P8106: Final Project

Joe LaRocca

May 8, 2025

Exploratory Data Analysis

We first analyzed the distributions of both the quantitative and categorical variables in the dataset, which contained 1000 total observations, a binary response indicating whether a patient developed severe flu within six months of vaccination, and eleven predictor variables. Out of the 1000 patients, 25.3% developed a severe flu within six months of vaccination. Six of the predictors (age, height, weight, BMI, systolic blood pressure (SBP) and LDL cholesterol levels) were quantitative, while the other five (race, gender, smoking status, diabetes, and hypertension) were categorical. Since the medians are very close to the means for all six quantitative variables, we can see that they are each roughly symmetrically distributed (Table 1). 65.6% of participants were White, while 18.4%, 9.6%, and 6.4% were Black, Hispanic, and Asian, respectively (Figure 1). Slightly more than half (522, or 52.2%) of participants were female. Current smokers, former smokers, and nonsmokers represented 10.3%, 31.3%, and 58.4% of the sample, respectively. About one-seventh (14.5%) of the patients had diabetes, while slightly under half (46.4%) had hypertension.

Model Training

Training/Test Split

Before conducting our analysis, we randomly selected 80% of the data as training data and the other 20% as testing data. While we planned to select the model with a highest mean area under the curve (AUC) generated from ROC curves, as a higher AUC is associated with greater sensitivity and specificity, we also sought to use a more interpretable model, such as a generalized linear model, as our final model if there was not a large difference between simpler models and black-box models. For the data split and for each of our models, we set the seed value to 2025. Each of our models was trained using 10-fold cross-validation.

Model Selection

We aimed to start with simpler models and then proceed to more complex ones. First, we created a logistic regression model of severe flu on the training data, using all eleven predictor variables. Then, we created a boosting model using the 'caret' R package and tested the tree count, interaction depth, and shrinkage parameters using cross-validation. We tested the values (100, 200, 500, 1000, 2000) for the number of trees, (1, 2, 3, 4) for interaction depth, and (0.005, 0.01, 0.05) for the shrinkage parameter; we found the optimal values to be 100, 1, and 0.05 for the tree count, interaction, depth, and shrinkage parameters, respectively.

We then used two different methods to create models using support vector machines (SVMs). We first created a model with a linear kernel using the 'kernlab' R package, testing for cost parameters between e^{-5} and e^{2}. Finally, we created a model with a radial kernel, using the same range of cost parameter values. We selected the value for the sigma parameter, which determines the shape of the decision boundary, using the 'sigest' function from the 'kernlab' package. For the linear kernel model, we found that the optimal cost parameter was about 6.405. For the radial kernel model, interestingly, we found the optimal cost value to be about 0.016.

When choosing our final model, we wished to achieve an optimal balance of predictability and interpretability. In order to select the final model, we quantified the area under the ROC curve for each of the 10 folds under cross-validation (Figure 2). Out of our four models, the SVM model with a linear kernel achieved the highest average AUC, followed by the SVM model with a radial kernel, the logistic regression model, and the boosting model, in that order. However, even though two of the more advanced models provided superior predictive power to the logistic regression model on average, the margin was not particularly large, with the average AUC of the linear kernel SVM model and the logistic regression model only differing by about 0.03. Given the ease of interpreting a logistic regression model compared to black-box methods such as boosting and SVMs, we chose to proceed with the logistic regression model as our final model.

Results

Interpretation of Model Coefficients

Despite the other models' slightly stronger performance, we chose to use logistic regression to model the relationship between the incidence of severe flu following vaccination and a number of key demographic characteristics due to the model's interpretability. Though there were 11 predictor variables in the dataset, the model contains 15 coefficients including the intercept, as race and smoking status were divided into dummy variables with four and three separate values, respectively (Table 2).

Male gender was associated with a slightly higher risk of severe flu, though the relationship was only mildly significant (P = 0.121). Compared to participants of Asian race, there were no statistically significant associations between severe flu and race for any of the other three races, holding other characteristics constant. Both former smokers and nonsmokers, however, had significantly lower odds of contracting severe flu than current smokers (P = 0.003 and P = 0.002, respectively). Holding all other variables constant, former smokers and current smokers had $e^{-0.887} = 0.412$ times and $e^{-0.437} = 0.510$ times the odds of severe flu than current smokers, respectively. On average, patients with diabetes had higher odds of contracting severe flu (P = 0.005); the odds were multiplied by $e^{0.604} = 1.829$. Interestingly, however, there was no clear association between hypertension and severe flu (P = 0.487).

