Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
Data Analytics for Business

MASTER'S FINAL WORK
Dissertation

Forecasting Inflation with Machine Learning

Vladislava Piletska

Supervision:
João Afonso Bastos
Adriana Cornea-Madeira

Research Gap & Objective

Problem Statement:

Traditional econometric models often struggle with inflation forecasting, especially in volatile economic environments like the Russian Federation for the last couple years. These methods rely on linear relationships and historical data and may easily fail to adapt to unexpected economic and geopolitical changes.

Proposed Solution:

Combine macroeconomic indicators with text-based features extracted from economics-focused news articles to train regression-based statistical and machine learning models. The goal is to outperform a benchmark model and demonstrate the effectiveness of this technique in a real-world application.

Data: Final dataset

Type	Features	Data type
Macroeconomic	Average monthly salary	float
	Unemployment Rate	float
	Sanctions	bool
	...	
	Oil Price	float
Text	topic1_ratio	float
	topic2_ratio	float
	....	float
	topic27_ratio	float

Macroeconomic	Lags (Macroeconomic)	Text	Lags (Text)
16	18	27	27

Region: Russia

Date range: June 2011 - December 2023

Data granularity: monthly

Number of collected news articles: 96K articles

Total number of features: 88

Number of observations: 151

Note: inflation rate and interest rate of t-1 were used as lag_1 features

Data: Data preprocessing

Macroeconomic data

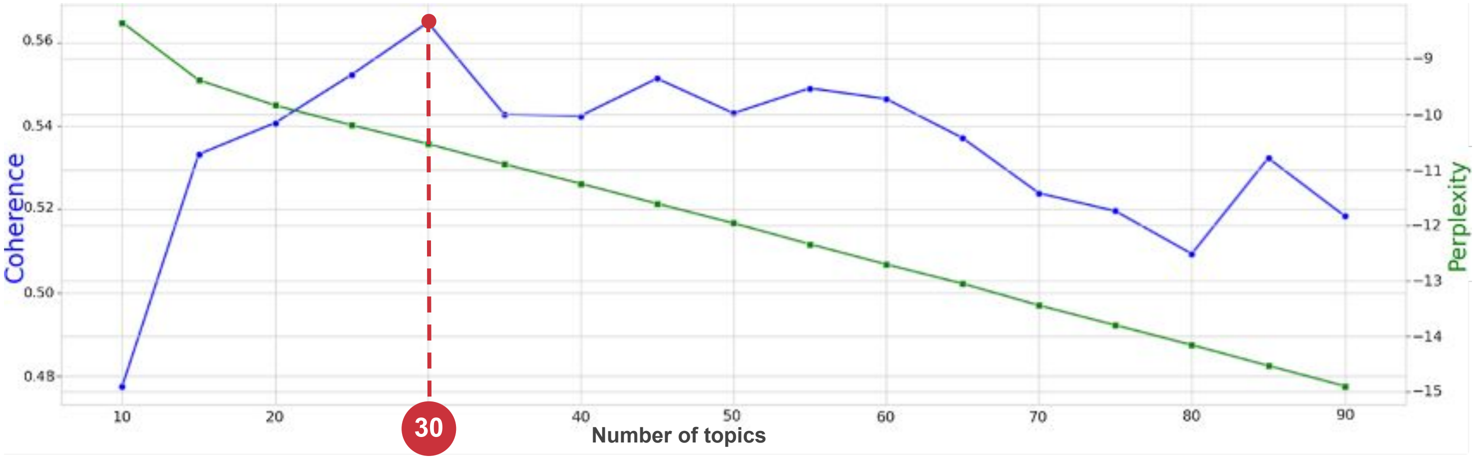
1. Data collection from various sources
2. Ensuring consistent granularity
3. Feeling in the missing values
4. Stationarity

Text data

1. Data collection from Lenta.ru
2. Text preparation
3. Text tokenization (BoW)
4. Topic extraction (LDA)
5. Transform into Time series
6. Stationarity

