Методы интеллектуального анализа данных: методология решения практических задач

Жуков Алексей Витальевич

м.н.с. ИСЗФ СО РАН

zhukovalex13@gmail.com

12.03.2019

План презентации

Задачи и цели семинара

Цель

Найти способы применения существующих эффективных решений на основе методов интеллектуального анализа данных из различных областей в текущих научных задачах.

Мотивация

- Существует множество решений подобных задач из иных областей (прогнозирование, компьютерное зрение и т.д.);
- Практические решения доступны для изучения (kaggle, github, etc);
- В рамках ИАД существует универсальная методологию решения задач;
- **Многие области науки не пользуются преимуществами ИАД.**

Что такое интеллектуальный анализ данных (ИАД)?

Интеллектуальный анализ данных (ИАД, Data Mining) - совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Решаемые задачи

- Классификация и ранжирование
- Восстановление регрессии и прогнозирование
- Кластеризация и сокращение размерности
- Поиск аномалий в данных
- Визуализация закономерностей
- ...

Классические задачи, но новые методы.

Индустриальная методология

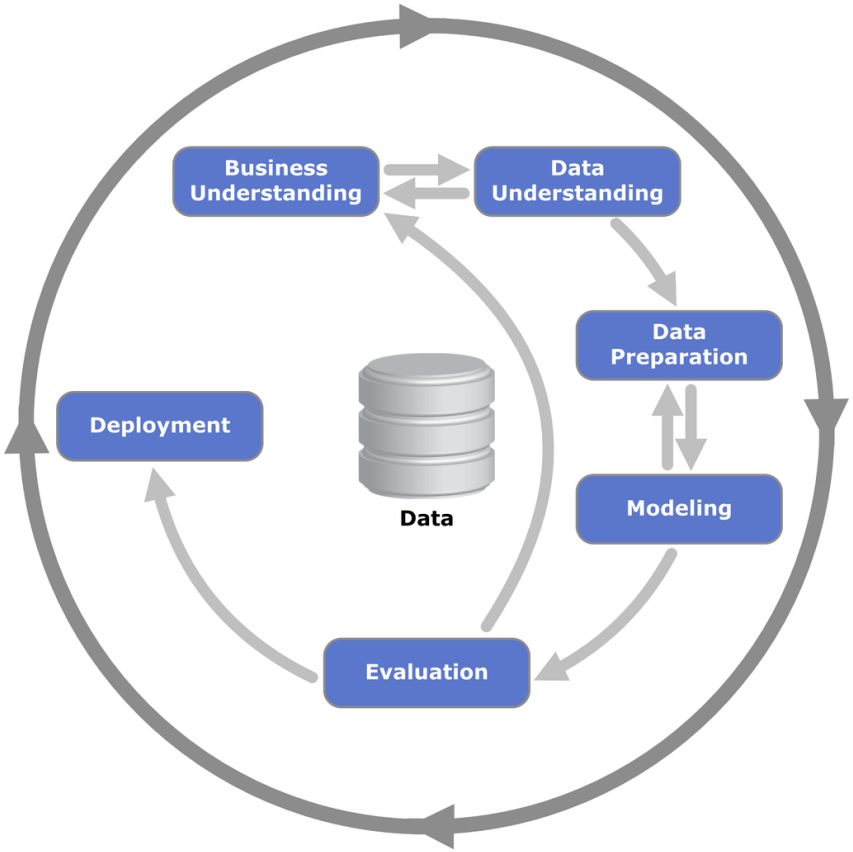


Figure: CRISP-DM (CRoss Industry Standard Process for Data Mining)

Постановка задачи и первичный анализ данных

Цель и постановка задачи

- Определить цель исследования
- Оценить ситуацию, существующие решения
- Определить цели анализа данных
- Составить план проекта

Сбор и исследование данных

- Собрать исходные данные
- Описать данные
- Исследовать данные (построить гипотезы о данных)
- Проверить качество данных (пропуски, выбросы)

Подготовка данных и моделирование

Подготовка данных

- Отобрать данные
- Очистить данные
- Сгенерировать производные данные
- Объединить данные
- Привести данные в нужный формат

Моделирование

- Выбрать методику моделирования
- Сделать тесты для модели
- Построить модель
- Оценить модель (метрика качества)

