Predicting Flu Vaccinations using Data from a 1-Million-Person Dataset

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

I ran each of the models below by using the training set to generate a model, then evaluating the model on the test set to calculate the AUC, misclassification error, and confusion table.

Then finally, I will pick the best classifier and run it on the validation dataset to see how well it performs.

OLS w/ Classifier

Notes: * Used an OLS regression model to predict the probability of receiving a flu vaccination within 30 days of treatment. * Used a threshold of 50% to calculate the predicted class (vaccination 30 days after treatment or not)

```
confusion_table <- structure(c(180113L, 0L, 24031L, 4L), dim = c(2L, 2L), dimnames = list(
Predicted = c("0", "1"), Actual = c("0", "1")), class = "table")

auc_ols <- 0.763

misspecification_error <- 0.117713619530929
```

Logistic Regression

Notes: * Used a threshold of 50% to calculate the predicted class (vaccination 30 days after treatment or not)

```
confusion_table <- structure(c(179014L, 1099L, 23192L, 843L), dim = c(2L, 2L), dimnames = list(Predicte

auc_ols <- 0.7624

misspecification_error <- 0.118987205360817
```

Relaxed LASSO with Logit

```
confusion_table <- structure(c(178570L, 1543L, 22987L, 1048L), dim = c(2L, 2L), dimnames = list(Predict

auc_ols <- 0.7404

misspecification_error <- 0.120157924642906
```

