A study of the influence of news reports and other contextual open-source information on the consumer behavior of bank card users

Student: Grigoryev Mikhail, J4133c

Supervisor: Chunaev Petr Vladimirovich
Associate Professor, Department of Digital Transformation

Consultant: Kovantsev Anton Nikolaevich

Goal, object and subject of research

Goal: to use open-source context information to increase consumption forecasting quality and to identify exogenous impact of this information on consumption

Tasks:
1) review relevant literature
2) scrape and process context info
3) implement forecasting model (preferrably dynamic)
4) evaluate model quality on real data

Object: impact of context information (news) on consumption of bank card users

Subject: correlations between transactions in different categories and context time series (news topic, macroeconomical data, etc.)

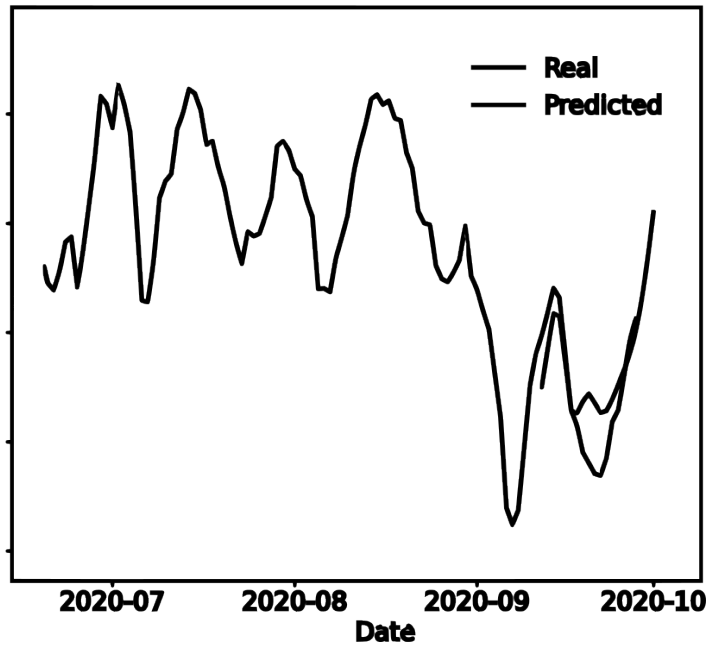
Application

NEWS

Causality?

