INTERNATIONAL
BUSINESS SCHOOL

2025

Advanced Forecasting

HOME ASSIGNMENT

BAYOUD EL MEHDI, POLIVKA DAVID, TAJIMOZAFARI RAZIEH

Table of Contents

List of figures

1. Introduction & Problem Context

The purpose of this report is to demonstrate, that in the field of recurring health problems the data-driven forecasting can be a game changer. By systematic measuring and the analysis of historical data, we enable ourselves to leave the reactive approach behind us, and instead of running after problems, switch to the new method, the predictive approach.

Therefore, we chose a fitting health issue, the seasonal influenza. Influenza remains still one of the most significant public health issues, which burdens the healthcare system in countless ways, in the form of: - hospital admissions, bed capacities, sick workers, number of vaccines shots, campaigns or just preventive measures. The above-mentioned burdens could be reduced only be knowing the characteristics of the given year's flue. Although the dynamics of the influenza is strongly seasonal, there are some year-to-year variabilities and changes, which influence the given year's outcome, which in our opinion could be forecasted.

Therefore, we devoted ourselves to answer a rather relevant question: Are we able to forecast the short-term the influenza activity by region, using historical influenza time series? So, the target variable became the ILI%, which stands for influenza-like illness. To address the target, we worked with a panel of regional time series, to design a forecasting pipeline, which enabled us to do systematic comparison of modelling approaches. Therefore, for benchmarking a baseline model was prepared, while later on some classical and advanced machine learning models were implemented for the given regions.