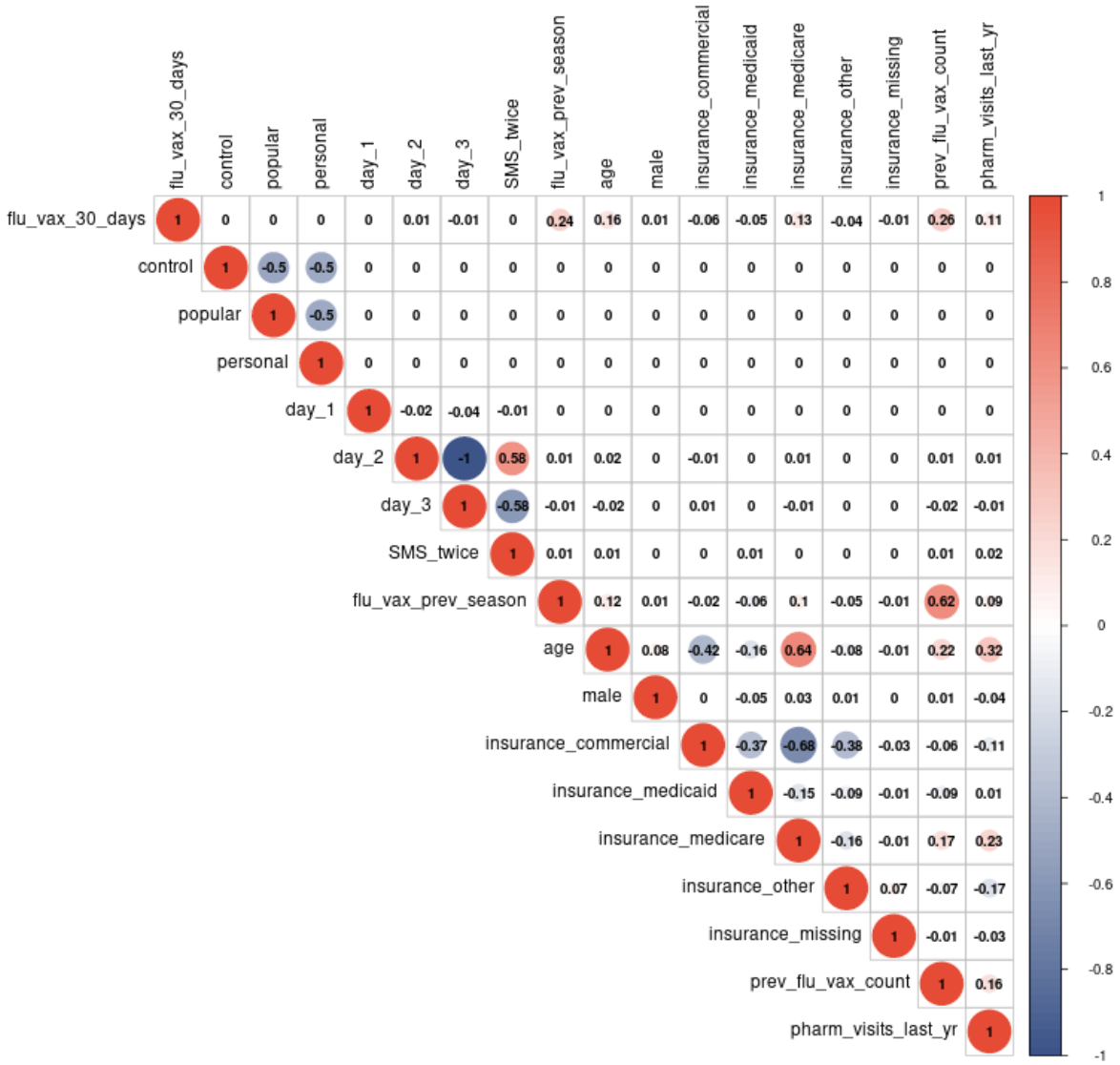

Figure 1: Spearman Correlation Plot of Key Variables

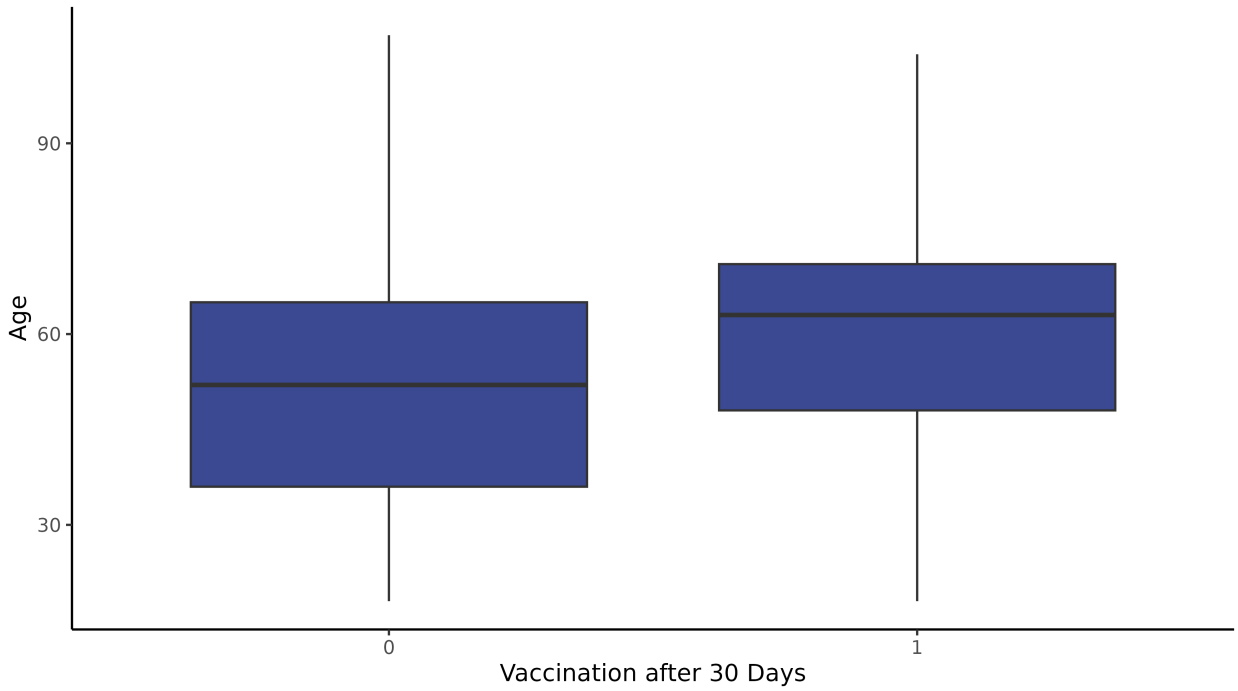

Figure 2: Boxplot of Vaccination (30 Days After Treatment) and Patient Age

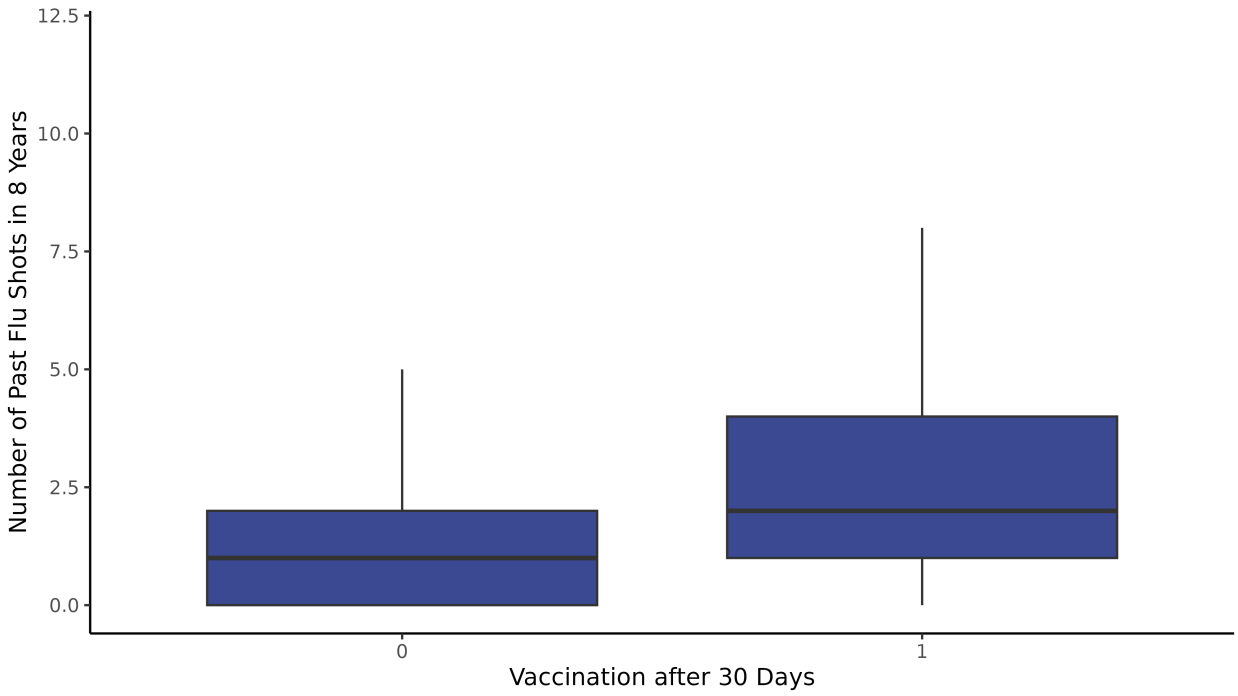

