95-828: Machine Learning for Problem Solving

Machine Learning for Public Health: COVID-19 Vaccination and Positivity in the U.S.

**Team #36**

Fanxing Bu

Xiao Yan

Diego Bermudez

Karla Mercado

April 27th, 2025

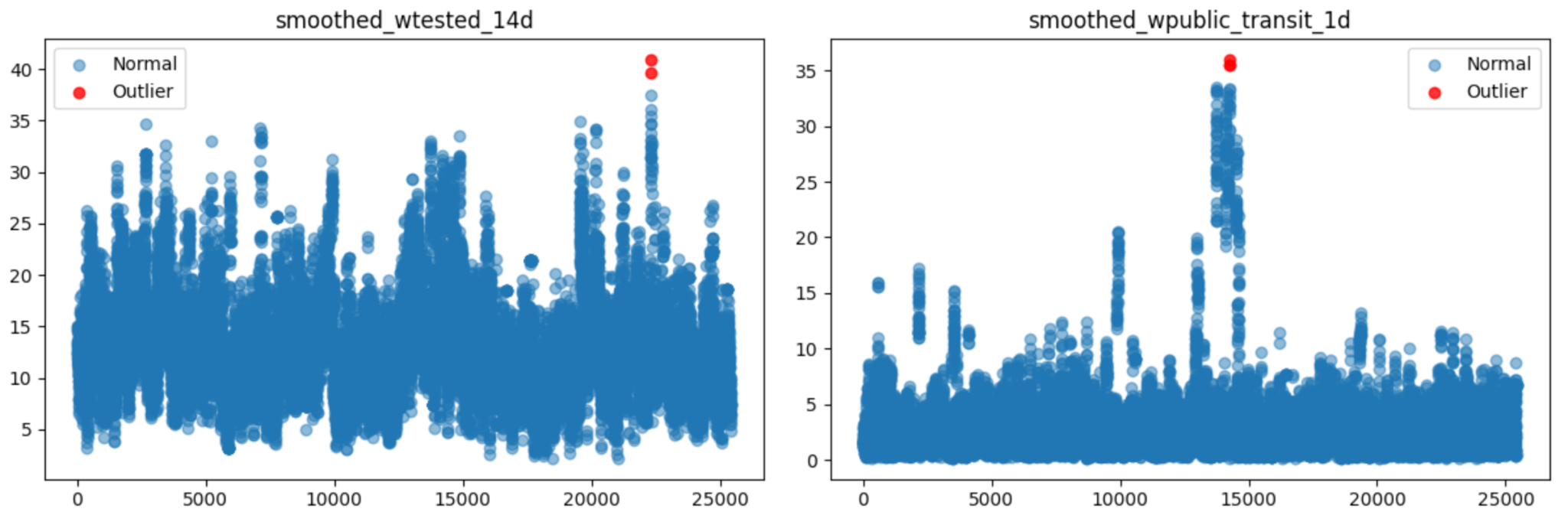
**Introduction**

During the COVID-19 pandemic, widespread fear arose regarding infection risks, driven by the virus’s high mortality rate. According to the Centers for Disease Control and Prevention (CDC), the COVID-19 death rate in 2020 was 93.2 deaths per 100,000 population [(Ahmad, 2023)](https://www.zotero.org/google-docs/?M0iyV9). The virus’s rapid spread and difficult-to-control nature prompted the urgent development of vaccines to mitigate its harmful effects. A critical challenge during this period was identifying the counties and zones most affected by the virus to prioritize preventive treatments and minimize its impact. To address this issue, this project aims to predict vaccination uptake and COVID-19 test positivity rates at the county level, leveraging self-reported behavioral and belief indicators.

The relevance of this project extends beyond the technical task of prediction. It contributes to the broader development of public health strategies in partnership with the CDC. By accurately forecasting vaccination behavior and test positivity trends, public health officials can design more targeted vaccination campaigns, allocate testing resources more efficiently, and enhance preparedness for future public health crises. Given the significant variation in vaccination acceptance and infection rates across communities during the pandemic, building robust predictive models is crucial for improving intervention efforts and reducing disease transmission.

**Data Analysis**

This analysis utilizes data from the COVID-19 Trends and Impact Survey (CTIS), conducted by the Delphi Group at Carnegie Mellon University. The dataset includes information for 687 counties over 37 days between January 7, 2021, and February 12, 2021. It captures a wide range of behavioral indicators, belief indicators, and COVID-19 activity indicators. The target variables for prediction are: (1) the percentage of individuals vaccinated, and (2) the COVID-19 test positivity rate. With its wide geographic and temporal coverage, the CTIS dataset provides a strong foundation for building predictive models and generating actionable insights to support public health decision-making.



*Figure 1. Outlier Detection Using DBSCAN on Selected COVID-19 Features.*

As with many large-scale survey efforts, CTIS data is inherently noisy and incomplete. In particular, the reliance on self-reported responses collected through Facebook led to missing information for certain counties on specific dates. To address this, we first assessed the extent of missingness by calculating the total frequency of missing values (MV) across variables. The test positivity rate exhibited the highest number of missing entries, with 25,236 instances, while explanatory variables had an average of approximately 4,000 missing values each. To impute missing data in the feature variables, we applied a forward fill followed by a backward fill, ensuring that information gaps at both ends of the time series were addressed. This method is well-suited for time-series structures, where neighboring values are highly correlated, allowing reasonable assumptions about the continuity of covariate behavior over time. Handling of missing values for the target variable, the test positivity rate, is discussed separately in the methodology section.

