Проект по дисциплине "Методы искусственного интеллекта в анализе данных"

Этап 2

Бобровских Глеб, Иванов Дмитрий, Угадяров Леонид
https://github.com/ugadiarov-la-phystech-edu/aimda-project

Группа 4

16 декабря 2020 г.

Набор данных и постановка задачи

- Рассматривается подвыборка за 2014 год из набора данных об убийствах в США Homicide Reports, 1980-2014 — https://www.kaggle.com/murderaccountability/homicide-reports
- Задача прогнозирования временных рядов — предсказание значений временного ряда (число убийств в США) для 24 месяцев (2013-ый и 2014-ые года) на основе исторических данных
- Метрика качества — RMSE для значений временного ряда
- Актуальность — возможность предсказывать количество преступлений в стране на будущий(ие) год(а) кажется вполне очевидно актуальной, поскольку качественное решение позволит планировать затраты на полицейские учреждения

Признаки

Number of cases - количество убийств на всей территории США, распределенное во времени

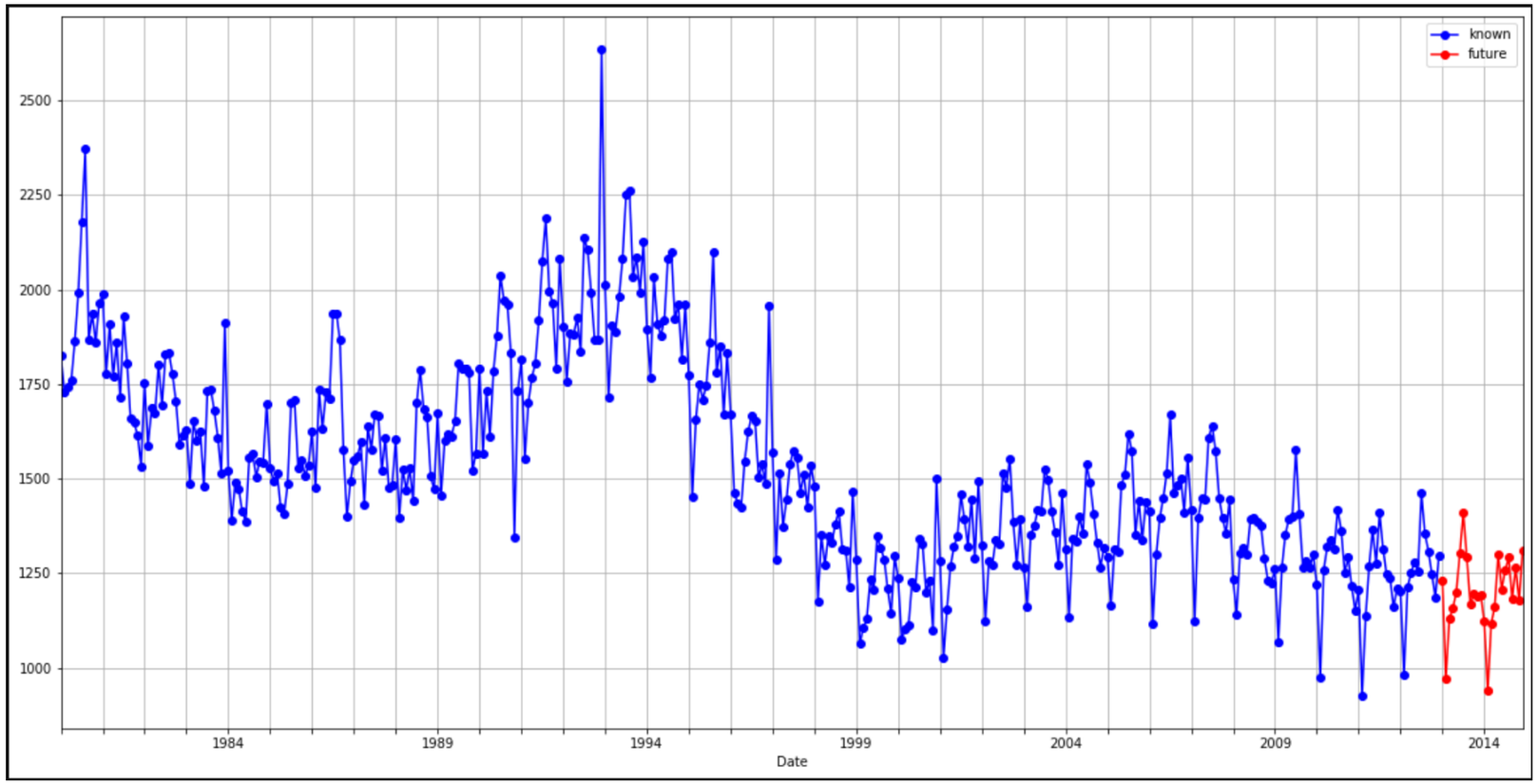
Количество объектов: 420 (месяцев)

Вклад:

- Бобровских Глеб — методы Linear Regression, AutoRegression, ARIMAX
- Иванов Дмитрий — нейронные сети
- Угадяров Леонид — методы ARIMA, SARIMA, ARIMAX

Иллюстрация данных

Временной ряд: (Ox - количество преступлений, Oy - даты в месяцах)



Также использовался временной ряд с логарифмированными значениями числа преступлений.

