МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра компьютерных технологий и систем

АНАЛИЗ НАБОРА ДАННЫХ «GERMAN CREDIT DATA»

Лабораторная работа номер 3

Минковского Виталия Викторовича
обучающегося 3 курса специальности
«информатика»

Минск, 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Целью данной лабораторной работы является проведение разведочного анализа (EDA) набора данных «German Credit Data» для закрепления навыков обработки информации на языке Python. Работа предполагает выполнение полного цикла анализа: от загрузки и предобработки «сырых» данных до визуализации результатов и организации хранения данных в реляционной базе данных.

Для выполнения поставленных задач были выбраны следующие методы и инструменты:

- **Библиотека Pandas:** для загрузки данных, очистки, обработки пропущенных значений и расчета описательных статистик.
- **Кодирование признаков (Label Encoding):** для преобразования категориальных переменных (таких как цель кредита, история, статус счета) в числовой формат, пригодный для корреляционного анализа.
- **Визуализация (Seaborn, Matplotlib):** построение гистограмм, ящиков с усами (boxplot) и тепловых карт корреляции для наглядного представления распределений и зависимостей.
- **SQL и SQLite:** создание локальной базы данных, экспорт обработанного датафрейма и выполнение аналитических SQL-запросов (выборки, агрегации, группировки).

В процессе анализа набора данных интерес представляли следующие вопросы:

1. Каковы основные характеристики заемщиков банка (возрастная структура, типичные суммы кредитов)?
2. Существует ли значимая корреляция между сроком кредита и его суммой?
3. Как цель кредитования (например, покупка автомобиля или развитие бизнеса) влияет на размер запрашиваемой суммы и разброс этих сумм?
4. Какие факторы (например, наличие собственного жилья) могут быть связаны с надежностью заемщика (переменная risk)?

ГЛАВА 1
ВЫПОЛНЕНИЕ ЛАБОРАТОРНОЙ РАБОТЫ

1.1 Загрузка и подготовка данных

Источник данных

Данные были загружены программно из репозитория UCI Machine Learning Repository.

- **Ссылка:** https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/german.data
- **Формат:** Текстовый файл без заголовков, разделителем значений является пробел.

Структура данных

После загрузки и присвоения именованных заголовков датасет представляет собой таблицу размером 1000 строк (наблюдений) и 21 столбец (признаков).

Список признаков включает:

- **Числовые:** duration (срок), credit_amount (сумма), age (возраст), installment_rate, residence_since, existing_credits, liable_people.
- **Категориальные:** checking_account, credit_history, purpose (цель), savings_account, employment, housing (тип жилья), job и др.
- **Целевая переменная:** risk (оценка кредитоспособности).

Обработка пропущенных значений

В ходе первичного анализа была выполнена проверка на наличие пропущенных значений (NaN) с использованием метода .isnull().sum().

Результат: Пропущенные значения в наборе данных не обнаружены. Дополнительная импутация (заполнение) или удаление строк не потребовались.

Предобработка целевой переменной

В исходном наборе данных переменная risk принимала значения {1, 2}. Для удобства дальнейшего анализа и интерпретации была произведена перекодировка:

- 1 (Good) → **1** (Кредит возвращен / Надежный заемщик)
- 2 (Bad) → **0** (Проблемный заемщик)

Кодирование категориальных переменных

Для проведения корреляционного анализа и построения тепловой карты была создана копия датафрейма, в которой все строковые (категориальные) признаки были преобразованы в числовой вид.

- **Метод:** LabelEncoder из библиотеки sklearn.preprocessing.
- **Принцип:** Каждой уникальной текстовой метке присваивается уникальное целое число (например, для признака housing: *'own'* → 1, *'rent'* → 2, *'free'* → 0).
- **Список закодированных признаков:** checking_account, credit_history, purpose, savings_account, employment, personal_status, debtors, property, other_installments, housing, job, telephone, foreign_worker.

1.2 Анализ данных

Описание числовых признаков

Для анализа были рассчитаны основные описательные статистики (метод .describe()).

