Отчет о проекте «Мастерская 1»

В проекте ставилась задача поработать с табличными данными, в которых представлена информация о стартапах, которые функционировали в период с 1980 по 2018 годы, и предсказать, какие из них закроются, а какие нет. Соревнование проводится на популярной платформе Kaggle, что позволит не только применить на практике знания в области анализа данных и машинного обучения, но и освоить работу с этой платформой.

Для этого требовалось решить две взаимосвязанные задачи: разработать модель машинного обучения для предсказания продолжения деятельности стартапа, провести полноценный разведочный анализ и сформировать рекомендации будущим создателям стартапов (какие факторы влияют на успешность стартапа).

Импортировал необходимые библиотеки и функции, задал константы. Использовал проверку директорий, после чего загрузил датафреймы. Создал ряд собственных функций для предобработки и исследовательского анализ данных

Провел предобработку данных с использованием собственных функций. Проверил категориальные столбцы на уникальные значения. Укрупнил категории типов стартапов путем выделения одного (первого) типа стартапа в мультитипе. Пропуски заполнены значением unknown. Название столбцов в датафреймах и текстовые значения в датафреймах привел к змеиному регистру, удалил пробелы в начале и конце. В столбце финансирование заменил пропуски на среднее значение.

Провел исследовательский анализ данных с помощью собственных функций. По результатам статистического и графического анализа распределения являются ожидаемыми и совпадают в тренировочных и тестовых данных, что означает их применимость для построения и проверки модели. Критичных выбросов не обнаружено, кроме одного выброса по объему финансированию, который был удален, т.к. негативно влиял на модель (определено эмпирическим путем).

По результатам корреляционного анализа определена коллинеарность признаков географического нахождения стартапа, а также начала и конца финансирования. В каждой группе оставлен только один из признаков. Созданы два синтетических признака: средняя периодичность этапов финансирования и отношение длительности финансирования к времени жизни стартапа. Повторный корреляционный анализ показал зависимость целевого признака от них, поэтому включили их в признаки для модели.

Создал пайплайн, в котором использовал два кодировщика, заполнение пропусков, перебирал несколько моделей и их гиперпараметры. Эмпирическим путем попарно сравнивал эффективность моделей. В итоге лучшей моделью по метрике F1 выбран метод опорных векторов (C=1, kernel='rbf', random_state=42). Применил модель к тестовым данным, сделал предсказание целевого признака и оценил качество. Метрика F1 на тестовой выборке: 0.86583.

Ввиду того, что на нелинейном ядре анализ важности признаков сделать проблематично, провел его на линейном ядре в SVC и сделал

Вывод: на успешность стартапа больше всего влияет срок его работы, предпочтительно постепенное финансирование в несколько раундов (предполагаю, таким образом инвесторы могут контролировать текущую успешность стартапа). Причем интервал между этапами финансирования не должен быть большим, модель показала отрицательное влияния больших промежутков между раундами. Менее выраженное влияние имеют география и тип стартапа.