TRANSACTIONS

PREDICTIONS



Literature overview

TOPIC ANALYSIS

- LDA
- NMF
- JASen
- Top2vec
- BERTopic

FORECASTING

- ARX
- RF
- BRNN
- (S)ARIMAX
- QRF
- RCNN
- VARX
- SVR
- LSTM
- NARX
- *k*NN
- LR/PCA

Plan of the experiment

NEWS HEADINGS

OTHER CONTEXT

TRANSACTIONS

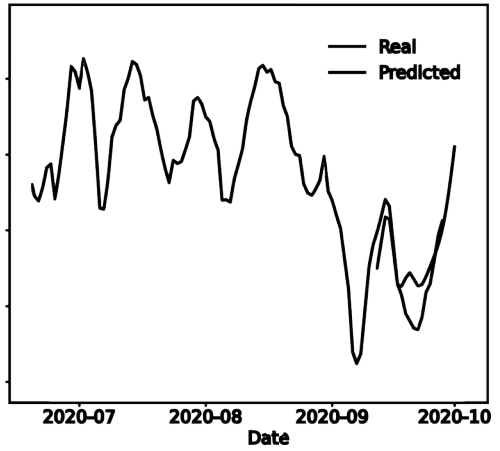
Lemmatization
BERTopic

Interpolation

