## **Документация**

## 

## 

## 

## **Дипломная работа по теме:**

## **“Анализ оттока клиентов банка (поиск инсайтов, составление рекомендаций стейкхолдерам, построение модели оттока)”**

**Профессия “Аналитик данных”, DA - 88**

**Романова А.А.**

**2024**

**Оглавление**

1. Введение……………………………………………………………………………………………...3

2. Описание исходного датасета и типов данных (14 столбцов)…………………………….…4

3. Разведочный анализ данных (EDA)………………………………….…………………………..5

3.1. Качество данных………………………………….………….……………………………………5

3.2. Зависимости ..……………………………….………………….…………………………………5

3.3. Рекомендации на основе анализа EDA ..……………………………….…………………….8

4. Построение модели логистической регрессии.. ………………………………….……………9

4.1. Результаты построение модели логистической регрессии прогнозирования оттока ....9

4.2. Комплексная оценка модели логистической регрессии и процесса обучения ..………11

5. Выявление инсайтов с помощью модели случайного леса и рекомендации ……………11

6. Обобщающие выводы анализа оттока клиентов банка…………………………………...…12

6.1. Корреляции признаков ..……………………………….………………………………………..12

6.2. Анализ модели ..……………………………….………………………………….…………...…12

6.3. Важность признаков………………………………….………………………………….…….…12

7. Общие рекомендации……………………………………………………….………………..……12

7.1. Целевая работа с возрастными группами……………………………………………………12

7.2. Финансовые стимулы и программы………………………………….………………..………13

7.3. Развитие программ лояльности и увеличение числа продуктов…………………………13

7.4. Исследование географических факторов………………………………….…………………13

7.5. Повышение активности клиентов………………………………….…………………..………13

8. Заключение………………………………….………………………………….……………………13

**1. Введение**

**Цели проекта:** Проанализировать данные клиентов банка для выявления факторов оттока и построения модели, прогнозирующей вероятность ухода, с целью разработки и внедрения эффективных стратегий удержания, улучшения клиентского опыта и поддержки принятия обоснованных управленческих решений для повышения лояльности и доходов банка.

**Бизнес-задача**: Анализ оттока клиентов банка для улучшения стратегий удержания и повышения лояльности.

**Актуальность**: Высокий уровень оттока клиентов может привести к значительным финансовым потерям и ухудшению репутации банка. Понимание причин и факторов оттока позволяет разработать эффективные меры по удержанию клиентов и оптимизации обслуживания.

**Гипотезы**:

* Возраст и финансовое состояние клиентов влияют на вероятность их ухода.
* Географические особенности могут быть связаны с различиями в уровне оттока.
* Активность клиентов и количество используемых продуктов могут снижать вероятность оттока.

**Стейкхолдеры отчета:**

* Руководство банка
* Команда по клиентскому обслуживанию и поддержке
* Аналитики и исследовательские группы, финансовые аналитики
* IT и технические специалисты

**2. Описание исходного датасета и типов данных (14 столбцов)**

Для исследования был взят датасет “[Churn for Bank Customers](https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers)” со статистикой операций клиентов банка.

