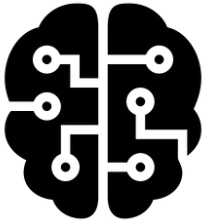
SKILLFACTORY

mail.ru group



Кредитный скоринг – классификация заемщиков по кредитному статусу

Дипломный проект Евгения Гринькина

Цели и задачи проекта

Цель

Построение модели машинного обучения по предсказанию кредитного дефолта

Задачи:

- Провести разведочный анализ данных
- Реаализовать библиотеку автоматического машинного обучения, включающую предобработку данных, ML с кросс-валидацией, подбор гиперпараметров и выбор оптимальной модели
- Построить и провалидировать модели на основе нейронных сетей
- Оценить качество оптимальной модели
- Максимизировать результаты на Kaggle

https://github.com/egrinkin/SKILLFACTORY_RDS/tree/main/thesis_project

Описание кейса

Когда клиент обращается в банк с заявлением о предоставлении кредита, банк принимает решение о выдаче кредита или об отказе в предоставлении кредита с использованием статистических моделей на основании информации о тех клиентах, которые уже брали кредит (кто-то из них выполнил свои обязательства по кредитному договору, а кто-то не выполнил). На вероятность возврата кредита может влиять много факторов, причем сложным образом, и для прогнозирования результатов по каждому отдельному случаю необходимо построить модель машинного обучения, которая на основании данных из заявления о выдаче кредита предсказывает, вернет ли заемщик этот кредит.

Данные

- **Источник данных**
  Бессрочное учебное соревнование по кредитному скорингу

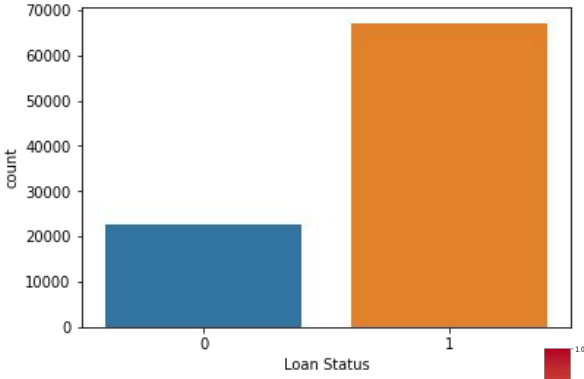
- **Датасеты**
  тренировочный набор данных credit_train.csv
  100000 записей о кредитах, относительно каждого из которых известно значение признака "Loan Status" - "Fully Paid" (погашен полностью) или "Charged Off" (не погашен)
  тестовый набор данных credit_test.csv, условно разделенный на две части:
  - публичная часть - 5000 записей о кредитах (участникам соревнования недоступна информация о значении признака "Loan Status")
  - закрытая часть - 5000 записей о кредитах; вычисление (участникам соревнования недоступна информация о значении признака "Loan Status").