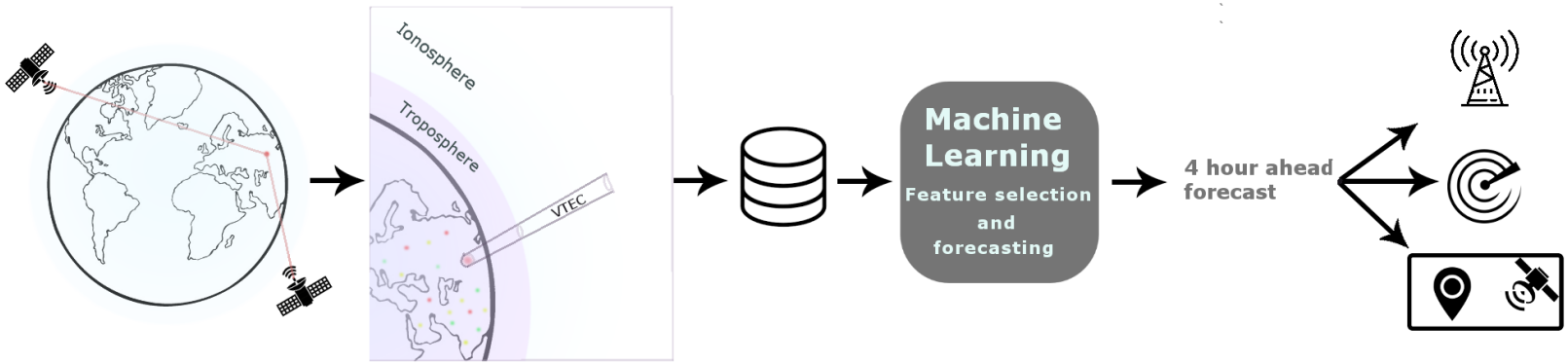
Оценка решения и внедрение/публикация результатов

Оценка решения

- Оценить результаты
- Сделать ревью процесса
- Определить следующие шаги

Внедрение/публикация результатов

- Запланировать развертывание
- Запланировать поддержку и мониторинг развернутого решения
- Сделать финальный отчет
- Сделать ревью проекта



Total electron content (TEC) is an important ionosphere parameter which can be used for ionosphere correction in radio systems.

Definition and Applications

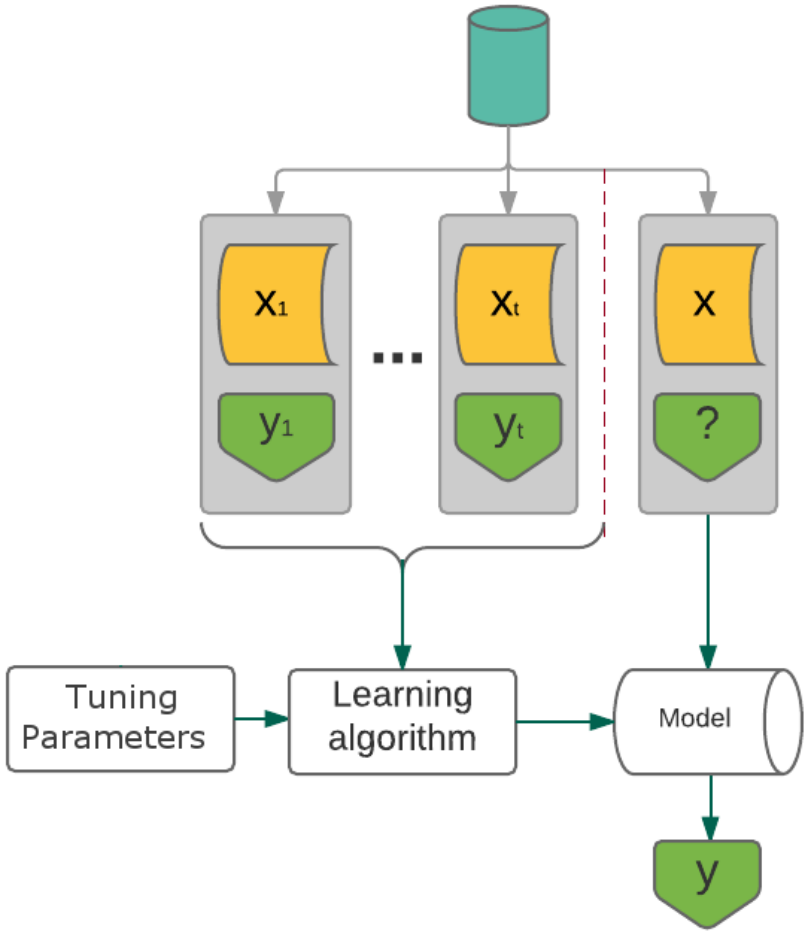
Machine learning provide generalized methodology for data processing which can be effectively applied to TEC nowcasting.

Machine learning problems

- classification
- **regression (nowcasting)**
- ...
- data discovery
  - **estimate variables relevance**
  - extraction of human-interpretable patters
  - data visualization
  - ...

Supervised learning problem for regression

Every x_i is a vector of input variables that measured or computed at specific time. Data input variables called **features**.