Figure 3: Boxplot of Vaccination (30 Days After Treatment) and Number of Past Flu Shots

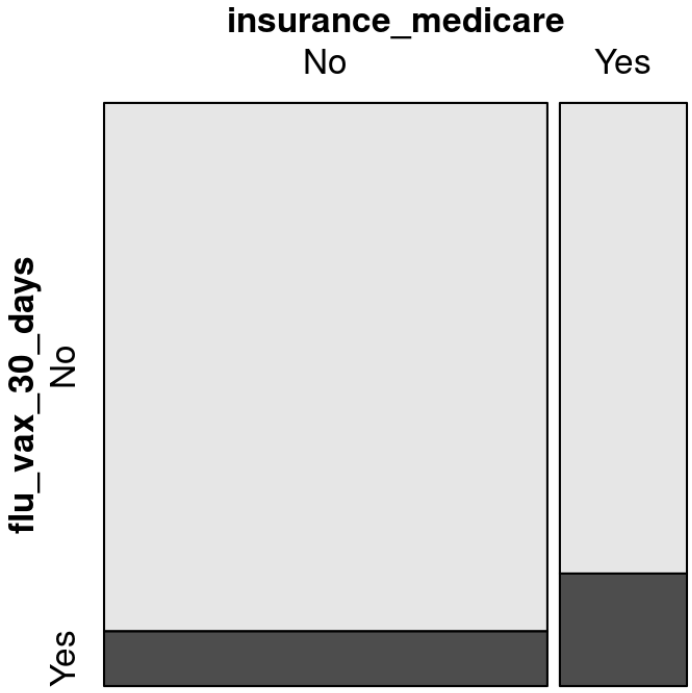

Figure 4: Mosaic Plot of Vaccination (30 Days After Treatment) and Medicare Insurance