Adjusting for other variables, there was a slight negative relationship between severe flu incidence and age, though this relationship was not statistically significant (P = 0.518). Controlling for other variables, systolic blood pressure also had a very weak negative association with severe flu (P = 0.37264), possibly due to the fact that hypertension – which is a direct function of systolic blood pressure – is also included in the model. There was a strong positive association between LDL cholesterol levels and severe flu (P = 0.049). Each 1-unit increase in LDL, accounting for other variables, was associated with $e^{0.009} = 1.009$ times higher odds of severe flu on average. Given that the mean LDL level among the training data was about 110, interpreting the LDL coefficient in terms of larger increases of LDL may be more appropriate; a 20-unit increase would be associated with $e^{20 * 0.011} = 1.197$ times higher odds of severe flu.

The variable with the largest effect size, and thus the greatest impact on the odds of severe flu according to the model, was BMI (P = 0.002). A 1-unit increase in BMI was associated, on average, with $e^{1.555} = 4.735$ times the odds of severe flu, even when holding other variables constant. A 3-unit increase in BMI would be associated with a $e^{5 * 1.555} = 106$ times higher odds of severe flu, representing a far stronger effect than any of the other predictors.

Predictive Risk Scores

Predictive risk scores, or the probabilities of having experienced severe flu given individual characteristics, can be calculated by inputting values of the models to calculate the log odds, exponentiating to convert to odds, and then further converting to probability using the formula $P = O/(1 + O)$, where P and O represent probability and odds, respectively. For example, consider a 62-year-old White man who has never smoked, with hypertension but no diabetes, with a height of 156cm, a weight of 84.4kg, a BMI of 34.9, and SBP and LDL levels of 140 and 116, respectively. According to the model, his probability of severe flu would be:

$$\text{Log odds(severe flu)} = -80.927 - 62(0.015) + 0.439 - 0.044 - 0.827 + 156 * 0.422 - 84.4 * 0.454 + 34.9 * 1.555 + 0.204 + 140 * 0.011 + 116 * 0.009$$

$$\approx 2.283$$

$$\text{Probability(severe flu)} \approx (e^{2.283})(1 + e^{2.283}) = \mathbf{0.907}\ (90.7\%)$$

This particular patient would be predicted to have a high probability of contracting severe flu following vaccination, likely in part due to his high BMI.

Model Performance

Our logistic regression model correctly predicted the severe flu status of 152, or 76%, of the 200 study participants randomly sorted into the testing set. The area under the curve (AUC) for our final model was about 0.674, slightly less than the mean AUC on the training set computed using cross-validation.

Conclusion

We aimed to create a model that would predict whether a patient developed a severe flu within six months of receiving a vaccination. We created four candidate models: a logistic regression model, a boosting model, an SVM model using a linear kernel, and an SVM model using a radial kernel. Though the SVM model with a linear kernel had the best predictive power of the four models by a small margin, we ultimately chose to proceed with the logistic regression model because of its interpretability, which can provide key insights into which patient characteristics are most strongly associated with higher or lower risk of severe flu. From our model output, we inferred that BMI had the largest effect size out of the predictor variables, with height, weight, LDL cholesterol levels, smoking status, diabetes status, and gender also having significant associations with contracting severe flu. Using the model parameters, predictive risk scores can easily be calculated as long as all relevant characteristics are known; these scores can be generated not only for participants involved in the study, but also for outside patients with different characteristics. Our model can hopefully be used in the future to inform patients about their potential risks of contracting severe flu after vaccination, and can inform healthcare providers about treatments and recommendations they can make to patients to decrease their risk of becoming sick.

Appendix

Variable	Min	Q1	Median	Mean	Q3	Max
Age	46	57	60	60.1	63	72
Height (cm)	151.5	165.2	169.7	169.7	174	191.9
Weight (kg)	59.1	75.1	80.1	80.0	84.8	103.7
BMI	20.1	25.9	27.7	27.9	29.6	36.7
Systolic BP	108	124	130	129.9	135	154
LDL Cholesterol	41	98	111	110.5	123	174

Table 1: Summary Statistics for Quantitative Predictor Variables. Five-number summaries, as well as the mean of each variable, are given.

