COVID-19 Behavior & Belief-Based Prediction Report

By Team !404 Team Found (Ian Hash, Kavi Sarna, Rishika Thiruvengadam Dwaraghanath and Vinayak Prasad)

1. Introduction

The COVID-19 pandemic exposed major disparities in test positivity and vaccination rates across U.S. communities, highlighting the urgent need for data-driven public health interventions.

In this project, acting as CDC data scientists, we address two tasks:

- Predict the 14-day COVID-19 test positivity rate at the county level.
- Predict the cumulative COVID-19 vaccination rate at the county level.

We use the COVID-19 Trends and Impact Survey (CTIS), a large-scale, county-level dataset collected by Carnegie Mellon University's Delphi Research Group in early 2021 (Carnegie Mellon University Delphi Research Group, 2021). The dataset contains 25,627 daily county-level records from January 7, 2021, to February 12, 2021, capturing self-reported behavioral indicators (e.g., mask-wearing, public gatherings, work location) and belief indicators (e.g., trust in vaccine recommendations).

The target variables are:

- smoothed_wtested_positive_14d: 14-day smoothed COVID-19 positivity rate.
- smoothed_wcovid_vaccinated: Cumulative percentage vaccinated.

Accurate predictions of these outcomes can guide public health resource allocation, vaccination outreach, and policy decisions, ultimately reducing pandemic impacts in vulnerable communities.

2. Data Analysis

2.1 Data Cleaning

We began by creating **two separate datasets** for the two prediction tasks:

- Dataset 1: Predicting Positive COVID-19 Test Rate (smoothed_wtested_positive_14d)
- Dataset 2: Predicting Vaccination Rate (smoothed_wcovid_vaccinated)

Target-Based Filtering:

- Dataset 1: Rows where smoothed_wtested_positive_14d was missing were removed.
- Dataset 2: Rows where smoothed_wcovid_vaccinated was missing were removed.

Feature Filtering:

- Dataset 1: For predicting positive tests, we removed belief indicators and low-correlation behaviors (e.g., vaccine trust, shopping, transit use).
- Dataset 2: For predicting vaccination rates, we kept only trust in friends (smoothed_wvaccine_likely_friends) and dropped other beliefs and low-correlation behavior features.

Handling Remaining Missing Values:

- Used **mean imputation** to fill missing values across all non-target features.
- **Rationale**: Mean imputation was chosen for its simplicity and effectiveness, given the low missing rates. This preserved the dataset size and distribution.

Final dataset sizes:

- Dataset 1: 3994 rows
- Dataset 2: 23024 rows

2.2 Feature Selection / Transformation

We analyzed the correlation between each feature and the target variable for both tasks and selected only strongly correlated features.

Feature Selection:

Dataset 1 (Positive Test Rate)

- Features retained: smoothed_wcli, smoothed_wlarge_event_1d, smoothed_wrestaurant_1d, smoothed_wspent_time_1d, smoothed_wwork_outside_home_1d, smoothed_wothers_masked, smoothed_wcovid_vaccinated
- Features dropped: Belief indicators (smoothed_wvaccine_likely_friends, smoothed_wvaccine_likely_who, etc.), smoothed_wshop_1d, smoothed_wtested_14d, smoothed_wpublic_transit_1d
- Visualization used: correlation bar plot *(Figure 2.2.1)*
- **Rationale**: Selected features had a correlation magnitude above ~0.15 with the target, with meaningful interpretation based on domain logic (e.g., illness and exposure behaviors linked to test positivity).

Dataset 2 (Vaccination Rate)

- Features retained: smoothed_wshop_1d, smoothed_wwork_outside_home_1d, smoothed_wspent_time_1d, smoothed_wrestaurant_1d, smoothed_wlarge_event_1d, smoothed_wvaccine_likely_friends, smoothed_wpublic_transit_1d, smoothed_wothers_masked, smoothed_wwearing_mask, smoothed_wvaccine_likely_govt_health
- Features dropped: Other belief-based features (smoothed_wvaccine_likely_who, smoothed_wvaccine_likely_politicians, smoothed_wworried_become_ill), Low-correlation behavior features (smoothed_wtested_positive_14d, smoothed_wcli, etc.)
- Visualization used: correlation heatmap *(Figure 2.2.2)*
- Rationale: Selected features showed positive correlation with vaccination rate and represented meaningful behavioral and belief indicators without introducing redundancy.

Feature Transformation:

Dataset 1 (Positive Test Rate)

Transformation Type	Feature(s) Transformed	Description
Feature Engineered	‘interaction_cli_vaccinated’ = 'smoothed_wcli' * 'smoothed_wcovid_vaccinated'	Created an interaction feature between COVID-like illness rates and vaccination rates to capture how vaccination reduces illness symptoms, enhancing predictive power.
Feature Engineered	'interaction_events_restaurant' = 'smoothed_wlarge_event_1d' * 'smoothed_wrestaurant_1d'	Combined attendance at large events and restaurant visits to model increased COVID exposure risk from compounded social activities.
Feature Engineered	'cli_squared' = ('smoothed_wcli’)^2	Squared the COVID-like illness percentage to emphasize the impact of high symptom levels on test positivity rates.
Feature Engineered	'vaccinated_cubed' = ('smoothed_wcovid_vaccinated')^3	Cubed the vaccination percentage to capture the stronger effect of high vaccination coverage on reducing test positivity.
Feature Scaling	All Features	Standardized all features to zero mean and unit variance to ensure consistent weighting during model training.

