МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский технологический университет «МИСИС»
СТАРООСКОЛЬСКИЙ ТЕХНОЛОГИЧЕСКИЙ ИНСТИТУТ ИМ. А.А. УГАРОВА
(филиал) федерального государственного автономного образовательного учреждения высшего образования
«Национальный исследовательский технологический университет «МИСИС»
(СТИ НИТУ «МИСИС»)

ФАКУЛЬТЕТ АВТОМАТИЗАЦИИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ КАФЕДРА АВТОМАТИЗИРОВАННЫХ И ИНФОРМАЦИОННЫХ СИСТЕМ УПРАВЛЕНИЯ ИМ. Ю.И. ЕРЕМЕНКО

Домашняя работа №1
по дисциплине: «Python для анализа данных»

Выполнил студент группы: АТ/МС-23Д, Небольсин Василий Дмитриевич
группа, ФИО полностью подпись

Проверил:
доцент, к.т.н., доцент кафедры АИСУ, Цыганков Юрий Александрович
Должность, звание, ФИО полностью подпись

Старый Оскол, 2023

1. Задание

Провести исследовательский анализ данных (EDA) предварительно очистив данные от нулевых значений, выбросов, также избыточных или идентичных строк данных. предобработать данные перед подачей на алгоритм машинного обучения. Протестировать модели регрессионного анализа и модели нейронных сетей в задаче прогнозирования временного ряда.

2. Введение в подготовку данных

Значительная часть любого проекта, связанного с данными, связана с предварительной обработкой данных, и ученые, работающие с данными, тратят около 80% своего времени на подготовку данных и управление ими. Предварительная обработка данных — это метод анализа, фильтрации, преобразования и кодирования данных, позволяющий алгоритму машинного обучения понимать обработанные выходные данные и работать с ними.

Проект обработки данных может быть успешным только в том случае, если данные, поступающие в машины, будут высокого качества. В данных, извлеченных из реальных сценариев, всегда есть шум и пропущенные значения. Это происходит из-за ошибок вручную, непредвиденных событий, технических проблем или множества других препятствий. Неполные и зашумленные данные не могут использоваться алгоритмами, поскольку они обычно не предназначены для обработки пропущенных значений, а шум нарушает истинную структуру выборки. Предварительная обработка данных направлена на решение этих проблем.

3. Цель предварительной обработки данных

После того, как правильно собраны данные, их необходимо изучить или оценить, чтобы выявить ключевые тенденции и несоответствия. Основными целями оценки качества данных являются:

- Обзор данных: общая структура и формат. Кроме того, просмотр статистики данных, как среднее значение, медиана, стандартные квантили и стандартное отклонение. Эти детали могут помочь выявить нарушения в данных.
- Определить недостающие данные: Встречаются в большинстве реальных наборов данных. Это может нарушить истинную структуру

и даже привести к еще большей потере данных, когда целые строки и столбцы удаляются из-за отсутствия нескольких ячеек в наборе.

- Выявить выбросы или аномальные данные: некоторые точки данных далеко выходят за рамки. Эти точки являются выбросами, и их, возможно, придется отбросить, чтобы получить прогнозы с более высокой точностью, если только основной целью алгоритма не является обнаружение аномалий.

- Удалите несоответствия: как и отсутствующие значения, реальные данные также содержат множество несоответствий, таких как неправильное написание, неправильно заполненные столбцы и строки, дублированные данные и многое другое. Иногда эти несоответствия можно устранить с помощью применения скриптов, но чаще всего они требуют ручной проверки.

После предварительной обработки данных и разделения их на обучающие/тестовые наборы переходим к моделированию. Модели — это не что иное, как наборы четко определенных методов, называемых алгоритмами, которые используют предварительно обработанные данные для изучения закономерностей, которые позже можно использовать для прогнозирования. Существуют различные типы алгоритмов обучения, включая контролируемое, полуконтролируемое, неконтролируемое и обучение с подкреплением.

Оценка модели. На этом этапе модели оцениваются с помощью конкретных показателей производительности. На их основе проводится оптимизация гиперпараметров для получения наилучшей модели.

Прогноз. Как получены результаты этапа оценки, мы переходим к прогнозам. Прогнозы делаются обученной моделью, когда она подвергается воздействию нового набора данных.

4. Результаты

В ходе проведения комплексного анлиза временных рядов получены следующие результаты (таблицы 1-2)

Table 1: Результаты тестирования реграссионных моделей

Модель	***LinearRegression***	*Lasso*	*Ridge*	*ElasticNet*
MSE	***111.75***	194.58	163.82	186.95
R2	***-1.01***	-2.51	-1.95	-2.37

Table 2: Результаты тестирования полносвязных моделей

Модель	1.0	2.0	2.1	2.2	2.3	$2FIN$	3.0
MSE	900.62	2218.01	324.78	374.66	***33.29***	81.27	45.88
R2	-15.23	-38.97	-4.85	-5.75	***0.4***	-0.46	0.17

Наилучший результат показала последовательная модель с Dense слоями и линейной функцией активации порядка 33.29 MSE. Структура из четырех полносвязных слоев с применением Dropout вероятностью 5% и Batchnormalization является хорошей связкой за счет предотвращения сложных коадаптаций отдельных нейронов на тренировочных данных и повышения производительности ИНС на этапе обучения. Также замечена зависимость между положением связки Dropout и Batchnormalization с результатами прогнозирования модели (таблица 2).

Замечено, что модели 2FIN и 3 могут быть улучшены как минимум на 10% посредством уменьшения размера пакета на этапе обучения.