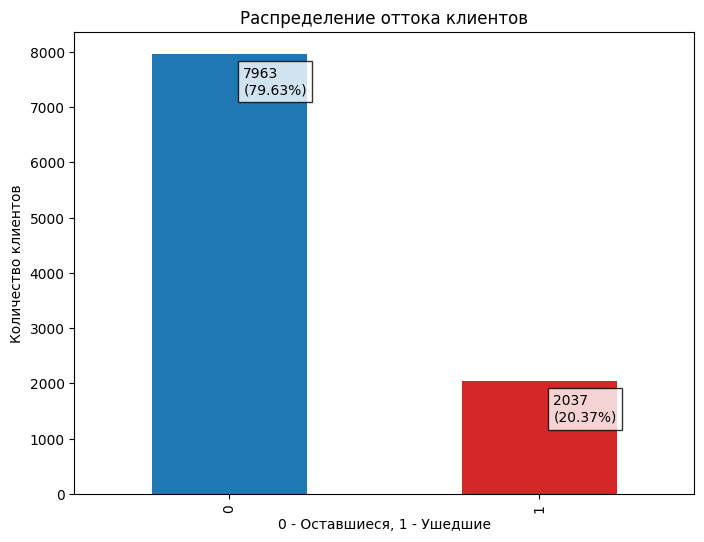
| **№** | **Имя Столбца** | **Описание** | **Тип данных** |
| --- | --- | --- | --- |
| 1 | **RowNumber** | Соответствует номеру записи (строки) и не влияет на результат. | int |
| 2 | **CustomerId** | Содержит случайные значения и не влияет на уход клиента из банка. | int |
| 3 | **Surname** | Фамилия клиента, не влияет на его решение покинуть банк. | object |
| 4 | **CreditScore** | Кредитный рейтинг клиента, может влиять на отток клиентов, так как клиенты с более высоким кредитным рейтингом менее склонны покидать банк. | int |
| 5 | **Geography** | Местоположение клиента может влиять на его решение покинуть банк. | object |
| 6 | **Gender** | Пол клиента, интересно исследовать, играет ли пол роль в уходе клиента из банка. | object |
| 7 | **Age** | Возраст клиента, это, безусловно, важно, так как пожилые клиенты менее склонны покидать банк по сравнению с молодыми. | int |
| 8 | **Tenure** | Стаж клиента в банке (в годах), обычно старые клиенты более лояльны и менее склонны к уходу. | int |
| 9 | **Balance** | Баланс на счету клиента, также хороший индикатор оттока клиентов, так как люди с более высоким балансом на счетах менее склонны покидать банк по сравнению с теми, у кого низкий баланс. | float |
| 10 | **NumOfProducts** | Количество продуктов, которыми пользуется клиент через банк. | int |
| 11 | **HasCrCard** | Наличие кредитной карты у клиента, это также важно, так как люди с кредитной картой менее склонны покидать банк. | int |
| 12 | **IsActiveMember** | Активный клиент менее склонен покидать банк. | int |
| 13 | **EstimatedSalary** | Предполагаемая зарплата клиента, как и баланс, люди с низкими зарплатами более склонны к уходу по сравнению с теми, у кого высокие зарплаты. | float |
| 14 | **Exited** | Ушел ли клиент из банка (0 - нет, 1 - да). | int |

**3. Разведочный анализ данных (EDA)**

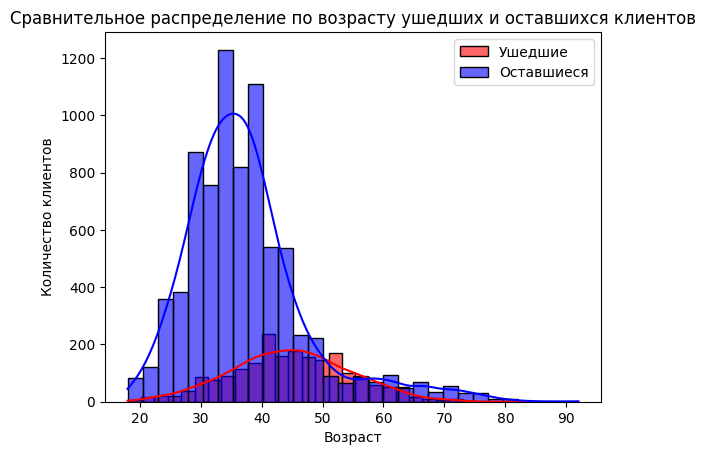
**3.1. Качество данных**: Проведен разведочный анализ данных (EDA-exploratory data analysis) - пропущенных значений нет, повторов нет, типы данных адекватны, значения категориальных данных адекватны, аномалии(нереалистичные значения) не выявлены.

Очистка или замена значений перед анализом не осуществлялась.