2. Description of Data and Exploratory Analysis.

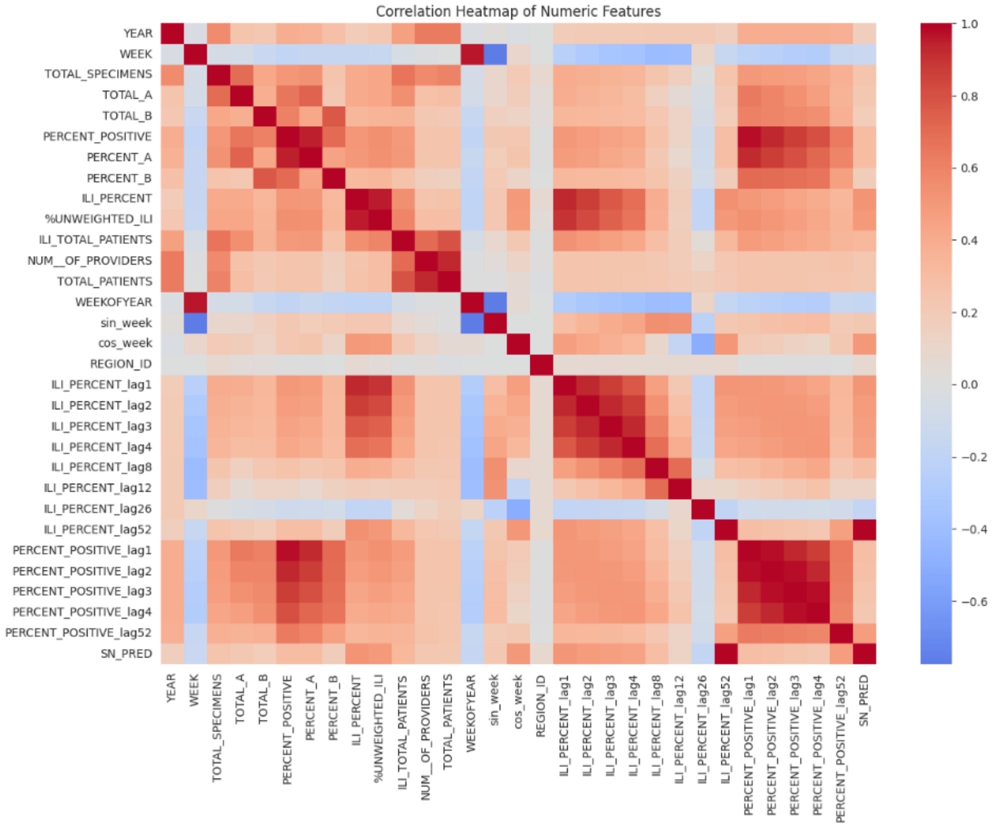


Figure 1 Correlation Heatmap

2.1 Data Sources and Variables

We were required to have a dataset that is a time series in which we could capture the spread of flu in various locations. It came to be an influenza surveillance data on a weekly basis at a regional scale. There is one region per row and the system provides both lab results and syndromic info.

The target variable in our models was ILI% the influenza-like illness percentage. It informs us about what the percentage of outpatient visits for a given region during a particular week can be attributed to flu-like symptoms, which is intended to predict one week. As one of the exogenous factors, we used PERCENTPOSITIVE, which is the percentage of the tested samples that returned positive as a result of a test of the influenza virus within the same system. This correspondence between the clinical observations and the real virus spread is quite relevant. We also stored region identifiers and extracted different calendar properties (week of the year, year, sinusoidal encodings) to reflect the high seasonality that we predict.

2.2 Panel Structure, Time Span and Coverage.

Instead of a single time series, the dataset is rather a panel, with different regions reoccurring with time. Therefore, areas appear in a number of years week after week. By grouping on region, we saw that each area has a long history of weekly data, covering multiple flu seasons in succession.

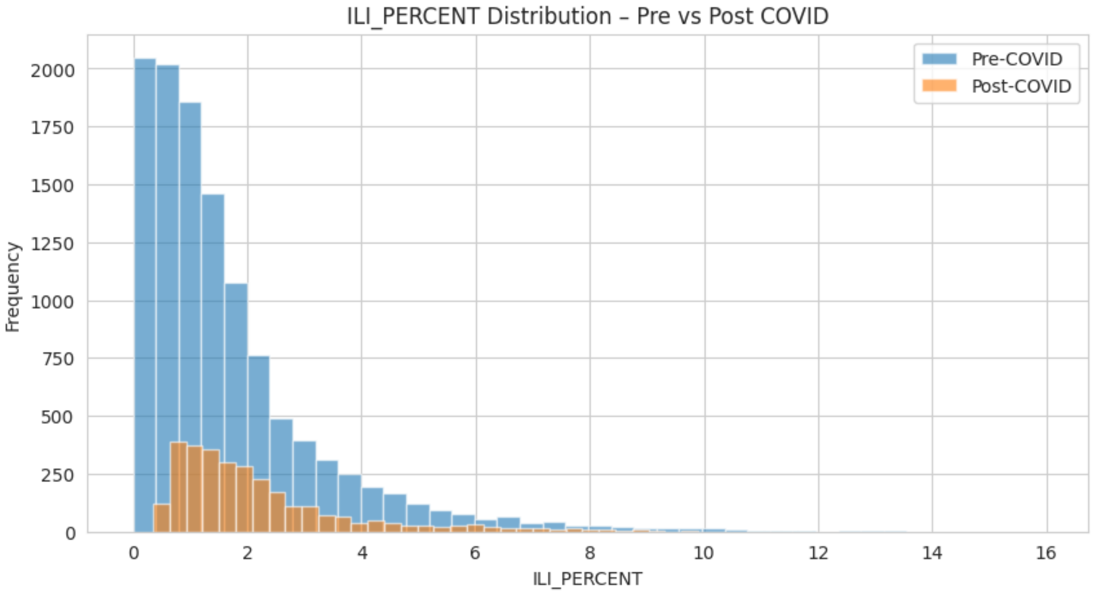
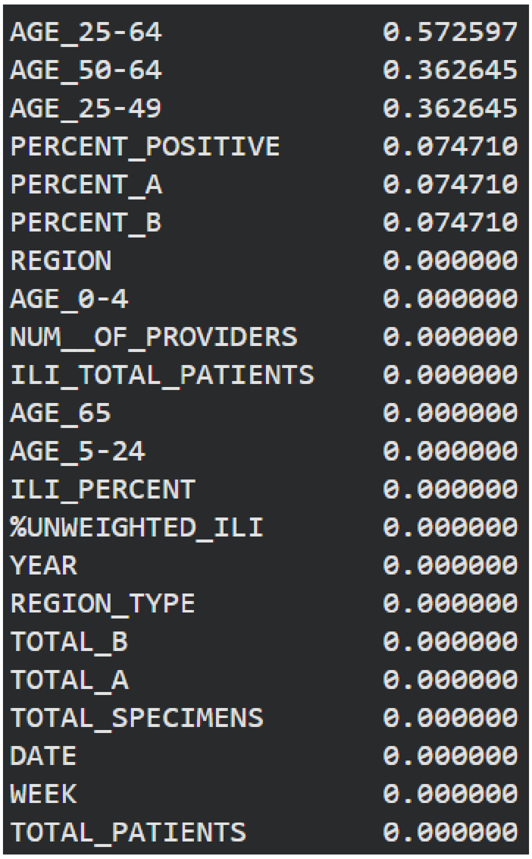


