

Forecasting and nowcasting

Towards predicting the next financial crisis

Mentor:

Alberto Americo

Participants:

Tiago, Vittorio, Georgios, Nikolaos, Robert, Francesco, Thomas

"Data modeling is where art meets science. It requires the precision of a mathematician, the creativity of an artist, and the judgment of an experienced craftsman to transform raw data into valuable insights."

Gepetto di Firenze

Data exploration

Mixed frequency data (daily, monthly, quarterly):

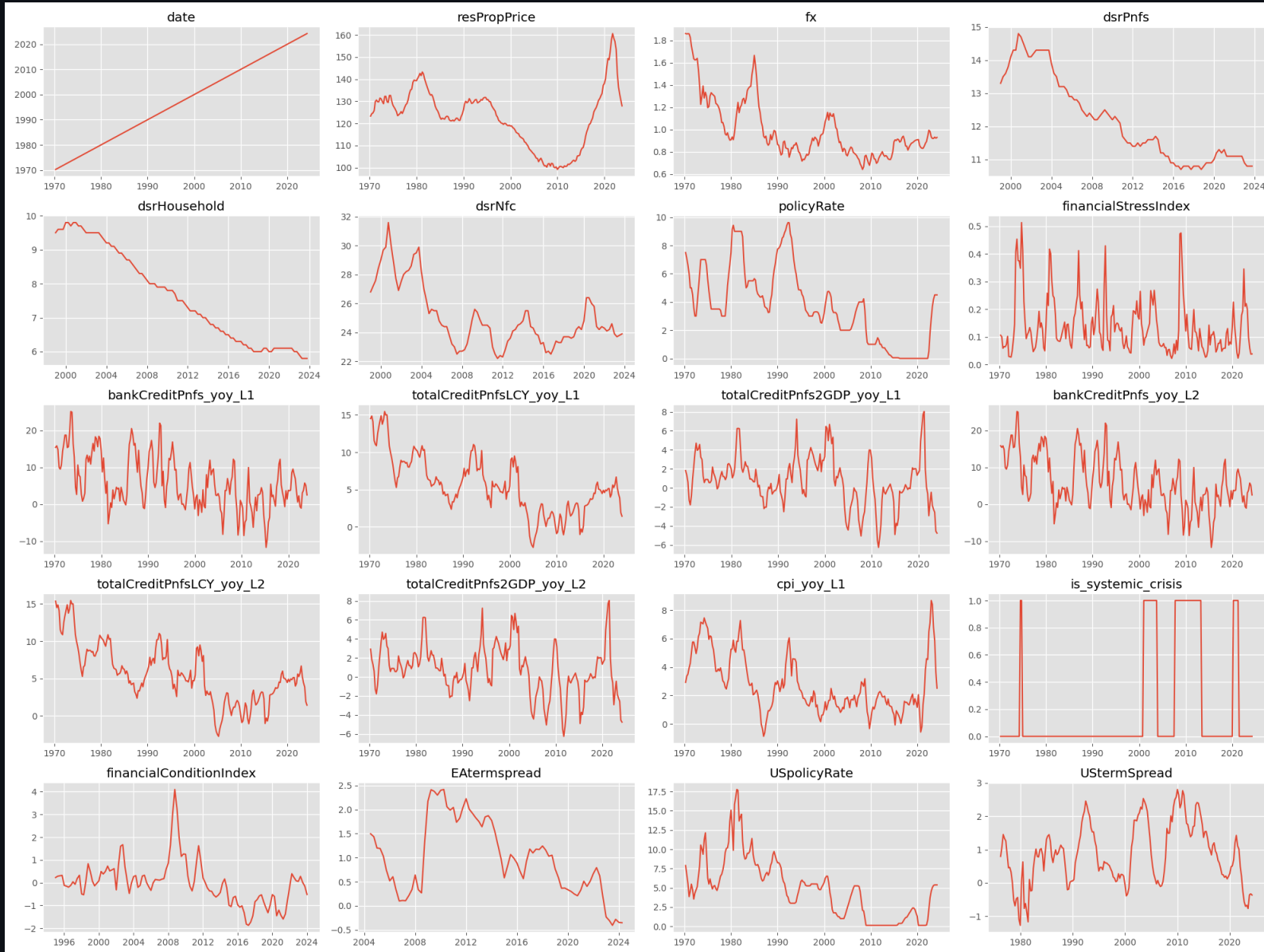
- Global Liquidity Indicators
- Bank Loans
- Real Residential Property Prices
- Consumer Prices
- Bank Credit to PNFS
- Total Credit to PNFS
- Debt service ratio
- Policy Rate
- Yield Curve data
- Exchange rates

