Data Cleaning:

The first decision we had to make as we prepared the data was what to do with some of our variables (time\_value, geologic, smoothed tested positive 14d). One method we tried removed the time value outright. Looking at federal policies, guidance, and recommendations during this timeframe, we assumed that the instances were independent. Though time surely has an effect, we follow this assumption that it will be noisy.

Next, we removed the 21,632 data points missing the target variable. In this cleaning process, we did not try to impute data from our known labels. This left us with 3994 instances of data with the target variable. The figure below shows the remaining features and how many instances of data are missing for each feature. Before filling in the missing features, we use histograms and table summaries, mainly skewness and kurtosis, to see which features are approximately normal.

Figure #

A screenshot of a computer

AI-generated content may be incorrect.

This brought us to our second decision point when preparing the data. Most of the data has some form of distribution that you could transform to make approximately normal, except geo\_value. As we were in the design stages, we were split on the end goal. We debated whether to have the local policy for these geo\_values or a generalized one. Keeping geo\_values could lead to overfitting with the training data. There are also confounding variables like population density and local or state policy we could not account for with the data provided. So for this preprocessing method we removed geographic values.

Figure #

A graph of a graph

AI-generated content may be incorrect.

After removing geo\_values, we decided to remove messy instances. These instances are missing out on over half of the feature values. This decreases the number of total instances from 3994 to 3881. We do this because we do not want to simulate over half of an instance, especially when we do not know how all the variables influence each other. The next part of our preprocessing is understanding the correlation matrix and noticing which variables could lead to multicollinearity issues in the modeling process.

Figure #

A screenshot of a graph

AI-generated content may be incorrect.

Before transforming the features, I impute some of the missing data by using a random value centered around the median or mean, depending on whether the data’s distribution is approximately normal or has a presence of outliers. We determine this by checking if the skew value is greater than |1| or the kurtosis value is greater than 3.

Figure #

A screenshot of a computer

AI-generated content may be incorrect.

Now, with the missing data filled, most of the heavily skewed and high kurtosis value features receive a log transformation. After the log transformation to reduce the skewness, we use a MinMaxScaler to scale the values while retaining their relationship. This works for most of the features to make them approximately normal, except for smoothed\_wpublic\_transit\_1d. This feature retained its heavy kurtosis using the previously mentioned transformation. To address this issue, we shifted the values to make them all positive before applying a log transformation. Then we used an extreme transformation by winsorizing the data before using a RobustScaler, which tamed the data.

Figure #

A graph with numbers and a blue line

AI-generated content may be incorrect.A graph of a blue line

AI-generated content may be incorrect.

After transforming the data, we examined the variance inflation factor and correlation between the various features and the two target variables. This gives us an idea of the variables that have some collinearity issues.

Figure #

A screenshot of a computer

AI-generated content may be incorrect.A screenshot of a computer program

AI-generated content may be incorrect.A graph of a bar graph

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A graph with different colored bars

AI-generated content may be incorrect.

Base line model:

For the baseline linear model we separate the data into two different tables two answer the different questions. We split each into train and test splits and send it directly to a linear model. Next, we did a cross-validation on the baselines. This showed a small increase in performance for the smooted wtested positive 14d model but a larger decrease in smoothed wcovid vaccinated.

Figure Baseline Results #

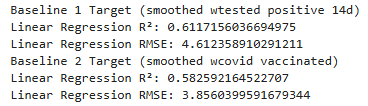
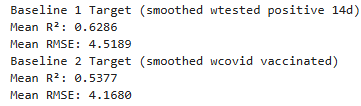
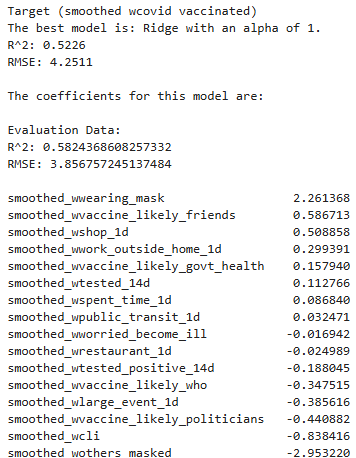
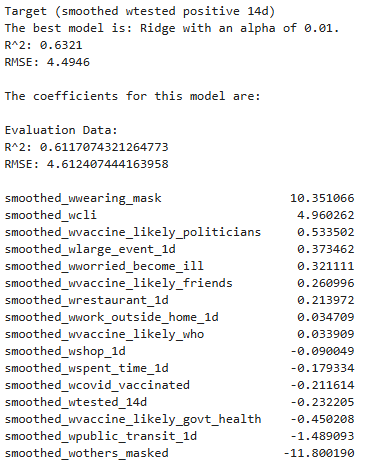