Figure 2 ILI_PERCENT Distribution – Pre vs Post COVID

2.3 Values and Preprocessing missing.



```
AGE_25-64              0.572597
AGE_50-64              0.362645
AGE_25-49              0.362645
PERCENT_POSITIVE       0.074710
PERCENT_A              0.074710
PERCENT_B              0.074710
REGION                 0.000000
AGE_0-4                0.000000
NUM__OF_PROVIDERS      0.000000
ILI_TOTAL_PATIENTS     0.000000
AGE_65                 0.000000
AGE_5-24               0.000000
ILI_PERCENT            0.000000
%UNWEIGHTED_ILI        0.000000
YEAR                   0.000000
REGION_TYPE            0.000000
TOTAL_B                0.000000
TOTAL_A                0.000000
TOTAL_SPECIMENS        0.000000
DATE                   0.000000
WEEK                   0.000000
TOTAL_PATIENTS         0.000000
```

Figure 3 Missing Values

We needed to verify the quality of our data before modelling. We obtained a quick view of missing values and noticed that there are several variables that were suffering from the lack of consistency. The ILI indicators based on age were highly missing and biased in terms of geographic coverage as well as years. Their inclusion would have entailed complex imputation and would have distracted our purpose of short-term forecasting. We therefore discarded the age-stratified variable, dropped the age column and instead considered aggregates.

For the key variables, ILI% and PERCENT_POSITIVE, the missingness was limited. The lab positivity fields were interpolated within each region to smooth the early sparse years. After that, we created the calendar features (YEAR, WEEKOFYEAR, and their sine/cosine encodings) and assigned numeric IDs to the regions. Finally, we generated the lag features for ILI% and PERCENT_POSITIVE, and at this stage dropped rows where lag values were missing. This mostly removed the first 52 weeks of each region, leaving a clean and complete modelling dataset.

2.4 Seasonality, Trends and Structural Breaks.

The second step was to observe the behavior of ILI% as time progressed. When we plotted a series for each region, clearly, a seasonal effect appeared: namely, sharp peaks in flu season and very low values the rest of the year.

Nevertheless, a structural break, which is linked to the COVID-19 pandemic, can be observed. At the beginning of COVID-19 when the typical flu peaks should have occurred, it