Neural networks for TEC nowcasting

Previous works

(Leandro, Santos, 2007) and (Habarulema, McKinnell and Opperman, 2011) used neural networks to analyze spatial and temporal variations of total electron content. The most popular ANN architecture for TEC nowcasting is Multilayer Perception (MLP).

General MLP shortcomings

- data normalization needed
- instability to data outliers (noise)
- many tuning parameters
- computationally expensive (in case of many layers)

Supervised machine learning methods

Random Forest (RF)

Ensemble machine learning method which uses randomized decision trees built at different part of training set.

Gradient Boosting Trees (GBM)

Ensemble machine learning method that construct ensemble in additive gradient-descent like manner.

Support Vector Machine (SVM) Regression

This method transform learning task to finding of the optimal approximation hyperplane in high dimension space (solving it in the form of a quadratic programming problem).

TEC nowcasting model construction

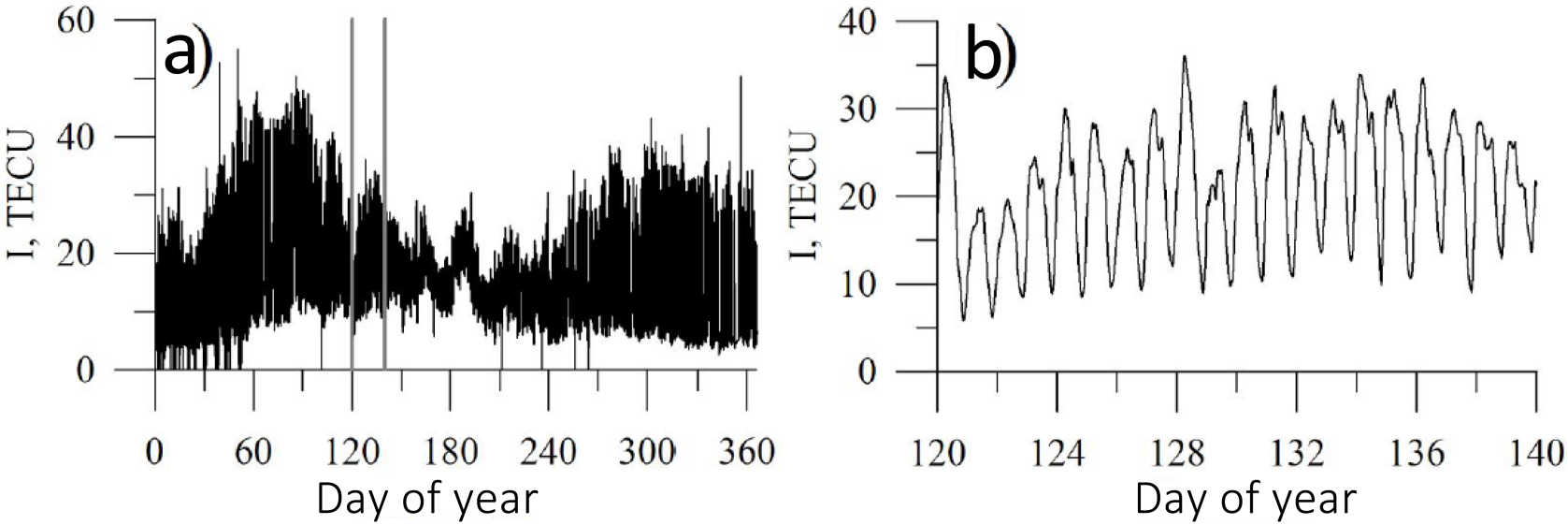
Goal

Development of an effective nowcasting model of absolute vertical total electronic content using machine learning (ML) techniques.

Tasks

- Build machine learning based nowcasting models for the absolute vertical total electronic content.
- Identify relative (important) input variables for building of nowcasting model
- Compare the obtained models with the naive models

Initial data



The time series include absolute vertical TEC which is computed on the basis of two-frequency joint phase and pseudorange measurements [1], its first and second time derivatives, the solar activity index F10.7, the geomagnetic activity indices SYM / H, AE. TEC and its derivatives were obtained using GPS/GLONASS station IRKJ (52 N, 104 E) data for 2014 with 30-min. resolution.

[1] Yasyukevich, Y. at al. 2015. Estimating the total electron content of the absolute value of the GPS / GLONASS data, Results in Physics Vol.5

Machine learning based nowcasting model building pipeline

- Dataset formation and preprocessing
- Feature extraction
- Feature selection
- Model building and performance assessment

Data formation

90% - used for model building and estimation of features importance.

10% — validation set
Final model quality check and visualization.