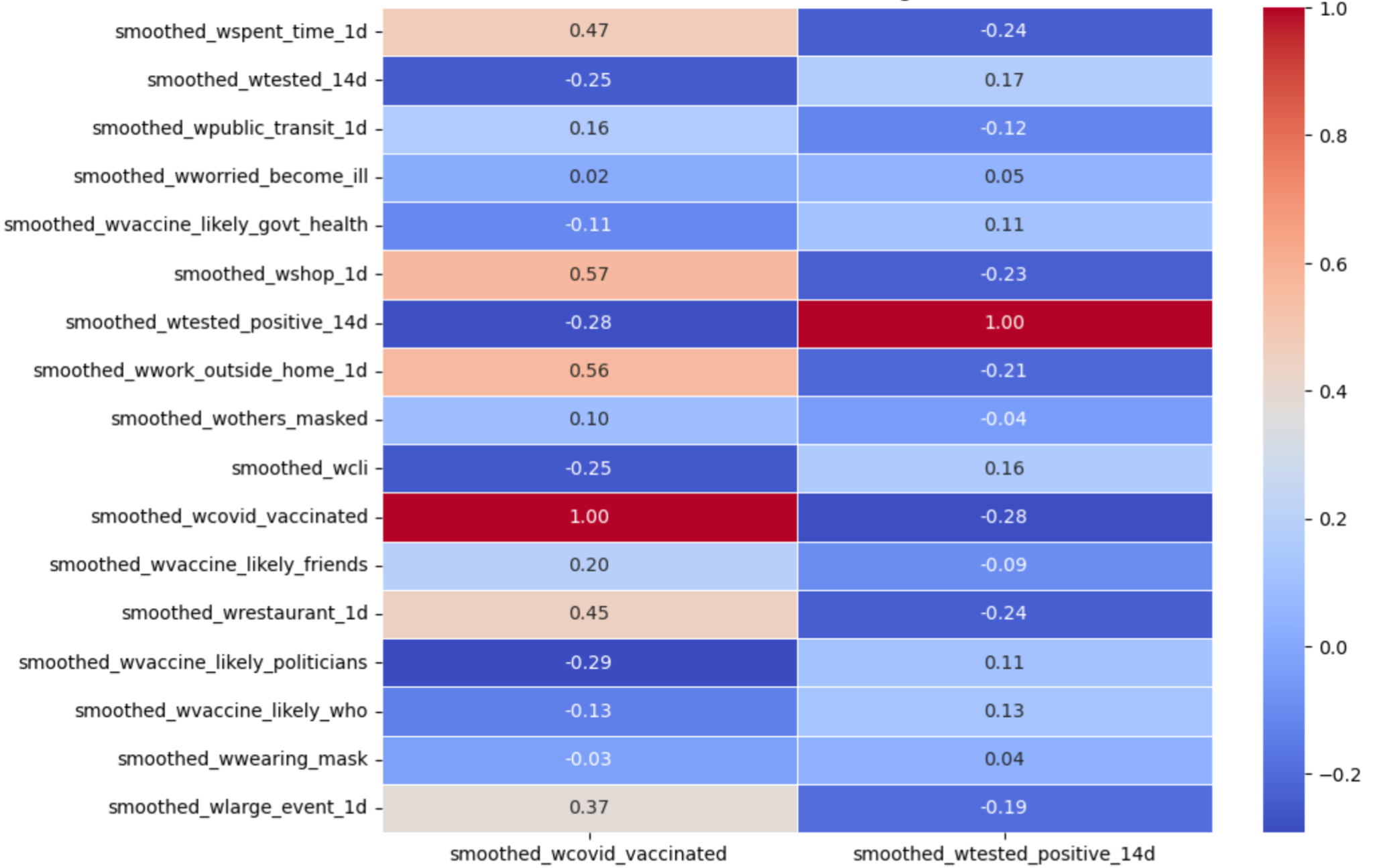
In addition to addressing missingness, we implemented outlier detection to further improve data quality and ensure stability in model performance. Key features—such as COVID-19 test positivity rates and behavioral indicators—were examined using DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Outlier detection was conducted after missing value imputation to enhance accuracy and ensure that DBSCAN evaluated the complete dataset. Although alternative approaches, such as Winsorization or IQR-based filtering, were considered, these methods are less effective at capturing the nonlinear patterns commonly observed in health data [(Gostic et al., 2020)](https://www.zotero.org/google-docs/?xyXNEL). As illustrated in Figure 1, outliers detected by DBSCAN are highlighted in red; these observations were subsequently removed to reduce noise and enhance model performance and stability.

**Methodology**

Although a time-series model could, in principle, be trained using prior values of the target variable, the limited number of data points available for each county presents a significant constraint. To address this, we rely on the available feature set to ensure that the model captures information from both previous and concurrent covariates. Our objective is to learn a function 𝑓 such that:

where is the value of the target variable (either test positivity rate or vaccine uptake) at the next time step, and is a matrix of feature variables at time . To approximate this function, we first identified covariates that exhibit stronger relationships with the target variables. This selection was informed through exploratory analyses, including scatter plots and histograms, as summarized in Figure 2. These visualizations reveal how each feature relates to the target and its significance in explaining outcome variation. We further supported this process with an extended heatmap analysis and feature perturbation methods to examine overall correlations. Notably, features such as the level of worry about becoming ill and the perceived prevalence of mask-wearing among others were found to have low predictive power for both vaccine uptake and test positivity rates. In contrast, outdoor activity levels appeared positively associated with vaccination rates, possibly reflecting a greater perceived need for protection among those engaging in higher-risk activities. For the test positivity rate, feature relevance remained stable across variables, with activities related to spending time outside showing moderate, albeit counterintuitive, associations with increased test positivity.

Despite the presence of strong pairwise correlations among some features, we chose to retain the complete set of covariates for both prediction tasks to maximize information usage. Some of the models employed incorporate regularization techniques that mitigate multicollinearity, while others, such as tree-based methods and neural networks, are naturally robust to correlated inputs. As a result, we opted not to perform dimensionality reduction at this stage, preserving the richness of the original feature space for model development.