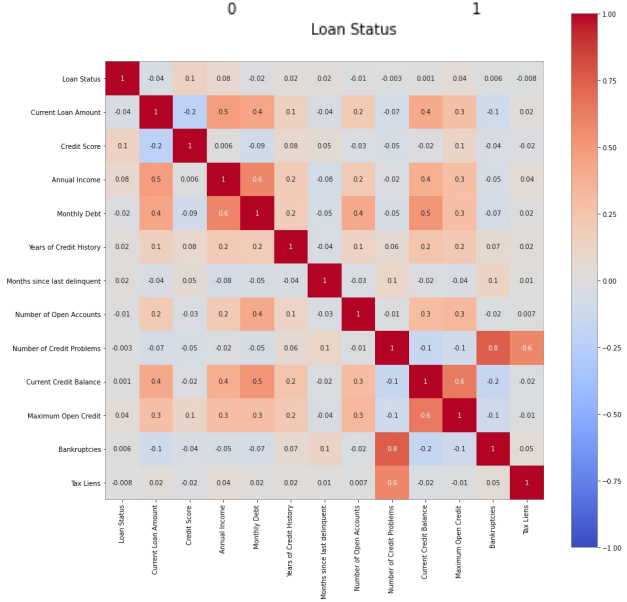
Описание полей

- **Loan ID**: уникальный идентификатор кредита;
- **Customer ID**: уникальный идентификатор клиента;
- **Loan Status**: категориальный признак (**целевая переменная**)- кредит погашен ("Fully Paid") или не погашен ("Charged Off");
- **Current Loan Amount**: размер кредита;
- **Term**: срок кредита;
- **Credit Score**: кредитный рейтинг - число от 0 до 800;
- **Years in current job**: стаж на текущем месте работы;
- **Home Ownership**: статус недвижимости - собственность ("Own"), ипотека ("Home Mortgage") или аренда ("Rent");
- **Annual Income**: годовой доход;
- **Purpose**: цель кредита;
- **Monthly Debt**: размер ежемесячного платежа по текущим кредитам;
- **Years of Credit History**: количество лет кредитной истории;
- **Months since last delinquent**: количество месяцев с последнего нарушения условий кредита;
- **Number of Open Accounts**: количество открытых кредитных карт;
- **Number of Credit Problems**: количество кредитных проблем;
- **Current Credit Balance**: суммарный текущий долг;
- **Maximum Open Credit**: максимальный кредитный лимит из всех источников;
- **Bankruptcies**: количество банкротств;
- **Tax Liens**: количество нарушений налогового законодательства.

Exploratory Data Analysis (EDA)

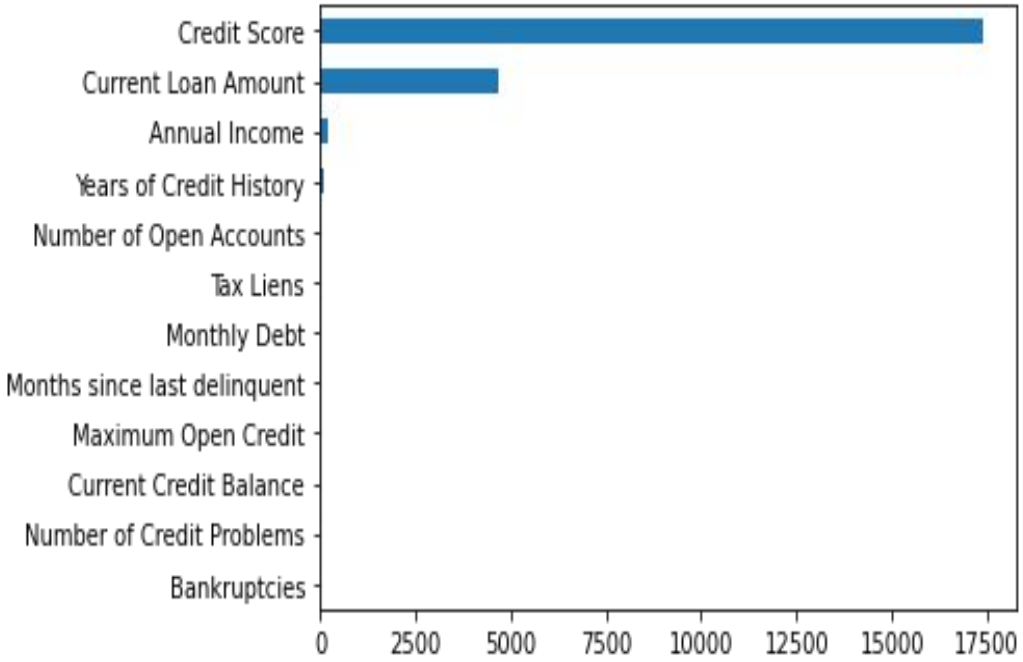
- В тренировочном датасете выявлены полные дубликаты в данных.
- Показано, что наблюдается невысокая степень несбалансированности классов (приблизительно 1 : 3).
- Множество значений таких признаков как "Кредитный рейтинг", "Годовой доход", "Стаж на текущем месте работы" и "Количество месяцев с последнего нарушения условий кредита" не заполнены.
- Такие атрибуты как "Размер кредита", "Кредитный рейтиг", "Годовой доход", "Суммарный текущий долг" и "Максимальный кредитный лимит из всех источников", содержат выбросы.
- С помощью статистических методов оценена значимость признаков. Показано, что среди числовых признаков самыми значимыми являются кредитный рейтинг и размер кредита, а среди категориальных - срок кредита.
- Между некоторыми признаками выявлена линейная корреляция Пирсона с коэффициентом, достигающим 0.8. Наблюдаемые линейные статистические взаимосвязи не противоречат бизнес-смыслу коррелирующих показателей.



