Пояснительная записка.

Группе поступила задача по поиску алгоритма обучения для предсказания оценки одного из вузов на основании характеристик статей, содержащихся в БД SCOPUS при формировании заказов ПАО «Газпромнефть» по НИОКР. Перечень статей (в общей сложности 297 477) и их характеристик, содержался в наборе из 36 файлов в формате csv. Позднее они были трансформированы в единый файл. Также была предоставлена база с перечнем оценок вузов по одной из отраслей науки (геофизики).

Методы машинного обучения делятся на 2 основные группы: «обучение с учителем» и «обучение без учителя». «Обучение с учителем» предполагает возможность анализировать данные на основании имеющегося набора признаков-факторов и признаков-результатов, а также прогнозировать (обучать) по новым единицам наблюдения, чьи признаки-факторы известны.

«Обучение без учителя» («кластеризация») в рамках данной задачи позволило бы разбить статьи на группы с определёнными признаками, но это не сможет принести практической пользы для бизнес-заказчика.

Было принято решение решить задачу как задачу как задачу «обучения с учителем», где «учителем» выступают экспертные оценки, предоставленные в файле «Geophisics-T». Создан отдельный столбец данных, в который взяты только те единицы наблюдения (строки из файлов БД SCOPUS), для которых в файле «Geophisics-T» существуют экспертные оценки. Эти оценки добавлены и называются score. Данные переменные являются целевыми.

Группа решила использовать различные методы машинного обучения в целях предсказания целевой переменной.

В качестве критерия оценки было принято решение использовать MAE (mean absolute error)

,

где – фактические значения целевой переменной score,

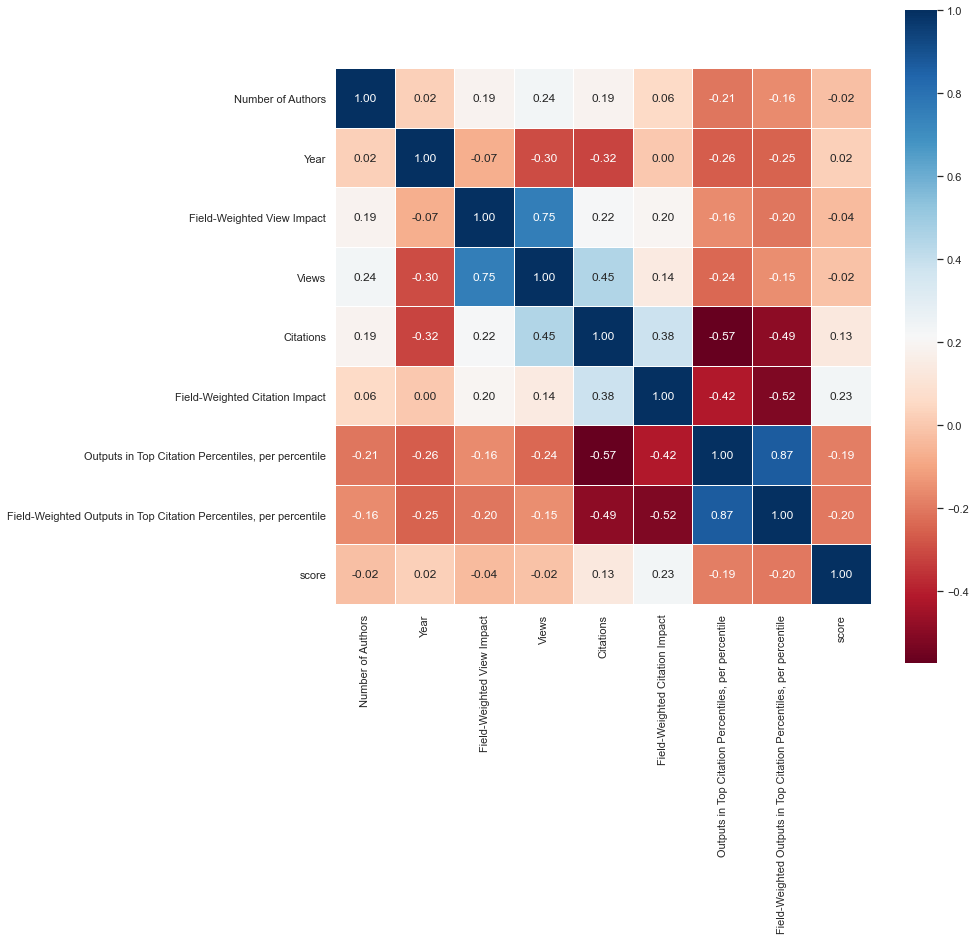
– прогнозные значения целевой переменной score,

n- число единиц наблюдения.

Существует другая распространённая метрика MSE (mean squared error) , но предпочтение было отдано MAE , так как значения переменной score лежат в интервале [0,1) и ошибка между фактическими и прогнозными значениями меньше 0. В результате при расчёте MSE при возведении в квадрат MSE будет занижаться (например, при ошибке 0,5 MSE станет равным 0,25, в то время как MAE останется 0,5).

Группа провела предварительную очистку данных: были удалены столбцы (Title, Authors, Scopus Author Ids, Scopus Source title, Reference, Abstract, DOI, EID, Scopus Affiliation IDs, Scopus Affiliation names, Country/Region, All Science Journal Classification (ASJC) code, All Science Journal Classification (ASJC) field name, Topic Cluster name, Topic Cluster number, Topic name, Topic number, University, Institutions), а столбцы year и Publication type были преобразованы в категориальные переменные (создаётся набор дополнительных переменных, признаков факторов, где придаётся значение 1 существующему типу данных (например, article) и 0 всем остальным возможным значениям.

Построена корреляционная матрица.



Корреляционная матрица – это набор попарных корреляций между признаками-факторами и признаками-результатами. Благодаря этим значениям возможно убрать из модели признаки-факторы, которые имеют тесную корреляцию между собой, определить, какие признаки тесно коррелируют друг с другом. Было принято решение убрать показатель Outputs in Top Citation Percentiles, per percentile, так как он связан с переменными Citations и Field-Weighted Outputs in Top Citation Percentiles, per percentile.

Одновременно с этим показано, что целевая переменная слабо (менее 0,3) связана с признаками-факторами.

Далее признаки-факторы были масштабированы посредством инструмента MinMax Scaler, а существующий набор данных был разделён на тренинговую и тестовую выборки в соотношении 80% данных передаются в тренинговую выборку (на которой будет строиться модель) и 20% на тестовую выборку, на которой будет проверяться качество модели как по значению MAE, так и на предмет переобучения.

Были построены следующие модели:

1. Множественная линейная регрессия.
2. Полином 4 степени
3. Полином 3 степени
4. Полином 2 степени
5. Случайный лес
6. Дерево решений
7. Бэггинг

Во всех случаях средняя абсолютная ошибка либо:

1. на тренинговой выборке ниже, чем на тестовой, что говорит о переобучении модели. Иными словами, модель будет работать только на тестовой выборке, а на иных данных она не будет показывать достоверного результата
2. либо слишком большая, что говорит о невозможности её практического применения.

В связи с этим мы пришли к выводу о том, что на основании предоставленных данных методы машинного обучения не могут построить статистически значимую модель. Для построения прогнозного значения необходимо использовать другие методы.