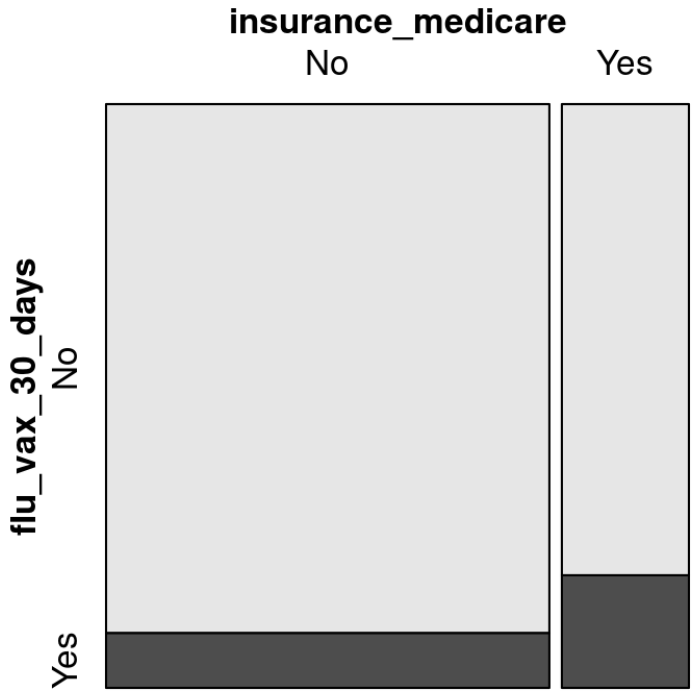

Figure 5: Mosaic Plot of Vaccination (30 Days After Treatment) and Medicare Insurance

