## Отчет о проекте «Мастерская 2»

В проекте ставилась задача поработать с табличными данными для интернет-магазина, который собирает историю покупателей, проводит рассылки предложений и планирует будущие продажи. Для оптимизации процессов необходимо в исследование определить пользователей, которые готовы совершить покупку в ближайшее время.

В исследовании последовательно решены задачи предобработки и изучения данных, разработки синтетических полезных признаков, создания модели для классификации пользователей и ее улучшения путем перебора гиперпараметров в пайплайне с максимизацией метрики roc_auc.

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 покупок путем выделения одного (первого) типа мультитипе. Провел обработку дубликатов. Синтезировал новые признаки, объединил данные в один датафрейм с таргетом.

Провел исследовательский анализ данных с помощью собственных функций. По результатам статистического и графического анализа распределения являются ожидаемыми, что означает их применимость для построения и проверки модели. Критичных выбросов не обнаружено.

По результатам корреляционного анализа определена коллинеарность некоторых признаков, которые удалены из дальнейшего использования в модели.

Создал пайплайн, в котором использовал заполнение пропусков, перебирал несколько моделей и их гиперпараметры. В итоге лучшей моделью по метрике **roc-auc** выбран CatBoostClassifier с метрикой 0,736 на тренировочных данных (требование заказчика метрика больше 0,7 выполнено). На заключительном этапе gприменил модель к тестовым данным, сделал предсказание целевого признака и оценил качество. Метрика **roc-auc** на тестовой выборке: **0.745**.

Построил матрицу ошибок, по которой определен снос предсказаний значений к нулю. Это обусловлено выбором метрики (которая считает площадь под кривой ROC и на которую неравномерно влияют несбалансированные классы) и малой долей положительных значений в таргете.

**Вывод:** Несмотря на выполненное по условиям заказчика исследование рекомендовано пересмотреть метрику для получения более практичных результатов моделирования и прогноза.