****

*Figure 2. Correlation Heatmap Between Selected Features and Target Variables.*

To preserve the shape of the time series for each county, we applied a MinMaxScaler to the features and vaccine uptake. As mentioned before, health data and survey information are noisy, and to reduce the chances of getting unstable predictions, we scale the features so they lie within the range from 0 to 1, where 0 is the lowest original point and 1 is the highest. In summary, this compresses the time series, reducing noise while maintaining variability in the process.

We adopted different data splitting strategies based on the model type: random split and K-fold cross-validation were used for traditional models. At the same time, TimeSeriesSplit was applied for temporal models. The first step to applying this approach was to treat the dataset as panel data, where each observation represents a county on a specific date. Instead of only considering the time for all the counties, we separated them for training and testing and made predictions based on the 37 days of data. By doing this, we ended up with 80% of our counties in the training set and 20% in the test set. The primary purpose of taking this approach is to prevent data leakage from occurring. This may happen due to partial information appearing for a single county. Suppose we use a traditional splitting method without this consideration. In that case, we may end up with data from two random days for a given county in the test set, while the training set may appear as if the county does not have information for those days, leading to unstable predictions.

1. **Baseline Models**

Lasso Regression was used as the baseline model to predict vaccine uptake.. By applying L1 regularization, Lasso selected the key features by shrinking less informative coefficients to zero, improving model interpretability. The regularization parameter () was optimized using GridSearchCV over a logarithmic range, with the best value identified as 0.1. The Lasso model achieved an average R² of approximately 0.323 and an average MSE of 29.32, indicating moderate predictive ability in this context. In terms of feature selection,eight predictors were consistently selected across all folds: smoothed\_wcovid\_vaccinated\_or\_accept, smoothed\_wothers\_masked, smoothed\_wshop\_1d, smoothed\_wspent\_time\_1d, smoothed\_wtested\_14d, smoothed\_wvaccine\_likely\_govt\_health, smoothed\_wvaccine\_likely\_who, and smoothed\_wwork\_outside\_home\_1d. These features represent critical behavioral and belief indicators associated with vaccine uptake. Overall, Lasso provided a strong, interpretable, and efficient baseline against which the performance of more complex models could be compared.

Linear Regression and Ridge Regression were used as baseline models to predict the COVID-19 test positivity rate (smoothed\_wtested\_positive\_14d). Linear Regression, using fit\_intercept=True, served as a simple benchmark and achieved a cross-validation R² score of 0.6296. To improve model generalization and reduce potential overfitting, Ridge Regression was applied with L2 regularization. Hyperparameters for Ridge were optimized using GridSearchCV with 5-fold cross-validation, identifying the best parameters as alpha=10.0 and solver='lsqr'. The Ridge model achieved a cross-validation R² of 0.6298 ± 0.0180, a cross-validation MSE of 20.4333 ± 0.2069, and a cross-validation RMSE of 4.5203 ± 0.3216. When evaluated on the testing set, Ridge Regression achieved an R² of 0.6646 and an RMSE of 4.2517, indicating slightly improved robustness compared to the unregularized Linear Regression model. These baseline results established a reliable reference point for assessing the performance gains offered by more advanced ensemble methods and neural network models explored in subsequent stages of the analysis.

## **Methods**

## **Vaccine Uptake**

To complement the baseline Lasso Regression model, a variety of additional methods were chosen to model vaccine uptake, each addressing different potential patterns in the data. Gradient Boosting Regressor, XGBoost, Random Forest, and LightGBM were chosen because they are tree-based ensemble methods capable of modeling complex, non-linear relationships between behavioral indicators and vaccination rates, offering greater flexibility than linear models. These models are also robust to multicollinearity among predictors and can automatically capture important feature interactions. Additionally, a Recurrent Neural Network (RNN) was included to explore deep learning techniques specifically designed for sequential data, allowing the model to learn dependencies across time in a more dynamic way. Collectively, these methods were selected to provide a comprehensive comparison across linear, tree-based, statistical time series, and deep learning approaches to best understand the factors influencing vaccine uptake.

**Test Positivity Rates**

|  |  |
| --- | --- |

*Figure 3. Model comparison Figure 4. Model comparison*

A linear regression model was initially selected as the baseline, followed by the evaluation of more advanced models including Random Forest, XGBoost, and Neural Networks. GridSearch was used for hyperparameter tuning and K-fold cross-validation ensured robust performance assessment. These models were selected for their complementary strengths: Random Forest offers robustness to multicollinearity and nonlinear relationships; XGBoost achieves high predictive accuracy through effective regularization; and Neural Networks are capable of capturing complex, high-dimensional feature interactions. Together, these approaches are well-suited to address the heterogeneous nature of COVID-19 test positivity and vaccine uptake prediction.

Following model training, feature importance was analyzed using model coefficients (for linear regression), built-in importance scores (for Random Forest and XGBoost), and SHAP values (for Neural Networks). Both individual and grouped feature importances—categorized into behavioral, belief, and COVID activity indicators—were visualized to enhance interpretability and enable cross-model comparison of key factors. To address missingness in the target variable, LSTM and Transformer models were additionally implemented to capture sequential patterns and generate pseudo-labels. These pseudo-labeled samples were combined with the original labeled data to retrain Random Forest, XGBoost, and Neural Network models, facilitating a more comprehensive and resilient predictive analysis.