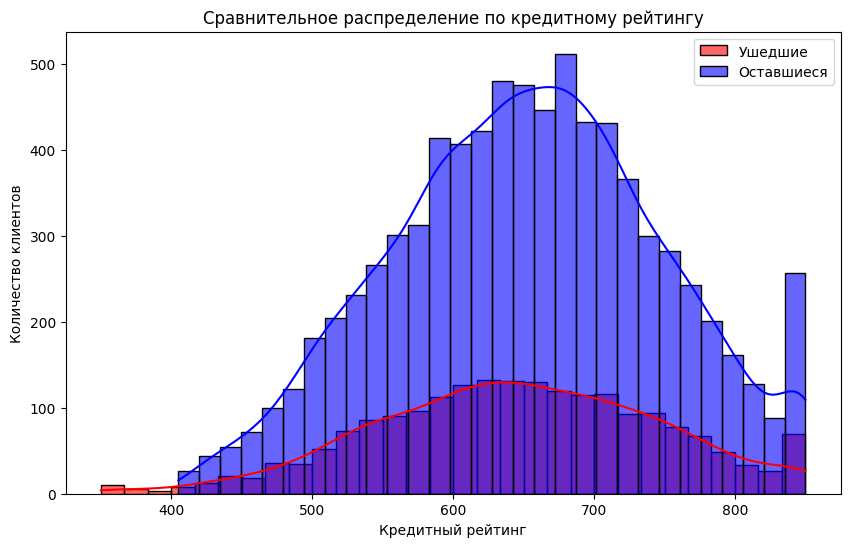
**3.2. Зависимости**: Проведен анализ корреляций между признаками и целевой переменной Exited. Визуализированы данные для поиска инсайтов.

****

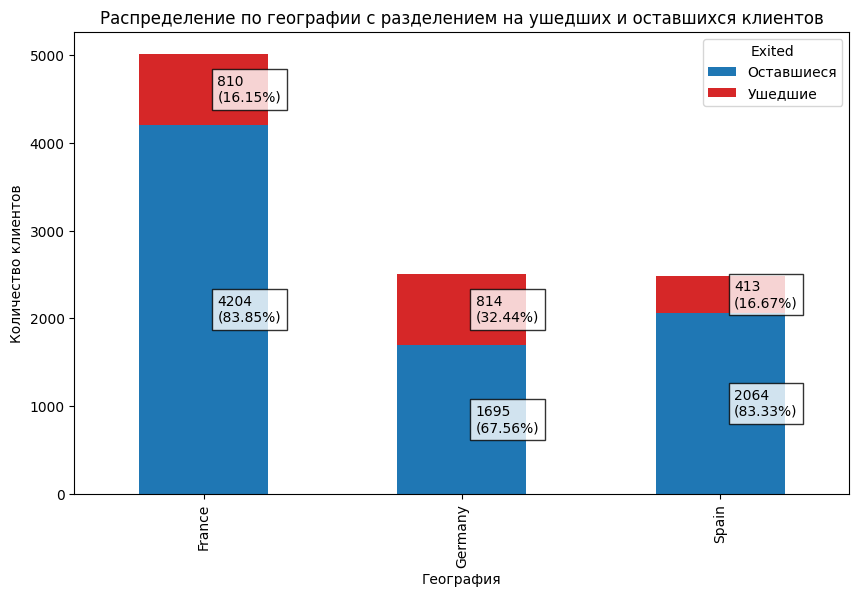
**Вывод:** Гистограмма показывает, что из 1000 клиентов за рассматриваемый период осталось 79,63%, ушло - 20,37%.

****

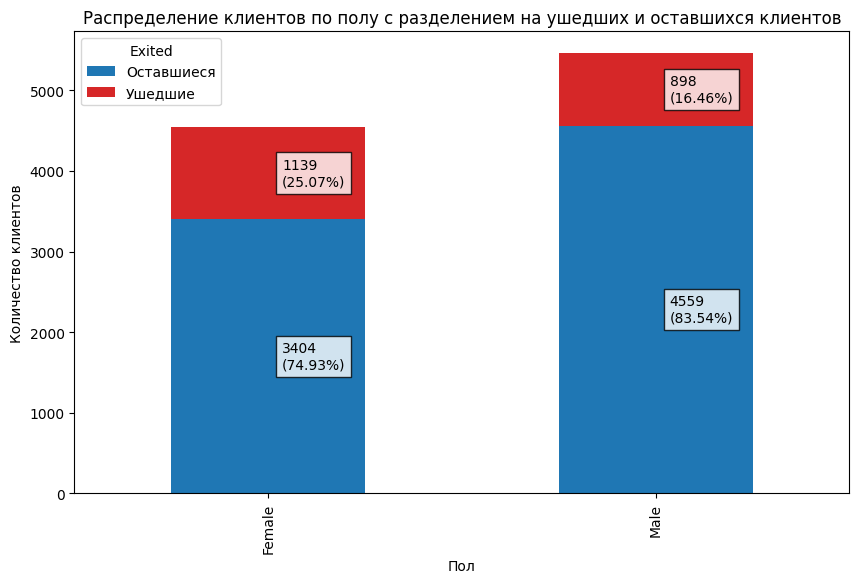
**Вывод:** Видим, что большее количество действующих клиентов банка находятся в возрасте примерно от 28 до 45 лет, а существенная доля тех, кто принял решение уйти из банка, находятся в возрасте примерно от 40 до 50 лет.

