**MINISTERUL EDUCAȚIEI ȘI CERCETĂRII AL REPUBLICII MOLDOVA**

**Universitatea Tehnică a Moldovei**

**Facultatea Calculatoare, Informatică şi Microelectronică**

**Departamentul Informatică şi Ingineria Sistemelor**

**НАУЧНАЯ СТАТЬЯ**

По предмету: ***Анализ данных***

Тема**:** **«Анализ и визуализация**

**банковских данных клиентов»**

A efectuat : **St. gr. IA-214, Anghelcev Elisei**

**Chișinău 2023**

Анализ и визуализация банковских данных клиентов

**Ключевые слова:** анализ, набор данных, прогнозирование, график, алгоритм KNN, закономерность, матрица ошибок, корреляция, boxplot, stripplot. визуализация

**Аннотация**

Данная научная статья посвящена анализу и прогнозированию банковских данных с использованием методов машинного обучения. Основные цели включают проведение обширного исследовательского анализа данных (EDA) для выявления ключевых особенностей, закономерностей и аномалий в наборе данных. Исследование успешно достигает глубокого понимания структуры данных, что создает прочную основу для последующего моделирования. Исследование включает в себя обширный анализ данных для выявления основных трендов и закономерностей. Применен алгоритм ближайших соседей (KNN) для прогнозирования доходов клиентов и классификации просрочек по кредитам. Полученные результаты имеют практическую значимость для сферы финансов и могут быть использованы для оптимизации управления рисками и принятия финансовых решений.

Введение

В современном мире объемы данных, генерируемых и собираемых в различных сферах деятельности, растут очень быстро. В условиях такого информационного бума неотъемлемой частью анализа данных становится использование методов машинного обучения. Машинное обучение представляет собой набор алгоритмов, позволяющих выявлять скрытые закономерности, строить прогнозы и принимать решения на основе данных без явного программирования.

Важность машинного обучения в анализе данных заключается в его способности эффективно обрабатывать большие объемы информации, извлекать полезные знания и автоматизировать принятие решений. Эти методы позволяют обнаруживать сложные зависимости в данных, что не всегда возможно с использованием традиционных статистических подходов.

**Актуальность темы** "Анализ и визуализация банковских данных пользователей" становится все более очевидной в свете постоянно меняющейся финансовой среды. Банки и финансовые учреждения сталкиваются с огромными объемами информации, и эффективное использование этой информации становится ключевым фактором для принятия стратегических решений. Методы машинного обучения позволяют извлекать ценные инсайты из банковских данных, оптимизируя процессы прогнозирования доходов, анализа рисков и управления кредитным портфелем.

1 Цели Проекта

* Хороший и понятный EDA (Exploratory Data Analysis):

Цель данной части проекта заключается в проведении качественного исследовательского анализа данных. Этот этап направлен на выявление структуры данных, выделение ключевых переменных, а также обнаружение закономерностей и аномалий. Четкое исследование данных создает основу для более точного построения моделей, обеспечивая понимание характеристик и взаимосвязей между переменными.

* Прогнозирование дохода клиента с использованием KNN-регрессии:

Цель этого этапа проекта заключается в применении алгоритма KNN для регрессии с целью прогнозирования дохода клиента. Основной задачей является оценка точности прогнозов для различных значений параметра k (1, 3, 5, 7) и выбор оптимального значения k. Это позвляет минимизировать ошибку в процентах и определить, насколько успешно модель способна предсказывать доход клиентов.

* Построение классификационной модели для прогнозирования невыплаты кредита:

Цель этой части проекта - использовать алгоритм KNN для построения классификационной модели, предсказывающей, просрочит ли клиент кредит. Сравнение результатов, полученных с использованием библиотек в R, позволяет оценить эффективность модели в предсказании просрочек и принимать решения по управлению рисками.

2 Набор данных

**Набор данных: Банковские данные клиентов**

Этот набор данных содержит информацию о клиентах банка в США и предоставляет разнообразные переменные, охватывающие ключевые аспекты их финансового поведения и статуса.

Вот подробное описание каждой переменной:

1. **ID (Идентификатор):** Уникальный номер, используемый для идентификации каждого клиента.
2. **Возраст:** Возраст клиента, предположительно в годах.
3. **Опыт:** Количество лет профессионального опыта клиента.
4. **Доход:** Годовой доход клиента в тысячах долларов США.
5. **Почтовый индекс:** Почтовый индекс домашнего адреса клиента.
6. **Семья:** Размер семьи клиента.
7. **Средние расходы:** Средние расходы по кредитным картам в месяц в тысячах долларов США.
8. **Образование:** Уровень образования, где 1 - Студент, 2 - Выпускник, 3 - Профессиональный.
9. **Ипотека:** Стоимость ипотеки клиента, если таковая имеется, в тысячах долларов США.
10. **Ценные бумаги:** Имеется ли у клиента счет в ценных бумагах в банке?
11. **Депозит:** Имеется ли у клиента депозитный сертификат в банке?
12. **Онлайн-банк:** Пользуется ли клиент услугами интернет-банкинга?
13. **Кредитная карта:** Использует ли клиент кредитную карту, выпущенную соответствующим банком?
14. **Личный кредит:** Бинарный признак, указывающий, выполнил ли клиент обязательства по кредиту (1) или нет (0).