Dataset 2 (Vaccination Rate)

- Standardization: Used the StandardScaler to scale to a standard normal distribution by removing the mean from them and dividing by their standard deviation. This helps make the model converge faster and give equal importance to all the features.
- Transformations: None

2.3 Correlation Analysis

Method:

- Pearson correlation coefficients
- Visualized with: heatmap / bar plot (Figures 2.2.1 and 2.2.2)

Positive Test Rate Correlation Highlights:

- Strong positive: smoothed_wcli (corr ≈ 0.64)
- Strong negative: smoothed_wothers_masked (corr ≈ -0.42)
- Multicollinearity issues: smoothed_wvaccine_likely_who and smoothed_wvaccine_likely_govt_health (corr ≈ 0.80 based on general belief feature behavior)
- Resolution: Dropped belief-based features to avoid multicollinearity; retained only masking and vaccination features with caution.

Vaccination Rate Correlation Highlights:

- Top positive correlations: smoothed_wshop_1d(corr≈0.40), smoothed_wwork_outside_home_1d(corr≈0.36)
- Weak/Unexpected: smoothed_wvaccine_likely_govt_health (corr ≈ -0.07)
- Multicollinearity handled via: Dropping redundant belief-based features (retained only smoothed_wvaccine_likely_friends to represent social influence)

3. Baseline Modeling

3.1 Positive Test Rate Prediction

- **Model:** Linear Regression (no regularization)

Linear regression was selected as a simple, interpretable baseline for comparison against more complex models. It provides clear insights into feature relationships, helping public health decision-makers understand predictors of COVID-19 positivity.

As reported in section **3.1.Metrics** below, the final R^2 value of 0.47 indicated that the model explains 47% of the variance of the target variable explained by the model.

- **Split:** 70/30(15/15) Train/Test(Validation/Test)

A 70/15/15 split into Train/Validation/Test sets was used. Stratified sampling based on quartiles of the target variable (smoothed_wtested_positive_14d) ensured that all subsets retained similar distributions of COVID-19 positivity rates. This helped prevent bias, maintained representativeness across splits, and improved model generalizability.

- **Cross-validation:** k-fold (k = 5)

We applied 5-fold cross-validation during model training and evaluation. Using k=5 strikes a balance between bias and variance: a smaller k would increase bias, while a larger k could increase variance. With ~4,000 records available, 5-fold CV ensures that the model is tested on diverse subsets without overfitting to any specific fold.

- **Metrics:**

In Sample Error				Out Sample Error			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
MAE	4.1271	±	0.0992	MAE	3.9982	±	n/a
RMSE	5.2297	±	0.0984	RMSE	5.1177	±	n/a
R^2	0.4794	±	0.0249	R^2	0.5246	±	n/a

MAE: Mean Absolute Error (MAE) measures the average magnitude of prediction errors, treating all deviations equally, making it intuitive for public health applications where both under- and over-estimation can affect resource allocation.

RMSE: Root Mean Square Error (RMSE) penalizes larger errors more heavily than MAE, offering a sensitivity check for extreme mistakes that could have amplified consequences during a public health crisis.

R-Squared: R-Squared (R^2) indicates that approximately 48% of the variability in positive test rates is captured by the model, providing a useful baseline for evaluating improvements through more complex methods.

3.2 Vaccination Rate Prediction

- **Model:** Linear Regression (no regularization)
  Reference section 3.1 for the reasoning behind using this model.
- **Split:** 70/15/15
  Reference section 3.1 for the reasoning behind using this split.
- **Cross-validation:** k-fold (cv = 5)
  Reference section 3.1 for the reasoning behind using this technique.
- **Metrics:**
  Reference section 3.1 for the reasoning behind using these metrics.

In Sample Error				Out Sample Error			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
RMSE	5.8069	±	n/a	RMSE	5.8118	±	0.0564
MAE	4.5477	±	n/a	MAE	4.5513	±	0.0481
R^2	0.2418	±	n/a	R^2	0.2401	±	0.0089

4. Improved Modeling

4.1 Positive Test Rate

4.11 Best Performing Model - Random Forest Optimized (Third and Fourth Iteration)

- **Model:** Random Forest Optimized

Building on the baseline, we implemented a Random Forest model, which significantly improved performance. Random Forest, as an ensemble of decision trees, is well-suited to capture complex, non-linear relationships among behavioral and belief-based features. Its robustness to outliers and multicollinearity made it an appropriate choice for this dataset, which included interaction and polynomial features.