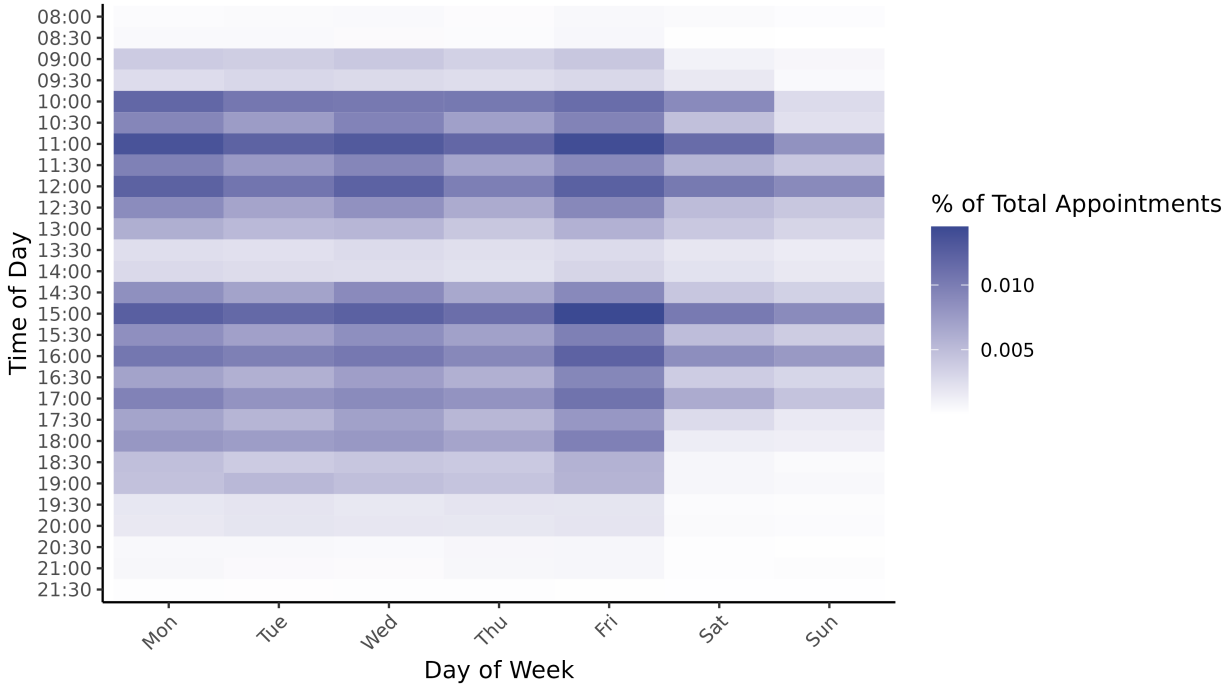

Figure 6: Heatmap of Last Vaccination Times

Relaxed LASSO with OLS

```
confusion_table <- structure(c(180040L, 73L, 23970L, 65L), dim = c(2L, 2L), dimnames = list(Predicted =

auc_ols <- 0.7444

misspecification_error <- 0.117772400415385
```

Random Forest

manually tuned the model (because r packages were not available on the secure server)

mtry = 4, ntree = 500

```
confusion_table <- structure(c(179382L, 731L, 235411, 494L), dim = c(2L, 2L), dimnames = list(Predicted

auc_ols <- 0.7489

misspecification_error <- 0.118894135627094
```

Neural Network

Ran a neural net using the nnet package in R - uses a logistic activation function - neural network with 1 hidden layer with 10 nodes - 100 iterations (the most the server could take - it was very slow)

```
confusion_table <- structure(c(180113L, 24035L), dim = 1:2, dimnames = list(Predicted = "0", Actual = c

auc_ols <- 0.7644

misspecification_error <- 0.117733213159081
```

Conclusions

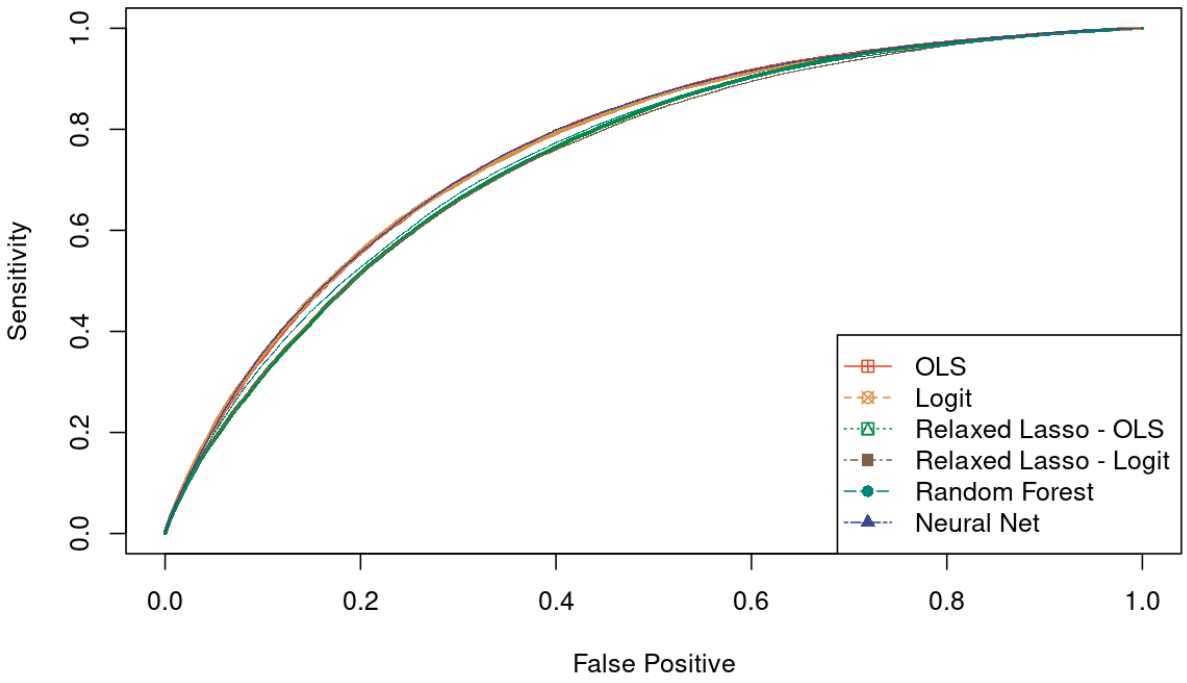

Figure 7: ROC Curve Comparison of Different Models