**Results**

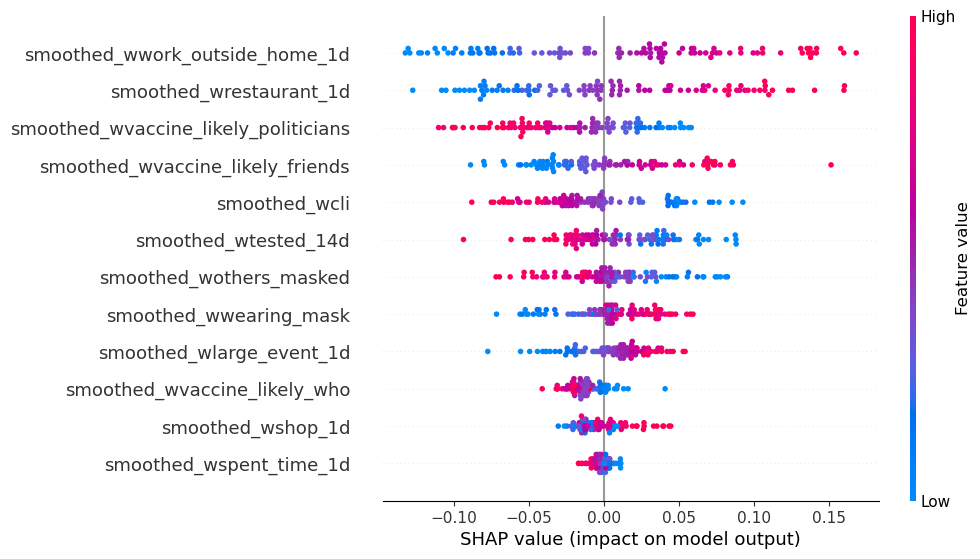
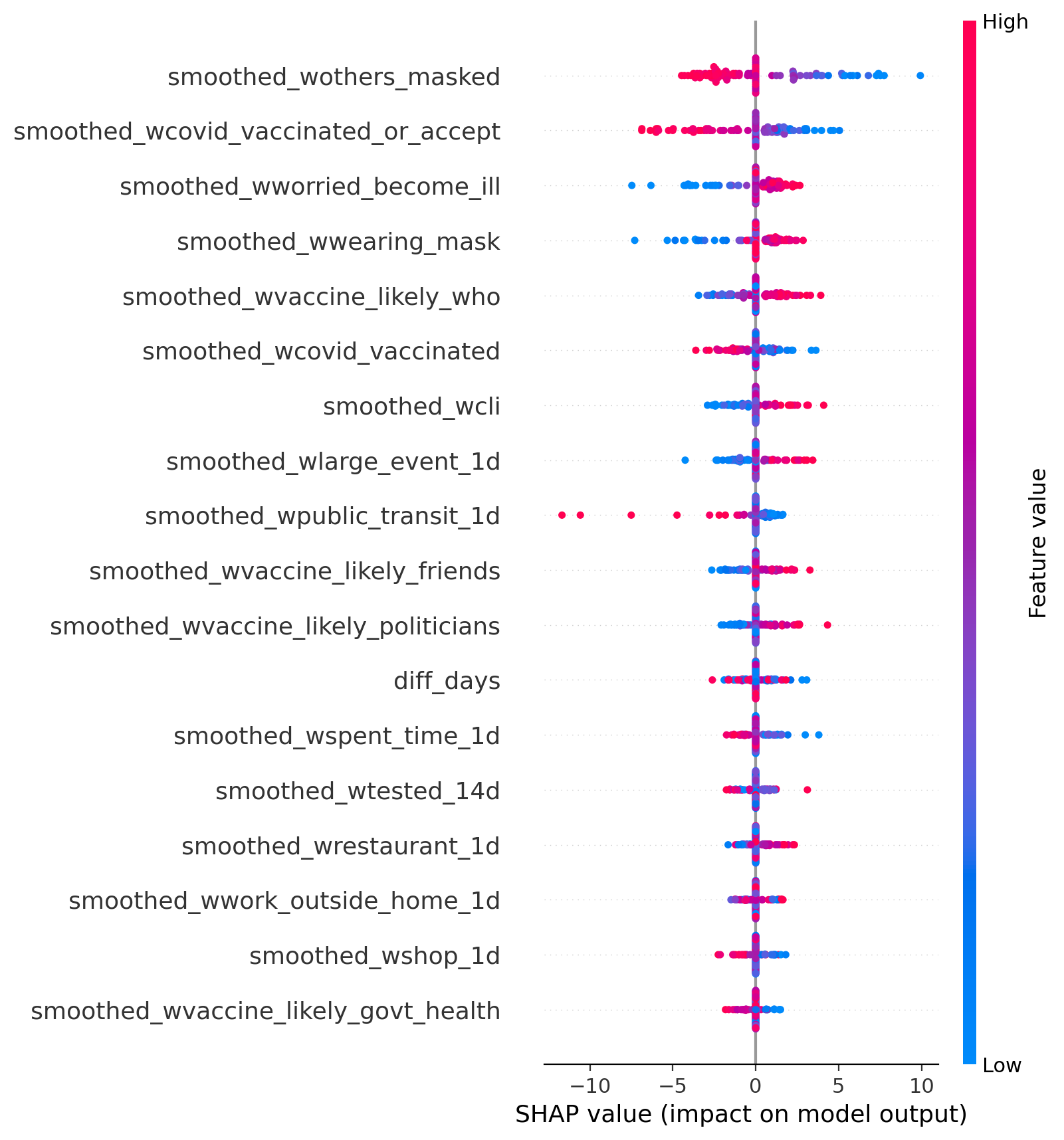
R², MSE, RMSE, and MAE were selected to comprehensively evaluate the diverse models used.

R² captures the overall predictive strength, suitable for comparing linear and nonlinear models.

MSE and RMSE emphasize large errors, aligning with the high-accuracy goals of Random Forest and XGBoost.MAE provides a robust measure less sensitive to outliers, complementing the error profiles of Neural Networks.Together, these metrics ensure consistent, interpretable assessment across models with different architectures and error behaviors.