Hyperparameter Tuning:

We applied RandomizedSearchCV to optimize key parameters such as n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features, achieving a stronger, more generalized model. (Full grid search details are available in Appendix 7.)

- **Best Model Rationale**

The optimized Random Forest model explained approximately 74% of the variance in COVID-19 positive test rates (test/held-out data), a substantial improvement from the baseline. Both MAE and RMSE decreased significantly, indicating fewer and smaller errors across predictions.

Feature Importance: The top predictors influencing positivity rates were: COVID-like illness symptoms (smoothed_wcli, cli_squared), Reported mask-wearing by others in public (smoothed_wothers_masked), Social activities like restaurant visits (smoothed_wrestaurant_1d). These findings highlight actionable areas for public health interventions, such as promoting mask usage and encouraging symptomatic individuals to self-isolate.

Understanding these key features provides actionable guidance for public health strategies. For example, policies could focus on encouraging individuals with COVID-like symptoms to quarantine and reinforcing indoor mask mandates. The feature importance analysis offers strong quantitative support for such interventions, strengthening the foundation for evidence-based decision-making.

- **Split:** 70/30(15/15) Train/Test(Validation/Test)

The same split was used for all Vaccination Rate models. Please see section 3.1 for rationale.

- **Cross-validation:** k-fold (k = 5)

The same cross-validation was used for all Vaccination Rate models. Please see section 3.1 for rationale.

- **Metrics:**

Best Model: Random Forest Optimized (Fourth Iteration):

In Sample Error				Out Sample Error			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
MAE	3.0336	±	0.0823	MAE	2.8949	±	n/a
RMSE	4.0314	±	0.0821	RMSE	3.7867	±	n/a
R^2	0.6908	±	0.0106	R^2	0.7397	±	n/a

Random Forest (Third Iteration - Non-Optimized):

In Sample Error				Out Sample Error			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
MAE	3.2178	±	0.0912	MAE	3.0581	±	n/a
RMSE	4.2078	±	0.1002	RMSE	3.9442	±	n/a
R^2	0.6634	±	0.0075	R^2	0.7176	±	n/a

The same metrics were used for all Vaccination Rate models. Please see section 3.1 for rationale.

4.12 Ridge Regression (Second Iteration)

- **Model:** Ridge Regression (Regularized Linear Model)

As a second modeling iteration, Ridge Regression was applied to introduce L2 regularization, aiming to prevent overfitting and improve model generalization. Ridge also helps mitigate multicollinearity by penalizing large feature coefficients.

Hyperparameter Tuning: A grid search over α values {0.01, 0.1, 1.0, 10.0, 100.0} identified an optimal alpha of 1.0, balancing model bias and variance.

Interpretation: Despite regularization, Ridge Regression showed no significant improvement compared to the baseline Linear Regression. This suggests that overfitting and multicollinearity were not major concerns in this dataset, and additional model complexity did not provide performance gains.

- **Split:** 70/30(15/15) Train/Test(Validation/Test)

The same split was used for all Vaccination Rate models. Please see section 3.1 for rationale.

- **Cross-validation:** k-fold (k = 5)

The same cross-validation was used for all Vaccination Rate models. Please see section 3.1 for rationale.

- **Metrics:**

In Sample Error				**Out Sample Error**			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
MAE	4.1273	±	0.0987	MAE	3.9969	±	n/a
RMSE	5.2295	±	0.0978	RMSE	5.1163	±	n/a
R^2	0.4794	±	0.0249	R^2	0.5249	±	n/a

The same metrics were used for all Vaccination Rate models. Please see section 3.1 for rationale.

4.2 Vaccination Rate

4.2.1 Best Performing Model - RandomForestRegressor

For predicting vaccination rates, Random Forest outperformed other models such as Linear Regression, Ridge, Lasso, Decision Trees, and Gradient Boosting. As an ensemble method, Random Forest captures complex, non-linear patterns in the data while remaining robust to overfitting, making it well-suited for this task.

Hyperparameter Tuning:

We conducted a grid search over parameters including n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, and bootstrap options. (Full tuning details are listed in Appendix 7.) **The best hyperparameters identified were:** *n_estimators: 300, max_depth: 20, min_samples_split: 2, min_samples_leaf: 1, max_features: 'sqrt', bootstrap: False*

Cross-Validation: A 5-fold cross-validation was performed on the training set (70%) to ensure model robustness and to avoid overfitting on any particular subset of the data.

Performance:

Below is the table depicting the performance of the model for the training and validation samples, which helped us drill down and figure out that it is the best model. It was able to achieve an RMSE value of 4.7552 and R2 of 0.4945 on the test dataset. (Figure 4.2.1 and Figure 4.2.2)

In Sample Error				Out Sample Error			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
RMSE	0.9783	±	n/a	RMSE	4.9221	±	0.0798
MAE	0.5484	±	n/a	MAE	3.7371	±	0.0498
R^2	0.9785	±	n/a	R^2	0.4550	±	0.0033

Interpretation: The optimized Random Forest model explained approximately 46% of the variance in vaccination rates, showing about a 90% improvement over the baseline linear models. MAE and RMSE were also substantially lower, indicating more accurate and stable predictions across different counties.