Data: Topic extraction with LDA

Coherence and Perplexity evaluation



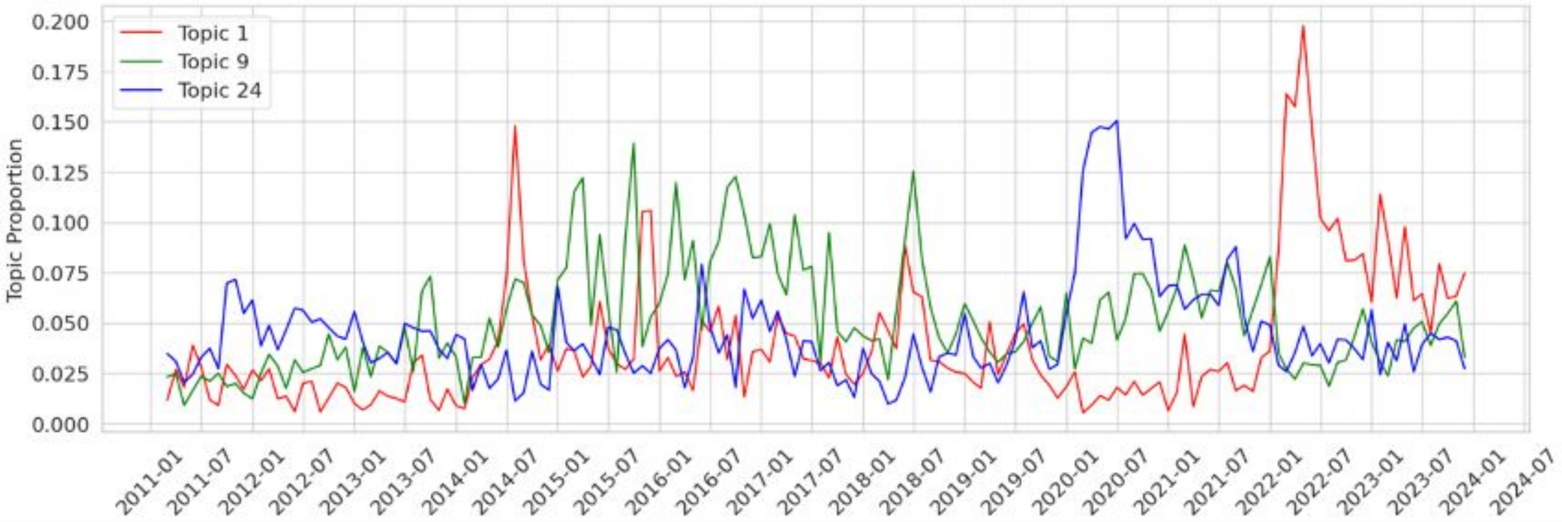
Coherence score measures the interpretability of the topics generated by the model. Higher coherence scores indicate that the topics are more meaningful.

Perplexity evaluates how well the model predicts unseen data, with lower perplexity values indicating better generalization and model performance.

$$\text{The topic's weight (or ratio) for a given month} = \frac{\text{Number of articles assigned to a specific topic in a given month}}{\text{Total number of topics presented in a given month}}$$

Data: Time series creation

Top-3 the most discussed topics frequency over time



Topic 1: Discussions on sanctions and limitations

Topic 9: Pension payment amounts

Topic 24: Economic relations between Russia and the United States

Data: Stationarity

- ❖ To achieve stationarity, a differencing transformation was applied.
- ❖ An order of differencing check was performed using two statistical tests, the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test.

<u>Macroeconomic indicators used in this study</u>

<u>Stationarization via Differencing</u>

Methodology: Models

Lisbon School of Economics & Management
Universidade de Lisboa

The following models were used:

- Random Forest (RF)
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net (ENet)
- Partial Least Squares (PLS)

Benchmark model:

- Autoregressive Model of Order 4 (AR(4))

Methodology: Concept

Lisbon School of Economics & Management
Universidade de Lisboa

Train-test split

85% of the dataset (130 out of 151 observations) was used as the initial training set.

Rolling window approach

Using window approach, each model produce forecasts of the inflation rate at t+3, t+6, t+9, and t+12 horizons.

At each time step (t+3, t+6, t+9, or t+12), the model was re-trained using a fixed window of past data, with hyperparameter tuning performed via grid search.

