

Цель проекта: предугадать вероятность выдачи кредита на основании нескольких факторов, выявить наиболее значимые связи, что влияет на одобрение банком кредита.

Описание набора данных: Датасет содержит 58645 записей, 12 признаков (числовые и категориальные):
Целевая переменная - Loan_Status (Y/N) показывает, выдали кредит (Yes), или нет (No).

Описание данных:

- Id – идентификатор клиента.
- Person_age – возраст клиента (в годах)
- Person_income – доход клиента (в долларах в год)
- Person_home_ownership – статус владения жильем (аренда, ипотека и др.)
- Person_emp_length – стаж работы клиента (в годах)
- Loan_intent – цель кредита (образование, медицина и др.)
- Loan_grade – рейтинг кредита (A, B, C и др.)
- Loan_amnt – сумма кредита (в долларах)
- Loan_int_rate – процентная ставка кредита (в процентах)
- Loan_percent_income – доля дохода, занимаемая кредитом (в процентах)
- Cb_person_default_on_file – наличие дефолта в кредитной истории (да/нет)
- Cb_person_cred_hist_length – длина кредитной истории (в годах)

Визуализация: Гистограмма дохода, box-plot дохода по статусу кредита и другие закономерности

Пропуски и выбросы: Пропусков не обнаружено, выбросы в доходе есть, но оставлены для точности предсказаний

Выбрала модель MLP, потому что она простая, базовая нейронная сеть. Подходит для нашей задачи.

Три скрытых слоя позволяют улавливать нелинейные зависимости между признаками (например, как доход и стаж работы влияют на одобрение кредита).

Функции активации:

- ReLU, потому что с ней обучение приходит быстрее
- Sigmoid на последнем слое, потому что у нас задача бинарной классификации и нужно посчитать вероятность каждого варианта. Она для этого подходит

Гиперпараметры:

30 эпох, потому что это оптимально, чтобы модель не переобучилась, но смогла выявить нужные связи, и это не заняло много времени. Если эпох будет много, то это займет очень много времени. На 50 эпохах она начинает переобучаться

Батч – 32, потому что в датасет 58к записей (он не очень большой), это хороший вариант

Метрики:

Accuracy для валидации: 0.8393, порог - 0.5.

ROC-AUC для валидации: 0.8963.