Data preparation pipeline

```
COUNTRY = 'DE'  
FILE = './data/data_input_quarterly.csv'  
TIME_INTERVALL = "quarterly"  
  
df = read_data(FILE, COUNTRY)  
df = get_processed_df(df, COUNTRY, TIME_INTERVALL, verbose=True)  
df = subselect_data(df)
```

Data preparation pipeline

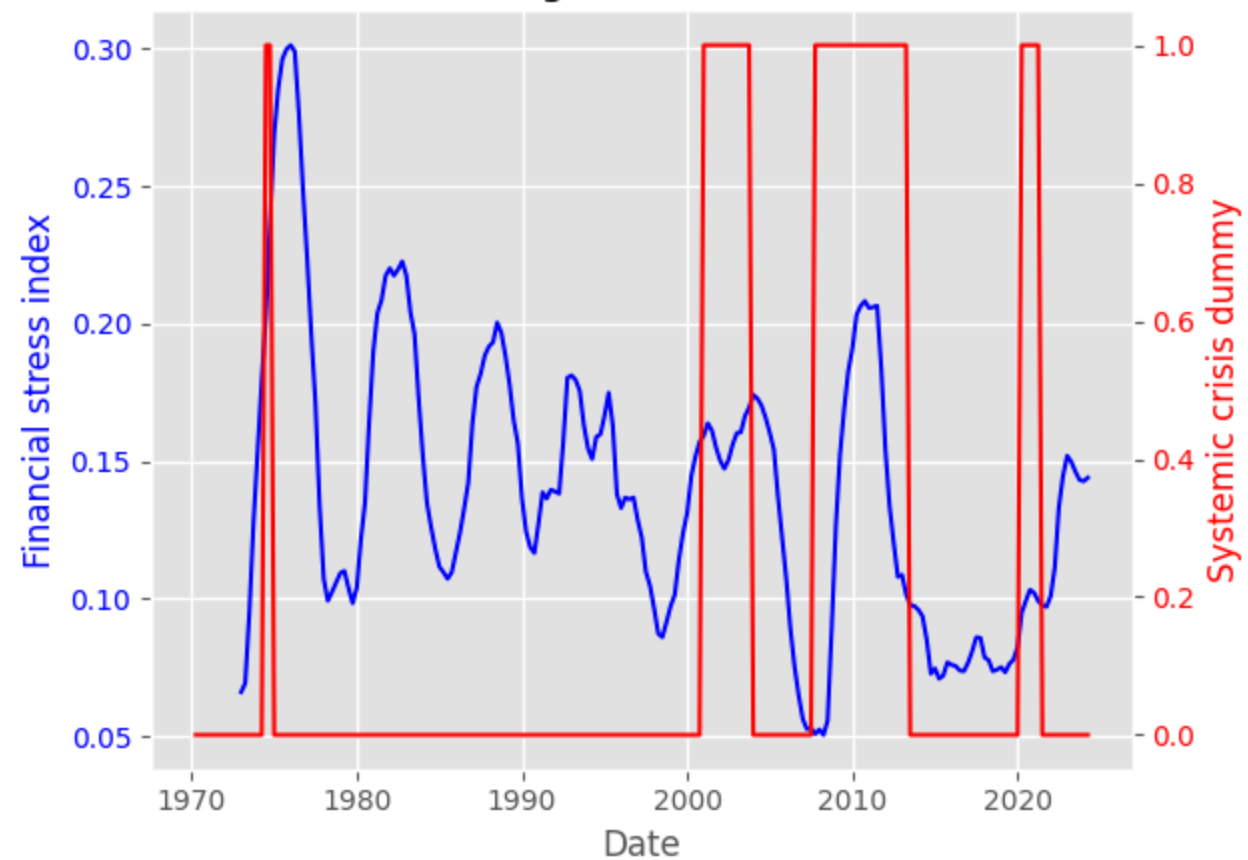
```
df = give_sliding_window_volatility(df, 4, "fx")
df = calculate_growth_rates(df, yoy_variables)
df = get_lagged_variables(df, 2, lag2_variables)
df = add_missing_variables(df, country)
df = add_systemic_risk_dummy_with_df(df, df_dummies, country)
```



