Создадим 12 дополнительных шумовых признаков, элементами которых будут некоррелируемые случайные числа с нормальным, равномерным и Лапласовым распределениями. Параметры каждого из распределений берутся случайным независимым образом.

давайте ещё раз по методологии пройдёмся

для оценки значимости используются 2 критерия в практической части - коэффициенты регрессии в задаче линейной регрессии, и значимости из случайных лесов

показываете, что если в исходный датасет добавлять некоррелированные признаки с разными видами распределений, то значимости у них по этим обоим критериям оказываются низкими, как и ожидается

Все кросс-валидации – 5-кратные перекрестные с сохранением частоты появления классов в выборках. Используемые далее обозначения: train\_mean\_scores – математическое ожидание оценки точности классификатора при проверке на тренировочной выборке, test\_scores – на тестовый выборке.

Самым важным признаком для случайного леса [10] стал fnlwgt. Это можно интерпретировать как то, что самым важным признаком того, что человек зарабатывает больше определенного порога является количество людей с такими же характеристиками. Происходит это потому, что как деревья решений, так и случайные леса могут выдавать смещённую оценку важности признаков [6]. Притом, чем выше риск переобучения модели, тем выше риск получить более высокое смещение, поэтому доверять оценкам таких моделей надо с осторожностью.

В отличие от случайного леса, регуляризация позволила сразу определить низкую важность параметра fnlwgt, что хорошо согласуется с выбором метода прямого последовательного отбора признаков. Отметим также, что модели оказались не переобучены (точности классификатора как на тестовой, так и на тренировочной выборках оказались равными 82,8%)

Несмотря на 12 добавленных шумовых признаков, точность модели при валидировании выросла как в среднем. Кроме того, все шумовые признаки имеют относительно высокую важность, сравнимую с двумя оригинальными. Очевидно, что наша модель переобучена. Здесь мы видим отчетливый пример того, как влияет переобучение на важность признаков модели – значения важности шумовых/маловажных признаков увеличивается. Поэтому часто бывает сложно подобрать постоянное пороговое значение важности признаков для исключения их из модели.

После добавления дополнительных признаков модель не переобучилась, к тому же указанные выше признаки имеют более низкие коэффициенты, чем оригинальные. Отметим, что коэффициенты линейных моделей зависят от способа нормализации или масштабирования признаков

Такой результат был показывает, что маловажными оказались шумовые признаки и признак fnlwgt, который при первоначальной оценке случайным лесом был самым значимым для модели. Однако из всех оригинальных признаков он имел наименьшее значение оценочной функции как для линейных моделей с регуляризацией.

На основе подобных наблюдений был сформирован модифицированный метод оценки важности признаков на основе перестановок. Схемы оригинального и модифицированного алгоритмов представлены на рисунке. Прогностической моделью был выбран случайный лес.

В отличие от обычной перестановочной важности признаков, её модифицированная версия склонна к увеличению важности параметров, которые были оценены низко другими метриками. Так же было сохранено отношение порядка, полученного оригинальным методом, что показывает согласованность данной разработки с показанными ранее методами.