**Московский авиационный институт**

**(Национальный исследовательский университет)**

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Искусственный интеллект»

**Лабораторная работа № 0**

Тема: Получение и обработка данных.

Студент: Чекушкин Д.И

Группа: М80-304Б

Москва, 2019

1. Постановка задачи

Требуется сформировать/получить два набора данных соответствующие следующим критериям:

Один из датасетов должен представлять собой корпус документов. Язык, источник и тематика произвольна

Второй датасет должен содержать категориальные, количественные признаки. Для данного датасета определить предсказываемые признаки (для задачи регрессии и классфикации). Если такого признака нет, спроектировать

По каждому датасету построить распределения признаков (в случае корпуса документов – построить распределение слов) и объяснить имеющуюся картину. Вычислить статистические характеристики признаков. Обнаружить и решить возможные проблемы с данными. Если решить данную проблему невозможно, объяснить почему.

1. Решения задачи

ЯП: python 3.5.2

Библиотеки:

matplotlib - библиотека для визуализации графики

numpy - высокоуровневые математические функции

pandas - обработка и анализ данных

statistics - статистические функции

fix\_yahoo\_finance - работа с API Yahoo - finance

1. Руководство по использованию программы

Файл ai1.py содержит код программы, которая обрабатывает данные из файла BLK2.csv, строит распределения признаков и вычисляет статистические характеристики признаков(цены акций на момент закрытия).