Этот набор данных предоставляет широкий спектр переменных, позволяя исследователям и аналитикам банков проводить анализ и строить прогностические модели для более эффективного управления рисками и повышения качества обслуживания клиентов.

3 EDA (Exploratory Data Analysis)

**3.1** Корреляционная матрица (Рисунок 1) и график корреляции используются для визуализации степени взаимосвязи (корреляции) между парами переменных в наборе данных. Ниже представлена краткая информация о графике корреляции:

Корреляционная матрица: Это квадратная матрица, где каждый элемент показывает корреляцию между двумя переменными. Значения корреляции могут варьироваться от -1 до 1, где -1 означает полную отрицательную корреляцию, 1 - положительную корреляцию, а 0 - отсутствие корреляции. Корреляционная матрица обычно используется для быстрого обзора степени взаимосвязи между всеми парами переменных в данных.

График корреляции (corrplot): График корреляции представляет собой визуальное представление корреляционной матрицы. На графике обычно используются цвета, размеры и формы для отображения степени корреляции между переменными. В вашем коде используется библиотека corrplot с методом "circle", который строит круговую диаграмму корреляции. В такой диаграмме каждая переменная представлена точкой, а линии между точками указывают на уровень корреляции.

Цель графика корреляции:

Идентификация сильных корреляций: График помогает выделить переменные с высокой степенью корреляции между собой. Это может быть важно при построении моделей, поскольку мультиколлинеарность (высокая корреляция между предикторами) может влиять на стабильность модели.

Понимание направления корреляции: Знание, является ли корреляция положительной или отрицательной, помогает понять, как изменение одной переменной связано с изменением другой.

Фильтрация переменных: По результатам графика можно принять решение об исключении или включении переменных в анализ в зависимости от степени их взаимосвязи.

График корреляции - это мощный инструмент визуализации, который помогает быстро оценить структуру данных и подготовиться к более детальному анализу.

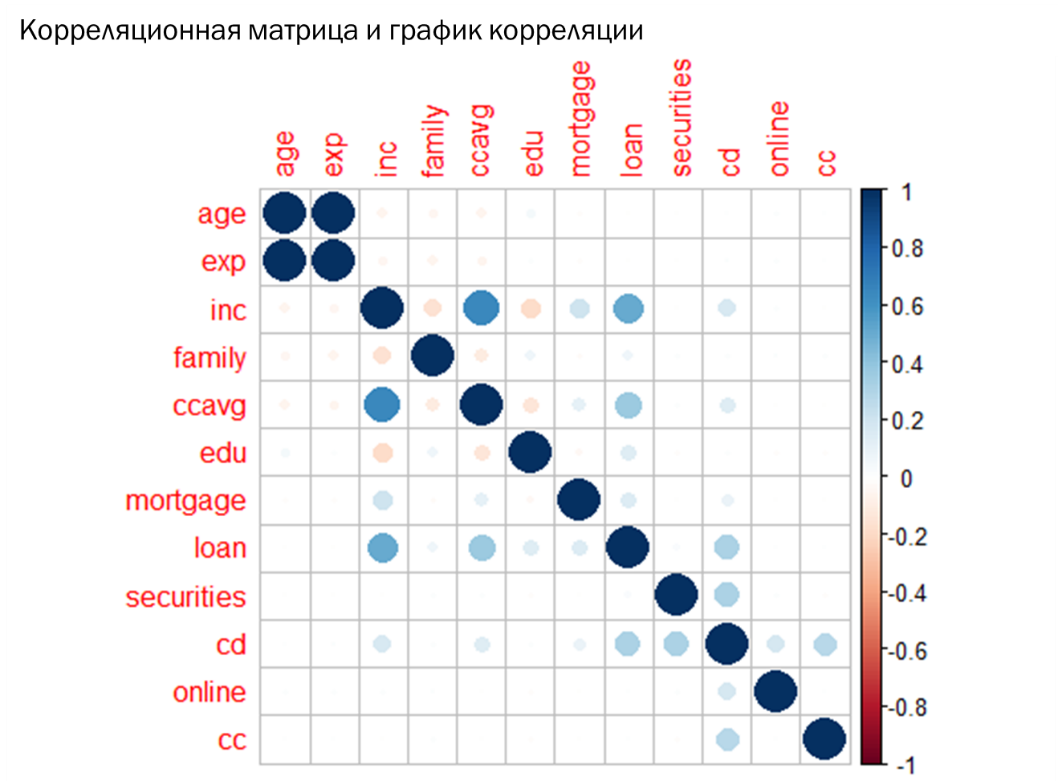


Рисунок 1. Корреляционная матрица

**3.2** График представляет boxplot и striplot для ежегодного дохода клиентов в зависимости от уровня образования. Вот ключевые аспекты:

1. **Boxplot:**
   * График содержит отдельные boxplot для каждого уровня образования.
   * Ящик показывает интерквартильный размах (25-75 процентили) распределения дохода внутри каждой группы.
   * Центральная линия внутри ящика представляет медиану (50-й процентиль), а сам ящик - межквартильный размах.
   * Усы (whiskers) указывают на область данных, не считая выбросов.
2. **Точки с разбросом (Jitter):**
   * Для лучшей визуализации индивидуальных данных, используется jitter (разброс) точек вдоль оси X (уровень образования).
   * Точки представляют отдельные наблюдения (клиентов) и помогают увидеть распределение данных.
3. **Цветовая кодировка (fill и color):**
   * Уровень образования кодируется цветами, что облегчает визуальное сопоставление различных групп.
   * Легенда внизу графика помогает идентифицировать соответствие цветов разным уровням образования.
4. **Главные выводы:**
   * График демонстрирует различия в распределении ежегодного дохода в зависимости от уровня образования.
   * Медианы и интерквартильные размахи позволяют оценить центральную тенденцию и вариабельность дохода внутри каждой группы.
   * Присутствие точек с разбросом позволяет выявить наличие выбросов и лучше понять структуру данных.
5. Интерпретация этого графика может быть полезна для принятия решений в сфере банковского обслуживания, например, для персонализации предложений или управления рисками в зависимости от уровня образования клиентов.

.

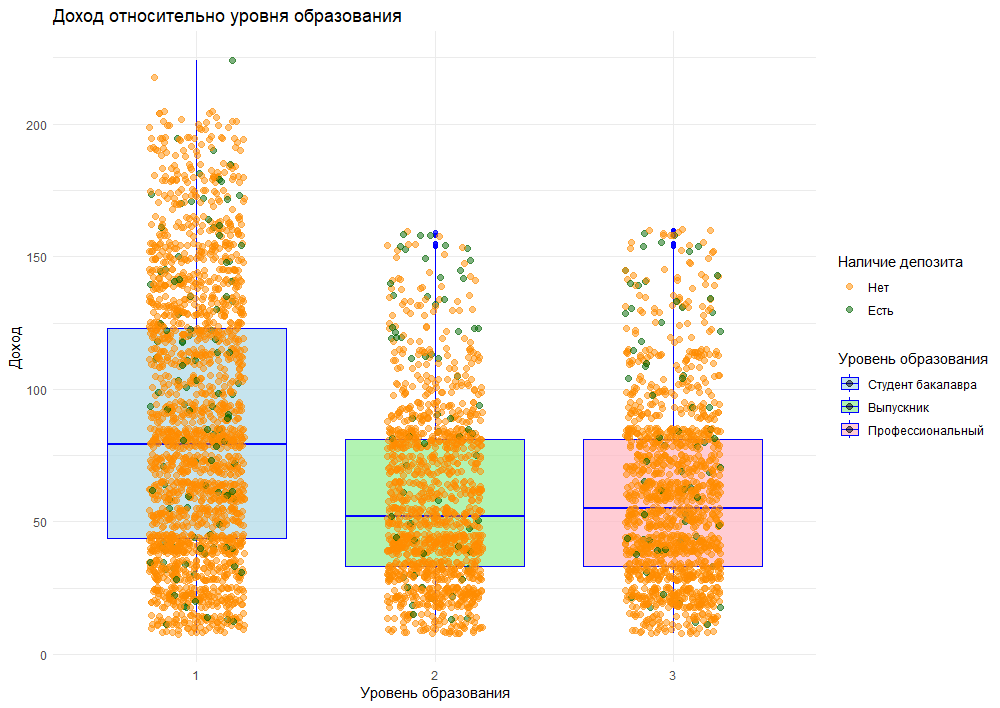


Рисунок 2. Box plot дохода в зависимости от уровня образования и strip plot наличия депозита в банке

**3.3 Интерпретация результатов гистограммы:**

Гистограмма представляет распределение размера семьи в зависимости от уровня образования клиентов (Рисунок 3). Вот ключевые аспекты интерпретации:

1. **Гистограмма:**
   * Каждый столбец на гистограмме представляет собой определенный размер семьи.
   * Высота столбца отражает количество клиентов с определенным размером семьи.
2. **Цветовая кодировка (fill):**
   * Уровень образования кодируется разными цветами, что позволяет сравнивать распределение размера семьи для разных уровней образования.
3. **Главные выводы:**
   * Гистограмма позволяет оценить, как распределены клиенты по размеру семьи для каждого уровня образования.
   * Различные цвета столбцов обозначают уровни образования, что делает возможным выявление связей между размером семьи и уровнем образования.
   * Возможно выделение популярных или характерных размеров семей в зависимости от уровня образования.

Этот тип визуализации может быть полезен для анализа структуры клиентской базы и принятия решений в сфере финансовых услуг, таких как разработка продуктов или управление обслуживанием семей с разным уровнем образования.