- Сумма кредита (credit_amount): Среднее значение составляет 3271 DM, при этом медиана существенно ниже — 2319 DM. Это указывает на наличие "тяжелого хвоста" справа (небольшое количество очень крупных кредитов). Максимальная сумма достигает 18 424 DM.
- Длительность (duration): Средний срок кредитования — около 21 месяца. Диапазон варьируется от 4 до 72 месяцев.
- Возраст (age): Средний возраст заемщика — 35.5 лет. Самому молодому клиенту 19 лет, самому пожилому — 75.

Распределения признаков

Анализ показал, что ключевые числовые признаки (Возраст и Сумма кредита) имеют ярко выраженную правостороннюю асимметрию (positive skewness). Данные не распределены нормально: большая часть наблюдений сгруппирована в области низких значений (молодые люди, небольшие суммы), с длинным "хвостом" в сторону увеличения.

Анализ категориальных признаков

- Цели кредита (purpose): Самыми популярными целями являются покупка бытовой техники/электроники (код A43) и покупка новых автомобилей (A40).
- Жилье (housing): Большинство клиентов (более 70%) являются собственниками жилья (own), что является позитивным фактором для скоринга.

Наиболее информативным признаком для оценки риска может служить сочетание суммы кредита и цели. Асимметрия распределений подсказывает, что для применения линейных моделей машинного обучения (например, логистической регрессии) в будущем потребовалось бы логарифмирование этих переменных.

1.3 Визуальный анализ

Для выявления скрытых зависимостей были построены три типа графиков. Ниже приведена их интерпретация.

Тепловая карта корреляций (Heatmap)