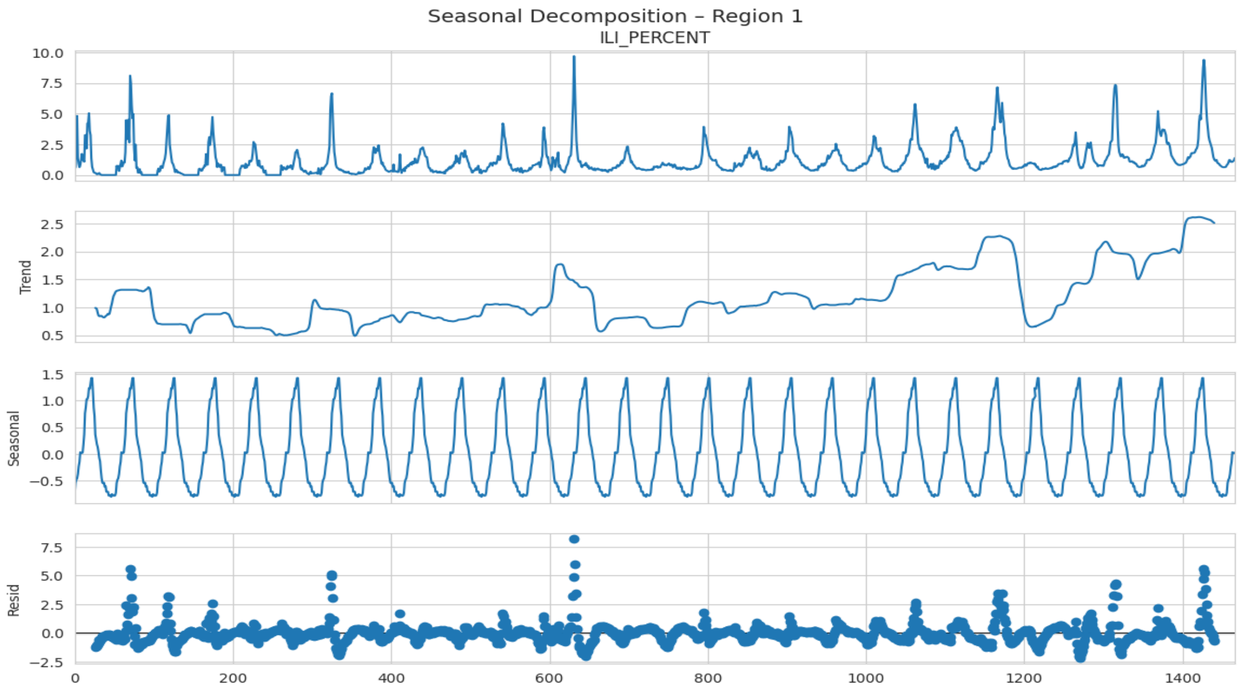


Figure 4 Seasonal Decomposition – Region 1

3. Forecasting Problem Formulation and Evaluation Set-Up

Having presented the overall concept, we had to decide on the matter of what we want to forecast and on what terms. The variable in which we were interested in, was the weekly, flu by region, in terms of ILI% (ILIPERCENT). This value informs us about the proportion of outpatient visits for a given region and week, which can be explained by influenza-like symptoms. Our objective is to predict this ILI percentage one week in advance hence we have a short-term forecasting problem. As we are not only considering the past values of ILI, but also other data, such as lab positivity rates and calendar effects, this issue can be called a multivariate panel forecasting problem with exogenous variables as a number of regions are predicted simultaneously with advanced models being learnt on all data. To make our assessment sound and reasonable, we have chosen to divide the data in the plain chronological manner. The period of the training incorporates all the observations as of 2016-06-30. The validation period is 2016 07 01 2020 06 30, whereas the test period includes weeks later than 2020 06 30. This system ensures that the models are only trained and tuned to view what has happened in the past and there is no information flowing backward to the future. In comparing the various methods, we have applied three standard measures of error: Mean Absolute Error (MAE), Root Mean Squared error (RMSE) and Mean Absolute Percentage error (MAPE) which we did in our compute measurements method. Test window is of primary interest to us, as it demonstrates the performance of the models out-of-sample, and we also examine by-region results to observe how models vary in the performance of various geographic regions.