****

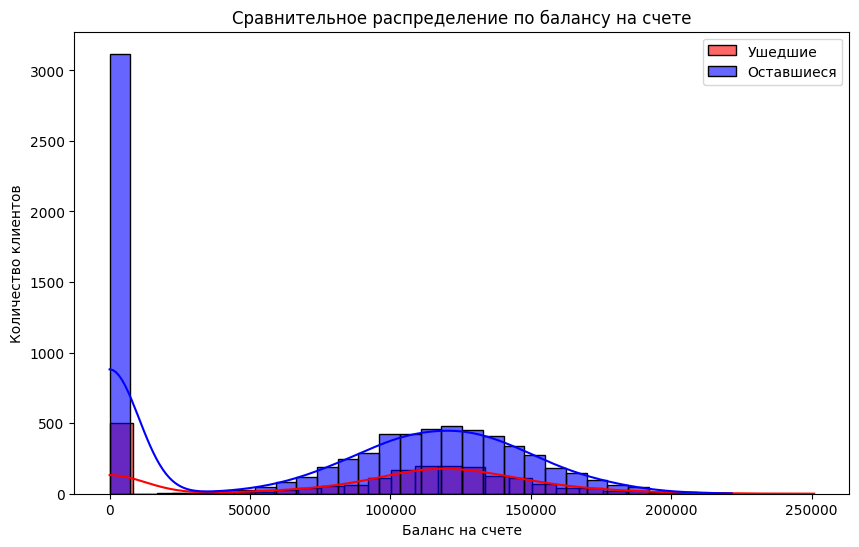
**Вывод:** Видим, что средний рейтинг ушедших в среднем похож, но все же немного ниже.

****

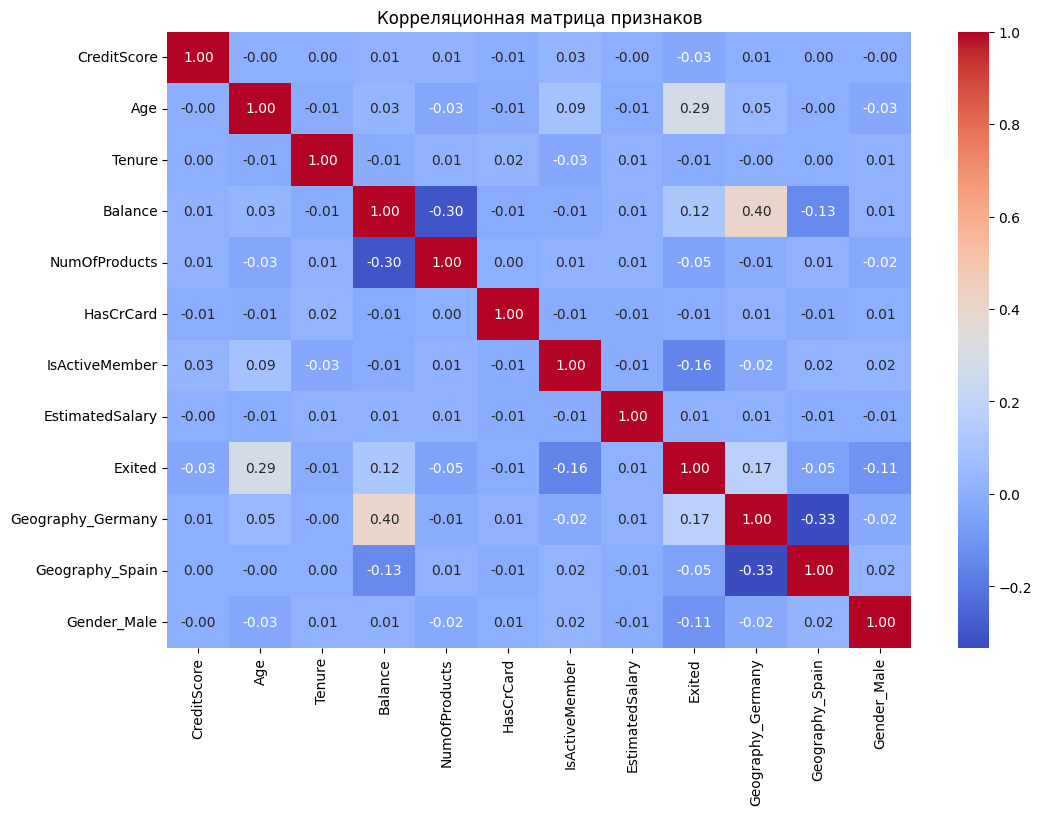
**Вывод:** Гистограмма показывает, что большее количество клиентов располагается во Франции и доля ушедших клиентов там составляет ~16%. Количество клиентов как в Германии, так и в Испании в два раза меньше, чем во Франции, но отток разный ( ~16% в Испании и ~32% в Германии).