Target value

- Systemic crisis (*dummy*)
- Systemic stress *continuous*
- Inflation ?

Target variables



Benchmarking (with MLflow)

- define train and test datasets
- define forecast horizon
- define set of regressors

Models

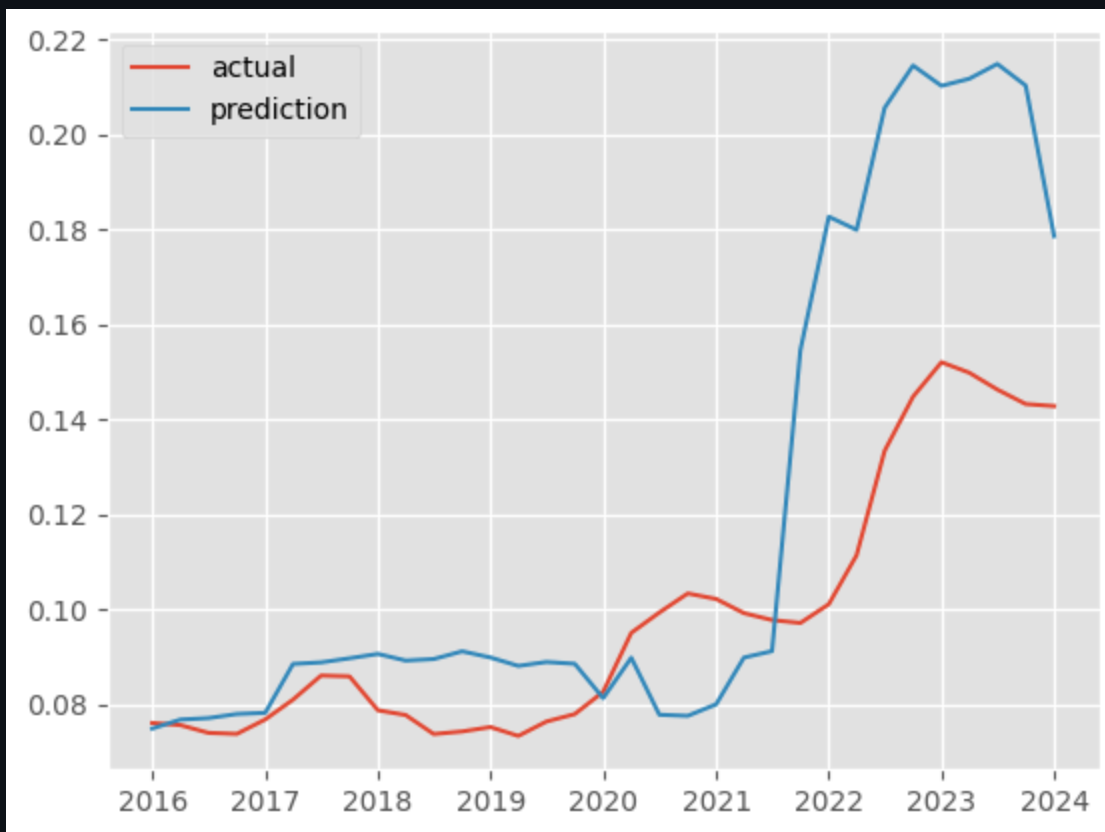
- OLS
- Random Forest
- Ridge
- XGBoost
- Logistic Regression

