SKILLBOX

Профессия Data Scientist PRO

Итоговый проект

Модель кредитного риск-менеджмента.

Выполнил: Стучилин Леонард Валерьевич

В рамках проекта необходимо решить задачу — оценить риск неуплаты клиентом по кредиту.

Сервис на базе обученной модели, позволит банку или другой кредитной организации оценить текущий риск по любым выданным займам и кредитным продуктам. И с большой долей вероятности предотвратить неисполнение кредитных обязательств клиентом. Таким образом, банк меньше рискует понести убытки.

Техническое задание:

1. Оценить важность признаков.

2. Сгенерировать новые признаки.

3. Собрать итоговый датафрейм. состоящий из признаков для обучения модели.

4. Сделать предсказания на тестовом датасете.

5. Подготовить автоматизированный пайплайн, который по вызову fit будет готовить данные и обучать модель, а по вызову predict — делать предсказания на заданном наборе данных.

6. Обучить пайплайн подготовки данных и обучения модели и сохранить результат обучения в бинарном формате pickle.

Краткое описание проекта:

Проект реализован в PyCharm.

Предусмотрено обращение для предсказаний через api либо запуск проекта в AirFlow.

Реализовано логирование каждого этапа работы проекта.

Предусмотрен контроль метрик модели.

Подготовлена документация по проекту.

Для предсказаний применяется функция main.predictor.

def predictor:

Принимает на вход:

данные для предсказания,

в случае df=None - данные для предсказания берутся из - data.to_predict, в случае нескольких файлов, данные объединяются в один запрос.

Файл с предсказаными значениями записывается в - data.predictions.

На выход:

Подается df - содержащий id клиента и процентную вероятность дефолта.

Особенности подготовки данных:

1. Первоначальные данные собираются и агрегируются из 12 файлов parquet, к ним добавляется целевая переменная.

2. Удаляются признаки с низкой корреляцией.

3. Для удаления признаков с нулевой значимостью применяются возможности библиотеки shap.

4. Удаление дубликатов большего класса.

Создание новых признаков:

1. Отношение планового количества дней до закрытия кредита к количеству просроченых платежей - 'planned_to_zero_loans_530'.

2. Сумма просроченых платежей от 30 до 60 дней и от 60 до 90 дней - 'suspended_loans_3060_6090'.

3. Процент использования кредита при отсутствии просрочек - 'utilization_no_overdue'.

4. Сумма просрочек по каждому типу кредита - 'overdue_by_credit_type'.

5. Отношение максимальной просрочки к использованию кредита - 'max_overdue_to_utilization'.

6. Сумма кодов платежей - 'enc_payment_sum'.

Конвейер и моделирование:

Для моделирования применяются:

XGBClassifier

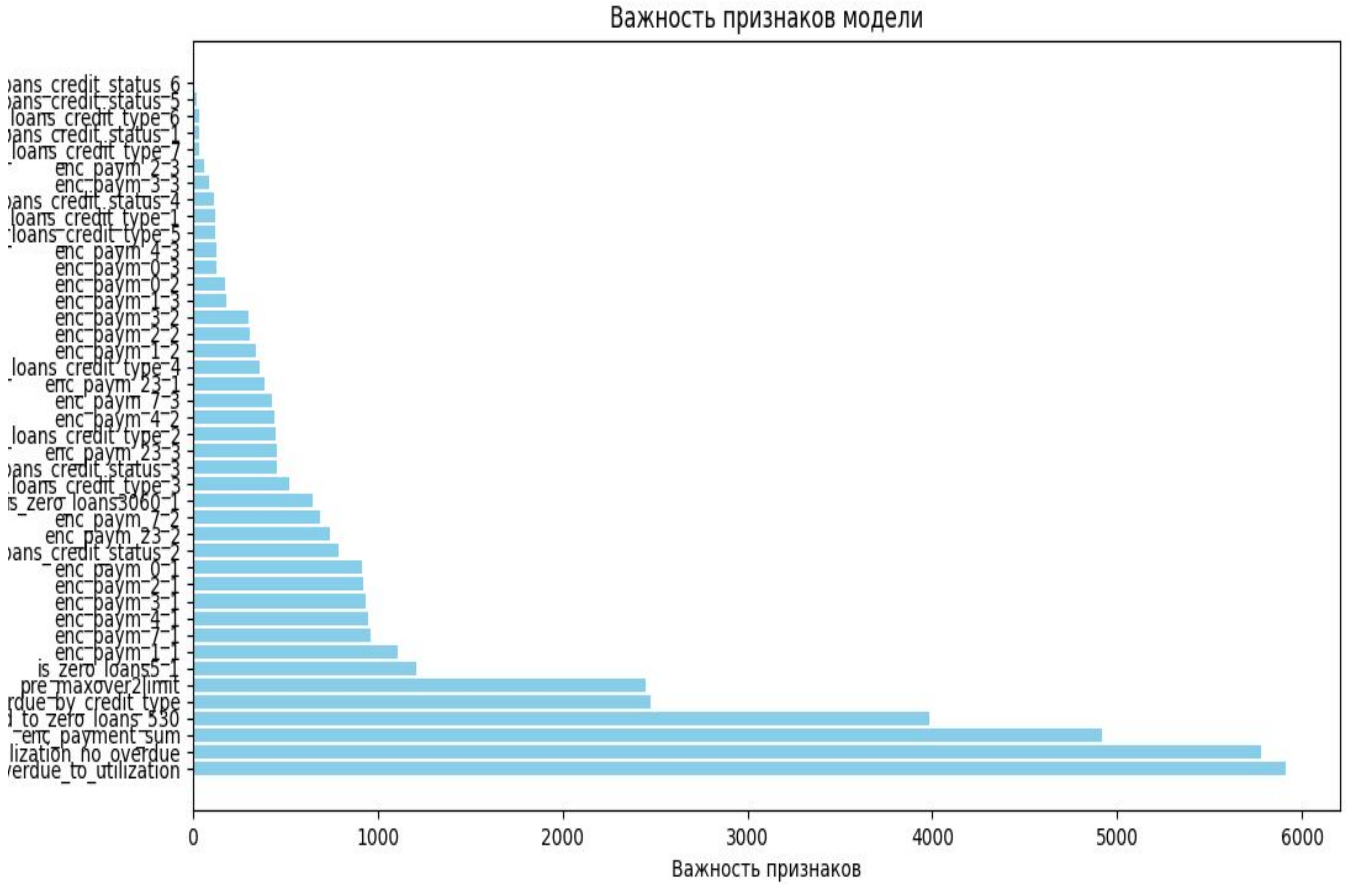
CatBoostClassifier

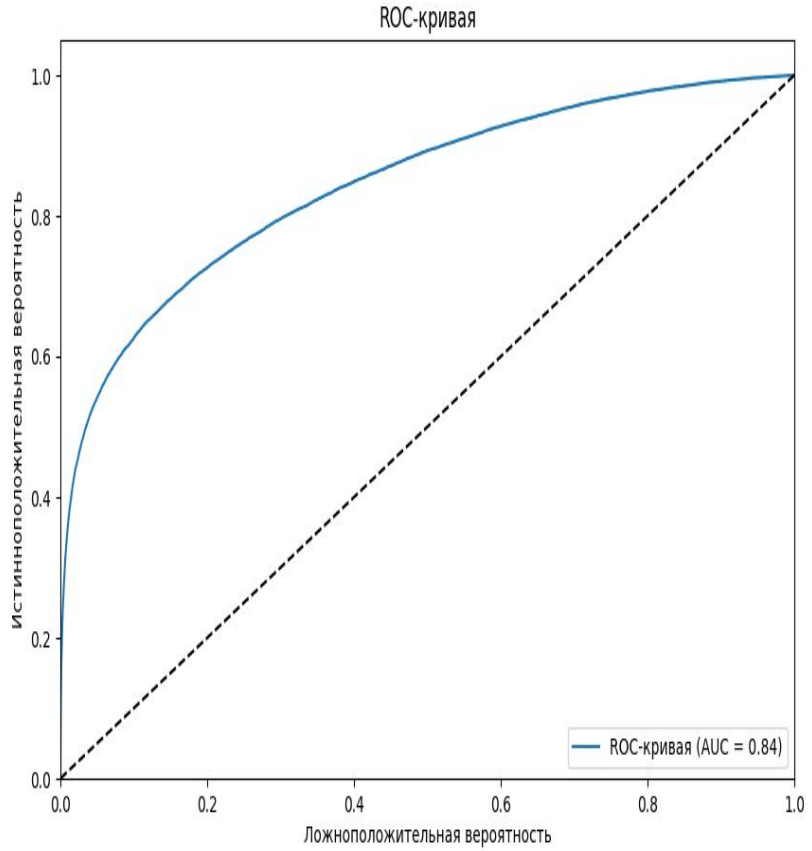
LGBMClassifier

Для выбора лучшей модели и подбора гиперпараметров применяются возможности библиотеки optuna.

Результаты работы:

model_name,	roc_auc_train,	roc_auc_test,	roc_auc_cv
LGBMClassifier,	0.880421033265016 2,	0.844686669135828 1,	0.844109319996376 2





Выводы:

Реализация данного проекта, позволит банку или другой кредитной организации оценить текущий риск по любым выданным займам и кредитным продуктам. И с большой долей вероятности предотвратить неисполнение кредитных обязательств клиентом. Таким образом, банк меньше рискует понести убытки.

Спасибо
за внимание!