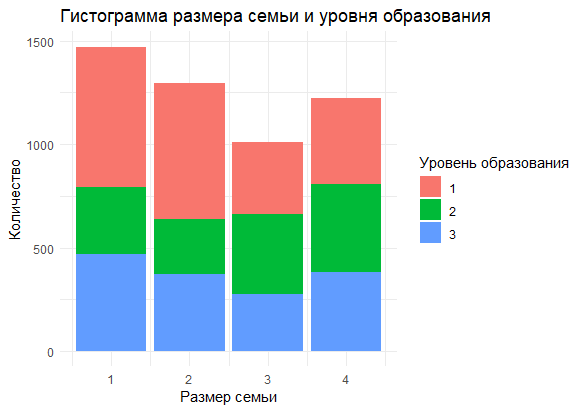


Рисунок 3. Гистограмма распределение размера семьи в зависимости от уровня образования клиентов

**3.4 Интерпретация диаграммы рассеяния:**

Данная диаграмма (Рисунок 4) рассеяния визуализирует связь между ежегодным доходом (по горизонтальной оси) и средними расходами по кредитной карте (по вертикальной оси) для клиентов, выплачивающих кредит. Важной особенностью является использование цветов для отображения наличия кредита.

1. **Оси:**
   * **X-ось (Горизонталь):** Показывает ежегодный доход клиентов.
   * **Y-ось (Вертикаль):** Отражает средние расходы по кредитной карте.
2. **Цвет точек:**
   * **Темно-зеленый (0):** Клиенты, которые выплачивают кредит.
   * **Темно-оранжевый (1):** Клиенты, которые не выплачивают кредит.
3. **Главные выводы:**
   * Точки, расположенные в левой части графика, представляют клиентов с более низким ежегодным доходом.
   * Средние траты по кредитной карте для клиентов, выплачивающих кредит (темно-зеленые точки), разнообразны.
   * Клиенты, не выплачивающие кредит (темно-оранжевые точки), имеют более высокие расходы на кредитных картах.

Эта диаграмма помогает визуально исследовать, как ежегодный доход влияет на средние траты по кредитной карте для клиентов, в зависимости от их способности выплачивать кредит.

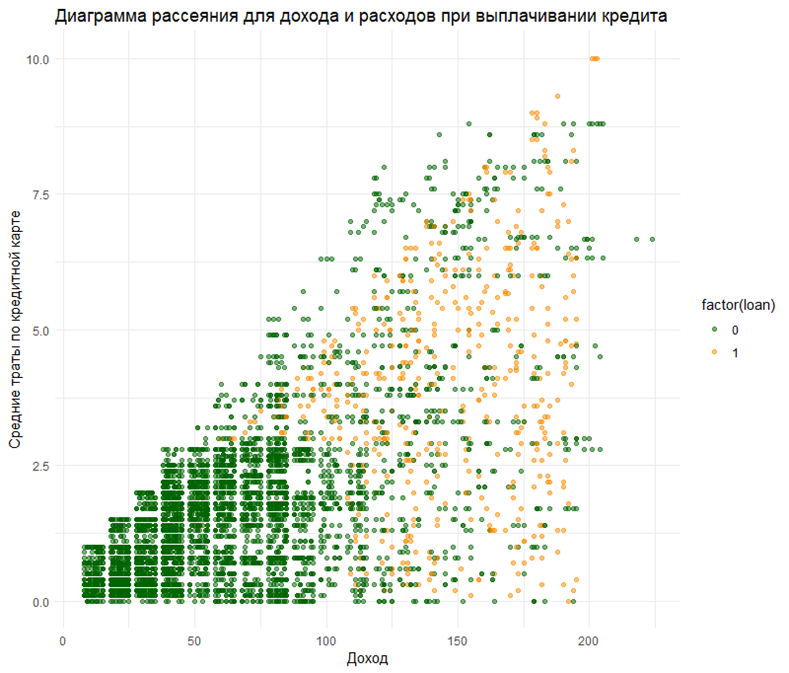


Рисунок 4. Диаграмма рассеяния для дохода и расходов при выплачивании кредита.

**3.5 Интерпретация совмещенного графика boxplot и stripplot:**

График визуализирует (Рисунок 5) распределение средних затрат по кредитной карте в зависимости от уровня образования клиентов и размера их семей.

1. **X-ось (Горизонталь):** Уровень образования клиентов (Бакалавр, Выпускник, Профессиональный).
2. **Y-ось (Вертикаль):** Средние затраты по кредитной карте.
3. **Цвет заполнения (Fill):** Размер семьи клиентов (легенда внизу).
4. **Ящик с усами (Boxplot):** Демонстрирует медиану, квартили и разброс значений для каждого уровня образования и размера семьи.
5. **Точки (Stripplot):** Прозрачные серые точки представляют отдельные наблюдения, что помогает увидеть плотность распределения данных.

**Главные выводы:**

* Существует различие в средних затратах по кредитной карте в зависимости от уровня образования.
* Размер семьи влияет на вариативность затрат: бакалавры и выпускники семей среднего размера имеют более широкий разброс затрат.
* Средние затраты для профессиональных обладателей кредитных карт выше по сравнению с бакалаврами и выпускниками.