****

**Вывод:** Гистограмма демонстрирует, что общая доля клиентов мужчин больше и что они реже уходят.

****

**Вывод:** Распределения баланса ушедших и оставшихся очень похожи.

****

**Вывод** (Анализ корреляционной матрицы, корреляция с целевой переменной Exited):

Age: Корреляция с Exited 0.29 -> c возрастом клиенты могут чаще покидать банк.

Geography\_Germany: Корреляция с Exited 0.17 -> клиенты из Германии чаще уходят.

Balance: Корреляция с Exited 0.12 -> клиенты с высоким балансом реже уходят, но корреляция слабая.

IsActiveMember: Корреляция с Exited -0.16 (отрицательная) -> активные клиенты реже покидают банк.

Другие признаки имеют еще меньшую корреляцию с целевой переменной -> имеют еще меньшую значимость в контексте оттока клиентов.

Нет высоких корреляций (больше 0.8) между признаками -> низкий уровень мультиколлинеарности -> можно более точно оценить важность каждого признака.

# 

# **3.3. Рекомендации на основе анализа EDA**

На основе анализа корреляционной матрицы можно сделать следующие выводы и рекомендации:

Возраст клиентов: Поскольку возраст является значимым фактором, рекомендуется проводить целевые кампании для удержания клиентов определенного возраста. Возможно, стоит предложить специальные продукты и услуги для различных возрастных групп.

Географическое расположение: Высокий уровень оттока клиентов в Германии (~32%) по сравнению с другими регионами, такими как Франция (~16%) и Испания (~16%). Возможно, стоит провести локальные опросы или исследования для выявления специфических особенностей, которые могут влиять на уровень оттока в Германии, и разработать отдельные стратегии удержания для клиентов из этого региона.

Пол клиентов: Так как мужчины составляют большую долю клиентов и имеют меньшую вероятность ухода по сравнению с женщинами, можно рассмотреть возможность создания специализированных предложений и акций для женщин, чтобы улучшить их удержание. Анализ причин более высокой склонности к уходу среди женщин может помочь в разработке эффективных стратегий по удержанию.

Баланс на счету: Несмотря на слабую корреляцию, клиенты с высоким балансом менее склонны к оттоку. Возможно, стоит разработать программы лояльности или привилегии для клиентов с высоким балансом, чтобы еще больше уменьшить вероятность их ухода.

Активность клиентов: Активные клиенты реже уходят. Это указывает на важность вовлеченности клиентов. Можно предложить программы поощрения за активное использование банковских услуг.