CONTEXT TIME SERIES

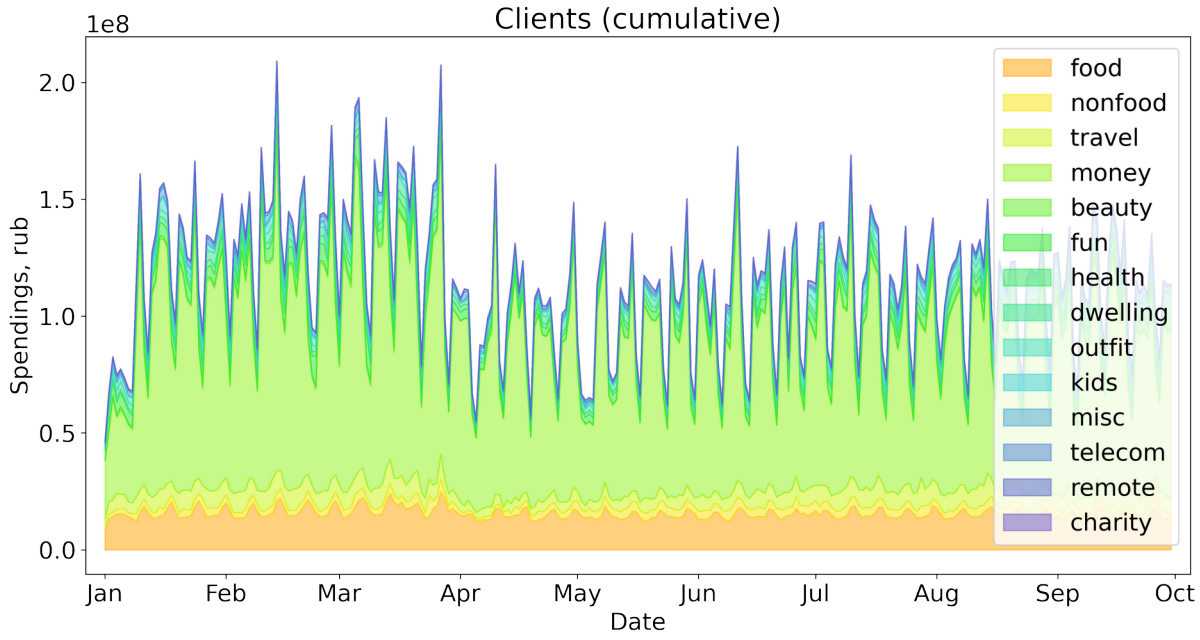
ARIMAX MODELS

PREDICTIONS

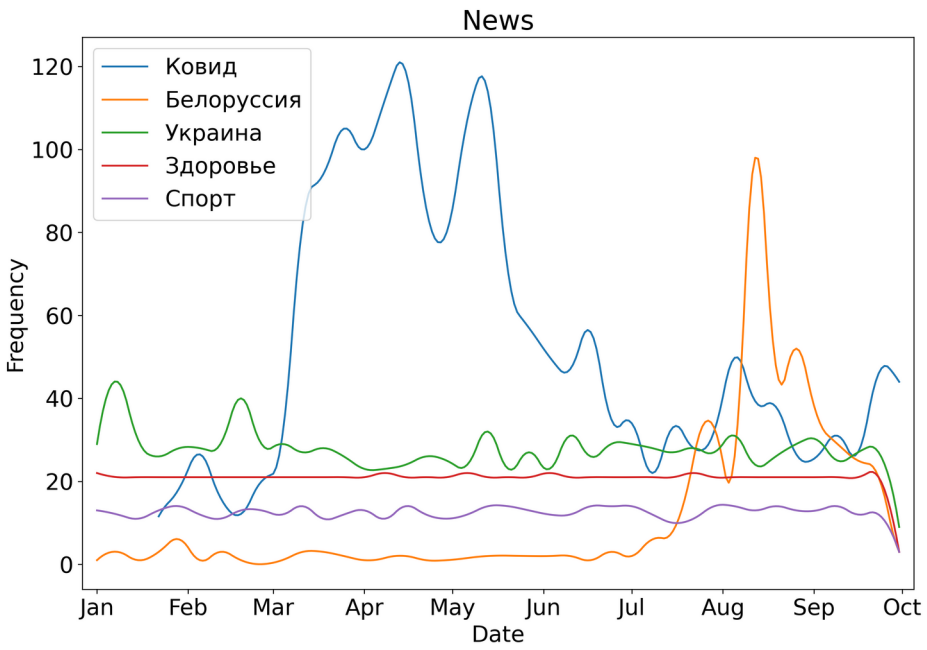


Data overview

Synthetic transactions

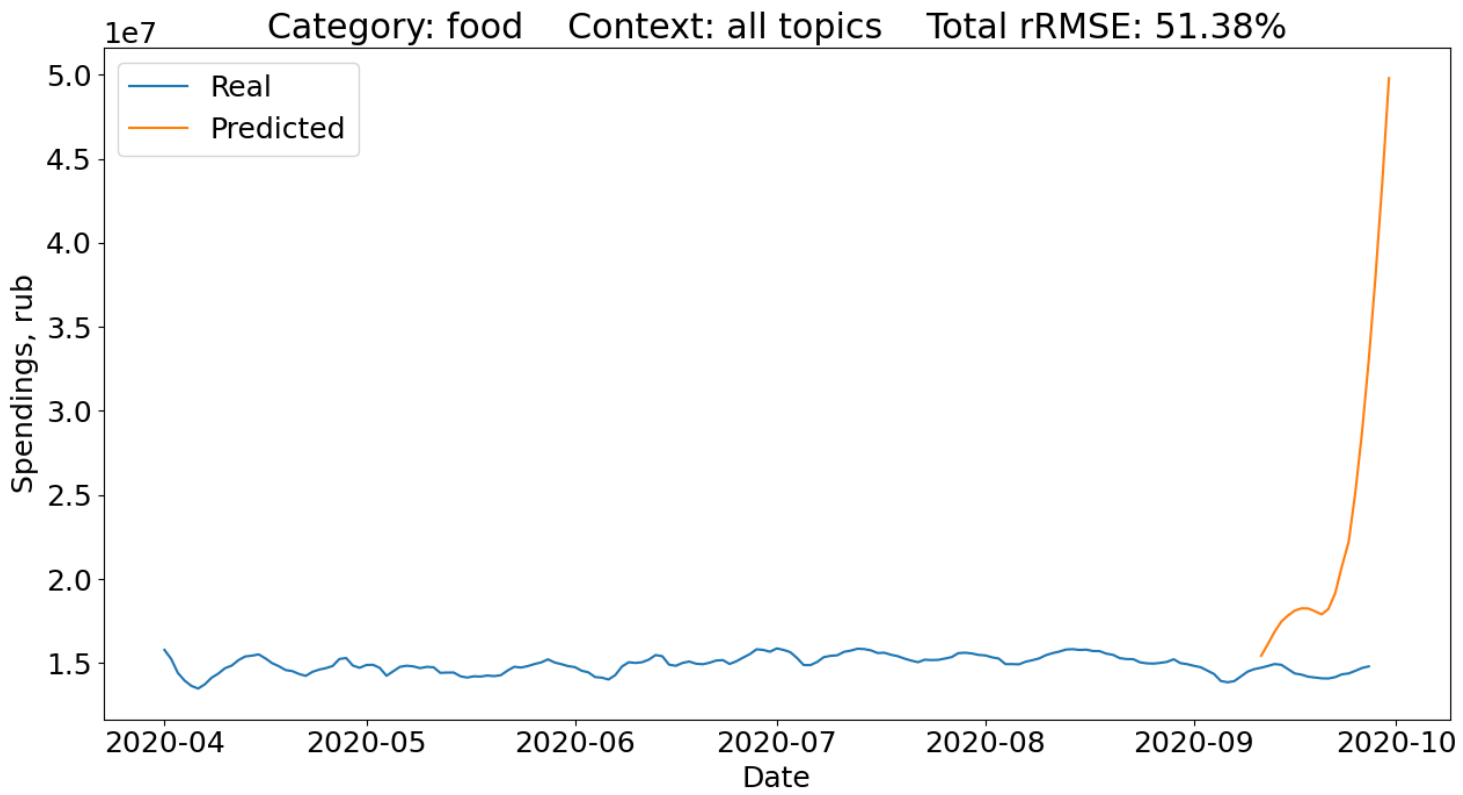


News after topic analysis



Additionally, **20** macroeconomical and epidemiological **time series** were added to the context.