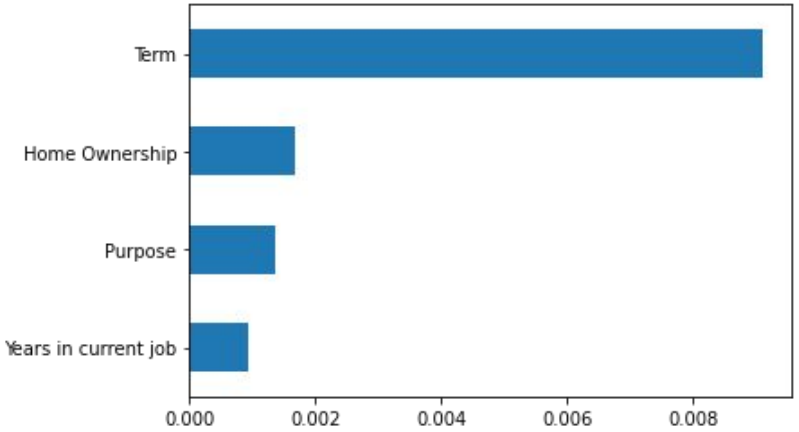


Значимость признаков

Числовые признаки



Категориальные признаки



Machine Learning

Реализована библиотека, автоматизирующая последовательные части конвейера машинного обучения:

- заполнение пропусков в данных
- кодирование категориальных переменных
- масштабирование числовых переменных
- подбор признаков
- выбор модели
- настройку гиперпараметров.

Оптимальная модель

Методом случайного решетчатого поиска с кросс-валидацией выбраны оптимальные гиперпараметры машинного обучения:

- без масштабирования числовых признаков
- алгоритм заполнения пропусков - медиана
- количество выбранных для обучения модели признаков - 41 шт.
- алгоритм ML - "случайный лес"
- количество деревьев - 275 шт.
- критерий информативности - Джини.

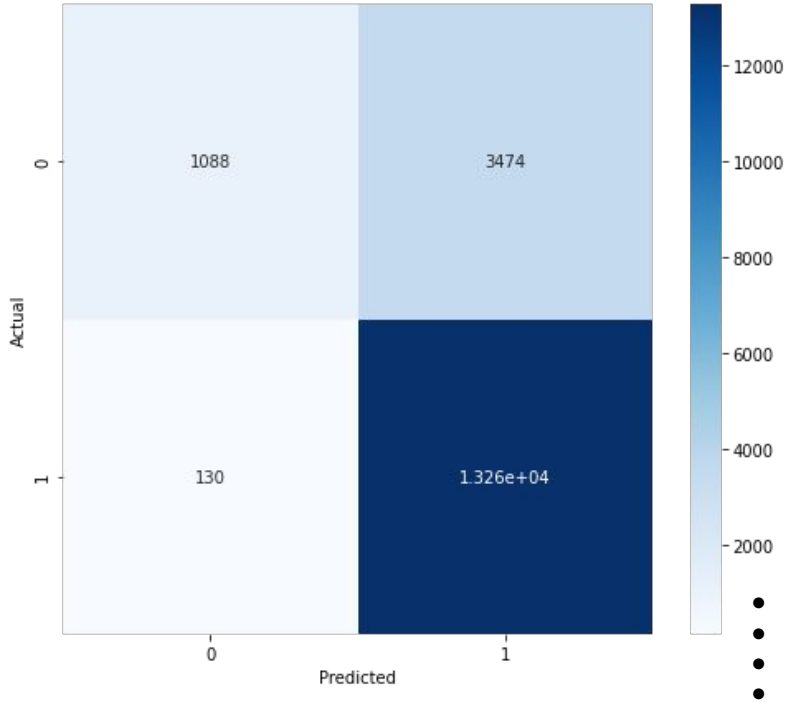
Обучение указанной модели ML позволило достичь **Accuracy,** составляющей более **0.8.**