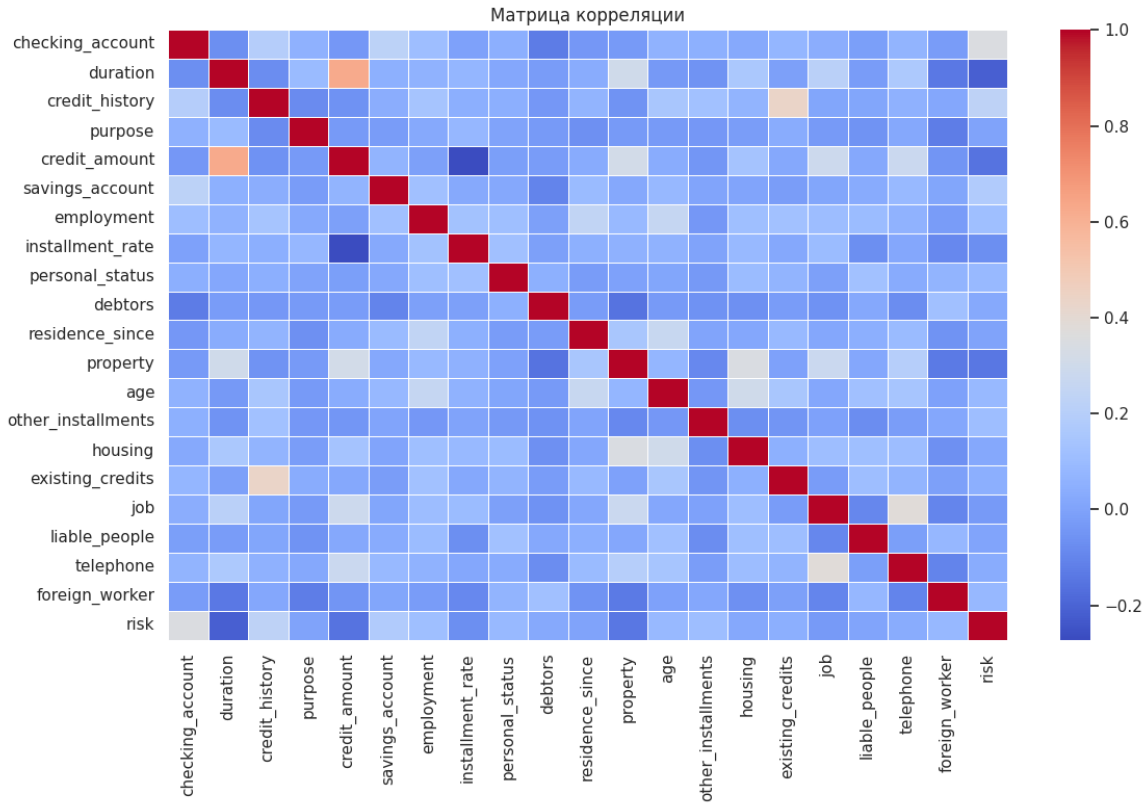

Рисунок 1. Матрица корелляции

На графике четко выделяется зона высокой положительной корреляции (оранжево-красный квадрат) между переменными credit_amount (сумма) и duration (срок). Коэффициент корреляции близок к 0.6–0.7. Это логично: чем больше сумма займа, тем на более длительный срок он выдается. Остальные признаки демонстрируют слабую корреляцию,

что говорит об отсутствии мультиколлинеарности (дублирования информации) в данных.

Гистограммы распределения

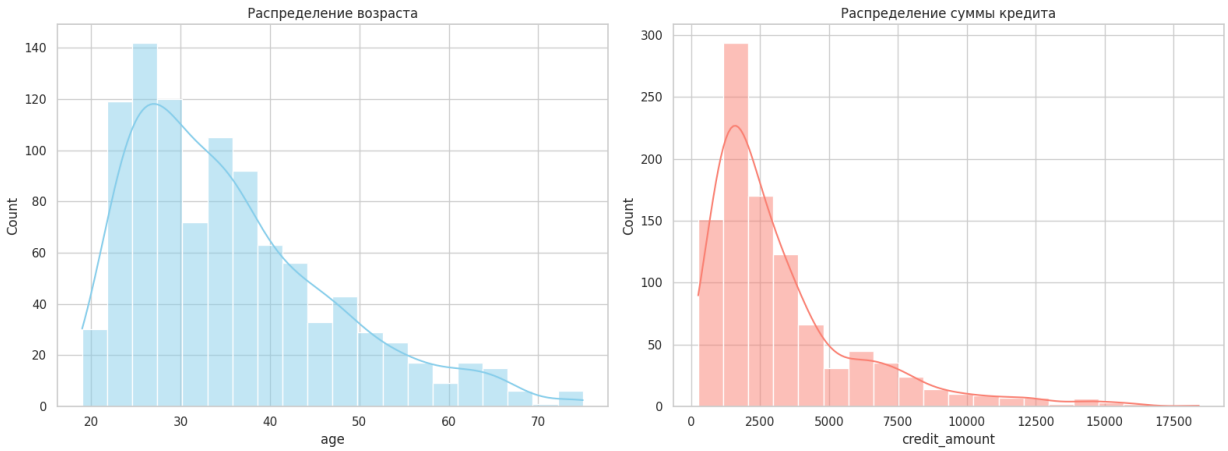

Рисунок 2. Гистограммы распределения

График подтверждает, что основная аудитория банка — люди от 25 до 35 лет. После 40 лет количество заявок резко падает. График показывает экстремальный пик в районе 1000–2000 DM. Кредиты свыше 10 000 DM являются редкостью и могут рассматриваться как аномалии или VIP-сегмент.