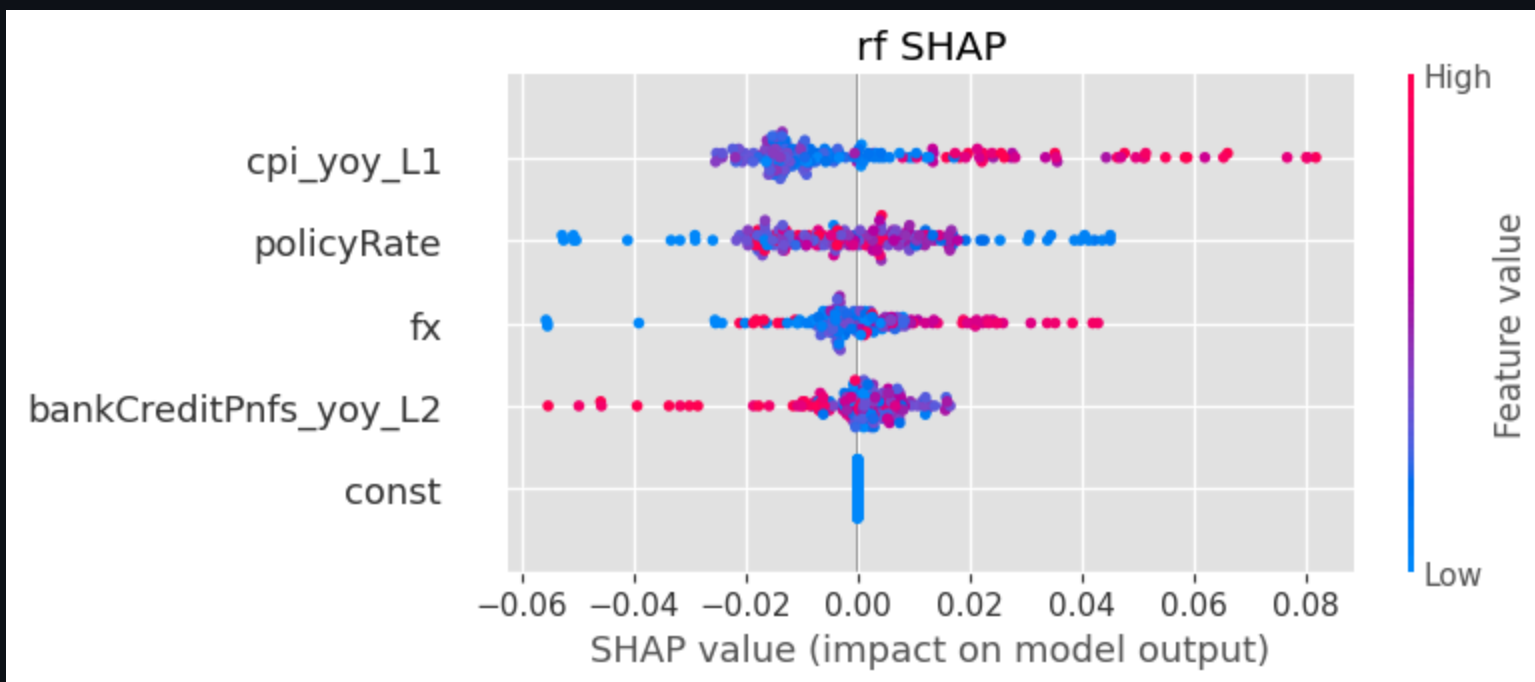
Hyperparameters tuning

- GridSearchCV with TimeSeriesSplit

Model interpretability and evaluation

- Shapley values
- Permutation importance
- RMSE
- F1-score
- ROC AUC
- R-squared (pseudo)





PCA intuition

- Principal component analysis (PCA) reduces the number of dimensions in large datasets to principal components
- Retain original information.
- Transforming potentially correlated variables into a smaller set of variables, called principal components.