4.2.2 Model - Gradient Boosting Regressor

For predicting vaccination rates, Gradient Boosting Regressor is another good model and it too outperformed other models such as Linear Regression, Ridge, Lasso, and Decision Trees. As an advanced ensemble method, it builds trees sequentially and reduces overfitting and gives superior results. Even though it is a good ensemble method and performed equally good as Random Forest, it still was not able to beat its results.

Hyperparameter Tuning: We conducted a grid search over parameters including n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf, max_features, and subsample options. (Full tuning details are listed in Appendix 7.) **The best hyperparameters identified were:** *learning_rate: 0.05, n_estimators: 300, max_depth: 9, min_samples_split: 2, min_samples_leaf: 2, max_features: 'sqrt', subsample: 0.9*

Cross-Validation: A 5-fold cross-validation was performed on the training set (70%) to ensure model robustness and to avoid overfitting on any particular subset of the data.

Performance:

In Sample Error				Out Sample Error			
Metric	**Value**		**Standard Deviation**	**Metric**	**Value**		**Standard Deviation**
RMSE	2.4296	±	n/a	RMSE	4.9320	±	0.0532
MAE	1.7941	±	n/a	MAE	3.7730	±	0.0368
R^2	0.8673	±	n/a	R^2	0.4528	±	0.0038

Interpretation: The optimized Gradient Boosting model explained approximately 45% of the variance in vaccination rates, showing about a 90% improvement over the baseline linear models. MAE and RMSE were also substantially lower, indicating more accurate and stable predictions across different counties.

5. Policy Recommendations

5.1 Analysis and Limitations

- **Assumptions**: The self-reported nature of survey responses introduces potential bias.
- **Limitations**: Aggregated county-level data may mask within-county variations.
- **Data Constraints**: Static snapshot from early 2021; behavior changes post-vaccine rollout are not captured.

5.2 Reducing Positive Test Rate

- **Key behaviors identified**:
    - Reporting COVID-like illness symptoms (smoothed_wcli, cli_squared)

    - Visiting restaurants, bars, or cafes (smoothed_wrestaurant_1d)
    - Mask-wearing behavior (smoothed_wothers_masked)
- **Recommendations**::
    - Strengthen public health messaging encouraging symptomatic individuals to stay home, seek testing, and self-isolate as needed, particularly in communities with high illness reporting but low testing or vaccination rates.
    - Reinforce indoor mask mandates in public spaces (restaurants, large events, gatherings) during periods of rising case counts.
- **Supporting data**:
    - smoothed_wcli had high feature importance in Random Forest, strong positive correlation with test positivity (exact corr not specified).
    - smoothed_wothers_masked had an important negative correlation with positivity rates (higher mask-wearing → lower positivity).

5.3 Improving Vaccination Rate

- **Key beliefs identified**: Trust in friends' likelihood to get vaccinated (smoothed_wvaccine_likely_friends)
- **Recommendations**
    - Launch community-based vaccine endorsement initiatives, encouraging vaccinated individuals to share their experiences within personal networks.
    - Promote "friends and family" campaigns leveraging social proof to build vaccine confidence and normalize vaccination behavior.
- **Supporting data**: Trust in friends' vaccination decisions was a top predictor of individual vaccination status, highlighting the role of social influence in vaccine uptake.

5.4 Additional Suggestions

- **Update survey data regularly** to track evolving behavior and beliefs, especially with emerging variants or new vaccines.
- **Incorporate mobility and hospitalization datasets** for richer predictive modeling and to better understand how behaviors translate to real-world health outcomes.
- **Conduct A/B testing** of intervention strategies (e.g., mask mandates, social campaigns) to directly measure impact on positivity and vaccination rates.

6. Conclusion

This project demonstrates the power of machine learning in uncovering key behavioral and belief-based drivers of COVID-19 outcomes. Through systematic data cleaning, feature selection, feature engineering, and model development, we accurately predicted both COVID-19 positive test rates and vaccination rates. Our results highlight the critical influence of self-reported illness symptoms, mask-wearing behaviors, and social activity levels on COVID-19 spread. Similarly, vaccination uptake correlated more strongly with trust in personal social networks than institutional authorities, suggesting that interventions must leverage both individual behaviors and social dynamics.

The optimized Random Forest models significantly outperformed simpler baselines like Linear Regression and Ridge Regression, emphasizing the value of capturing non-linear patterns in complex datasets. Proper model validation, hyperparameter tuning, and strict separation of test data were essential to achieve reliable results.

Several limitations were acknowledged. Reliance on self-reported survey data introduces potential biases, and the static snapshot from early 2021 may not reflect evolving behaviors or policies. County-level aggregation could also mask local heterogeneity. Future work could integrate mobility tracking, hospitalization records, or variant-specific data to strengthen models. Time-series forecasting and more complex techniques like Gradient Boosting or Neural Networks could further improve predictive power, albeit with reduced interpretability.