Ящик с усами (Boxplot) по целям кредита

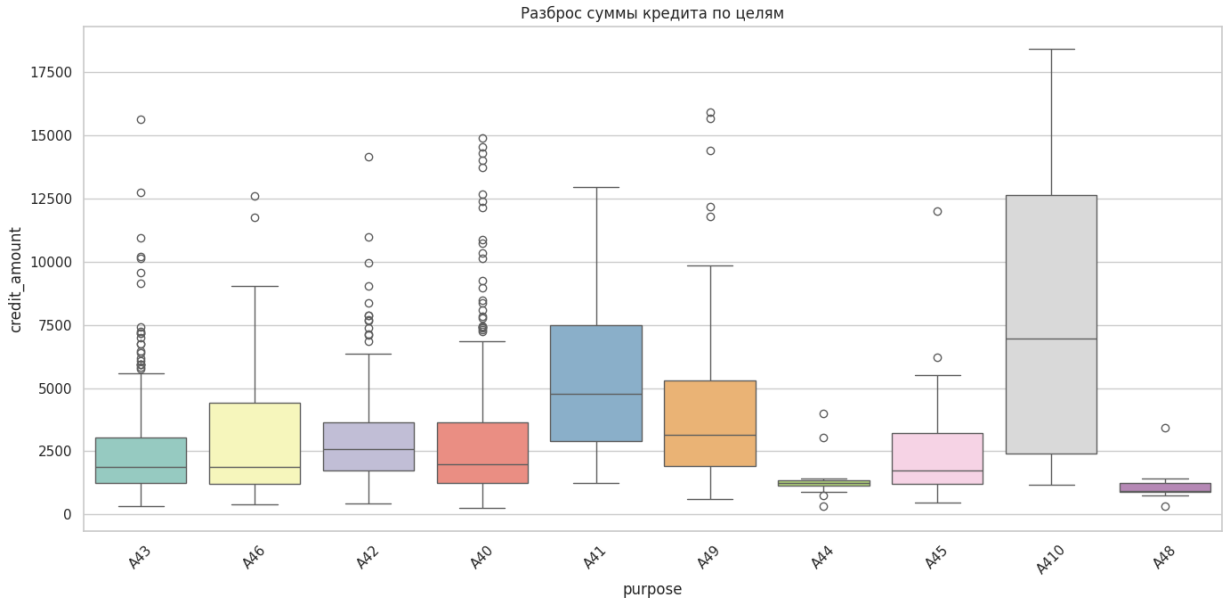

Рисунок 3. Разброс суммы кредита по целям

- Категория A410 (Others / Прочее): Имеет самый высокий медианный чек и огромный разброс (до 18 000+ DM). Это самая непредсказуемая категория.

- Категория A49 (Business): Также характеризуется высокими суммами.
- Категория A43 (Radio/TV): Самая массовая категория с низким средним чеком, однако на графике видно большое количество выбросов (точек сверху). Это означает, что хотя обычно технику берут недорогую, существуют единичные случаи покупки очень дорогостоящего оборудования.
- Вывод: Цель кредита существенно влияет на запрашиваемую сумму и профиль риска.

1.4 Работа с базой данных SQLite

Создание базы данных и таблицы

Для организации хранения и структурированного доступа к данным была создана локальная реляционная база данных german_credit.db с использованием библиотеки sqlite3. В базе данных была сформирована таблица credits, структура которой полностью соответствует предобработанному датафрейму.

- Количество столбцов: 21.
- Типы данных: Целочисленные (INTEGER) для возраста, суммы, срока и целевой переменной; Строковые (TEXT) для категориальных признаков.

Вставка данных

Наполнение таблицы данными производилось программно с использованием метода библиотеки Pandas:
df.to_sql('credits', conn, if_exists='replace', index=False)
Это позволило автоматически экспортировать все 1000 записей из оперативной памяти в SQL-таблицу с сохранением корректных типов данных.

Выполненные SQL-запросы и результаты

В ходе работы были выполнены три типа запросов для решения аналитических задач:

1. **Выборка с фильтрацией (Поиск рисковых сделок)**

Задача: Найти топ-5 самых крупных кредитов, выданных на срок более 24 месяцев, которые оказались проблемными (не были возвращены).

```
SELECT purpose, duration, credit_amount, age
FROM credits
WHERE risk = 0 AND duration > 24
ORDER BY credit_amount DESC
LIMIT 5;
```

Результат: Запрос выявил, что самые крупные невозвратные кредиты (сумма от 14 000 до 18 424 DM) были взяты на цели A410 (Прочее) и A49 (Бизнес).

2. Агрегация и группировка (Статистика по целям)

Задача: Рассчитать средний чек и максимальный возраст заемщика для каждой цели кредитования.

```
SELECT purpose, COUNT(*) as count, ROUND(AVG(credit_amount), 2) as avg_amount, MAX(age) as max_age
FROM credits
GROUP BY purpose
ORDER BY avg_amount DESC;
```

Результат:

- Самая «дорогая» цель — A410 (Прочее) со средним чеком 8209 DM.
- Самая популярная цель — A43 (Радио/ТВ), 280 заявок со средним чеком 2487 DM.