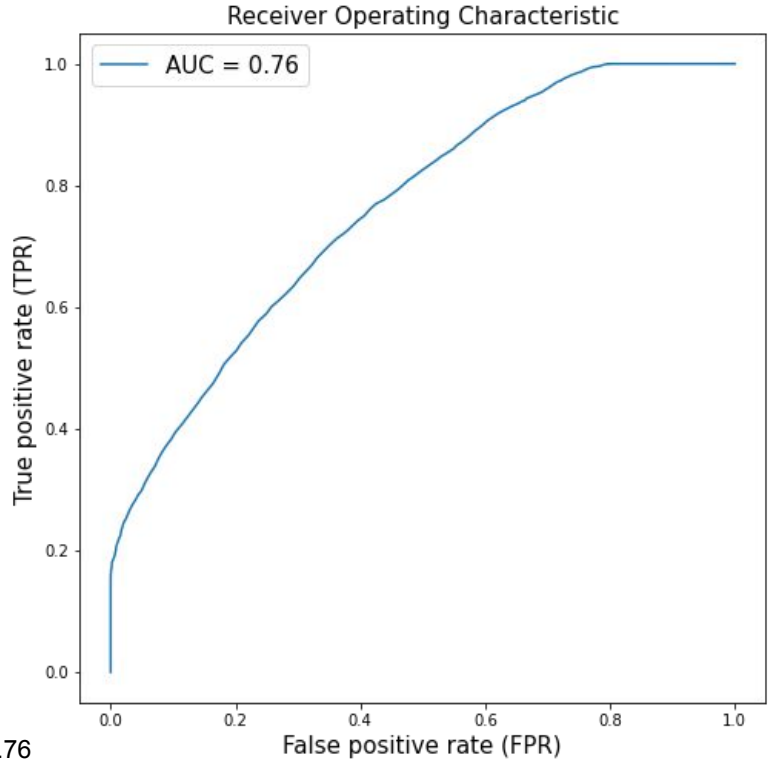
Тонкая настройка гиперпараметров не привела к улучшению результата. Многочисленные эксперименты по предварительному удалению выбросов из данных, генерации новых признаков, приведению непрерывных переменных к распределению, близкому к нормальному, понижению размерности признакового пространства и т.п. также не привели к улучшению метрики.

Deep Learning

- На тренировочных данных обучен многослоевой перцептрон
- Построен пайплайн, включающий заполнение пустых значений признаков, one-hot-кодирование категориальных признаков и стандартизацию числовых признаков, отбор значимых признаков и непосредственно обучение модели.
- Подбор гиперпараметров осуществлен с помощью случайного решетчатого поиска с кросс-валидацией на 5 фолдах.
- Значение метрики не превзошло достигнутого ранее результата с помощью "случайного леса".

Оценка качества модели





- F1-мера - 0.88
- Recall - 0.99
- Specificity: 0.24
- Precision: 0.79
- Ошибка I рода: 0.76
- Ошибка II рода: 0.0097

Матрица ошибок

Результаты на Kaggle

Nickname - Evgeniy Grinkin

- Accuracy - 0.82266
- TOP 7%
- 17 место из 269