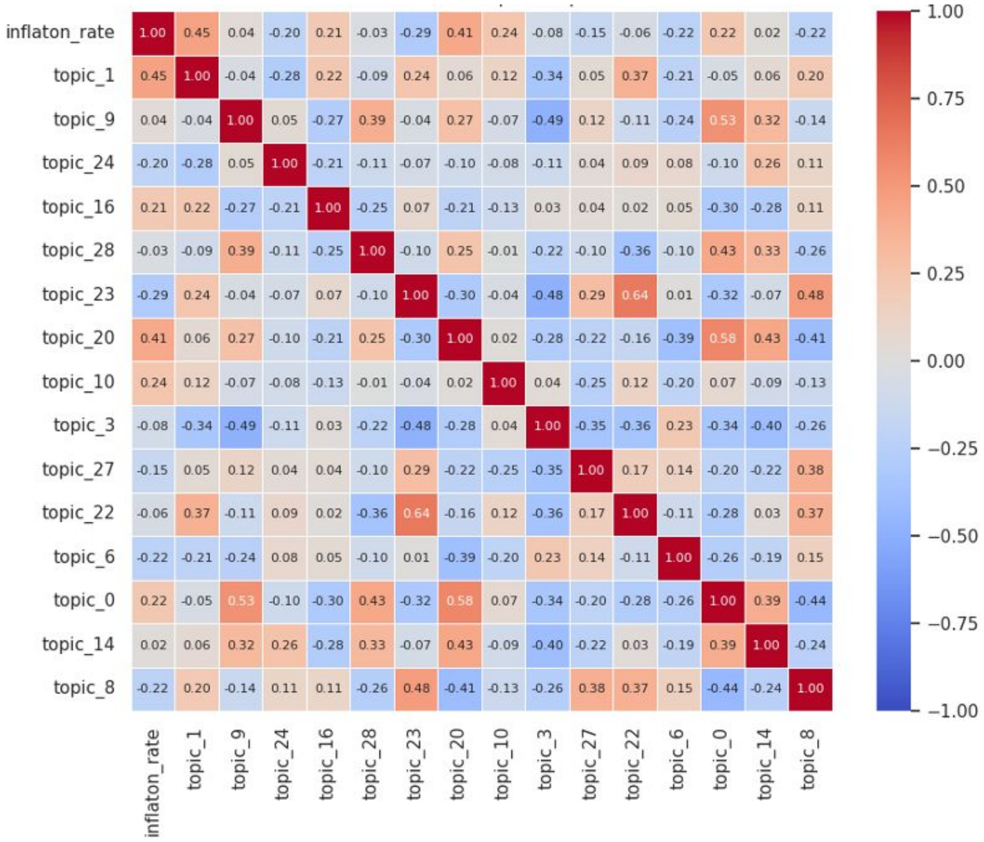
This concept was tested with three strategies:

- Macroeconomic data only (M);
- Narrative data only (N);
- Combined: Macroeconomic + Narrative data (M+N).

Methodology: Feature selection

- ❖ Text features with low correlation to the target variable (|correlation| < 0.05) were excluded.
- ❖ LASSO shrinks some coefficients to zero automatically.
- ❖ ENet selects some, regularizes others.
- ❖ RF ranks importance of all features.
- ❖ PLS doesn’t perform feature selection.

Correlation top-15 news topics with Inflation Rate



Lisbon School of Economics & Management
Universidade de Lisboa

Grid search

Model	*Tuning hyperparameter(s)*
LASSO	λ – penalty parameter
ENet	λ – penalty parameter l1_ratio – controls the balance between L1 (LASSO) and L2 (Ridge) regularization
PLS	n_components – number of latent factors (components)
RF	L – maximum depth of the trees N – number of features considered at each split

On each grid search iteration, performance metrics were calculated, and the best combination of hyperparameters for each model was identified.

Methodology:
Evaluating forecasting models

- RMSE
- rRMSE
- Diebold-Mariano (DM) test

The loss differential was defined as:

$$\boxed{d_t = e_{m,t}^2 - e_{AR4,t}^2}$$

where d_t is the loss differential at time t;

$e_{m,t}$ is the forecast error of the model being evaluated;

$e_{AR4,t}$ is the forecast error of the AR4 benchmark.

$$\boxed{RMSE_{m,t+1,t+h} = \sqrt{\frac{1}{T - T0 + 1} \sum_{t=T0}^{T} (\hat{e}_{t+1,t+h}^{m})^2},}$$

where $\hat{e}_{t+1,t+h}^{m} = \pi_{t+1,t+h} - \hat{\pi}_{t+1,t+h}^{m}$ is the forecasting error;

$\hat{\pi}_{m,t+1,t+h}$ is the forecasting value for the next h-month inflation rate made by model m;

$T - T0 + 1$ is the number of observations (days).