Feature extraction

Base input features

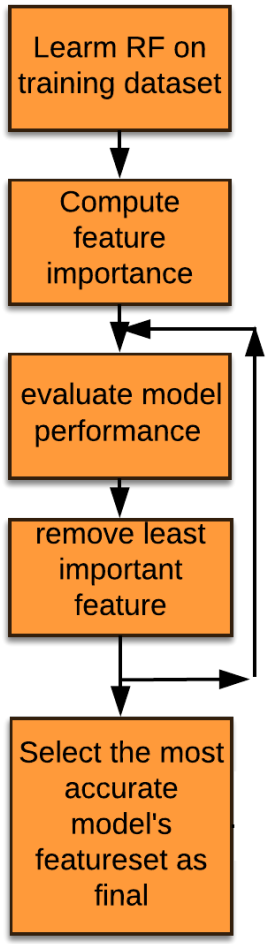
Feature type	Choosing technique	Used parameters
Lag	Correlation analysis	0.5, 12, 24, 48, 125, 360
Moving average	a priori knowledge	2, 3, 4, 12, 24, 48, 72, 96
Additional	a priori knowledge	

Additional features

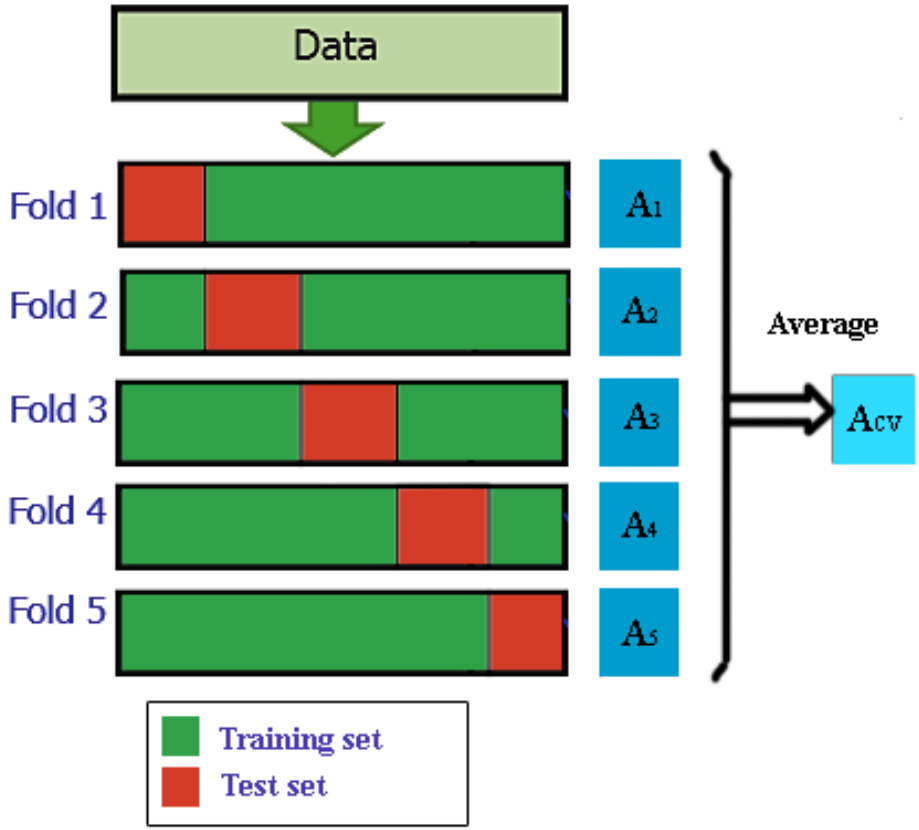
Based of information about periodical nature of considered time series, we added cosine from local time $(cos(2\pi * LT/24))$ to feature list.

Feature selection

We need to obtain optimal set of features (input variables) to reduce problem dimension and keep high accuracy.