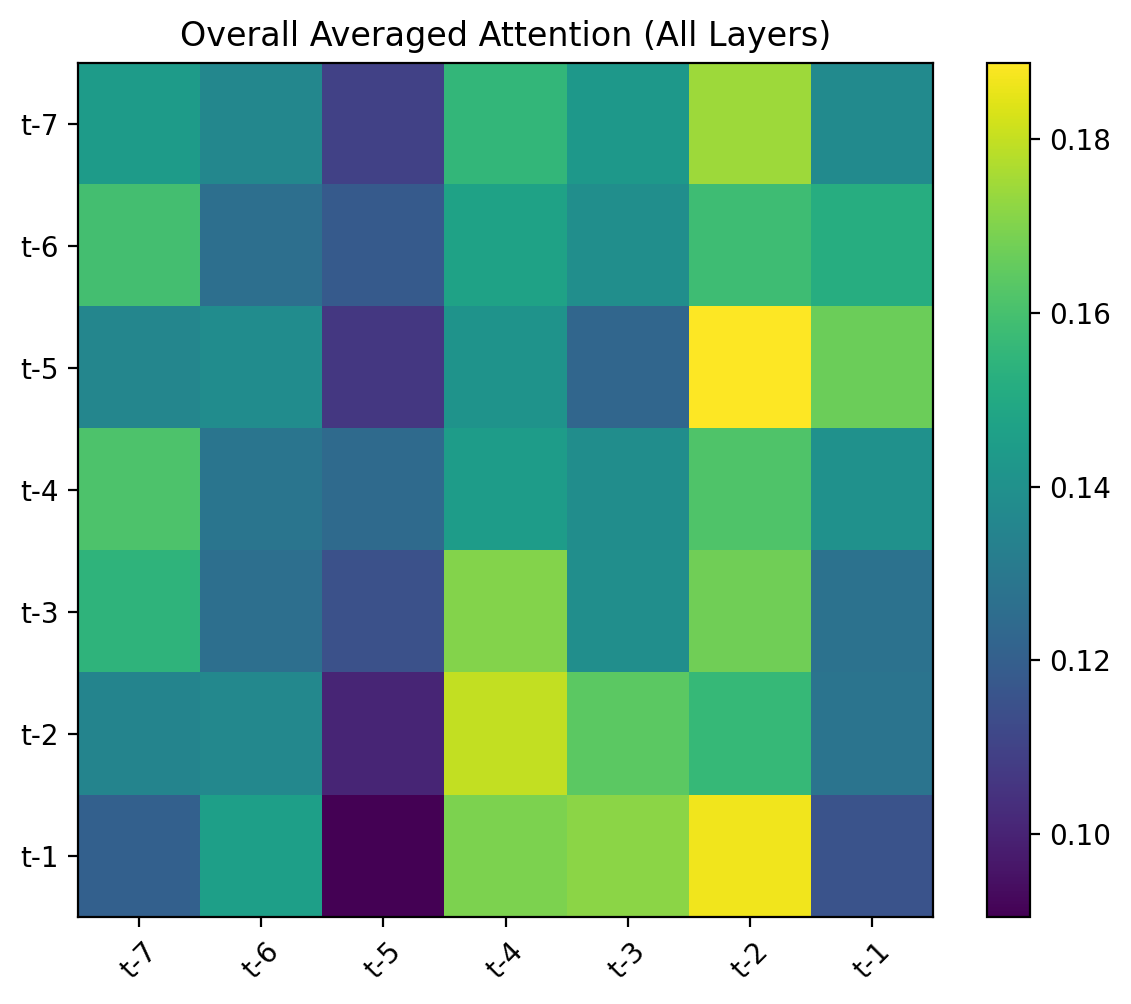
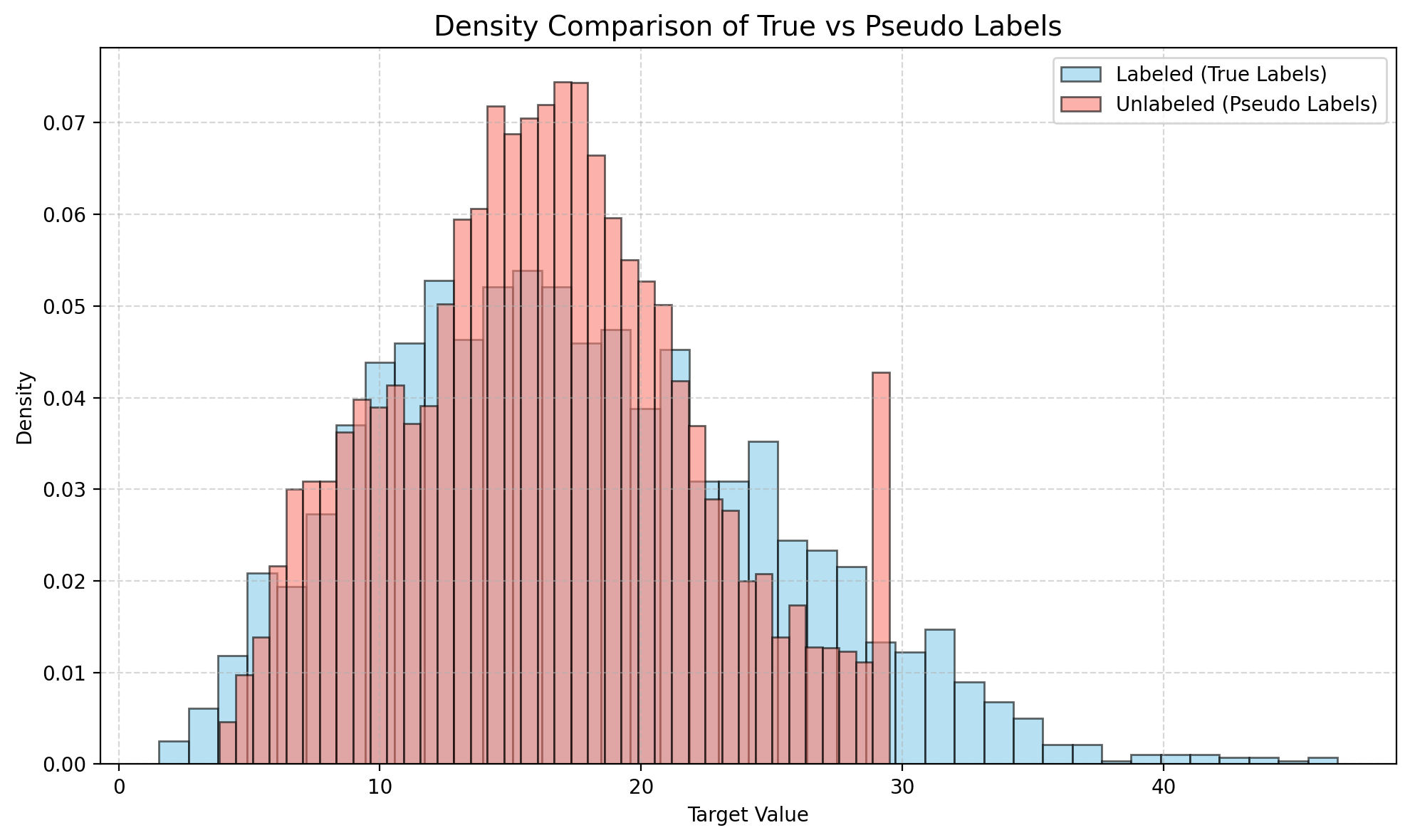
| **Vaccine Uptake** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Train RMSE** | **Test RMSE** | **CV RMSE  (mean ± std)** | **Train R²** | **Test R²** | **CV R²  (mean ± std)** |
| **Lasso Regression** | 0.2273 | 0.2223 | 0.2309 ± 0.0059 | 0.4732 | 0.2227 | 0.4545 ± 0.0187 |
| **XGBoost** | 2.243 | 6.059 | 6.0593 ± 1.4299 | 0.7104 | -0.7003 | -0.7004 ± 0.3571 |
| **LightGBM** | 2.251 | 6.000 | 6.0009 ± 1.4378 | 0.7321 | -0.6672 | -0.6673 ± 0.3611 |
| **GBR** | 0.2133 | 0.2203 | 0.2290 ± 0.0061 | 0.5360 | 0.4979 | 0.4631 ± 0.0214 |
| **RNN** | 0.1084 | 2.6651 | 0.1659 ± 0.0229 | 0.8364 | 0.8328 | 0.6112 ± 0.1019 |
| **Test Positivity Rates** | | | | | | |
| **Linear Regression** | 4.488 | 4.249 | 4.5212 ± 0.3152 | 0.636 | 0.665 | 0.6296 ± 0.0183 |
| **Ridge Regression** | 4.488 | 4.252 | 4.5203 ± 0.3216 | 0.636 | 0.665 | 0.6298 ± 0.0180 |
| **Random Forest** | 1.121 | 2.700 | 3.2045 ± 0.4843 | 0.977 | 0.865 | 0.8139 ± 0.0134 |
| **XGBoost** | 0.405 | 2.530 | 2.9926 ± 0.5780 | 0.997 | 0.881 | 0.8381 ± 0.0088 |
| **Neural Network** | 1.192 | 2.292 | 2.5876 ± 0.5502 | 0.974 | 0.903 | 0.8785 ± 0.0147 |