Ultimately, this project illustrates that data-driven models can guide targeted public health messaging, optimize vaccination efforts, and inform evidence-based pandemic response strategies, helping build more resilient communities.

References

Carnegie Mellon University Delphi Research Group. (2021). *COVID-19 Trends and Impact Survey (CTIS)*. Carnegie Mellon University.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Tools Used: Jupyter Notebooks (JupyterLab environment).

Appendix

Figures

Figure 2.2.1

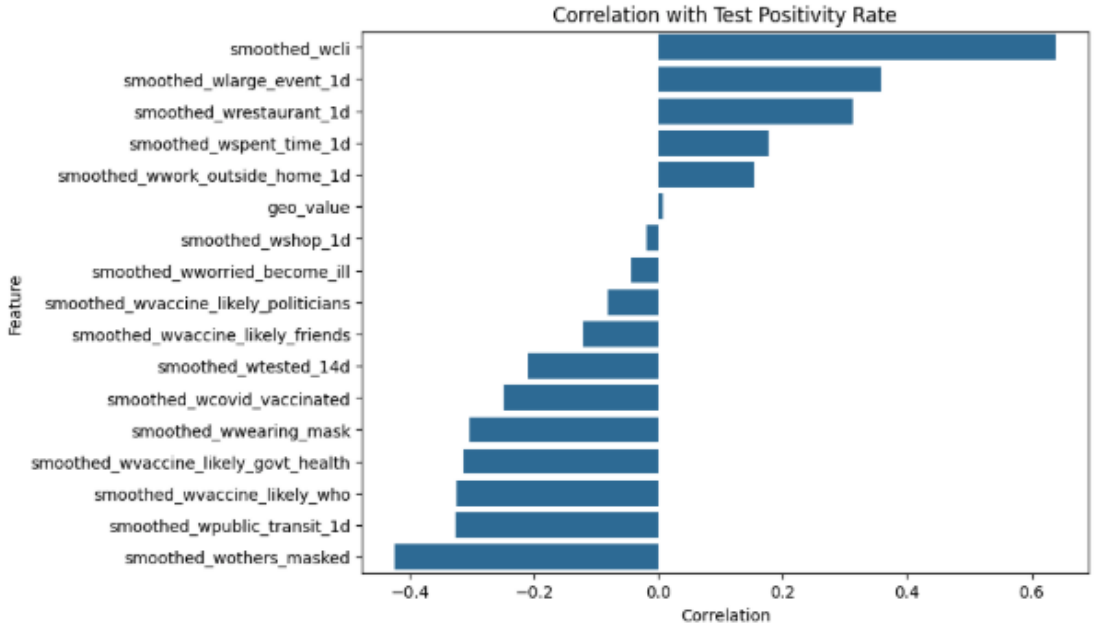


Figure 2.2.2

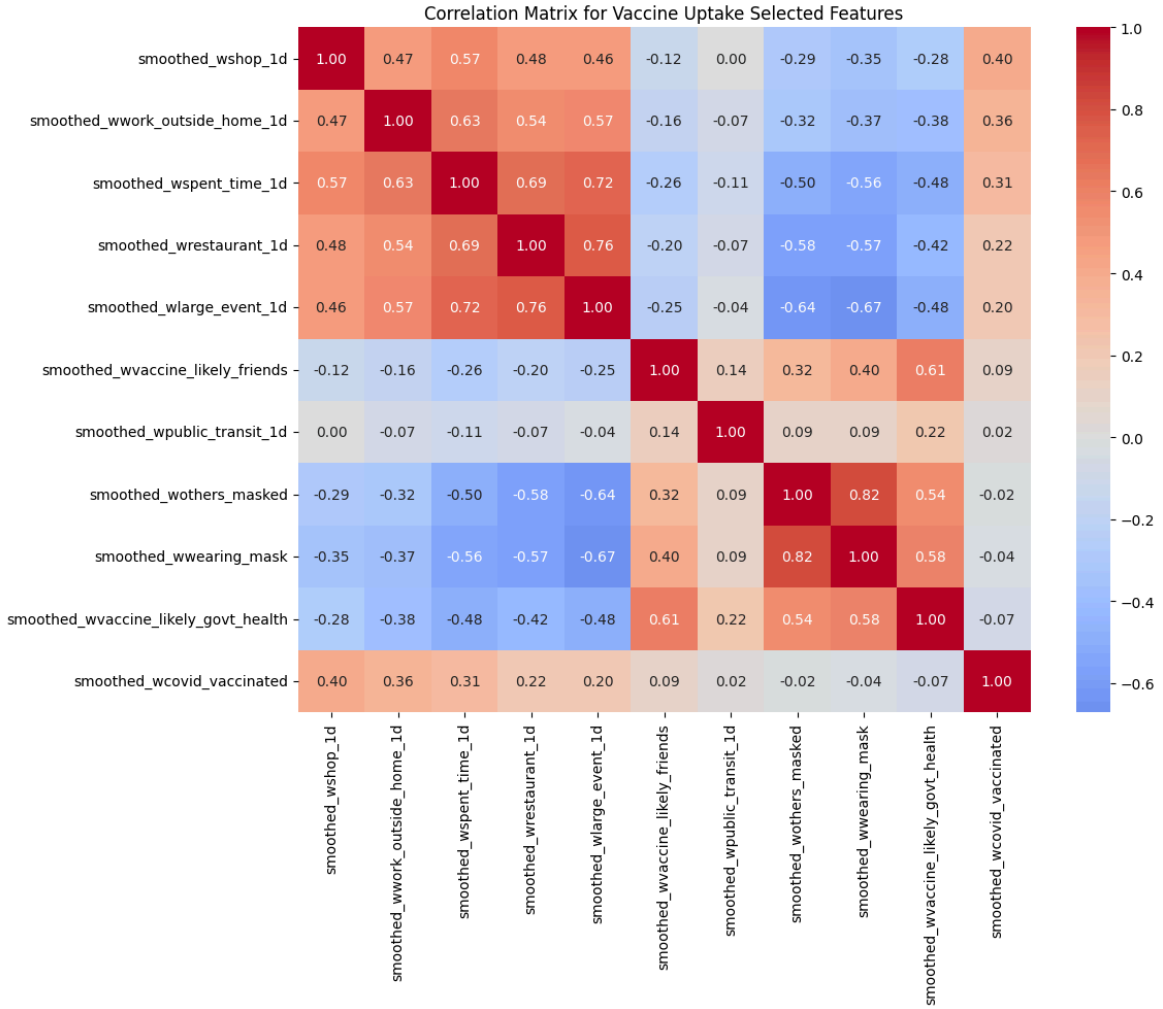


Figure 4.11.1

```
Python
  'n_estimators': [50, 100, 200],
   'max_depth': [None, 10, 20, 30],
   'min_samples_split': [2, 5, 10],
   'min_samples_leaf': [1, 2, 4],
   'max_features': ['sqrt', 'log2', None]
_____
Best hyperparameters: {'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'log2', 'max_depth': 30}
```