4. Methodology: Feature Engineering & Models

Once we had nailed down the problem and constructed the evaluation, we needed to construct a learning pipeline that was capable of learning the dynamics of the flu plots. Let us provide a brief explanation in this section on how we turned the features of the surveillance information and how we constructed the classical and the advanced models on top of the base.

4.1 Feature Engineering

Since flu is time-centered and possesses high seasons, my initial features were time-centered. We determined the observational date of the weekly observation through the weekly dates we constructed WEEKOFYEAR, which tells us within which epidemiological week we have made the observation. We further encoded the week number with the sine and cosine of the calendar so that week 1 and week 52 are brought near one another in the feature space. We also computed the calendar year for exploration, but in the final models we only used WEEKOFYEAR and its sine/cosine encodings as time features, not the raw year.

Then we paid attention to lag features, as the activity of the flu is very much dependent on its history. I applied my add group lags function that created lag individually per region and therefore the history of one region will not leak into another. The lags of ILI percent I used were 1, 2, 3, 4, 8, 12, 26 and 52 weekly. Short lags are immediate dynamics, medium lags are memory in the same season and long Lag is the pattern in a year. In the case of PERCENTPOSITIVE, I included fewer lags (e.g 1, 2, 3, 4, and 52 weeks) since I only expect that its recent trends and seasonal reoccurrence to be relevant.

4.2 Baseline Models

Two baseline models were constructed before the deep dive of advanced means. The former is an unsophisticated panel forecast: in each region, the forecast of next week is the final observed ILI% in that region. The second one is a seasonal naive model: the predicting of a particular week is, whenever possible, made on the basis of same week of the past season.

The classical model is a per-region SARIMA. We took a classical benchmark that was based on a SARIMA / SARIMAX. Then the statsmodels package was used, which fitted one model at a time. All the regional models were trained on the train + validation period and they were then utilized to make predictions on the test period of a region**.**

4.3 Global random forest: Advanced Model.

As the advanced approach, a RandomForestRegressor model was selected, as universal panel model. We did not use a single model per region, but we only trained a single Random Forest on the entire regions. The input, at every given time (t), is REGIONID, time features, and the lagged values of ILIPERCENT and PERCENT-POSITIVE; the goal is ILI percent at time t +1.

The justification is the fact that the Random Forests are non-linear and tree models that can deal with a great number of features and interactions without defined assumptions. Furthermore, a single global model was trained, allowing it to acquire patterns of cross-regional dynamics and general structures of lags.

5. Results & Model Comparison

Having assembled the models, we were interested in how much they can assist us when it comes to getting rid of the reactive mentality and adopting the predictive one. To complete the last throw-down.

5.1 Overall Performance on Test Set

All of the models run over the test period, which is made up of the COVID and post-COVID years. We have briefly included a table with MAE, RMSE and MAPE of the four main models in the report:

	Model	**MAE**	**RMSE**	**MAPE**
0	Naive	1.294771	1.978649	47.412533
1	SeasonalNaive	1.086166	1.770939	65.243284
2	SARIMA	0.993842	1.450021	55.241175
3	RandomForest	0.225478	0.372289	10.618856

By now it is not very hard to guess which one it is where. The naive and seasonal naive models are at the lowest point of the list, and the seasonal naive already receives a booster due to the annual recurrence of the flu seasons.

The SARIMA model was able to beat the baselines, sweeping off the short-term autocorrelation and selecting the seasonal pattern in a more organized manner.

The global Random Forest is, though, the best, as it hits the lowest error values on the three measures of the test window.

Relative to the naive baselines the Random Forest slices the RMSE and MAE in a pleasing manner, and it even manages to beat SARIMA by a sufficient margin. Although we are not narrowing down to precise numbers in this case, it is evident that the advanced model does provide a significant boost compared to the simple seasonal persistence and the old-fashioned time-series method.