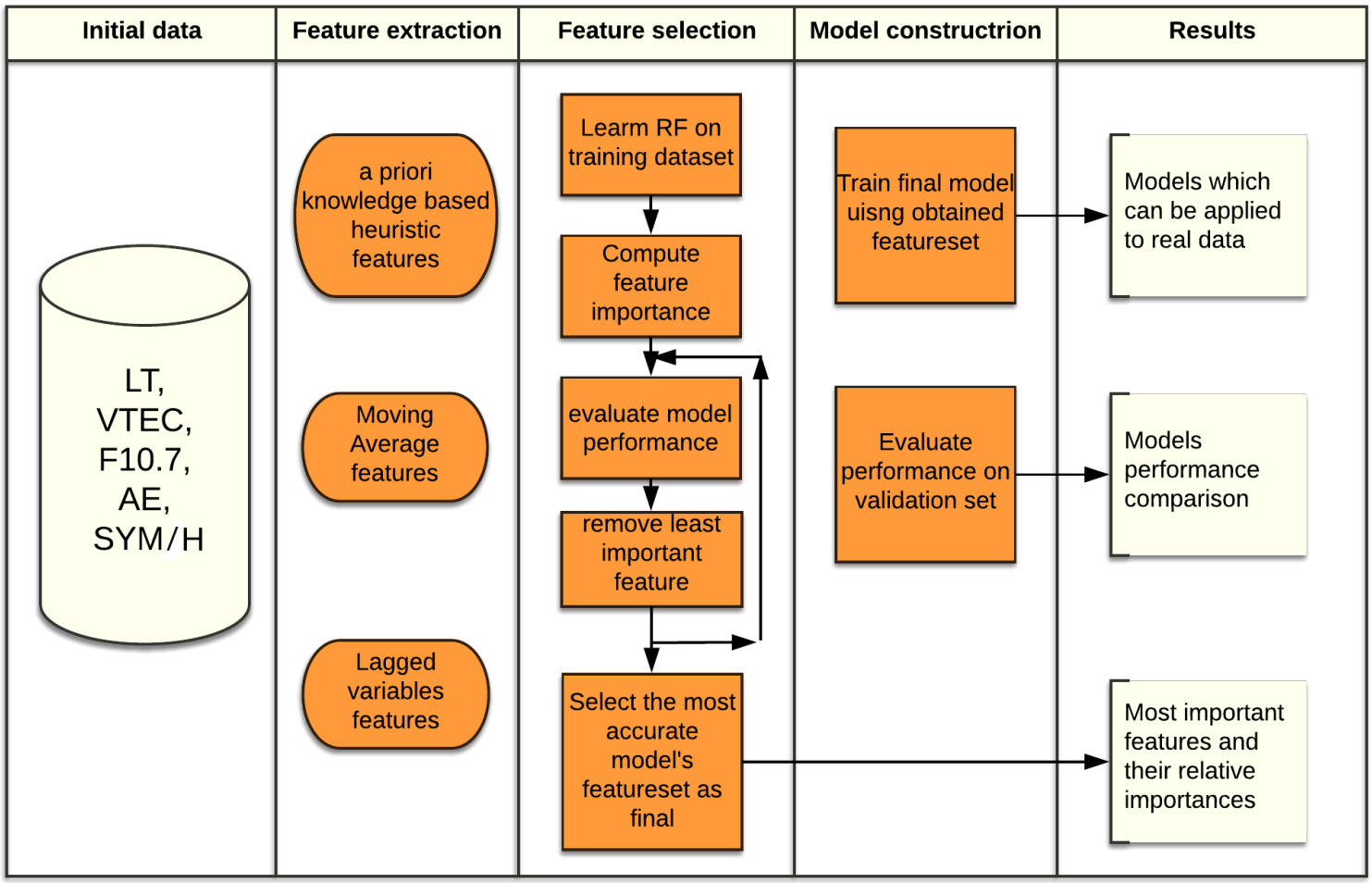


Model performance assessment



Cross-validation is a well-known ML method to obtain unbiased estimation of model error (generalization error).