Figure 4.11.2

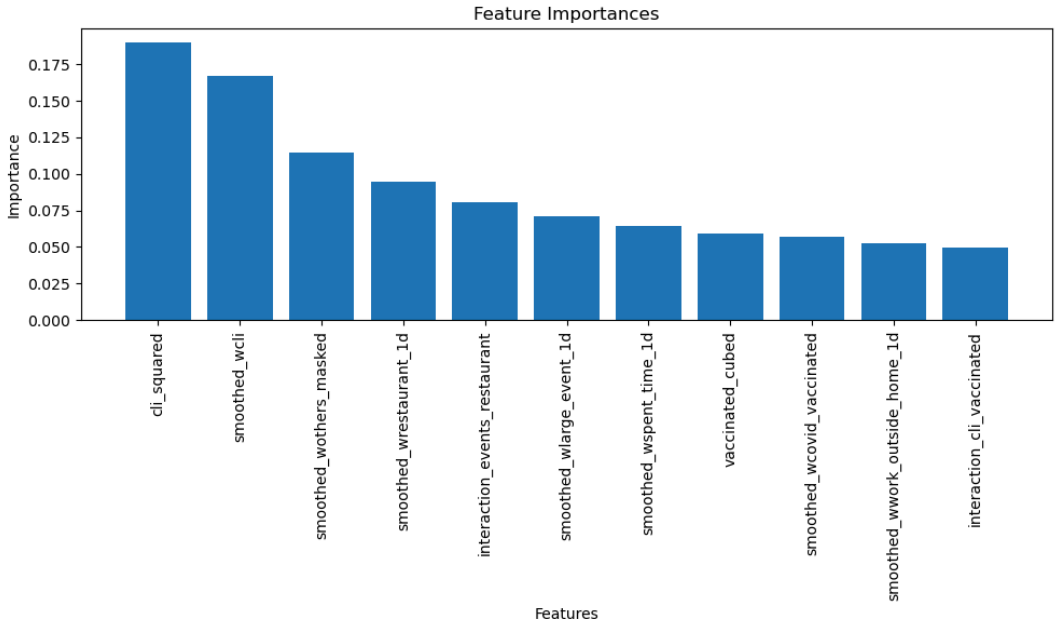


Figure 4.11.3

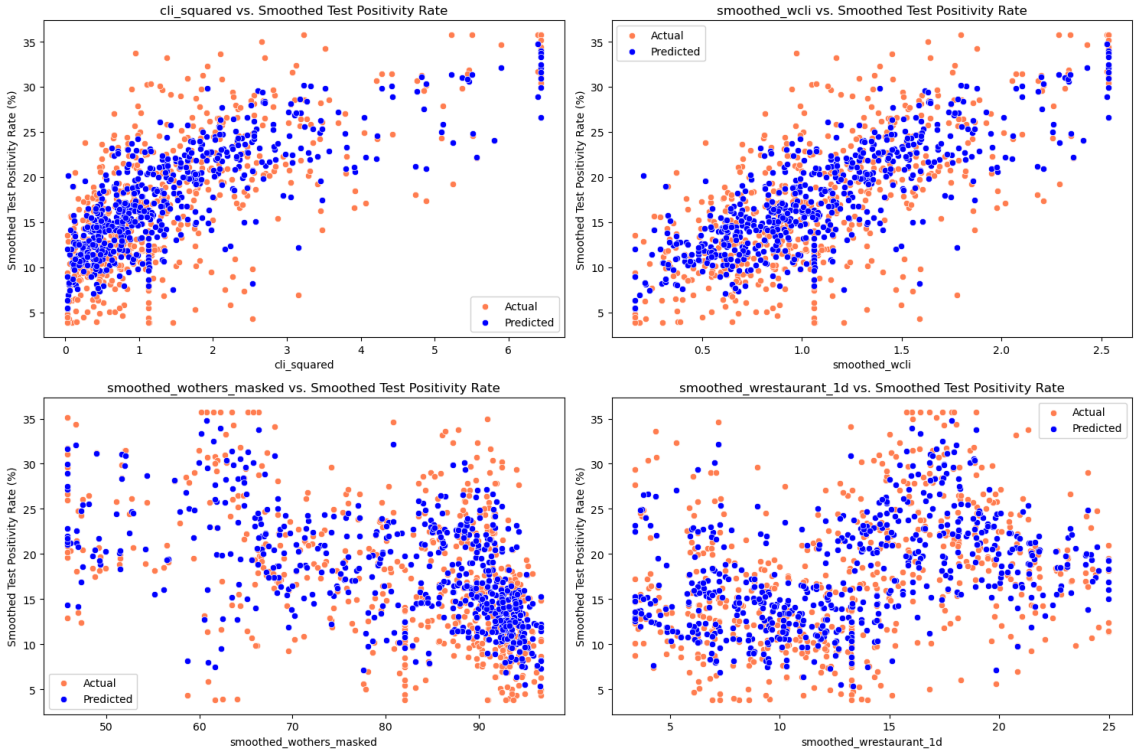


Figure 4.11.4

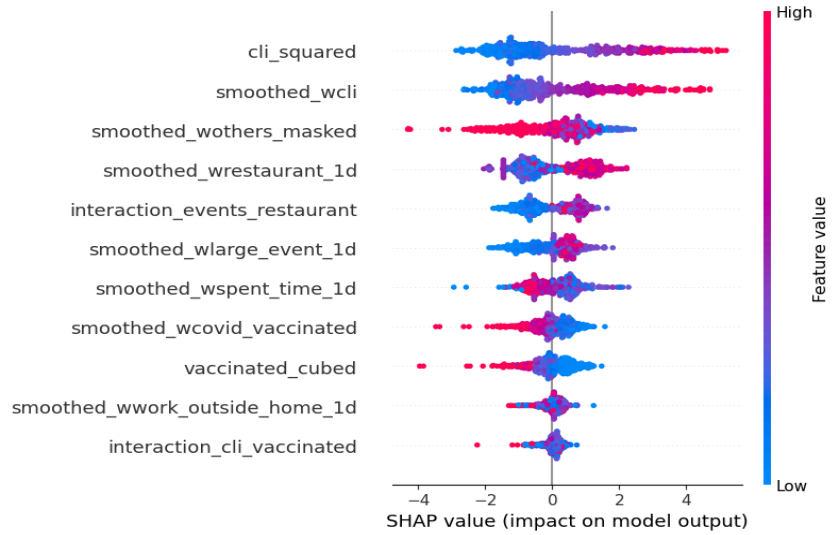


Figure 4.2.1

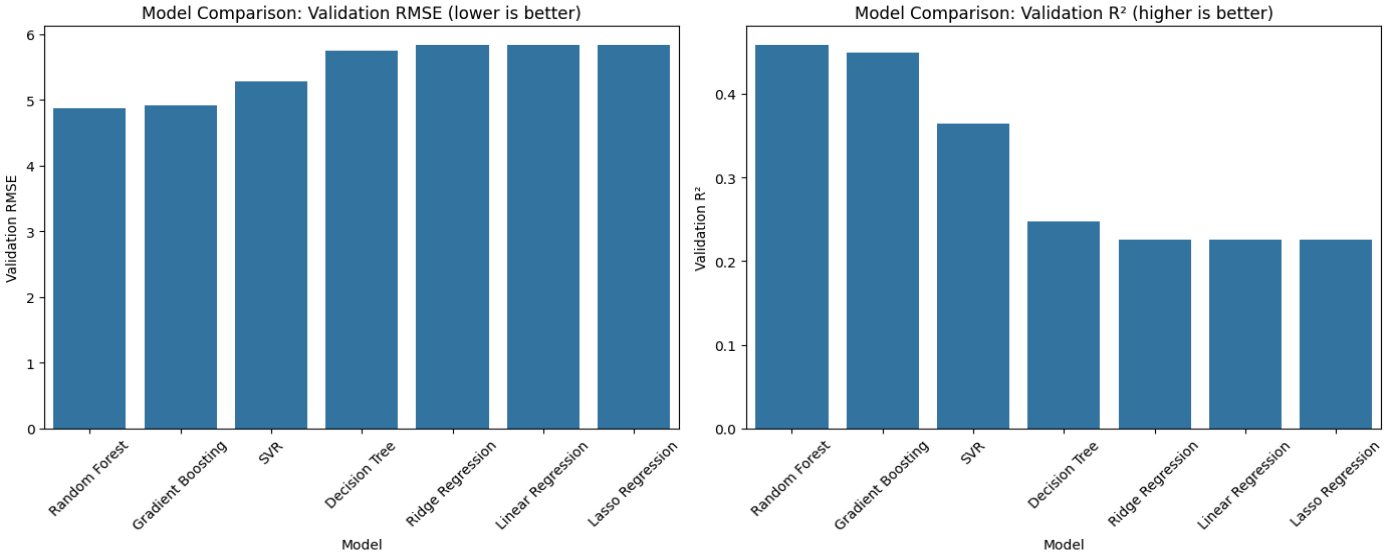