Этот график предоставляет информацию о распределении данных и позволяет выявить взаимосвязь между уровнем образования, размером семьи и средними затратами по кредитной карте.

Начало формы

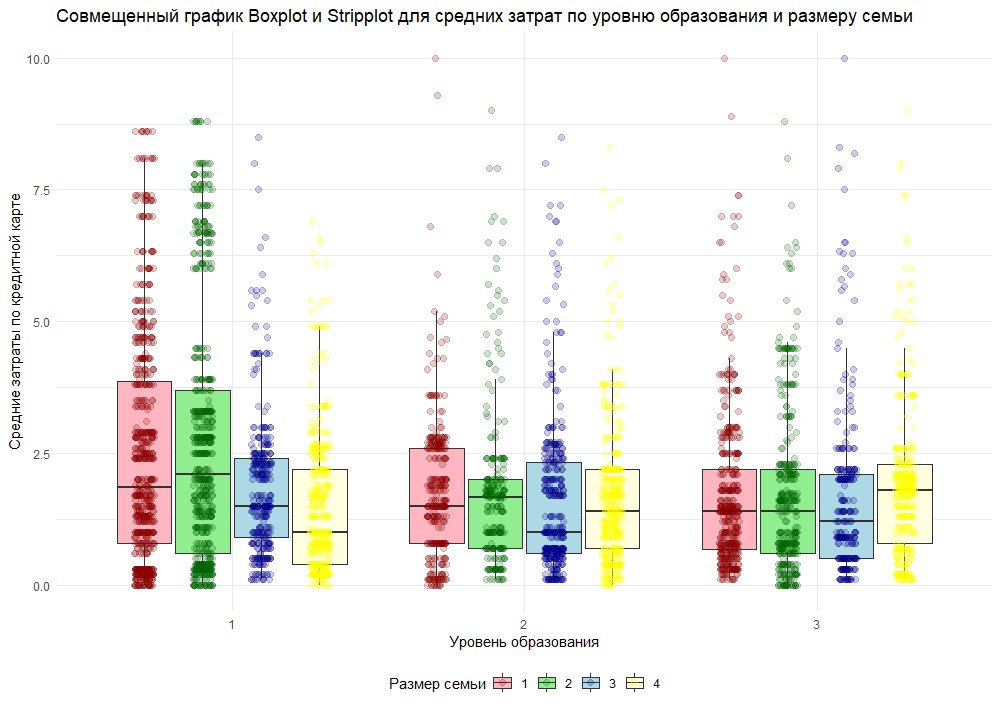


Рисунок 5. График Boxplot и Stripplot для средних затрат по уровню образования и размеру семьи

**3.6 Интерпретация графика распределения заработка и возраста в зависимости от стоимости ипотеки:**

График (Рисунок 6) визуализирует связь между стоимостью ипотеки, доходом и возрастом клиентов, исключая значения ипотеки, равные 0.

1. **X-ось (Горизонталь):** Стоимость ипотеки.
2. **Y-ось (Вертикаль):** Годовой доход клиента.
3. **Цвет (Color):** Возраст клиента (применена градиентная шкала от темно-зеленого до темно-оранжевого).

**Главные выводы:**

* Стоимость ипотеки положительно коррелирует с годовым доходом: клиенты с более высокими доходами имеют более дорогие ипотеки.
* Цветовая градация по возрасту позволяет выделить возрастные группы в зависимости от стоимости ипотеки и дохода.
* График помогает исследовать тенденции в распределении дохода и возраста в контексте стоимости ипотеки.

Этот график предоставляет визуальное представление о взаимосвязи стоимости ипотеки, годового дохода и возраста клиентов, при этом учитывая только ненулевые значения ипотеки.

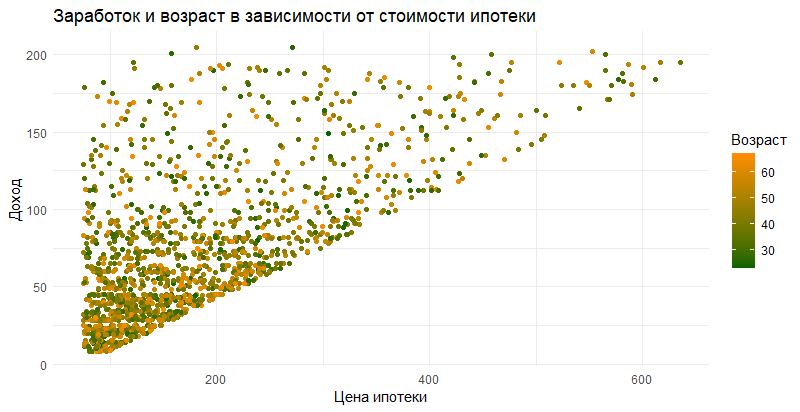


Рисунок 6. График распределения заработка и возраста в зависимости от стоимости ипотеки