5.2 Per-Region Performance

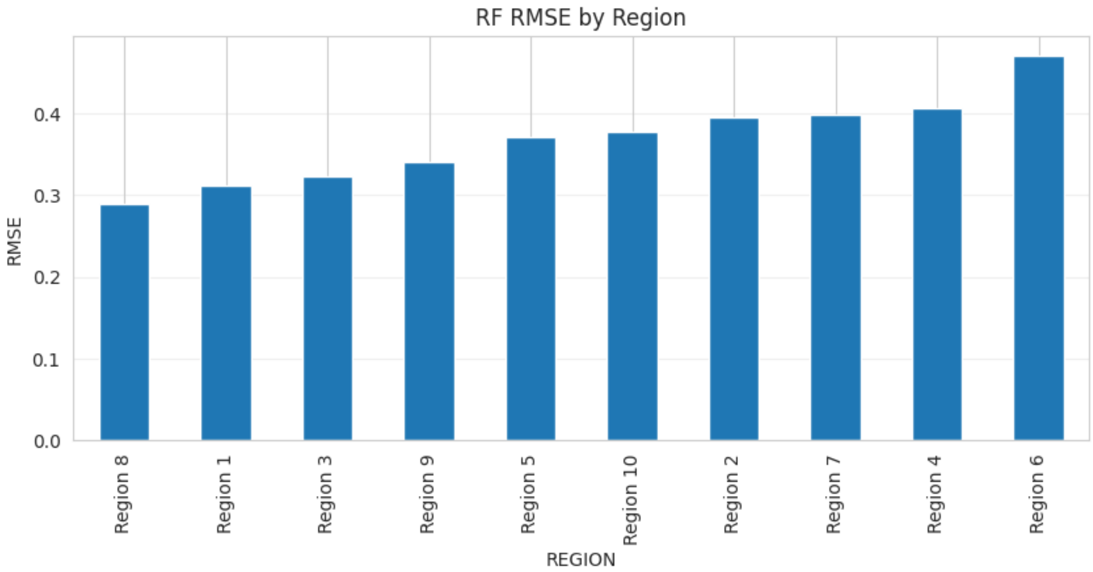


Figure 5 RF RMSE by Region

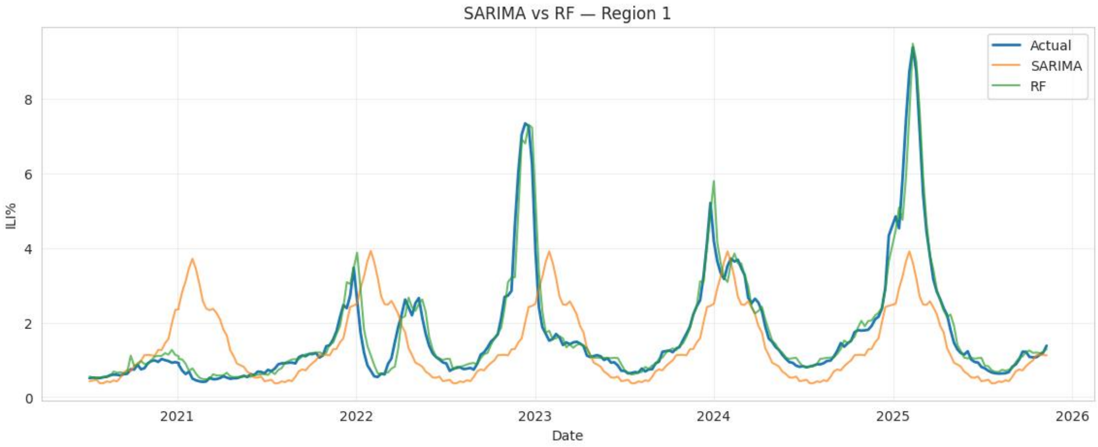


Figure 6 SARIMA vs RF — Region 1

We also monitored the performance of the models at regional level. With the results of the region metrics we observed that the Random Forest is steadily solid at most locations, but it goes without saying that absolute error also varies around depending on the activity and noisiness of various areas in terms of their flu history.

