**Отчет по проекту**

**Липатов Данила**

**Цикл I**

В течение первого цикла проекта были выполнены следующие шаги:

1. Написание скрипта для парсинга данных (репозиторий кредитных рейтингов)
2. Написание скрипта для решения капч при парсинге
3. Обучение CNN модели (сверточные нейронные сети, для классификации символов в капче, успех ~ 93%)
4. Парсинг данных
5. Парсинг данных (открытый источник) более ранних дат (до 2016 года)
6. Реализация когортного метода

Для каждого из пунктов более детальное описание:

Пункт 1:

Для написания скрипта использовался Python с библиотекой selenium.

Основные аспекты при парсинге были следующие: Возможность распарсить как общим целым, так и по отдельности (интересующих компаний).

Так как рейтинги могут пересматриваться, данный парсер работает от таймера с запуском в определенное время. Все данные автоматически подливаются в БД (база данных)

Пункт 2:

При «хождении» по репозиторию на каждый запрос необходимо решать captcha (капчу). Чтобы не пользоваться данной услугой платно (rucaptcha), было принято решение обучить CNN модель верно решать эти капчи. Для этого понадобилась небольшая реализация класса и сбор данных.

Скрипт писался так же на Python с библиотеками tenserflow / keras для обучения.

Пункт 3:

Данные так или иначе собирались через решатель каптч и генерацией рандомных каптч, схожих с теми, что были при запросах. В общей сложности, обучающая выборка составила 3591 каптчу на момент первого обучения. Результат первой версии модели составил 73,2%. Далее, по мере использования парсера (на каждый запрос) модель переобучалась и корректировала результат верных решений (лучший результат ~93%)

Пункт 4:

После того, как предыдущие шаги были выполнены и результаты парсинга были получены, сразу была произведена описательная статистика, то есть:

- кол-во собранных данных

- распределение по типам данных (компании / облигации)

- ко-во валидных данных (данные без учета не интерпретируемых показателей)

- колебания показателей в каждом агенстве

- обогащение данных за счет более старых показателей (в частности дефолты)

На основе этих результатов можно было примерно представить себе, с чем можно дальше работать.

Пункт 5:

Данный пункт был уже ранее реализован в работе бакалавриата. Особенность заключалась в том, что был собран дата-сет компаний и всевозможных наблюдений, но в нем было много мусора и приходилось данные «мержить» особым образом с теми, что имеются после парсинга репозитория

Пункт 6:

Данный пункт включает в себя реализацию базового подхода для оценки матрицы переходных вероятностей (матрицы миграции). Ничего сверх сложного в нем не было, оценка получена довольно интуитивно.

**Цикл II**

В течение второго цикла проекта были выполнены и планируются к выполнению следующие шаги:

1. Реализация метода Вальда для когортного метода
2. Анализ результатов для метода Вальда
3. Реализация метода дюрации
4. Сравнение методов
5. Начало реализации метода бутстрапа (bootstrap)

Пункт 1:

Метод Вальда может быть применен для построения доверительных интервалов для элементов матрицы переходных вероятностей Марковской цепи. В общем случае, данный подход прост в реализации, так как требует лишь оценки вероятностей через доверительный интервал.

Пункт 2: (реализован после пункта 1)

Пункт 3:

Реализован метод дюрации (получение матриц переходных вероятностей с учетом непрерывного времени, согласно David Lando). Данный алгоритм так же реализован на Python. Метод дюрации дает более приближенные оценки вероятностей к тем, что регламентируются в репозитории кредитных рейтингов.

Так как в общем случае когортный метод не оценивает нулевые переходы (т.е. фактических переходов нет, следовательно вероятность равна 0), метод дюрации позволяет получить ненулевую вероятность (местами она близка к 0 у очень высоких рейтингов, например ААА, АА+ и т.д.)

Пункт 4: Сравнение после реализации пункта 2

Пункт 5: Под конец цикла