(between us: just let Sklearn do the magic)

PCA Pipeline

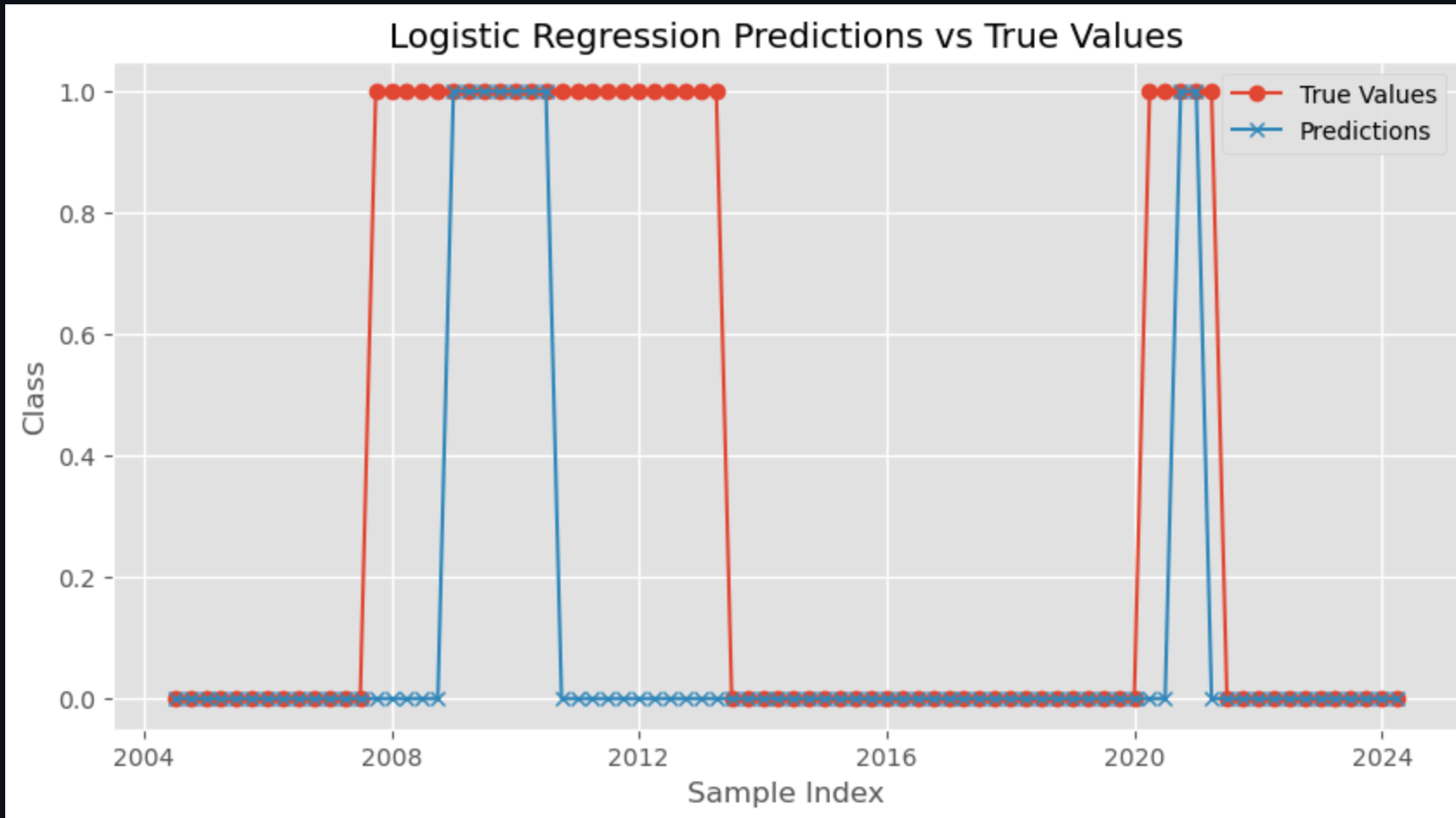
```
from sklearn.decomposition import PCA

pca = PCA()
principalComponents = pca.fit_transform(X_SCALED)
PCA_components = pd.DataFrame(principalComponents)

