confusion Matrix for Relaxed Lasso - Logistic Regression

	0	1
0	178570	22987
1	1543	1048

Relaxed LASSO with OLS

Applying LASSO to the OLS model leaves us with the following covariates: age, insurance, previous number of flu vaccinations received, and number of pharmacy visits in the last year. This is the same set of covariates returned by the logistic regression model with relaxed LASSO.

On the test set, the Relaxed Lasso with OLS model obtained an AUC of 0.744, indicating moderately good discrimination between vaccinated and unvaccinated patients, though not quite as high as the standard OLS with Classifier. The misclassification rate was 11.78%, so predictions were incorrect for just under 12% of patients.

The confusion matrix shows this model correctly predicted a high number of true negatives (180,040 patients) and a moderate number of true positives (65 patients). It had a fairly low false positive rate, only predicting

73 patients would get vaccinated when they did not. However, like the OLS with Classifier, it had a much higher false negative rate, incorrectly predicting 23,970 patients would not get vaccinated when in fact they did.

Table 4: Confusion Matrix for Relaxed Lasso - OLS Regression

	0	1
0	180040	23970
1	73	65

This suggests that while using the Lasso for feature selection can help yield a more parsimonious model, it may come at a slight cost to overall performance compared to using all available predictors. Both OLS and Relaxed LASSO w/ OLS seem to do better at identifying patients who will not get vaccinated than those who will, which could relate to the class imbalance in the data with most patients not getting vaccinated.

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.

In this model, age, the number of pharmacy visits in the last year, and the last vaccination time at the 30-minute granularity are the three most important variables (Figure 7). Similar to the logistic regression models, the number of pharmacy visits has high importance. Interestingly, the time at which an individual was vaccinated previously is also important. This may be an indication of overfitting or perhaps it is important conditional on the type of reminder text that the individual received.

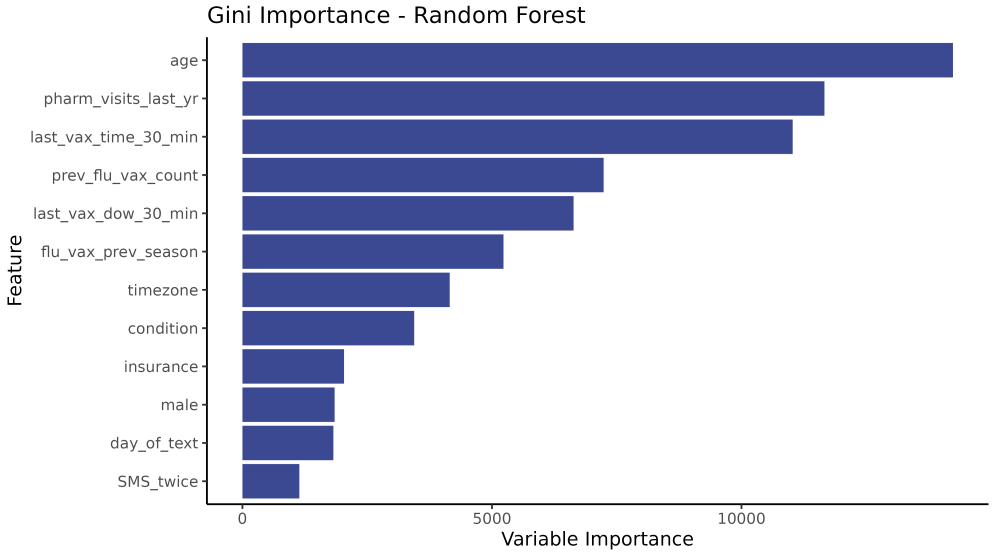

Figure 7: Random Forest Variable Importance - Gini

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 that 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.

Table 5: Confusion Matrix for Random Forest

	0	1
0	179382	235411
1	731	494

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 run the neural network model in R using the package `nnet`, as installing python packages on the secure server was not possible.

This architecture has a total of 751 parameters, and the architecture is included in the figure below. 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.