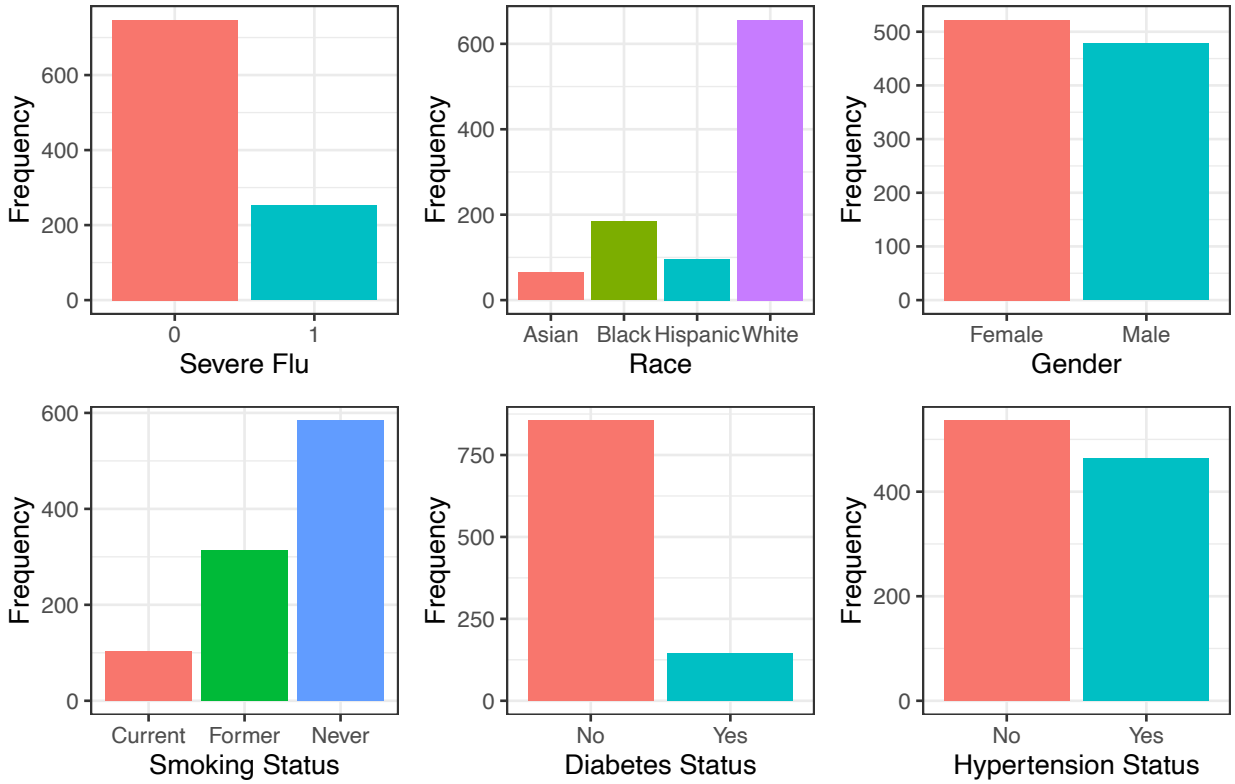

Figure 1. Distributions of Categorical Predictor Variables. Severe flu status indicates whether a patient contracted a severe flu within six months of vaccination.

