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

25 декабря 2020 г.

Набор данных и постановка задачи

- Рассматривается временной ряд количества убийств в США, совершённых за месяц, полученный агрегацией исходного набора данных Homicide Reports (1980-2014): https://www.kaggle.com/murderaccountability/homicide-reports
- Решается задача прогнозирования количества убийств на следующий месяц по историческим данным
- Метрика качества — RMSE для значений временного ряда
- Актуальность — качественное прогнозирование количества преступлений поможет эффективнее организовать работу полицейских учереждений

Целевая переменная и признаки

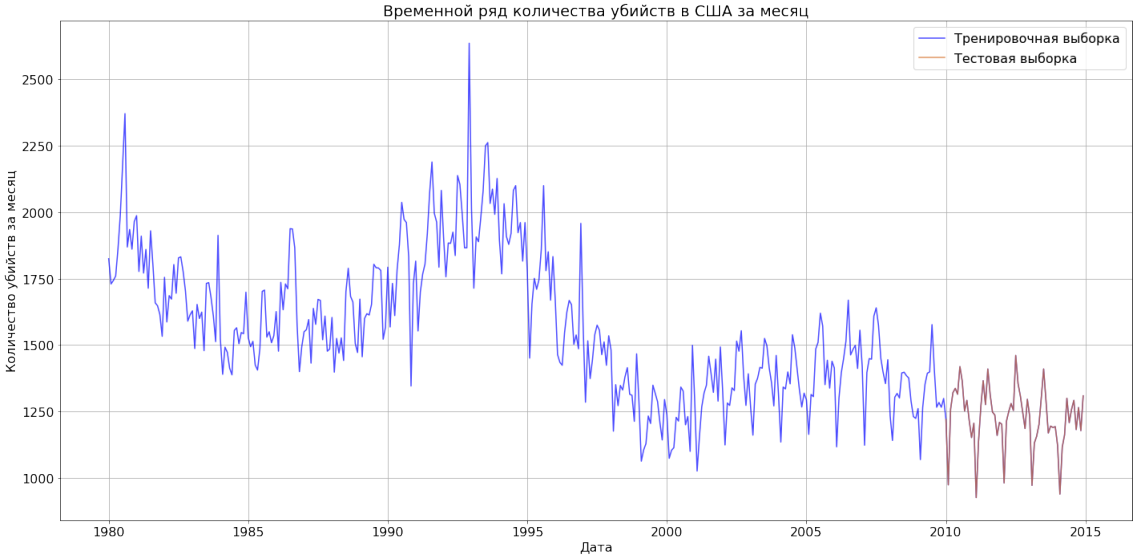
Целевая переменная: NumCases — количество убийств на всей территории США за один месяц
Признаки: значения NumCases за предыдущие 12 месяцев и значения 117 агрегированных признаков исходного набора данных за предыдущие 12 месяцев (всего 1416 признаков)
Также производились эксперименты с использованием данных за предыдущие 24 месяца (2832 признака)

Количество объектов: 420 (месяцев)
Тестовая выборка: 60 (месяцев)

Вклад:

- Бобровских Глеб — ARIMA, ARIMAX
- Иванов Дмитрий — RNN, LSTM
- Угадяров Леонид — ElasticNet, MLP

Иллюстрация данных



Применённые модели

ElasticNet – sklearn.linear_model.ElasticNet

Подбор гиперпараметров с помощью sklearn.linear_model.ElasticNetCV:
l1_ratio=[.01, .1, .5, .7, .9, .95, .99, .999, 1], n_alphas=1000, max_iter=10000.
Лучшие значения для прогноза по 12 месяцам: l1_ratio=0.1, alpha=3077.72.
Лучшие значения для прогноза по 24 месяцам: l1_ratio=1, alpha=56.68.

MLP – фреймворк PyTorch

Рассматривались архитектуры до четырёх полносвязных слоёв с батч-нормализацией и дропаутом. Лучшее качество достигнуто при использовании двухслойной архитектуры без батч-нормализации и дропаута.
Количество нейронов нейронов скрытых слоёв: 12, 24, 32, 64, 96.
Лучшее значение: 32. Функция активации: ReLU.