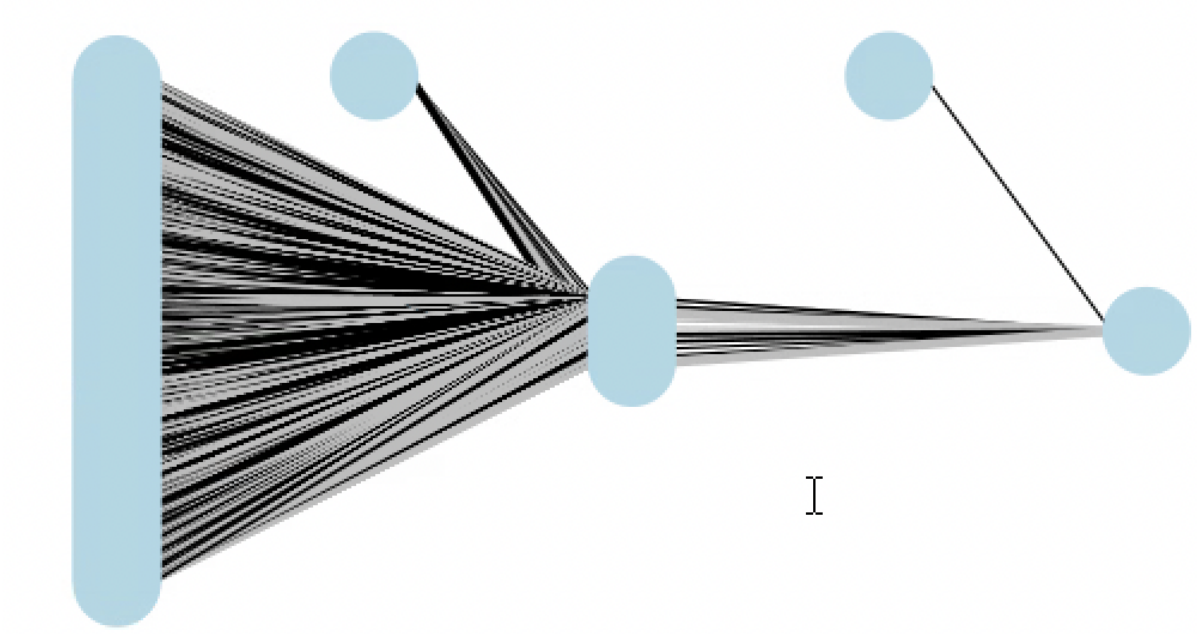

Figure 8: Architecture: 73-10-1 neural network with 751 weights

In this model, age, whether the individual received a flu vaccine the previous season, and the number of vaccines the individual has received are the three most important variables. As with our logistic regression models, history is an important consideration. Age is the most important variable once again, as was the case

with our random forest model. In our OLS and logistic regression models, age is a statistically significant covariate but the coefficients are small and positive in both cases. Future research can consider examining the role of age more thoroughly and how it impacts vaccination rates.

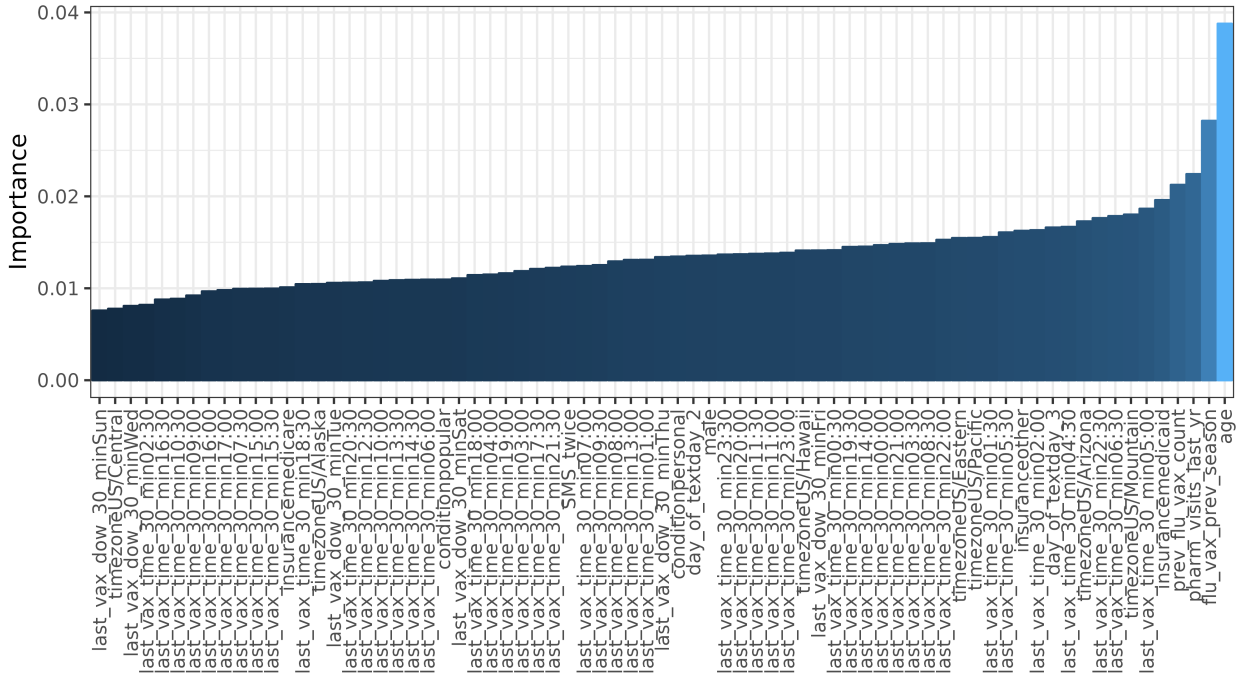

Figure 9: Neural Network Variable Importance - Garson

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.

Table 6: Confusion Matrix for Neural Network

	0	1
0	180113	24035
1	0	0

Holdout-Validation

We selected the OLS regression model as our final model to evaluate on the holdout validation set, as it has the highest AUC of all the models and is simpler to interpret than the neural network. The AUC on the holdout validation set is 0.7621, which is similar to the AUC on the training set (0.763). The misclassification errors are also similar, with the holdout validation set having a misspecification error of 11.9% (as opposed to 11.8%. We conclude that our model generalizes well to the holdout validation set, and that the OLS model does not suffer from overfitting.