Considering overall stability and generalization after training, we selected the **Neural Network as the final model**  for subsequent analysis and we used SHAP plot to visualize key predictive relationships between predictors and your outcomes.

**

*Figure 5. SHAP plot-Test Positivity Rates(NN) Figure 6. SHAP plot-Vaccine Uptake(RNN)*

Due to the high missing value presence in the test positive rate target, we implemented LSTM and Transformer models in addition to basic methods. Both showed clear improvements over traditional machine learning models. The LSTM achieved a Test MSE of 4.1410 and R² of 0.9185, while the Transformer achieved a Test MSE of 6.0310 and R² of 0.8896, demonstrating strong temporal modeling capabilities. Although data were randomly shuffled during training, which exposed models to both past and future information, this is acceptable for pseudo-label generation. Attention visualization in the Transformer revealed that **t-5** and **t-6** time steps received the highest attention weights, indicating that past behaviors approximately five days prior had the greatest predictive influence. This aligns with the known COVID-19 incubation period, 5 to 6 days [(World Health Organization, 2020)](https://www.zotero.org/google-docs/?P3y0pD).



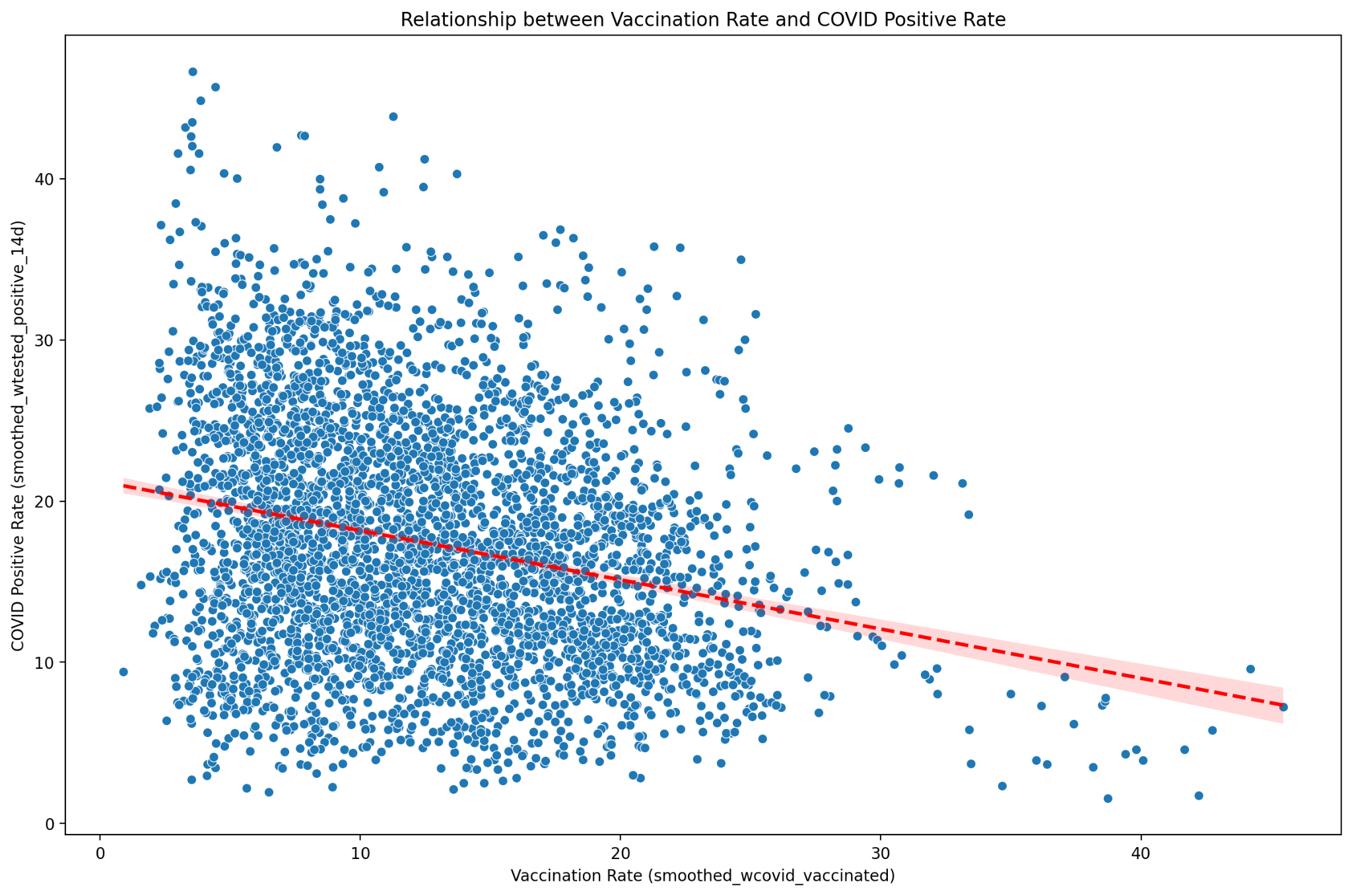
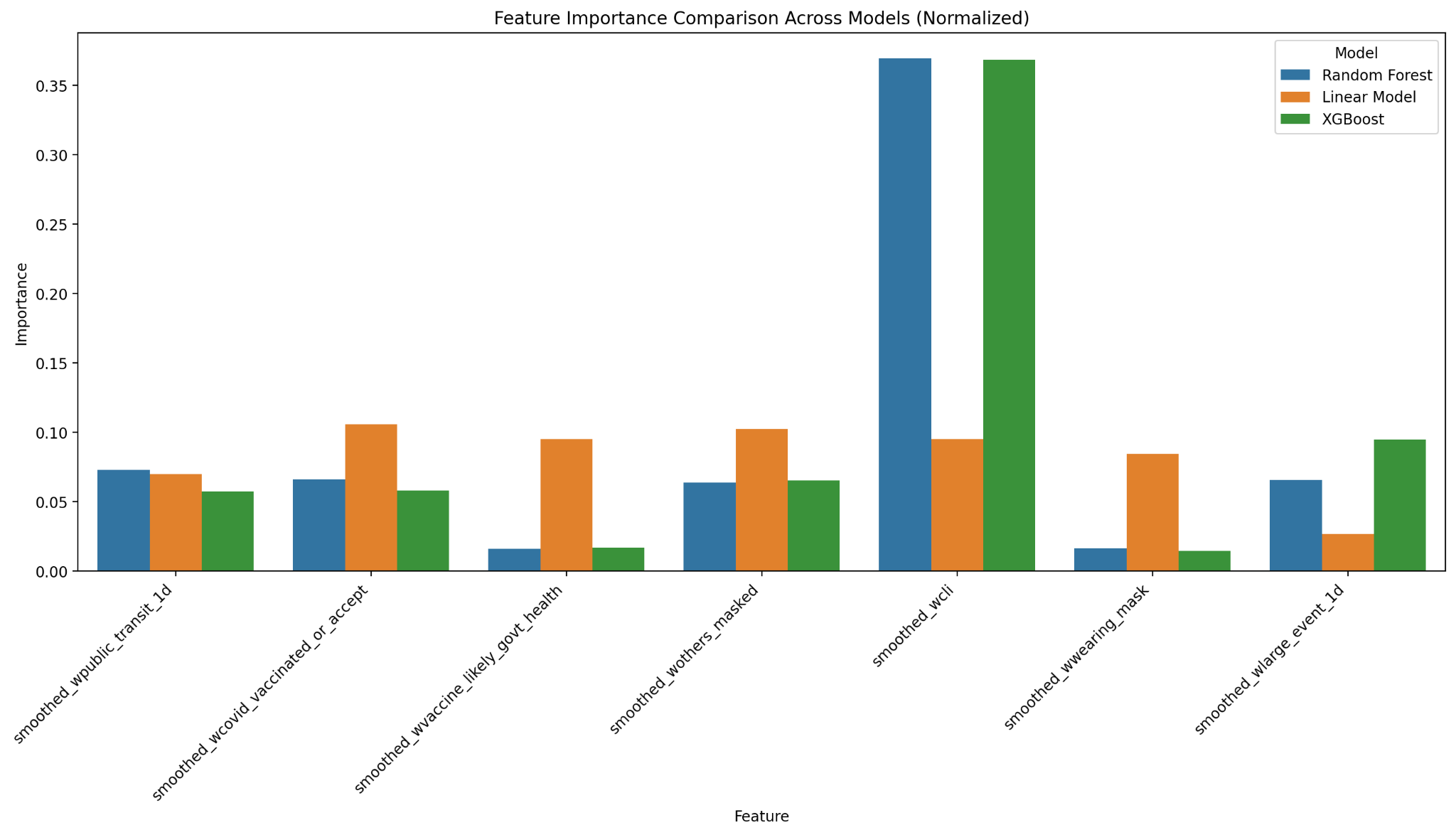
*Figure 4. Attention weights across time steps (left) and density comparison of true and pseudo-labeled targets (right).*