2. **Аналитический запрос (Оценка риска по типу жилья)**

Задача: Определить долю «хороших» заемщиков (risk=1) в зависимости от типа жилья.

```
SELECT housing,
   COUNT(*) as total,
   ROUND(AVG(risk) * 100, 1) as good_loans_percent
FROM credits
GROUP BY housing
ORDER BY good_loans_percent DESC;
```

Результат:

- Собственники жилья (own): Самые надежные, возвращают кредит в 73.9% случаев.
- Социальное/бесплатное жилье (for free): Наименее надежные, процент возврата составляет всего 59.3%.

1.5 Итоговые выводы

По результатам выполненной лабораторной работы и проведенного разведочного анализа (EDA) набора данных «German Credit Data» можно сделать следующие выводы:

Наиболее значимые признаки:

1. **Цель кредита (purpose):** Является ключевым фактором, определяющим размер запрашиваемой суммы. Категории «Бизнес» и

«Прочее» характеризуются наиболее высокими суммами и значительным разбросом значений.

2. **Тип жилья (housing):** Оказался важным индикатором надежности заемщика. Статистический анализ показал существенную разницу в проценте возврата кредитов между собственниками жилья и теми, кто проживает бесплатно.
3. **Сумма и Длительность:** Эти признаки имеют ненормальное распределение с «тяжелым хвостом», что необходимо учитывать при выборе методов машинного обучения (требуется логарифмирование или нормализация).

Обнаруженные взаимосвязи

- **Корреляция суммы и срока:** Подтверждена сильная прямая линейная зависимость (коэффициент корреляции > 0.6) между суммой кредита и его длительностью. Чем больше сумма, тем дольше срок возврата.
- **Связь риска и имущества:** Клиенты, владеющие собственным жильем, статистически являются более добросовестными заемщиками (73.9% "хороших" кредитов) по сравнению с клиентами, пользующимися социальным жильем (59.3%).
- **Аномалии в потребительских кредитах:** В категории массовых кредитов (бытовая техника) обнаружены выбросы — единичные заявки на аномально высокие суммы, которые требуют дополнительной проверки службой безопасности банка.

Рекомендации на основе данных

1. **Управление рисками:** Рекомендуется внедрить более строгую процедуру скоринга для заявителей категорий «Бизнес» (A49) и «Прочее» (A410), так как именно в этих категориях встречаются самые крупные невозвратные кредиты.
2. **Сегментация:** Клиентов с социальным жильем (housing='for free') следует относить к группе повышенного риска.
3. **Автоматизация:** Признаки checking_account и credit_history (после кодирования) показали достаточную вариативность, чтобы использоваться в качестве основных предикторов в моделях классификации (Logistic Regression, Random Forest).

ЗАКЛЮЧЕНИЕ

В ходе выполнения лабораторной работы была успешно достигнута поставленная цель — проведен комплексный анализ набора данных «German Credit Data» для закрепления навыков обработки информации. С помощью языка Python и библиотек Pandas, Matplotlib, Seaborn был реализован полный цикл работы с данными: от загрузки, очистки и кодирования категориальных признаков до углубленной визуализации и интеграции с реляционной базой данных SQLite.

В результате исследования были выявлены ключевые статистические закономерности и факторы риска. Анализ показал, что цель кредитования и тип жилья являются значимыми индикаторами надежности заемщика: собственники недвижимости статистически чаще возвращают кредиты, в то время как заявки на развитие бизнеса сопряжены с высокими суммами и рисками. Также была подтверждена сильная корреляция между суммой и сроком кредита, а распределение финансовых показателей продемонстрировало явную асимметрию.

Полученные результаты создают качественную основу для дальнейшего развития проекта. В перспективе работа может быть расширена за счет построения моделей машинного обучения для автоматического кредитного скоринга, а также внедрения методов нормализации данных для устранения выбросов. Таким образом, работа подтвердила практическую значимость разведочного анализа данных (EDA) для принятия обоснованных решений в сфере оценки кредитных рисков.