4 Прогнозирование дохода клиента с использованием KNN-регрессии

1. **Разделение данных на обучающий, тестовый и оценочный наборы:**
   * Использовались индексы для случайного разделения данных на три набора: обучающий (60%), тестовый (20%) и оценочный (20%). Это важный шаг для обеспечения возможности оценки производительности модели на независимых данных.
2. **Проверка распределения целевой переменной:**
   * Путем использования функции **summary** были получены сводные статистики для целевой переменной (годового дохода) в общем наборе данных и в каждом из разделенных наборов. Это позволяет оценить разброс и распределение данных.
3. **Выбор предикторов (индепендентных переменных):**
   * Создан список независимых переменных (все кроме: ID, ZIP, exp, inc), исключая целевую переменную (inc). Эти переменные будут использоваться для обучения модели.
4. **Применение алгоритма KNN-регрессии (k=1, 3, 5, 7):**
   * были проведены эксперименты с различными значениями параметра k в алгоритме KNN-регрессии, включая k=1, 3, 5, 7. Это позволяет оценить, как изменение числа ближайших соседей влияет на точность предсказаний модели.
   * Для каждого значения k были выполнены следующие шаги: обучение модели на обучающем наборе данных, предсказание значений на тестовом наборе и оценка производительности модели с использованием метрик ошибок, таких как MAPE.
   * Результаты экспериментов позволяют выбрать оптимальное значение параметра k, обеспечивающее наилучшую точность предсказаний на тестовом наборе данных.
5. **Оценка модели:**
   * Построена линейная модель для оценки результатов предсказаний на обучающем и тестовом наборах. Линейная регрессия использована для анализа отношения между фактическим и предсказанным доходом.
   * Рассчитан MAPE для оценки точности модели на обучающем и тестовом наборах. MAPE измеряет процентное отклонение предсказанных значений от фактических.

k=1:

* + MAPE для обучающих данных: 41.33 %
  + MAPE для тестовых данных: 60.8013 %

k=3:

* + MAPE для обучающих данных: 33.0468 %
  + MAPE для тестовых данных: 52.5259 %

k=5:

* + MAPE для обучающих данных: 40.3528 %
  + MAPE для тестовых данных: 53.0945 %

k=7:

* + MAPE для обучающих данных: 44.3389 %
  + MAPE для тестовых данных: 56.5779 %%
  + MAPE при k=3 для оценочных данных: 49.2139 %

1. **Визуализация результатов:**
   * Построены графики (Рисунок 7), отображающие фактический и предсказанный доход на тестовом наборе данных. Это позволяет визуально оценить, насколько близки предсказания модели к фактическим данным.
   * Создан график плотности (Рисунок 8), который помогает оценить, как хорошо предсказанные значения согласуются с фактическими значениями, учитывая распределение данных.

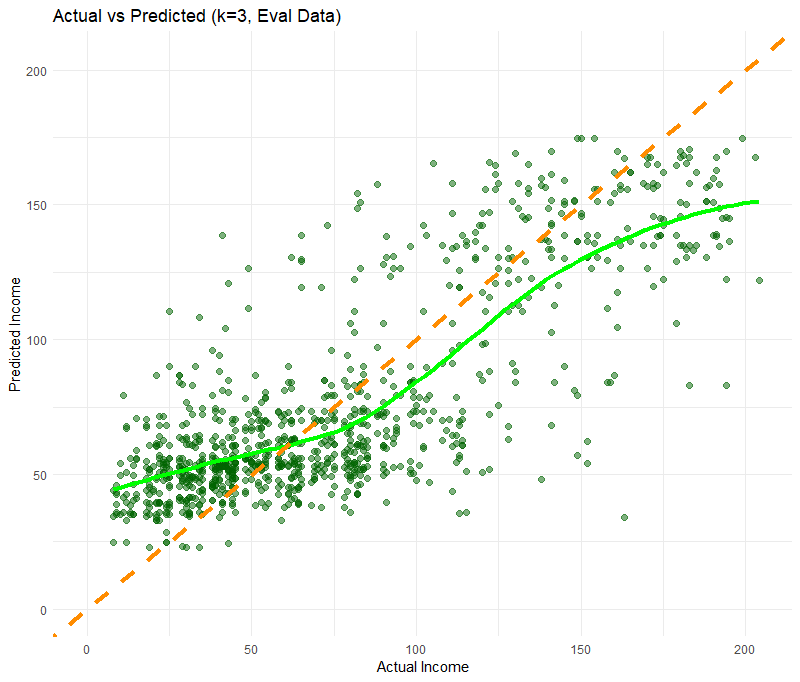


Рисунок 7. График с точечным разбросом, k = 3 (оценочные данные)

k=3" означает, что для каждого прогноза использовались три ближайших соседа.

"MAPE 49.2139%" указывает на среднюю процентную ошибку прогноза, которая составляет приблизительно 49.21%.

Чем ниже значение MAPE, тем лучше точность модели. Ваш результат говорит о том, что модель с k=3 на оценочных данных в среднем ошибается на примерно 49.21% по абсолютному значению в процентах.