Применённые модели

- Линейная регрессия – sklearn.linear_model, обучалась на количестве дней, расчитаных от начальной даты;
- Авто регрессия – statsmodels.tsa.ar_model.AutoReg, параметры lags = 12;
- ARIMA – statsmodels.tsa.arima_model.ARIMA, параметры order = $12, 0, 11$, предсказания строились пошагово для дифференцированного шага с параметром lag = 1;
- SARIMA – statsmodels.tsa.statespace.sarimax.SARIMAX, параметры order = (11,1,11), seasonal_order=(12,1,11,12), exog=None;
- ARIMAX – statsmodels.tsa.statespace.sarimax.SARIMAX, order = (1,1,12), seasonal_order = (0,0,0,0), exog = {one-hot признаки: месяц, день недели; бинарный признак: выходной день};
- Полносвязная нейросеть (FCN) – pytorch.nn.Linear, использовалось различное количество признаков на скрытом слое, в результатах указаны метрики для window_size = 24, n_hidden = 1024;
- Реккурентная нейросеть (RNN) – pytorch.nn.RNN, n_hidden = 1024, window_size = 24;
- LSTM – pytorch.nn.LSTM, n_hidden = 1024, window_size = 24;

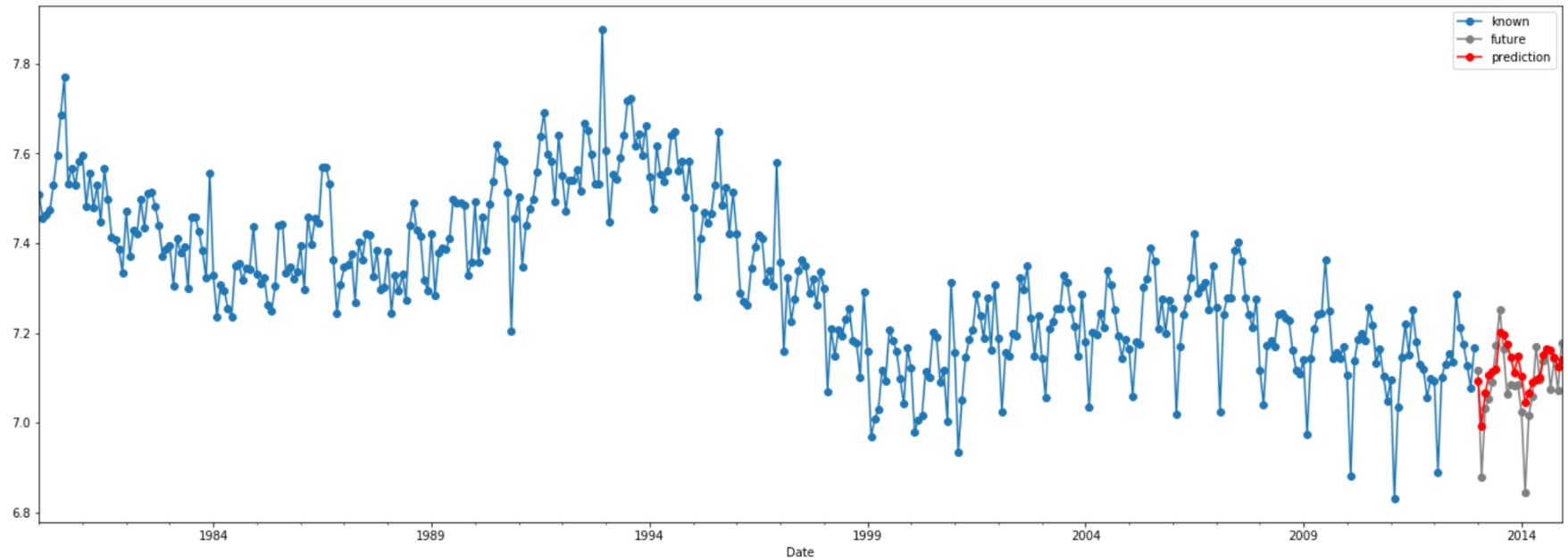
Результаты экспериментов

Таблица: Метрики качества классификации обученных моделей на Test

	RMSE	Log Data RMSE	Time Consumption Train (s)	Time Consumption Pred (s)
Linear Regression	1196.16	0.09	0.001	0.0004
AutoRegression	83.963	0.069	0.005	0.0015
ARIMA	184.55	0.144	99.447	0.0075
SARIMA	184.55	0.076	562.12	0.056
ARIMAX	92.809	0.08	5.88	0.019
FCN	109.66	0.08	0.422 (200 эпох)	–
RNN	273.82	0.175	16.357 (2000 эпохи)	–
LSTM	273.81	–	43.974 (2000 эпохи)	–

Спецификации рабочих машин:

- Измерение времени для линейной регрессии и прочих классических методов
  Процессор: 2,3 GHz Dual-Core Intel Core i5, Память: 8 GB 2133 MHz LPDDR3
- Измерения времени для нейронных сетей
  Процессор: Intel Core i-7-8750H CPU @ 2.20GHz, Память: 16834 Mb RAM, Видеокарта: NVIDIA GeForce RTX 2060



Auto Regression

Спасибо за внимание!

https://github.com/ugadiarov-la-phystech-edu/aimda-project