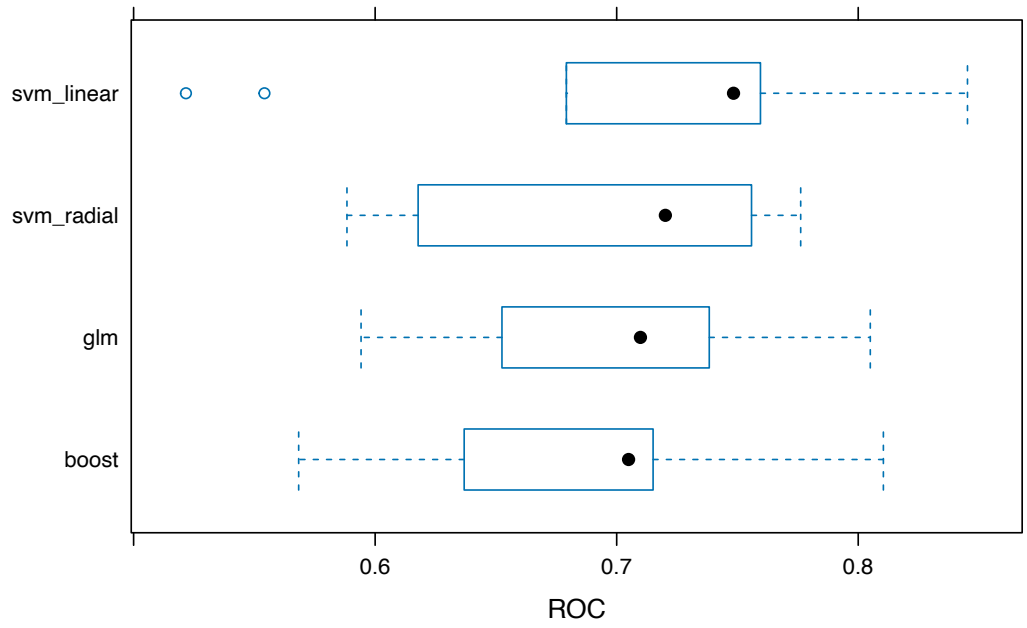

Figure 2. AUC Distributions for Candidate Models. AUCs were computed through ROC curves. 10-fold cross-validation was used to generate the distributions.

Parameter	Estimate	Standard Error	P-value
Intercept	-80.927	27.932	**0.004**
Age	-0.015	0.024	0.518
Gender = Male	0.439	0.178	**0.014**
Race = Black	-0.183	0.402	0.648
Race = Hispanic	0.326	0.43	0.445
Race = White	-0.044	0.356	0.901
Smoking = Former	-0.887	0.294	**0.003**
Smoking = Never	-0.827	0.273	**0.002**
Height	0.422	0.164	**0.010**
Weight	-0.454	0.171	**0.008**
BMI	1.555	0.49	**0.002**
Diabetes = Yes	0.604	0.237	**0.011**
Hypertension = Yes	0.204	0.294	0.487
SBP	0.011	0.019	0.572
LDL	0.009	0.005	**0.049**

Table 2. Parameter Estimates for Logistic Regression Model. Significant P-values are bolded and in **red**.

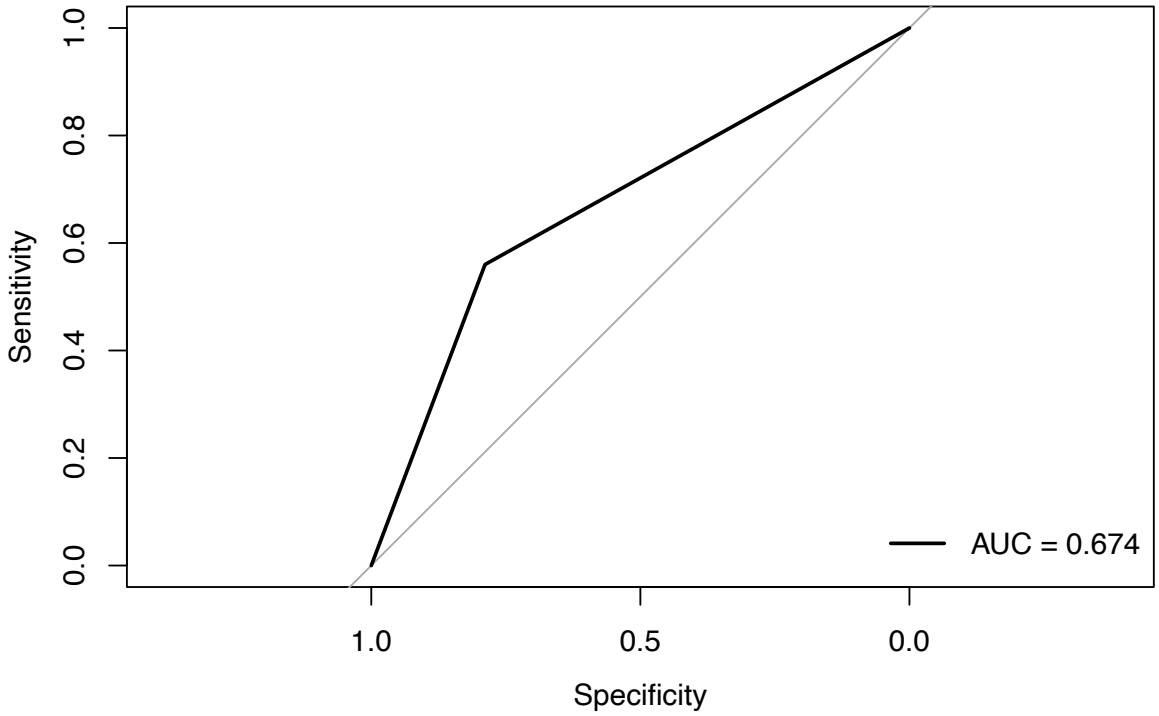

Figure 3. ROC Curve for Final Logistic Regression Model. Area under the curve (AUC) is shown in the bottom-right corner.