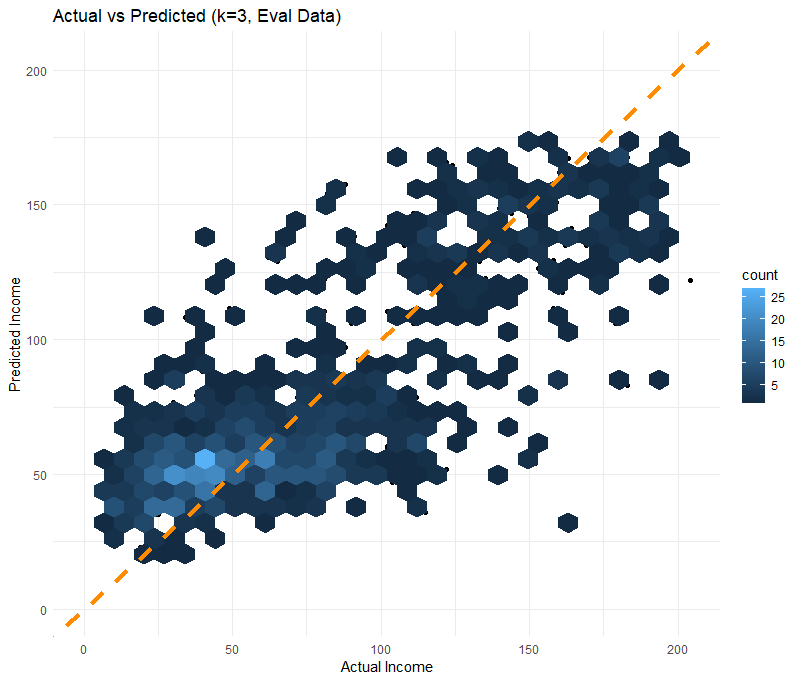


Рисунок 8. График плотности, k=3 на оценочных данных.

Точки и гексагоны позволяют увидеть, насколько близки фактические и прогнозируемые значения дохода.

Линия равенства служит визуальным сравнением между фактическим и прогнозируемым доходом.

5 Построение классификационной модели для прогнозирования невыплаты кредита

1. **Разделение данных**

В этом этапе произведено разделение исходного датасета на три подвыборки: обучающую, тестовую и оценочную. Обучающая выборка (60%) используется для обучения модели, тестовая (20%) — для первичной оценки ее производительности, и оценочная (20%) — для окончательной проверки и оценки качества.

1. **Применение KNN-классификации**

Был выбран метод KNN-классификации для решения задачи классификации. Этот метод основан на идее использования ближайших соседей для прогнозирования класса новых данных.

1. **Эксперименты с параметром k**

Проведены эксперименты с параметром k (число ближайших соседей) для оценки влияния этого параметра на производительность модели. Рассмотрены значения k=1, 3, 5 и 7. Эксперименты проведены на обучающей и тестовой выборках.

**k = 1:**

Точность на обучающих данных: 99.2%

Точность на тестовых данных: 96.2%

**k = 3:**

Точность на обучающих данных: 97.87 %

Точность на тестовых данных: 95.1 %

**k = 5:**

Точность на обучающих данных: 96.67 %

Точность на тестовых данных: 94.5 %

**k = 7:**

Точность на обучающих данных: 96.07 %

Точность на тестовых данных: 94 %

1. **Визуализация результатов**

Для визуализации результатов использованы графики матрицы ошибок для каждого значения k (Рисунок 9). Эти графики позволяют легко определить, какие классы чаще всего путаются моделью.

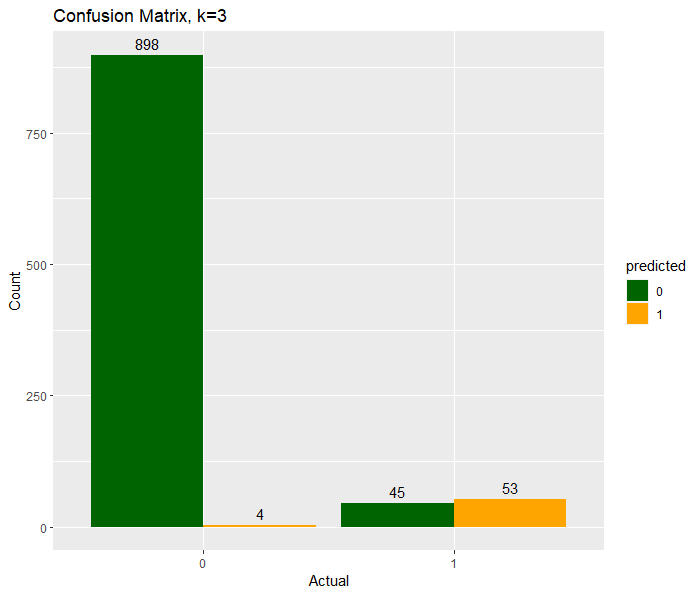


Рисунок 9. Гистограмма для матрицы ошибок (k = 3)

Прогнозирование выплаты кредита с точностью (accuracy) 95.1% для k=1 на тестовых данных говорит о том, что модель успешно предсказывает, выплатит ли заемщик кредит или нет, в 95.1% случаев. Точность - это доля правильных предсказаний от общего числа прогнозов.