ARIMAX using the whole context

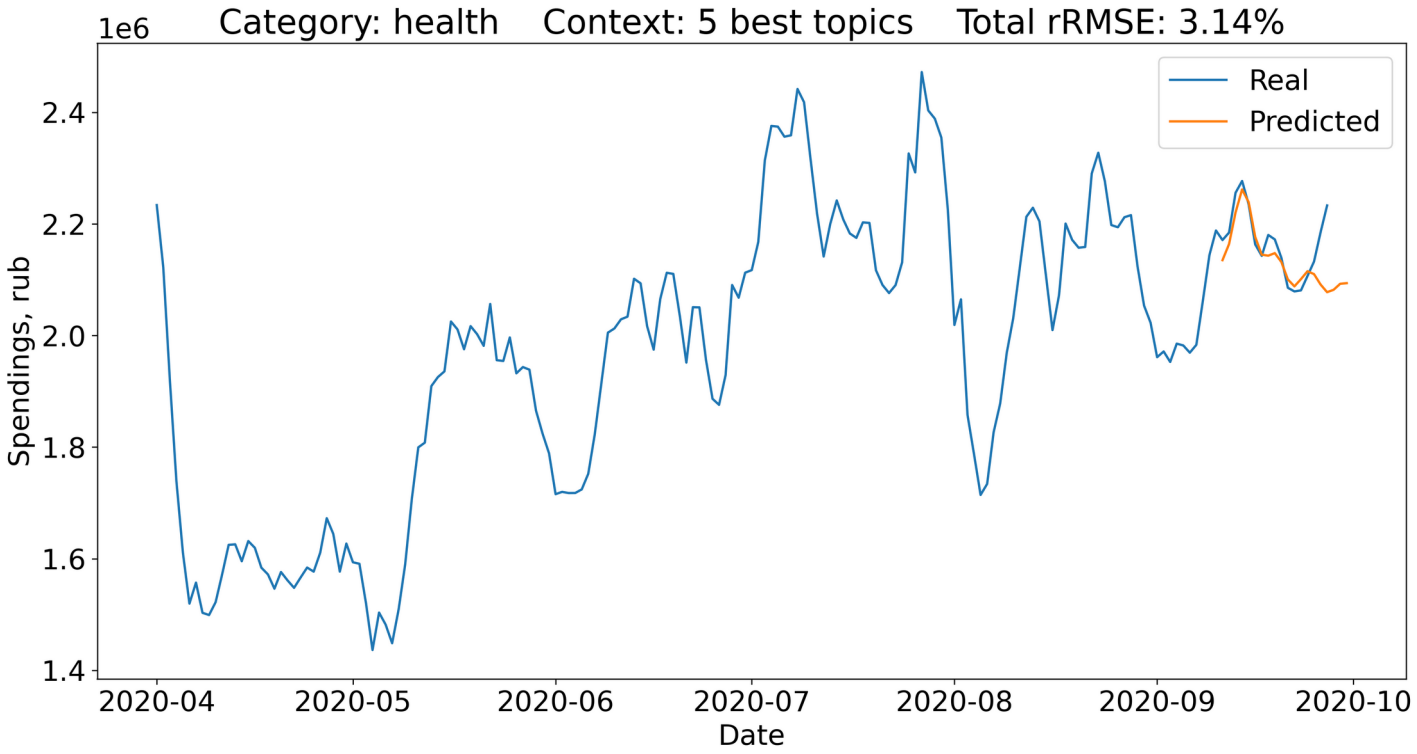


Error metrics are averaged across 10 runs with different prediction dates.

ARIMAX consumption forecast using the whole context is of low quality. ***Selection of exogenous variables is required.***