explained_variance_ratio = pca.explained_variance_ratio_
_target_variance = 0.80
_current_variance = 0.0
_num_features = 0

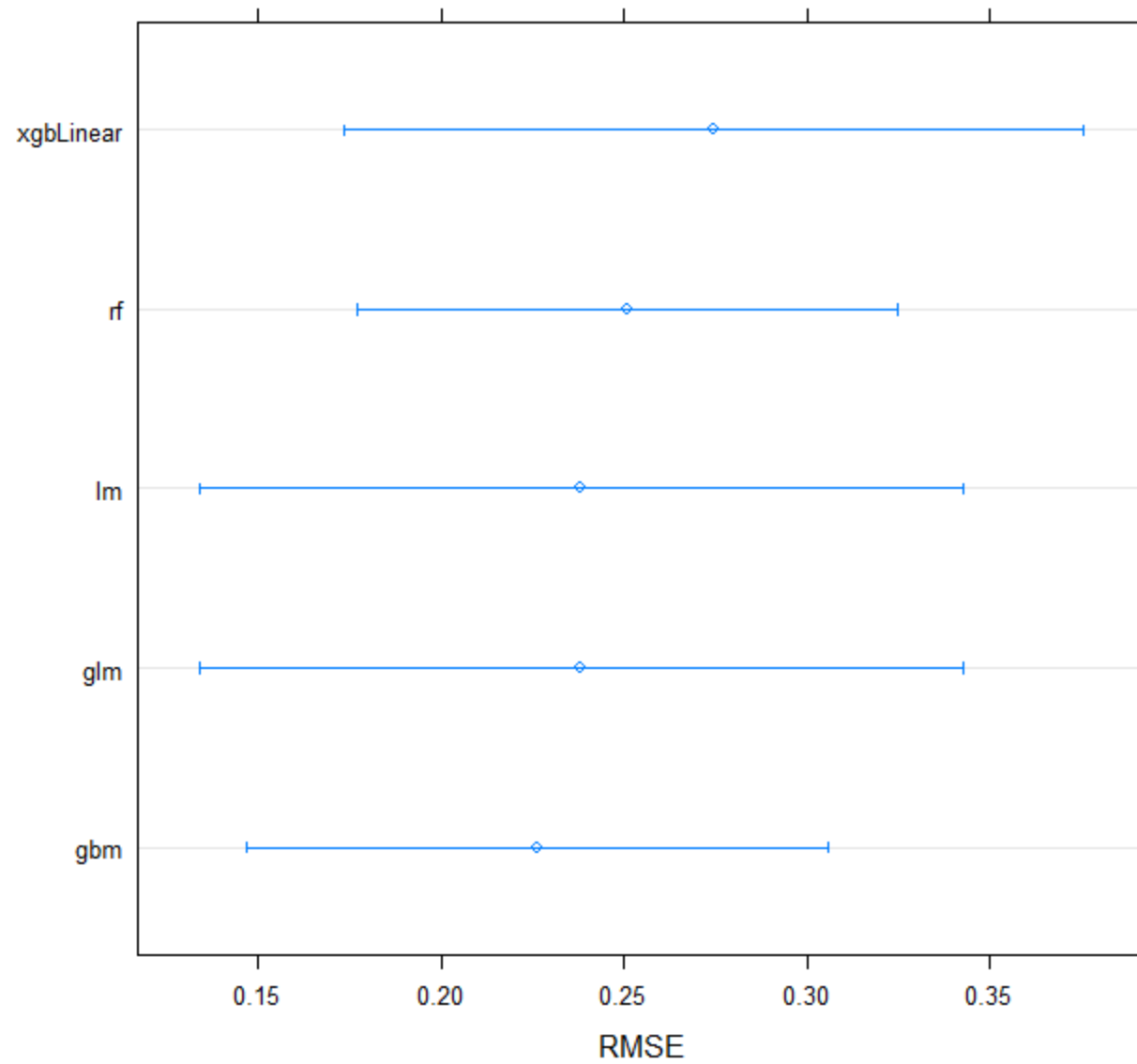
while _current_variance < _target_variance:
    _current_variance += explained_variance_ratio[num_features]
    _num_features += 1
```