Significance Levels

Symbol	*p-value Threshold*	*Significance level*
*	$p < 0.05$	5% (statistically significant)
**	$p < 0.01$	1% (highly significant)
***	$p < 0.001$	0.1% (very strong significance)

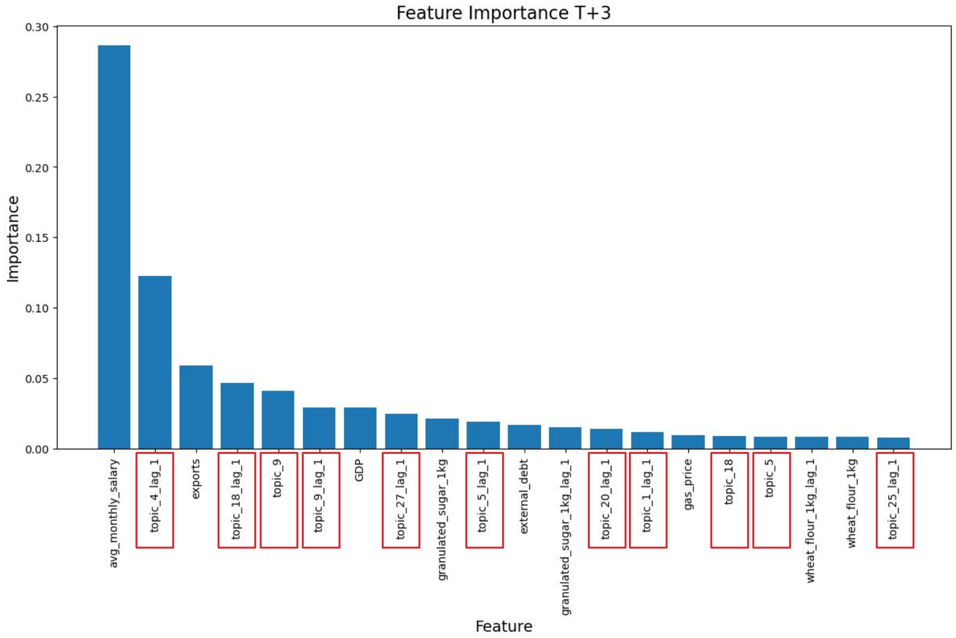
Results:

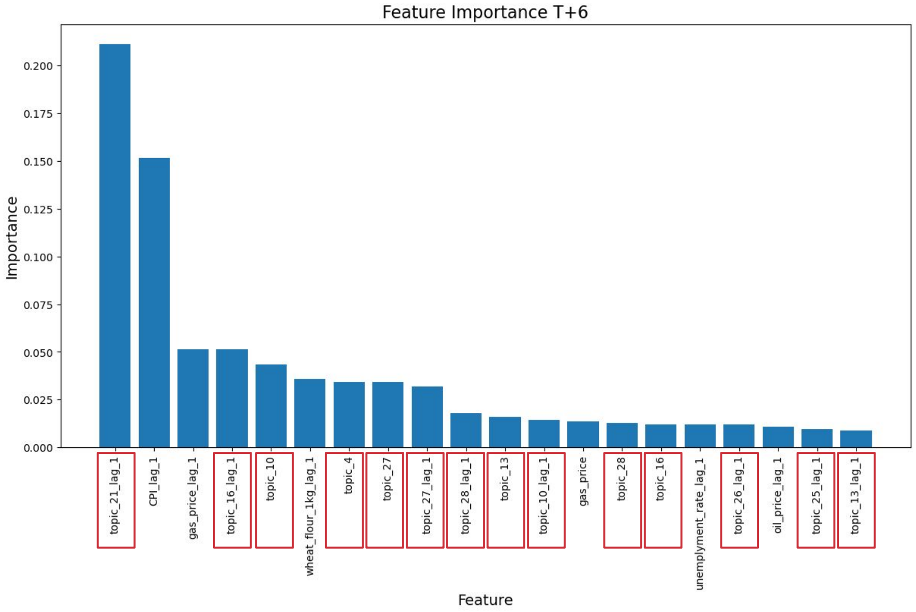
		rRMSE				DM-test				Significance			
Model	**Data**	**t+3**	**t+6**	**t+9**	**t+12**	**t+3**	**t+6**	**t+9**	**t+12**	**t+3**	**t+6**	**t+9**	**t+12**
RF	M	0,202	0,209	0,211	0,611	-5,350	-2,526	-2,199	-2,203	**	**	*	*
	N	0,281	0,327	0,362	0,698	-4,470	-2,298	-1,981	-1,666	***	*	*	*
	M+N	0,192	0,205	0,209	0,618	-5,172	-2,496	-2,183	-2,131	***	**	*	*
LASSO	M	0,242	0,231	0,241	0,240	-1,908	-2,153	-2,227	-2,012	*	*	*	*
	N	1,025	0,945	0,905	0,897	0,071	-0,273	-0,637	-0,810				
	M+N	0,242	0,231	0,241	0,240	-1,908	-2,153	-2,227	-2,012	*	*	*	*
ENet	M	0,314	0,291	0,292	0,291	-1,818	-2,135	-2,243	-2,026	*	*	*	*
	N	1,025	0,945	0,905	0,897	0,070	-0,273	-0,638	-0,811				
	M+N	0,314	0,291	0,292	0,291	-1,818	-2,135	-2,243	-2,026	*	*	*	*
PLS	M	0,529	0,483	0,467	0,469	-1,401	-1,980	-2,237	-2,032		*	*	*
	N	1,080	1,011	0,981	0,977	0,226	0,055	-0,123	-0,180				
	M+N	0,379	0,415	0,394	0,407	-1,557	-1,791	-2,060	-1,881		*	*	*

Results: Improvement with N data

Model	Data	rRMSE t+3	rRMSE t+6	rRMSE t+9	rRMSE t+12
RF	M	0.202	0.209	0.211	0.611
	N	0.281	0.327	0.362	0.698
	M+N	0.192	0.205	0.209	0.618
	% Improvement	-4.960	-1.930	-1.280	1.030
LASSO	M	0.242	0.231	0.241	0.240
	N	1.025	0.945	0.905	0.897
	M+N	0.242	0.231	0.241	0.240
	% Improvement	0.000	0.000	0.000	0.000
ENet	M	0.314	0.291	0.292	0.291
	N	1.025	0.945	0.905	0.897
	M+N	0.314	0.291	0.292	0.291
	% Improvement	0.000	0.000	0.000	0.000
PLS	M	0.529	0.483	0.467	0.469
	N	1.080	1.011	0.981	0.977
	M+N	0.379	0.415	0.394	0.407
	% Improvement	-28.270	-14.090	-15.770	-13.330

Results: Top-20 most important features across forecast horizons





Discussion

Lisbon School of Economics & Management
Universidade de Lisboa

- ❖ Combining macroeconomic indicators with narrative features derived via LDA topic modeling — capturing qualitative signals such as sanctions and geopolitical tensions — improved inflation forecasting accuracy, particularly over short- and medium-term horizons.

- ❖ The integration of narrative data led to a noticeable improvement in RF model accuracy, especially for short-term inflation forecasts (3–9 months).

- ❖ Regularization-based models (LASSO, Elastic Net) underperformed in the combined setting — possibly because their feature selection mechanisms excluded textual features.

Future research

Lisbon School of Economics & Management
Universidade de Lisboa

- Incorporate additional narrative sources (e.g., other news sources, social media, financial blogs) to reduce potential bias associated with relying on a single news source.

- Apply sentiment analysis to the news articles to complement topic modeling with emotional tone and polarity, potentially capturing market or public sentiment shifts that influence inflation.

- Experiment with deep learning-based neural forecasting models, which are capable of capturing complex temporal dependencies in time series data.

- Extend the framework to multilingual or cross-country datasets to forecast inflation across different markets.

Lisbon School of Economics & Management
Universidade de Lisboa

Thank you!