Figure Baseline cross validated #



Next, we try to create a model that performs better than the baseline. We start with by using Ridge, Lasso, and ElasticNet regression with different hyper parameters to see which model performs the best. They each go through a cross validation and the best performing training set we use the same parameters on the test sets.

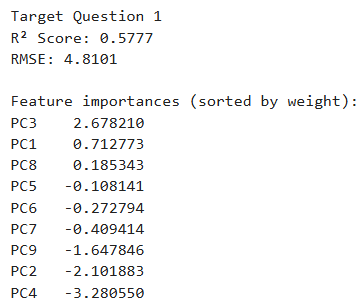
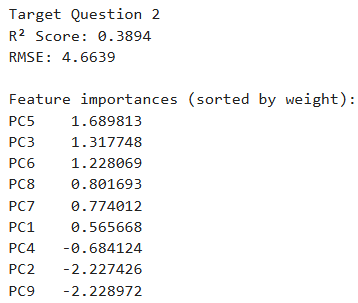
Figure Regularization #



The models did not perform better on the test dataset. This is expected considering the data is complex and not strictly linear. The next step we attempt is doing feature selection. We decide to retain 95% of the variance which decreases the number of features to 9. Reducing the feature space made the models perform worse on the data.

Figure Principal Components #



Random Forest Regressor Model

Using the random forest regressor model yielded significantly better results. After tuning the parameters for the tree, we used the model on the test set and achieved similar results. Here, we examined differences using both a small and large feature space. The small feature space performed better on the Covid Vaccine Data set. While the Positive Cases performed better using the larger feature space.

Figure # Random Forest Small Feature Space

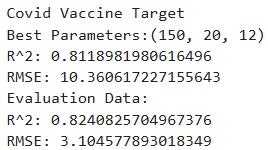
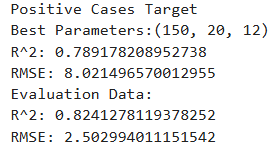
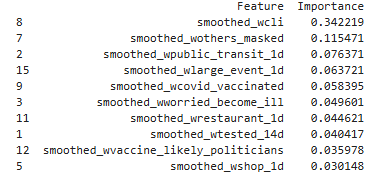
 

Figure # Covid Vaccine Feature Importance



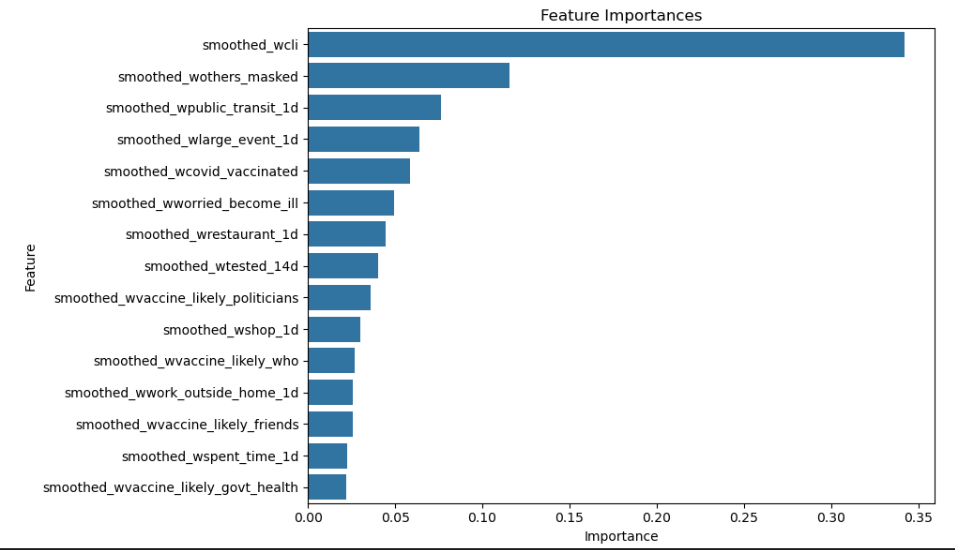
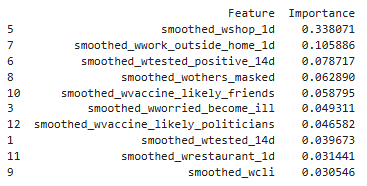