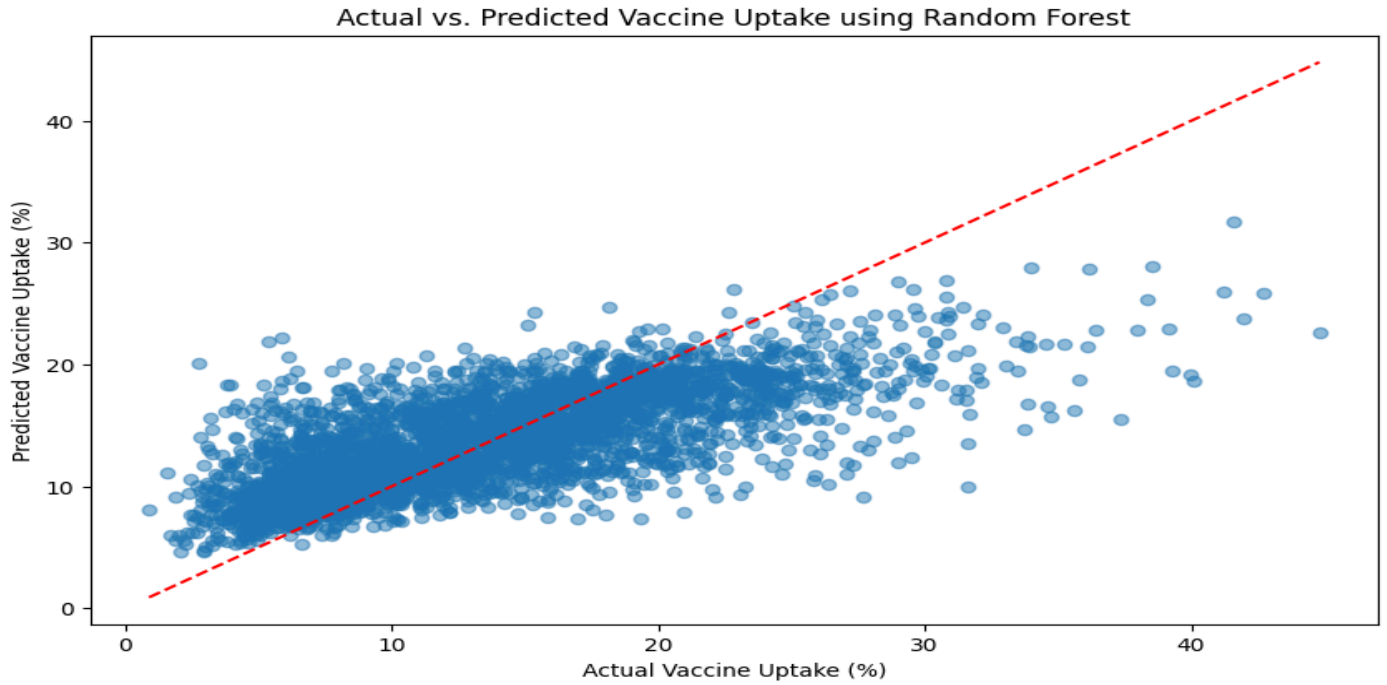

Model Comparison: Validation RMSE (lower is better)
Validation RMSE
Model
Random Forest
Gradient Boosting
SVR
Decision Tree
Ridge Regression
Linear Regression
Lasso Regression
Model Comparison: Validation R² (higher is better)
Validation R²
Model

Figure 4.2.2


Actual vs. Predicted Vaccine Uptake using Random Forest
Predicted Vaccine Uptake (%)
Actual Vaccine Uptake (%)