When we compared SARIMA to Random Forest region region by region, certain patterns were formed. Where flu has an extremely regular seasonal pattern, SARIMA can be so close to the Random Forest and in some cases the difference is small. Conversely, in areas where the trends are more variable or irregular or the COVID period introduced the wrench in the regular dynamic, the Random Forest clearly does better. That advantage is particularly visible when the lagged value of lab positivity (PERCENTPOSITIVE) is of central importance, since the Random Forest can more effectively leverage the lagged values when compared to SARIMA.

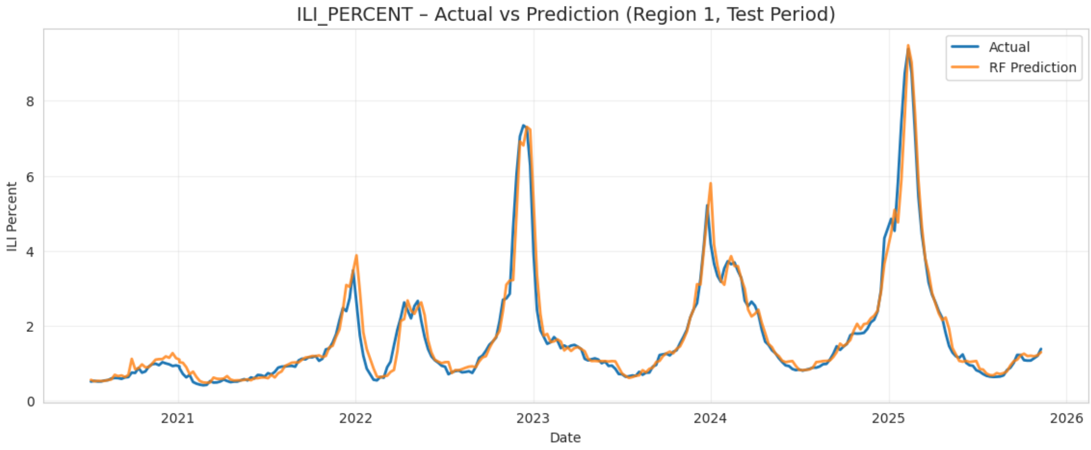


Figure 7 ILI_PERCENT – Actual vs Prediction (Region 1, Test Period)

6. Discussion, Limitations & Business Relevance

Here we are attempting to determine why some models had this one better than others, how concrete and transparent they are or not, and what we found may possibly convey to actual health-care settings.