Figure # Positive Cases Feature Importance



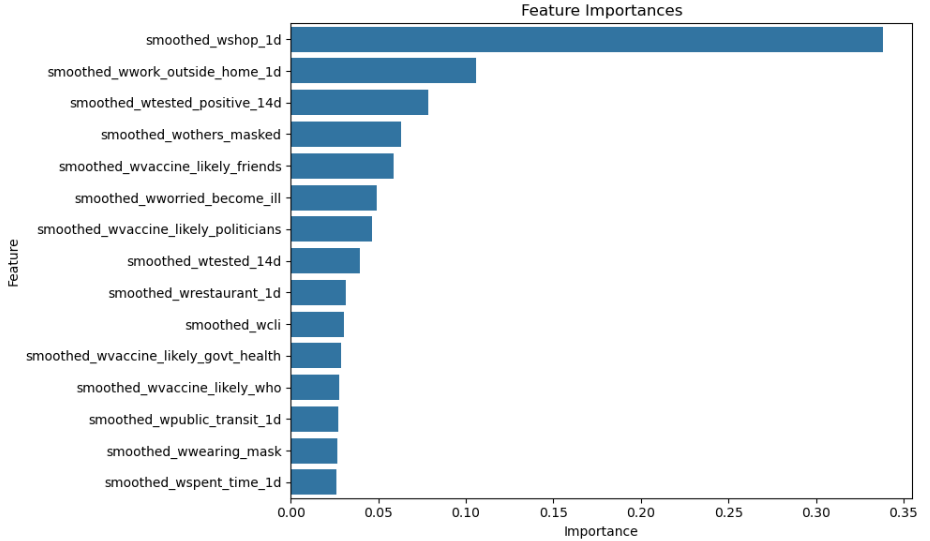
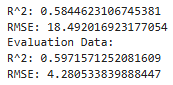
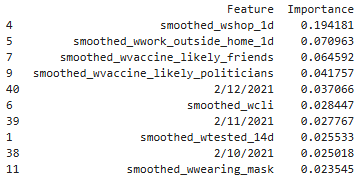
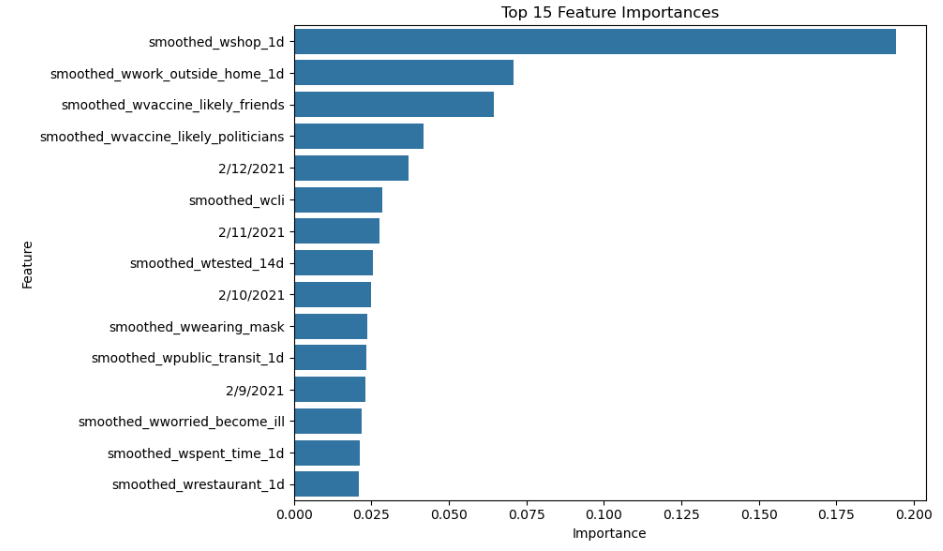


Figure # Random Forest Large Feature Space

Covid Vaccine







Positive Cases