PCA - preprocessed result



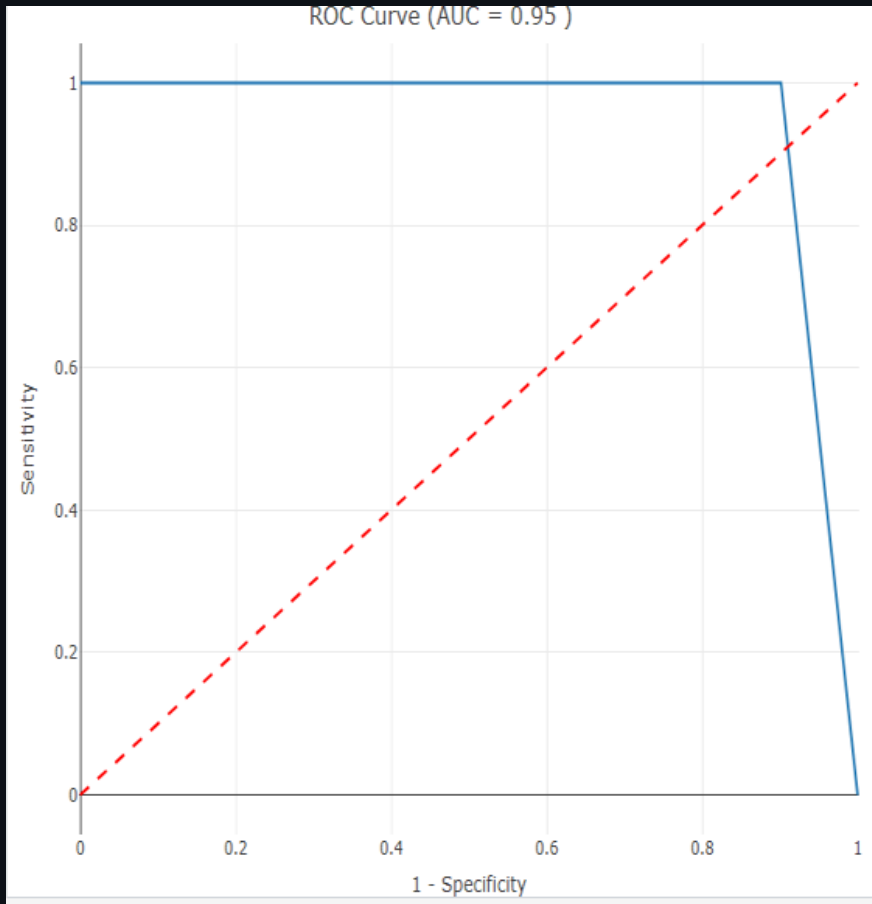
It seems to be too late, however, COVID-19 is exogenous and could not have been predicted. However, result is stable