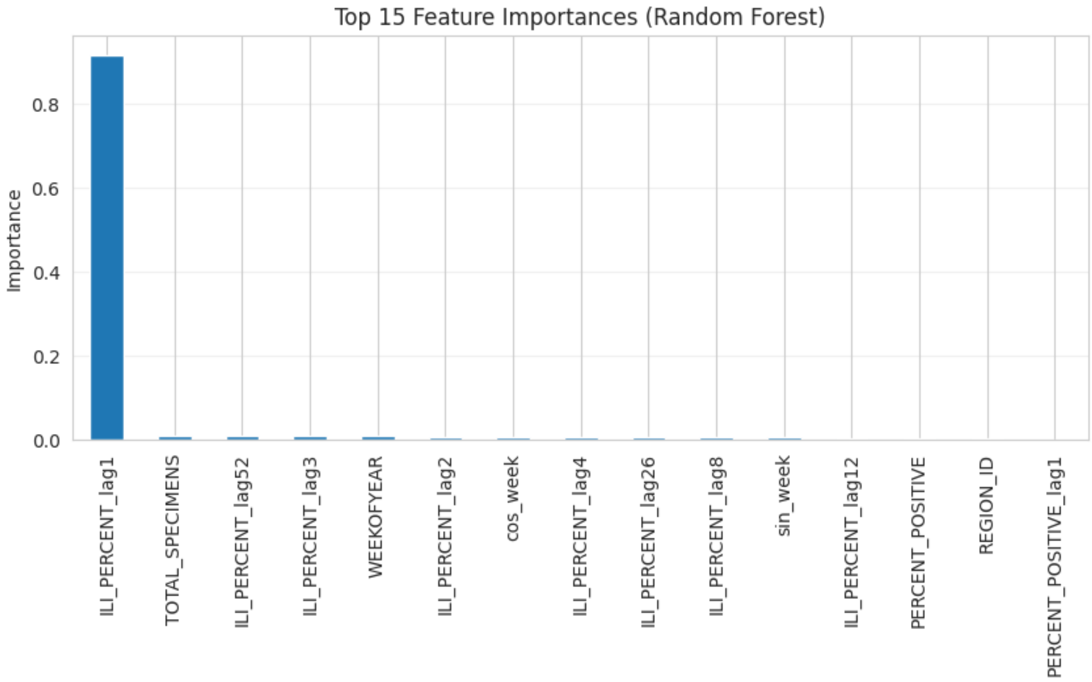


Figure 8 Top 15 Feature Importances (Random Forest)

6.1 Reason Why Not All Models Performed the same?

The random forest became a star in the show as it has the ability to multitask different kinds of information. It draws in fat lag structures to ILI% lagged lab positivity and seasonality, and it is sufficiently smart to teach non-linear interaction. Training on many places at once also confounds knowledge across locations, which also comes in very handy when individual time series are short or corrupt.

SARIMA models are winners in a case when the influenza is seasonally periodic. They are useful in explaining autoregressive and moving-average effects as well as annual seasonality in a clear and parametric manner. However, they are harder when we experience an abruptness of change, such as the COVID -19 pandemic, or when we have to harness complex, non-linear impacts of outside factors.

Both the naive and seasonal naive baselines are quite robust, in fact, this is easier than you would think, mere because flu is both persistent and seasonal. They have already stolen a large part of the signal. But since they are incapable of responding to structural changes and exploiting additional information, they inevitably fall behind those methods which are more sophisticated.

6.2 Business / Policy Relevance

When we put ourselves in the position of a manager of a hospital or a public health officer, the most efficient model in terms of performance, our global Random Forest, would be able to run a risk dashboard each week. That dashboard might report the predicted ILI % in each region next week with very easy to read cues of whether the activity is likely to remain low, increase or produce a peak. This may aid in staffing, bed capacity planning, and the communication campaigns.

7. Conclusion & Future Work

This project was also going to try to determine how much we can predict in the short-term going forward using historical surveillance data on a regional basis, and how far will we get away from an entirely reactive strategy to a purely predictive one. We have concluded that a global Random Forest model, constructed on lagged ILI percent, lagged lab positivity and seasonal terms best works during the test period, even in the challenging COVID years and the even more challenging post-COVID years. It is obviously better than naive predictions and simple as well as seasonal naive predictions and even better than classical SARIMA models. This highlights the importance of exogenous information and considering the problem as an international panel problem, as opposed to regional series issues.

As to the future, some natural sequences are available. One is to attempt LightGBM or XGBoost on the same set of features and find out whether it can squeeze any more profit. The other one is to widen the horizon, and do not make one-step-ahead forecasts, but multi-step forecasts, or even probabilistic forecasts which do give full distributions, not point estimates. Lastly, when appropriate data are available, the framework may be augmented with additional exogenous variables (such as mobility trends, weather information, or demand in vaccines) and move towards being the usable instrument of real-time influenza forecasting.

APPENDIX A Reproducibility / Collaboration GitHub.

Because this was a group project, GitHub was also our collaboration hub. Team members worked on different tasks such as exploratory analysis, feature engineering, model implementation, evaluation and writing, and combined their work through commits and pull requests. This workflow helped avoid overwriting each other's changes and preserved a clear history of how the project evolved.

The full notebook, scripts, and reproducible code used in this thesis are available at: https://github.com/ElMehdiBayoud/flu-forecasting-us.git