RNN и LSTM – фреймворк PyTorch

Размерность скрытого слоя: 16, 32, 64, 128, 256, 1024.
Лучшие значения для прогноза по 12 месяцам: 64. Лучшие значения для прогноза по 24 месяцам: 1024.

ARIMA и ARIMAX – пакет statmodels

Произведено дифференцирование исходного ряда. Стационарность ряда проверена критерием Дики-Фуллера. Для выбора лучших параметров использовался критерий Акаике.
ARIMA — order=(24, 1, 12). ARIMAX — order=(12, 1, 0), seasonal_order=(0, 0, 0, 0).

Результаты экспериментов

+features — модель использует агрегированные признаки исходного набора данных

Таблица: Метрики качества обученных моделей ElasticNet, MLP, LSTM на тестовой выборке

	RMSE	Время обучения, с	Время предсказания, с
ElasticNet+features (12 месяцев)	74.17	0.15	0.00012
MLP (12 месяцев)	70.16	32.1 (2092 эпохи)	0.0002
LSTM (12 месяцев)	76.45	44.1	0.001
ElasticNet (24 месяца)	61.02	0.0016	0.00012
MLP (24 месяца)	63.47	19.6 (1203 эпохи)	0.0002
LSTM (24 месяцев)	90.01	114.5	0.002

Таблица: Метрики качества обученных моделей ARIMA и ARIMAX на тестовой выборке

	RMSE	Время обучения, с	Время предсказания, с
ARIMA	64.92	2068.3	0.016
ARIMAX	62.15	1.37	0.0076

Спецификации рабочих машин:

- Измерение времени для ElasticNet и MLP: Intel Xeon E5-2699 v4 @ 2.20 ГГц, 12ГБ ОЗУ
- Измерения времени для LSTM и RNN: Intel Core i-7-8750H @ 2.20ГГц, 16 ГБ ОЗУ, NVIDIA GeForce RTX 2060
- Измерения времени для ARIMA и ARIMAX: Dual-Core Intel Core i5 @ 2.30ГГц, 8 ГБ ОЗУ

Иллюстрация предсказаний лучшей модели ElasticNet

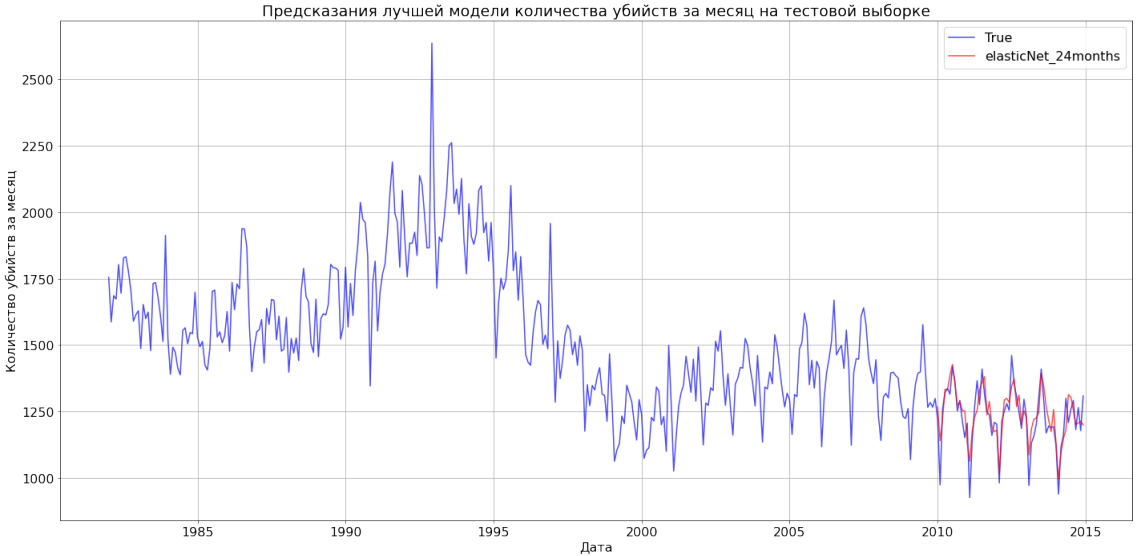


Иллюстрация предсказаний лучшей модели ARIMAX