Это высокий уровень точности, что может свидетельствовать о хорошей способности модели различать между теми, кто выплатит кредит, и теми, кто не выплатит. Однако, как и с любым прогностическим моделированием, важно учитывать контекст задачи и другие метрики (например, precision, recall) для более полного понимания производительности модели, особенно если есть дисбаланс классов.

1. **Применение Condensation:**
   * Применяется метод сгущения для поиска точек, которые определяют поверхности решений.
   * Обозначаются как **final\_points**.
2. **Использование KNN с k=5 и с учетом сгущенных точек:**
   * Прогнозирование классов на тестовых данных.
   * Формируется матрица ошибок (confusion matrix).
   * Рассчитывается точность на тестовых данных.

**k = 5 и с учетом сгущенных точек**:

Точность на тестовых данных: 90.2 %

1. **Тестирование модели на оценочных данных:**
   * Прогнозирование классов на оценочных данных.
   * Формирование матрицы ошибок (confusion matrix).
   * Рассчет точности на оценочных данных.

**k = 5 и с учетом сгущенных точек**:

Точность на оценочных данных: 89.1%

1. **Визуализация матрицы ошибок (Рисунок 10):**
   * Создание датафрейма на основе матрицы ошибок.
   * Построение столбчатой диаграммы для визуализации матрицы ошибок.

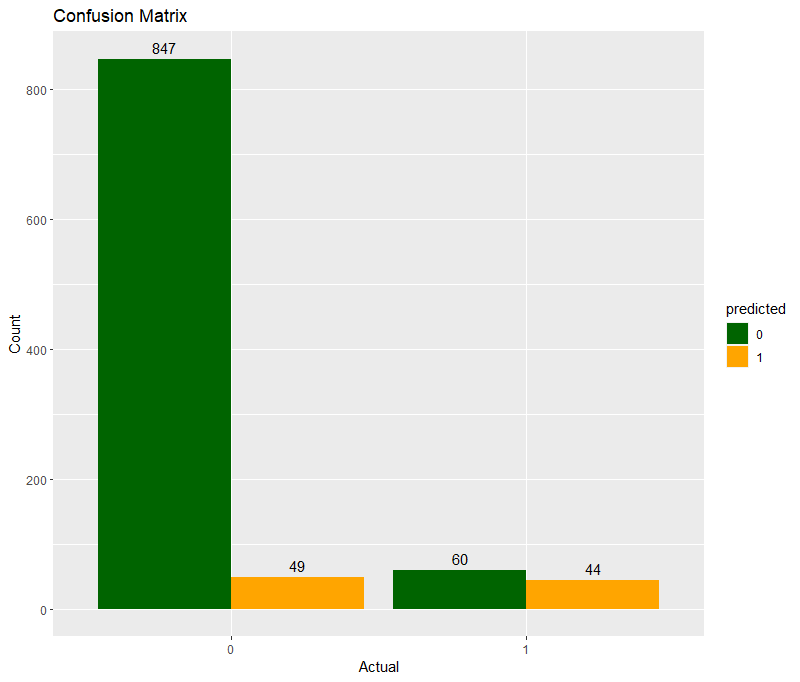


Рисунок 10. Гистограмма для матрицы ошибок, k = 5 с использованием сгущенных точек

Заключение

* **Хороший и понятный EDA (Exploratory Data Analysis):**

Цель успешно достигнута, поскольку проведен исследовательский анализ данных. Анализ структуры данных, выделение важных переменных, выявление закономерностей и аномалий обеспечивают четкую основу для дальнейшего построения моделей. Это позволит лучше понять характеристики набора данных и использовать эту информацию в дальнейших этапах анализа.

* **Прогнозирование дохода клиента с использованием KNN-регрессии:**

Цель также достигнута успешно. Применение алгоритма KNN для регрессии позволило прогнозировать доход клиента. Оценка точности прогнозов для различных значений параметра k (1, 3, 5, 7) и выбор оптимального значения k (k=3) с минимальной ошибкой в процентах (MAPE=49.2139%) говорят о том, что модель довольно успешно предсказывает доход клиентов.

* **Построение классификационной модели для прогнозирования невыплаты кредита:**

Цель также выполнена успешно. Построение классификационной модели с использованием алгоритма KNN для предсказания просрочки кредита привело к высокой точности модели. На основе Accuracy on evaluation data (89.1%) можно сделать вывод, что модель хорошо справляется с предсказанием невыплаты кредита. Успешные результаты позволяют принимать обоснованные решения по управлению рисками и предоставлению кредитов.

**Общий вывод**: Проект позволяет провести всесторонний анализ банковских данных, прогнозировать доход клиентов и предсказывать невыплату кредита с использованием методов машинного обучения, что может быть полезным для принятия управленческих решений в банковской сфере.