Table 7: Confusion Matrix for OLS on Validation Set

	0	1
0	172345	23217
1	4	2

Conclusions

In this paper, we analyze several different models and their abilities to predict whether certain individuals will get a flu vaccine this season. Specifically, we consider OLS models with and without LASSO, logistic regression models with and without LASSO, a random forest model, and a neural network. The OLS and neural network models achieve the highest performance; interestingly, both models essentially predict that no one will get vaccinated, indicating that a simple heuristic of guessing that nobody would get vaccinated would match the performance or outperform all of the models in our dataset. This gives us a baseline model to work with and can inform future research; that is, we should expect any predictive model to do at least as well as this simple heuristic that always predicts 0 for every individual.

Based on our models, we can see that history is highly predictive of future behavior. Whether an individual got vaccinated the previous year and the number of flu vaccines they have received in the past are statistically significant covariates with the highest magnitudes. This makes intuitive sense because we would expect that individuals who have received flu vaccines in the past would be more likely to receive another one, and the more they have the received, the more likely they would be to get vaccinated again this season. This is similar to the marketing policy that consider recency and frequency to be the most informative indicators of whether a customer is likely to remain loyal.

However, our best performing models cannot identify true positives with sufficient accuracy. This low specificity may be informative as it would suggest that either additional data is needed to accurately identify positive vaccinations, or that a theoretical limit on prediction accuracy may exist as a result of random shocks that affect the decision to follow-through on vaccination[3]. Thus, future research should test whether incorporating additional covariates which could improve the accuracy of models, or if vaccinations are driven by some unobservable randommness that cannot be captured in existing datasets.

[3]Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. Nature, 595(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Appendix

ROC Curve Comparison of Different Models

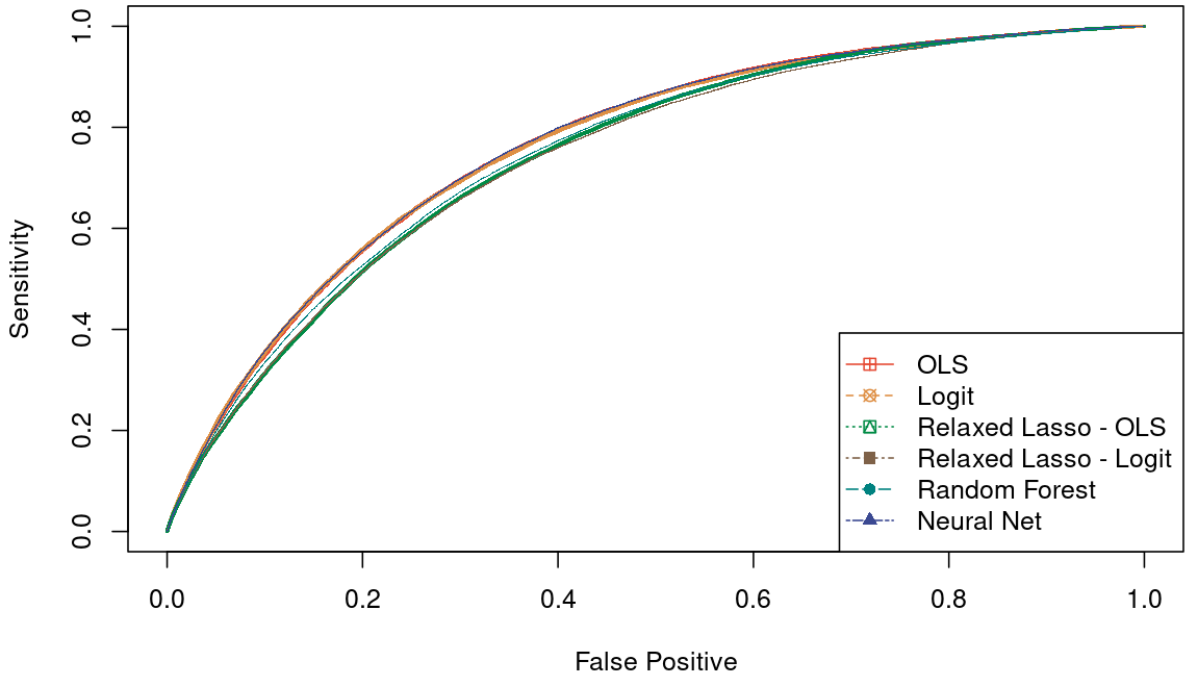

Figure 10: ROC Curve Comparison of Different Models

OLS Regression

Table 8: 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

term	estimate	std.error	statistic	p.value
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
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

term	estimate	std.error	statistic	p.value
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 9: 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 10: 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

term	estimate	std.error	statistic	p.value
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
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

term	estimate	std.error	statistic	p.value
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 11: 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 12: 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
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 13: 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 14: 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 15: 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