Model building process diagram



Feature importance

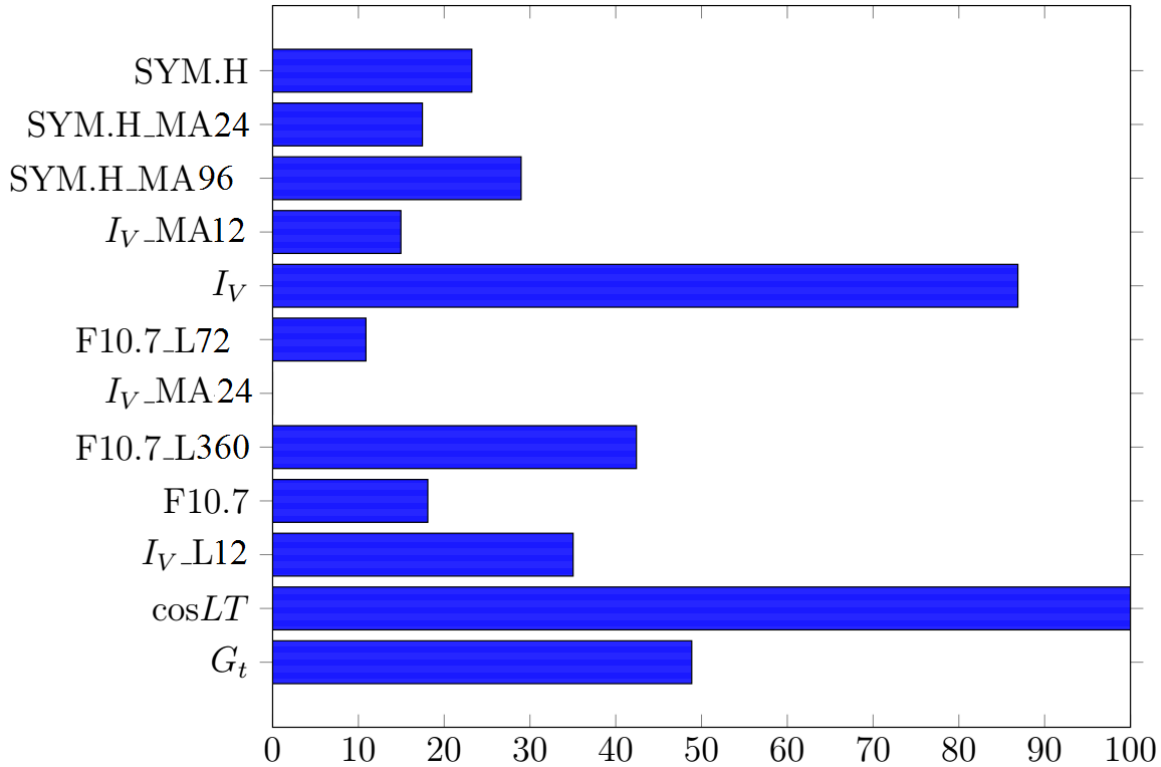


Figure: Relative features importance provided by Random Forest model

Results for naive models

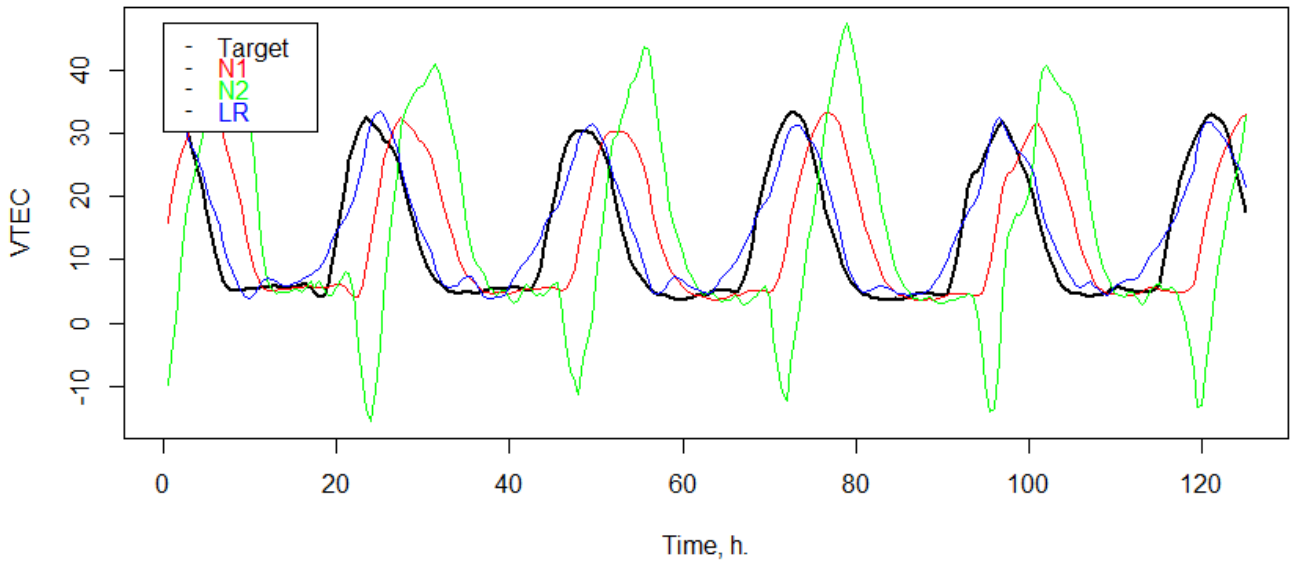


Figure: Naive and linear model's predictions for 4 hours ahead prediction.