ARIMAX predictions for the best categories

Health – 3.14%

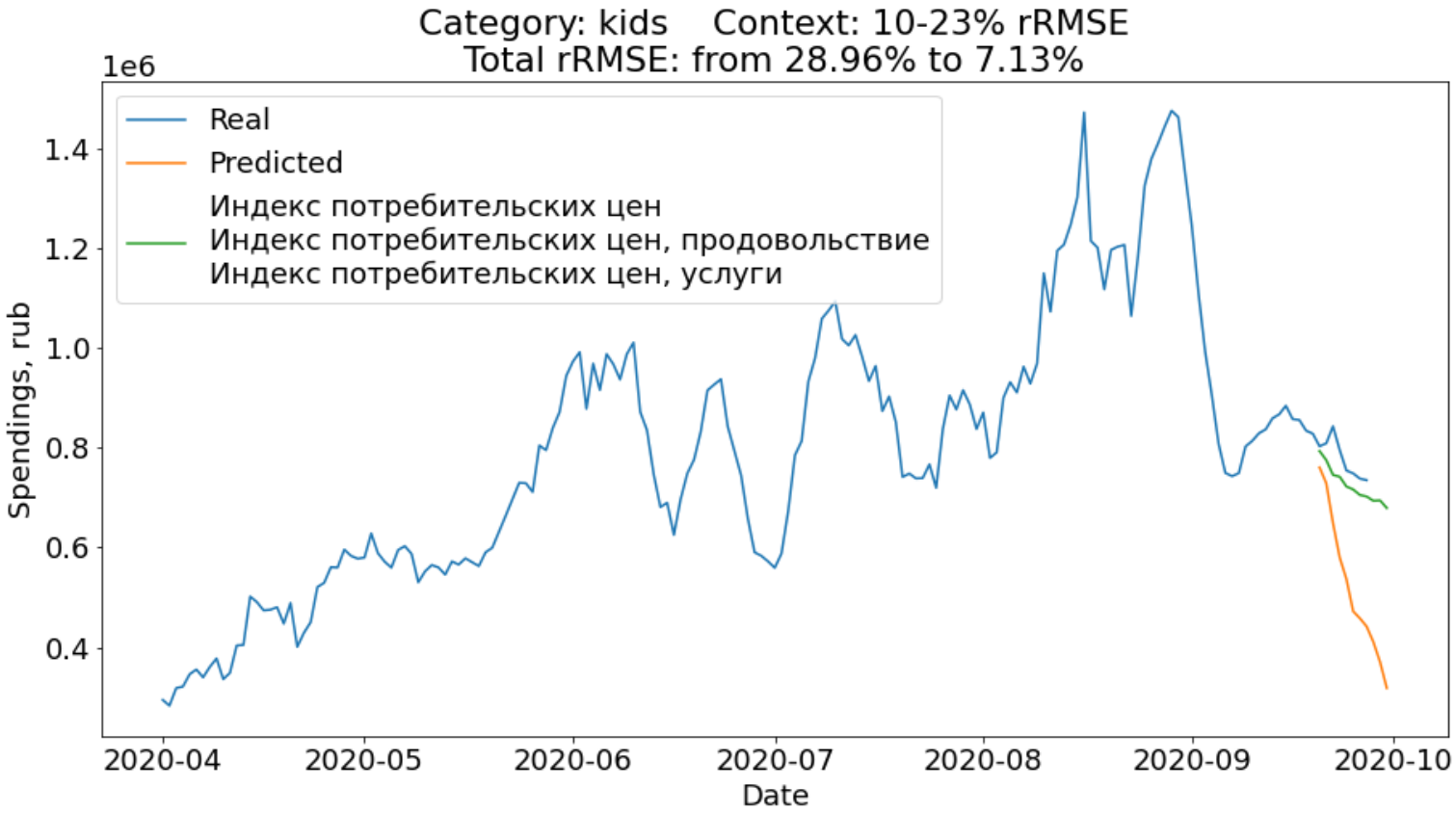


5 best topics, used for the model:

- Real estate affordability, Russia
- Real estate affordability, SPb
- COVID news
- GDP
- COVID-19 in Russia, new cases

ARIMAX prediction enhancement via adding exogenous variables

Kids – 29.96 → 7.13%



Added variables:

- Consumer price index
- Consumer price index, food
- Consumer price index, services