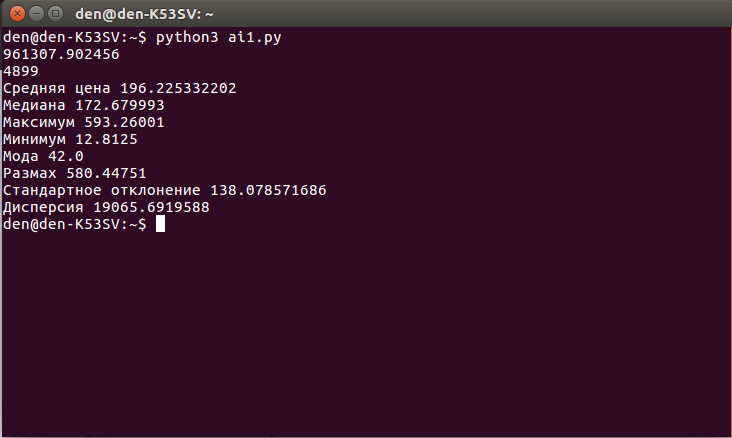


Рис. 1 - статистические характеристики цены закрытия

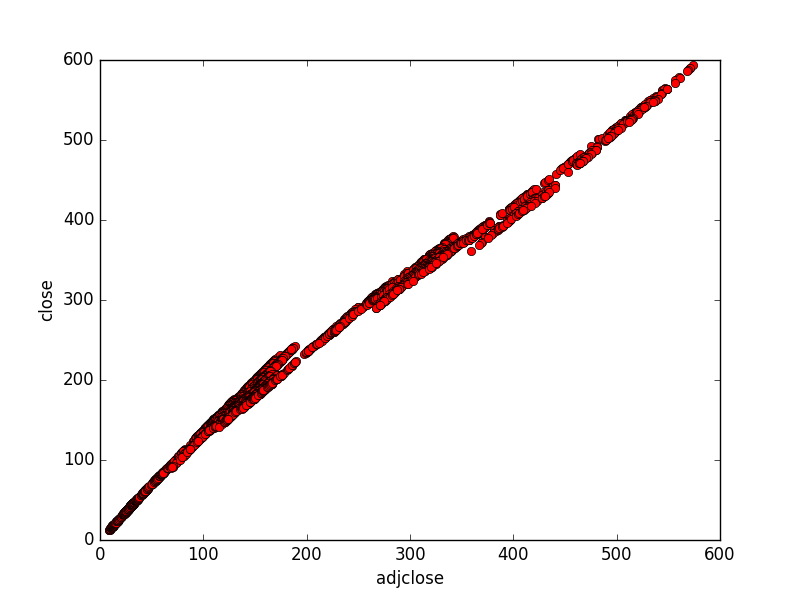


Рис. 2 Распределение close - adjusted close price

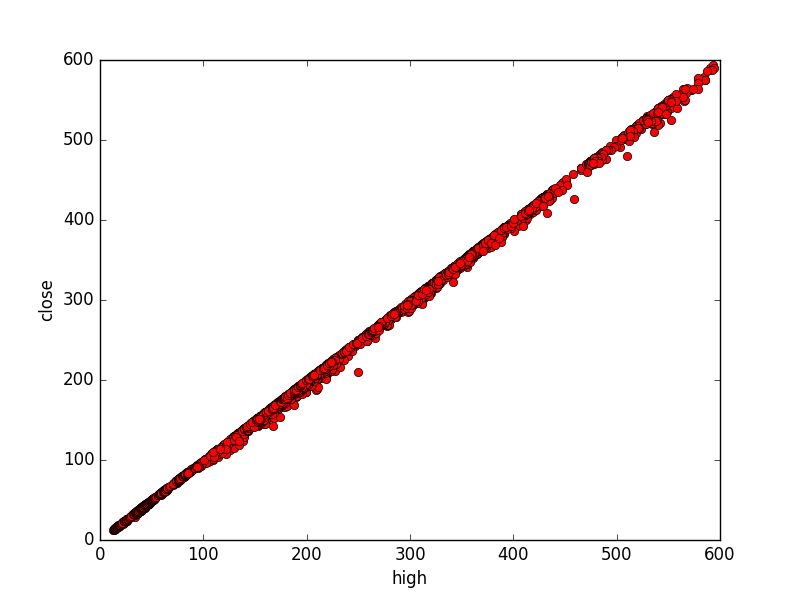


Рис 3. Распределение close - high price

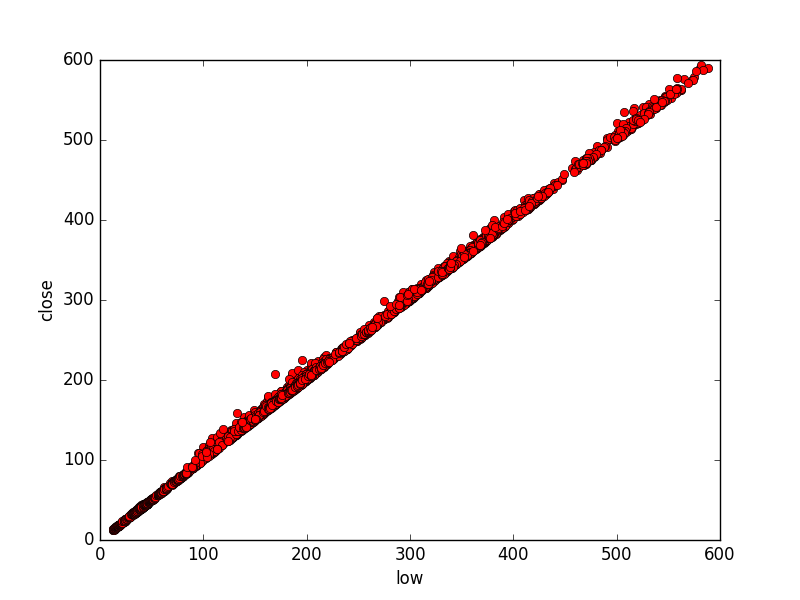


Рис. 4 Распределение close - low price

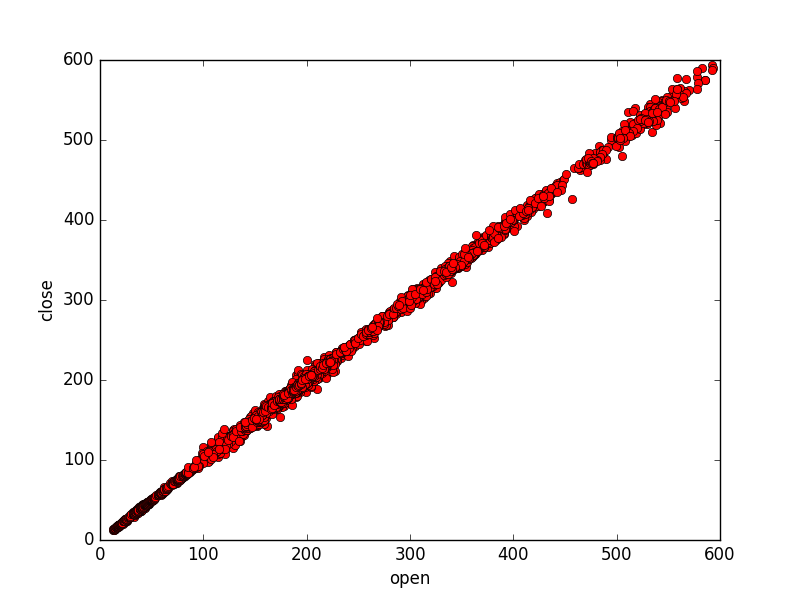


Рис. 5 Распределение close open price

Датасет 1:

История акций компании Blackrock, с момента выхода на биржу

https://finance.yahoo.com/quote/BLK/history?period1=941403600&period2=1553288400&interval=1d&filter=history&frequency=1d

Предсказываемые признаки: цена на момент закрытия торгов, максимальная/минимальная цена в торговый день.

Распределение признаков:

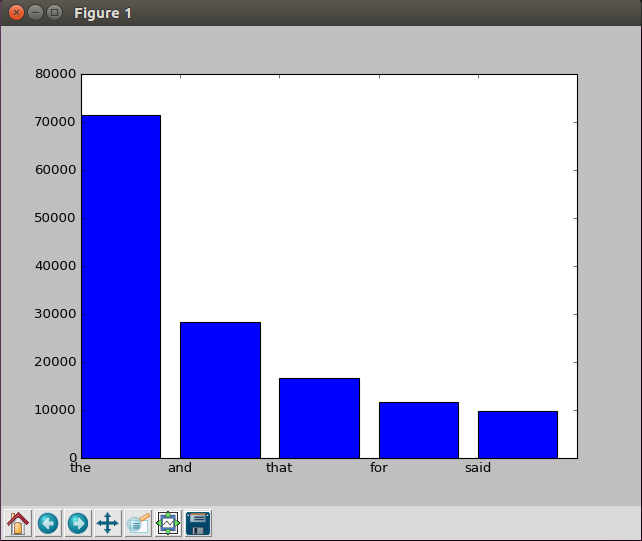
- "Цена на момент закрытия - скоректированная цена" практически идентичны, тк вторая отличается(не всегда) от первой стоимостью выплаченных дивидентов.

- "Цена закрытия - максимальная цена за торговый день". Цена на момент закрытия редко имеет сильное отличие от максимальной цены за день. Все выбросы это исключения из правил(падение рынка или локальные неудачи компании).

График "Цена закрытия - минимальная цена за торговый день". Представляет из себя примерно тоже самое, только с редкими выбросами вверх(резкий рост рынка или локальный успех).

- "Цена закрытия - цена открытия". Более «усредненная» версия предыдущих двух графиков.

Файл ai2.py содержит код программы, которая обрабатывает второй датасет — 1000 новостей(news.txt). И выводит гистограмму распределения слов.

 Рис. 6 диаграмма распределения слов

Ссылка: https://www.kaggle.com/snapcrack/all-the-news

- Больше всего артиклей, предлогов, местоимений и слова said. Во многих языках самыми употребляемыми являются именно предлоги и местоимения.

Вывод: благодаря проделанной работе, мне удалось освоить простую обработку информации, узнать об источниках датасетов для ml и вспомнить о том, что у меня есть гитхаб.

Ссылка на репозиторий github: https://github.com/israelcode/mai\_ai\_2019