N1 – first naive model: $I(t + \Delta h) = I(t)$
N2 – second naive model: $I(t + \Delta t) = I(t) + dI/dt \cdot \Delta t$

Nowcasting models based on RF, GB and SVM

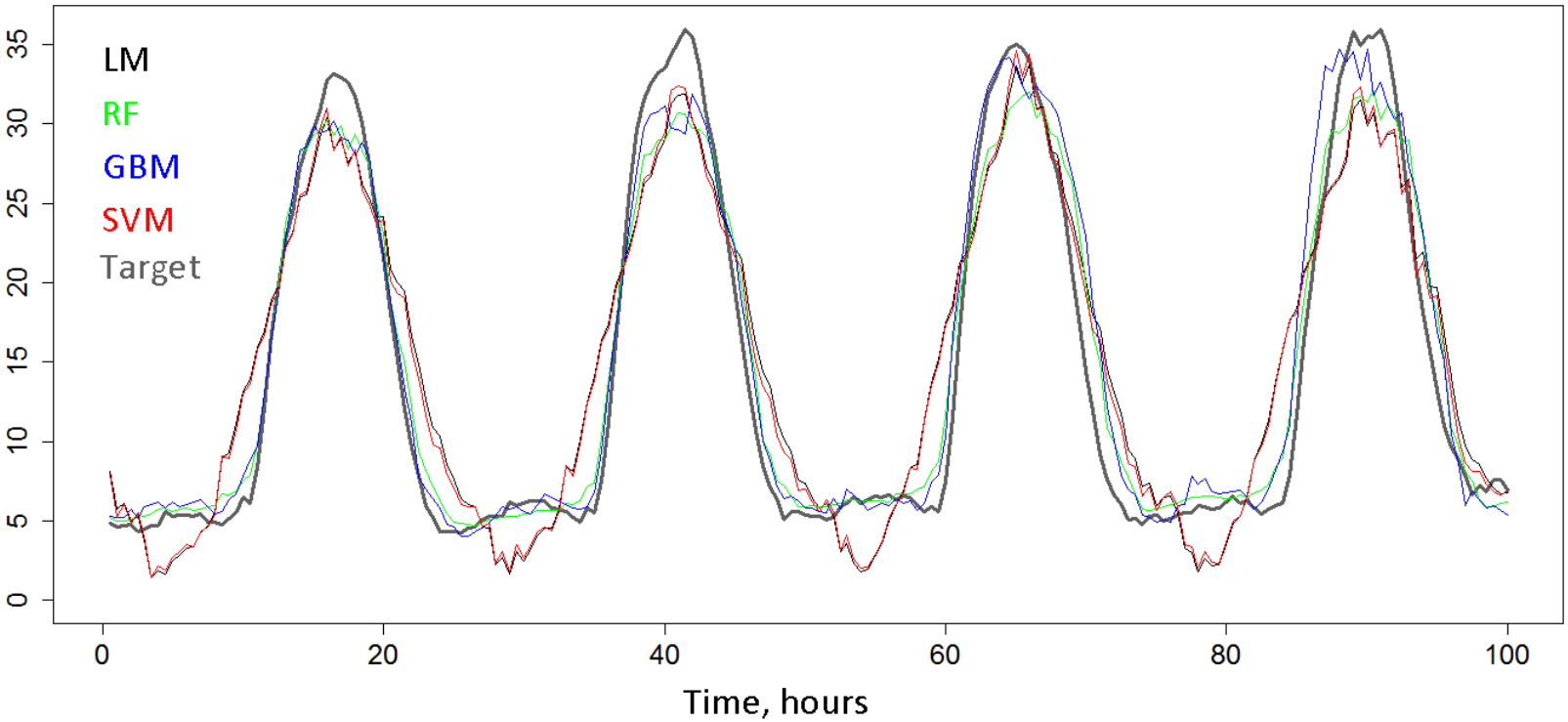


Figure: 4 hours ahead prediction for RF, GB and SVM

Results

	RF	GBM	SVM	N1	N2	LM
RMSE	3.49	3.30	4.49	9.14	16.51	4.61
MAE	2.49	2.35	3.50	6.74	12.23	3.66

Table: Errors for 4 hour ahead nowcasting models.

N1 – first naive model: $I(t + \Delta h) = I(t)$
N2 – second naive model: $I(t + \Delta t) = I(t) + dI/dt \cdot \Delta t$
LM – linear multi-parameter regression