Results of Robert alternative approach

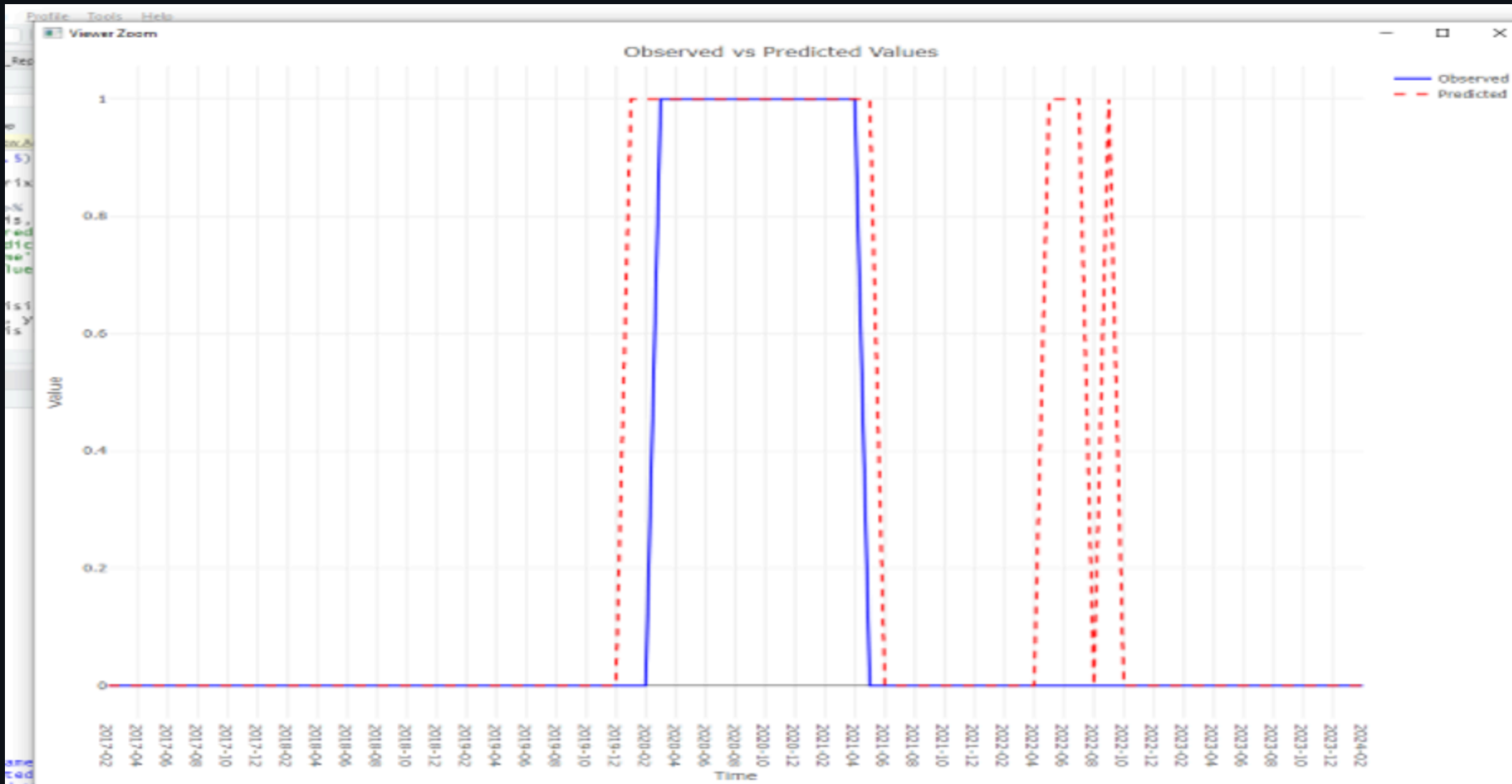


Confidence Level: 0.95

Out of the 5 models observed, it was built an ensemble of them all having an optimisation function that applies different weights in order to get a better RMSE for the ensemble built



In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.



Testing showed that the model was able to anticipate the crisis 3 months before (Covid period) and was also able to capture the geopolitical impact in comparison to the target variable

Next steps

- Expand country coverage
- Try aggregate model
- Increase data frequency
- Try different forecast horizons
- Improve feature selection process
- Benchmark against commercial forecast (eg Bloomberg, Turnleaf)
- Improve model interpretability
- Public GIT repo to share knowledge amongst institutions