Due to its greater stability and robustness across diverse geo\_value inputs, along with the favorable results observed in the density comparison between true and pseudo labels, we selected the Transformer as the optimal model for pseudo-label generation. After generating 14,404 pseudo-labels using the LSTM model and augmenting the dataset, we retrained Random Forest, XGBoost, and MLP Regressor with the previously selected optimal hyperparameters. As shown in Table, model performance declined due to the inherent noise introduced by pseudo-labels, with R² scores decreasing by approximately 12% to 17%.Despite this drop, models retained moderate predictive capabilities, and the expanded dataset (~18,000 samples) helped mitigate overfitting risks.

| **Model** | **Original** | **After Pseudo-Label Retraining** | **Change** |
| --- | --- | --- | --- |
| **Random Forest** | R² ≈ 0.86 | 0.7465 | ↓ 0.12 |
| **XGBoost** | R² ≈ 0.881 | 0.7283 | ↓ 0.15 |
| **Neural Network** | R² ≈ 0.903 | 0.7331 | ↓ 0.17 |

**Conclusions and Policy Recommendations**

Using our model predictions, the relationship between vaccine uptake and COVID cases is consistently significant across multiple modeling approaches. Feature importance analysis from both traditional and advanced models identified vaccination acceptance (smoothed\_wcovid\_vaccinated\_or\_accept) as one of the most influential predictors. In the Transformer model, this feature ranked second in importance with a ΔMSE of 5.378, indicating its substantial impact on predictions. The Neural Network model captured additional nuanced effects, showing that vaccination had a strong positive association with improved outcomes across different contexts. This consistent finding across linear models, tree-based approaches, and neural networks suggests a robust inverse relationship between vaccination rates and COVID case numbers, with higher acceptance of vaccination associated with decreased COVID incidence.

****

The scatter plot and statistical analysis confirm a significant negative correlation between vaccination rates and COVID-19 positivity.While the Pearson correlation coefficient (-0.2525) suggests a weak to moderate relationship, the extremely small p-value (1.6835e-57) indicates strong statistical significance. Areas with the highest vaccination rates had 60–70% lower positivity rates, consistent with model findings that ranked vaccination acceptance as a key predictor.Feature importance analysis across Ridge Regression, Random Forest, XGBoost, and Neural Networks showed consistent top factors: COVID-like symptoms (smoothed\_wcli), vaccination acceptance (smoothed\_wcovid\_vaccinated\_or\_accept), and community masking (smoothed\_wothers\_masked).Tree-based models further highlighted public exposure activities (e.g., transit use, large events), while Neural Networks captured more subtle non-linear effects, such as worry about illness.

Overall, health risk perception, protective behaviors, and public exposure emerged as the main drivers of COVID-19 positivity across models.

Based on these findings, I recommend the following vaccine policy strategies for the CDC to improve uptake and reduce new cases:

1. **Integrate Vaccination with Protective Behaviors:** Vaccination acceptance and community masking jointly influenced outcomes, as captured by the Neural Network with the highest R² (0.903). Linking vaccination to other protective behaviors would maximize impact.
2. **Focus on High-Exposure Settings:** Public transit usage and large event attendance ranked among top features in tree-based and Transformer models (ΔMSE = 6.578). Targeting these environments can reduce transmission risks.
3. **Enhance Health Risk Messaging:** Worry about illness showed a nuanced effect in the Neural Network and Transformer models (ΔMSE = 4.823), suggesting messaging should emphasize concrete health risk reduction.
4. **Address Political Trust:** Trust in politicians regarding vaccination was the fourth most important feature in the Transformer model (ΔMSE = 3.141). Campaigns should leverage trusted figures across the political spectrum.
5. **Maintain Consistent, Time-Sensitive Efforts:** Attention analysis in the Transformer model indicated time steps t-5 and t-6 were critical, highlighting the need for sustained and consistent campaign messaging.

**References**

[Ahmad, F. B. (2023). COVID-19 Mortality Update—United States, 2022. *MMWR. Morbidity and Mortality Weekly Report*, *72*. https://doi.org/10.15585/mmwr.mm7218a4](https://www.zotero.org/google-docs/?dHVmHF)

[Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., Salazar, P. M. D., Hellewell, J., Meakin, S., Munday, J. D., Bosse, N. I., Sherrat, K., Thompson, R. N., White, L. F., Huisman, J. S., Scire, J., … Cobey, S. (2020). Practical considerations for measuring the effective reproductive number, Rt. *PLOS Computational Biology*, *16*(12), e1008409. https://doi.org/10.1371/journal.pcbi.1008409](https://www.zotero.org/google-docs/?dHVmHF)

[World Health Organization. (2020). *Coronavirus disease 2019 (COVID-19): Situation report – 73*. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200402-sitrep-73-covid-19.pdf](https://www.zotero.org/google-docs/?dHVmHF)