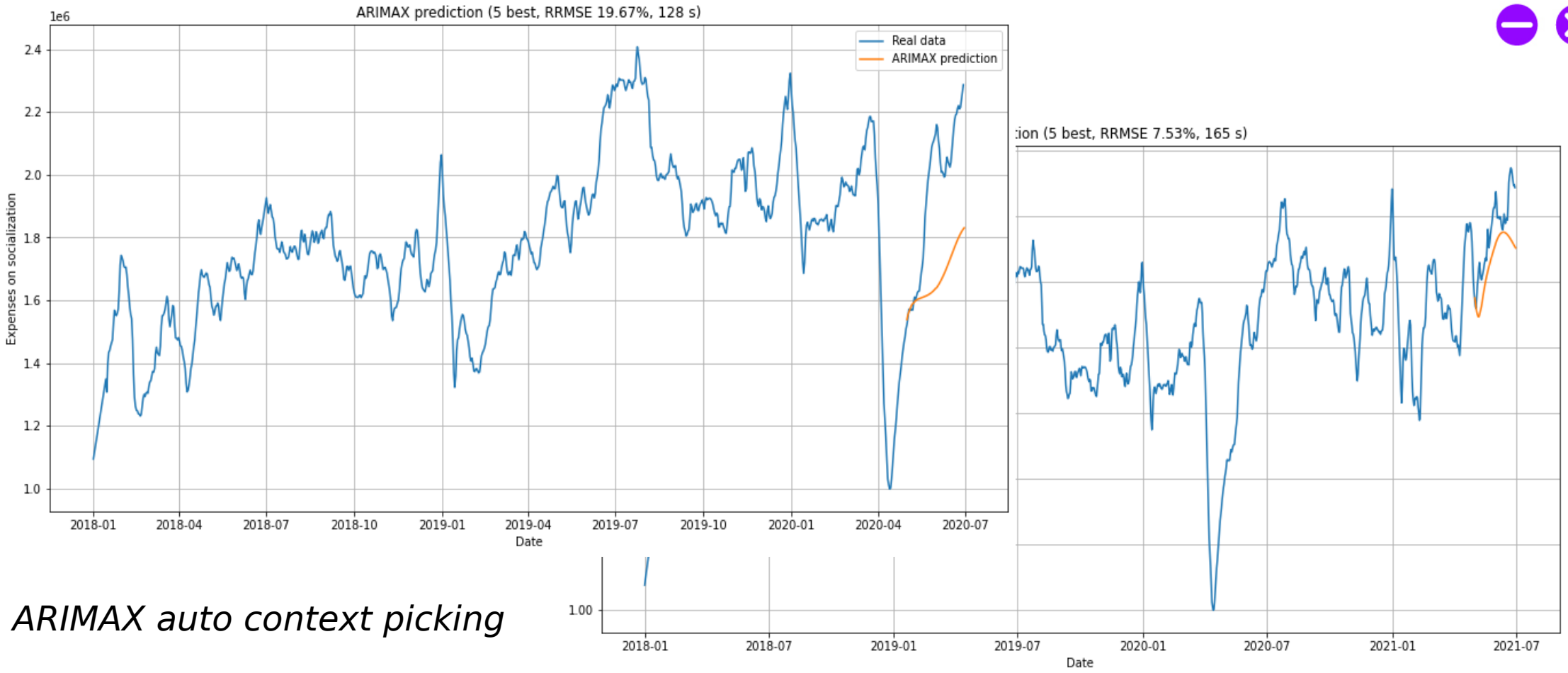
First two topics are negatively correlated with consumption with zero day lag.

Consumption and context cross-correlation

	food	travel	beauty	fun	health	kids
Unemployment	113, -0.81	**22, 0.94**	42, 0.95	26, 0.86	1, 0.83	0, 0.74
Covid-19 in Russia, total deaths	95, -0.80	0, 0.89	0, 0.95	**0, 0.87**	0, 0.73	0, 0.69
Key interest rate	103, 0.82	**18, -0.95**	36, -0.95	26, -0.88	0, -0.83	0, -0.73
COVID news	97, 0.81	**21, -0.93**	22, -0.91	18, -0.85	**0, -0.81**	0, -0.74
Travel news	94, -0.76	122, -0.66	84, 0.56	**35, 0.54**	35, 0.51	71, 0.65
Investment news	**39, 0.61**	117, 0.45	72, 0.61	116, 0.59	117, 0.65	93, -0.42
GDP, billions rub	83, -0.81	0, 0.82	0, 0.91	**0, 0.83**	**0, 0.69**	0, 0.66

Lag in days and Pearson cross-correlation

Context choice automation



ARIMAX auto context picking

Findings & conclusion

Findings:

- Mining and processing methods as well as ARIMAX models were implemented in Python3.
- Some consumption categories show dependency on context such as news. The best predictions are achieved when context is carefully selected.
- Time-lagged cross-correlation can be used to confirm or refuse causal hypotheses.
- Picking context can be automated.
- Autoregressive models seem to perform poorly on non-stationary data.

Conclusion:

- The Individual task was fully completed, although further research and implementation of non-autoregressive models is needed.

ИТМО

Thank you for attention!