Prediction accuracy for different periods

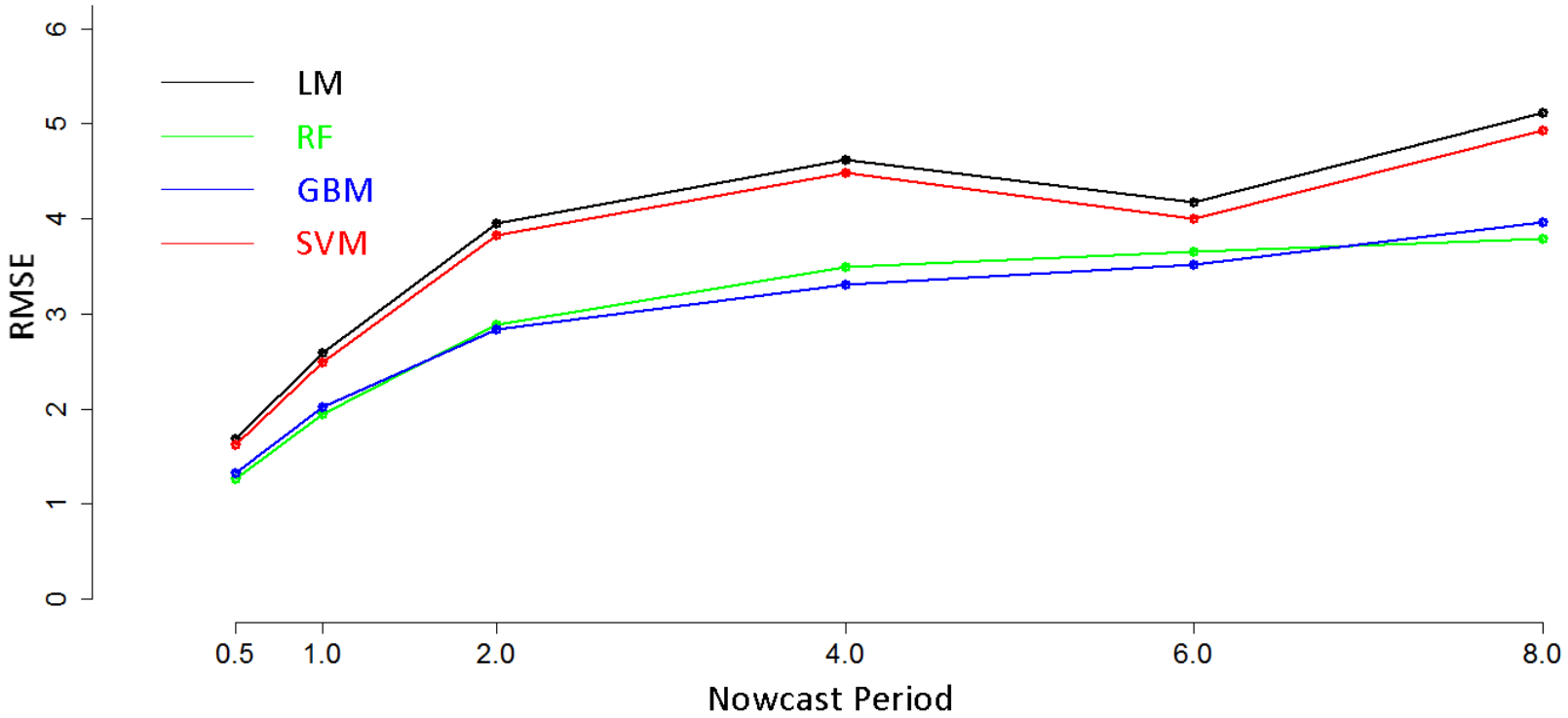


Figure: RMSE for different periods (in hours) and predictive models (RF, GB and SVM)

Results and conclusion

- Selected relevant input variables are current TEC value, first derivative of TEC, local time cosine, F10.7 and SYM / H, moving averages of TEC with periods of 12 and 24 hours and SYM/H with period 24 and 96 hours, as well as previously received data with lag, such as the TEC value with a delay of 12 hours, F10.7 with a delay of 3 and 15 days.
- A linear multi-parameter regression model have quite good results ($\sim$ 4.5 TECU error)
- Machine learning based models allowed to **significantly reduce** RMSE ($\sim$ 3.5 TECU)

Review and Discussion

Approach limitations and ways to overcome

- VTEC is a non-stationary data.
  - Offline model could be periodically updated (how often?).
  - Online models can be used.
- We need large data to get more robust model and discover the process more preciously.
- Do we have enough information provided by used features? (maybe automatic variable construction by deep learning can help)

Сочетаемость с традиционными подходами

Как учесть традиционные методы?

Возможно объединение разнородных традиционных моделей для получения лучшего качества. Сделать это можно на этапе моделирования или подготовки данных.

Необходимость в глубоком понимании задачи

Пример. Соревнования по анализу данных, задача рекомендации друзей в социальных сетях (SNA хакатон от Mail.ru). Десятки участников и множество современных сложных моделей машинного обучения.
Победившая модель:

$$w(x) = \frac{111}{\sqrt[3]{x + 50}} + 8 \tag{1}$$

Резюме

Итого

- Готовые решения ИАД могут быть успешно адаптированы к научным задачам;
- Применение универсальной методологии решения задач ИАД в сочетании с глубоким пониманием проблемы может быть крайне эффективно.

Вопросы

zhukovalex13@gmail.com