Другие факторы: Такие признаки как EstimatedSalary, Tenure, NumOfProducts и HasCrCard имеют низкую корреляцию с оттоком клиентов и, возможно, менее важны для прогнозирования оттока.

**4. Построение модели логистической регрессии**

Целью построения модели логистической регрессии в данной задаче является прогнозирование вероятности ухода клиентов из банка (оттока).

**4.1. Результаты построение модели логистической регрессии прогнозирования оттока**

Используемые метрики в данной оценке:

Основные метрики для классов:

Precision (точность): Доля правильно предсказанных положительных примеров от всех примеров, предсказанных как положительные.

Recall (полнота): Доля правильно предсказанных положительных примеров от всех реальных(фактических) положительных примеров.

F1-score: Гармоническое среднее между точностью и полнотой.

Support: Количество истинных экземпляров каждого класса.

Сводные метрики:

Accuracy (точность): Доля правильно классифицированных примеров от общего числа примеров.

Macro avg (среднее по классам): Среднее арифметическое для каждой метрики, взятое по всем классам.

Weighted avg (взвешенное среднее): Среднее значение метрик с учетом количества экземпляров каждого класса.

ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): Метрика для оценки качества бинарных классификаторов. В данной задаче позволяет понять, насколько хорошо модель способна различать между двумя классами (ушедшие и оставшиеся клиенты).

**Оценка на тренировочных данных (80%):**

Отчет о классификации (Classification Report):

**precision recall f1-score support**

**0 0.80 0.98 0.88 6356**

**1 0.38 0.06 0.10 1644**

**accuracy 0.79 8000**

**macro avg 0.59 0.52 0.49 8000**

**weighted avg 0.71 0.79 0.72 8000**

**ROC-AUC: 0.67**

**Оценка на тестовых данных (20%):**

Отчет о классификации (Classification Report):

**precision recall f1-score support**

**0 0.81 0.98 0.89 1607**

**1 0.45 0.07 0.12 393**

**accuracy 0.80 2000**

**macro avg 0.63 0.53 0.51 2000**

**weighted avg 0.74 0.80 0.74 2000**

**ROC-AUC: 0.67**

# **4.2. Комплексная оценка модели логистической регрессии и процесса обучения**

Точность модели (Accuracy) показывает, что модель правильно классифицирует примерно 79% тренировочных и 80% тестовых примеров. Это свидетельствует о том, что модель способна в целом адекватно различать между клиентами, которые остаются и которые уходят.

Из отчета классификации видно, что модель хорошо справляется с предсказанием класса 0 (не ушёл), т.к. высокие значения Recall (полнота) и F1-score (Гармоническое среднее между точностью и полнотой). Однако, точность предсказания класса 1 (ушёл) значительно ниже, т.е. модель менее эффективна в идентификации клиентов, которые уходят.

ROC-AUC(Receiver Operating Characteristic - Area Under the Curve) показывает способность модели различать между положительным и отрицательным классами. Значения 0.67 указывают на то, что модель лучше, чем случайное угадывание (0.5), но не близко к идеальной модели (1.0).

В целом, с учетом того, что в данных присутствует естественный (по логике целей бизнеса) дисбаланс, модель в комплексе хорошая, но при дальнейшем улучшении стоит обратить внимание на эффективность прогноза категории клиентов, которые уходят.

**5. Выявление инсайтов с помощью модели случайного леса и рекомендации**

**Модель случайного леса показывает ключевые факторы, влияющие на отток клиентов:**

Возраст / Age 0.236450

Предполагаемая зарплата / EstimatedSalary 0.146307

Кредитный рейтинг / CreditScore 0.142262

Баланс / Balance 0.141439

Количество продуктов / NumOfProducts 0.132736

Стаж / Tenure 0.082074

Активный ли / IsActiveMember 0.041785

Германия / Geography\_Germany 0.025716

Мужчина / Gender\_Male 0.018655

Наличие кредитной карты HasCrCard 0.018226

Испания Geography\_Spain 0.014349

**Вывод по наиболее значимым признакам:**

Возраст (Age) (0.236): Возраст клиента является самым значимым фактором, указывая на то, что клиенты в определенных возрастных группах могут быть более склонны к оттоку.

Финансовые показатели EstimatedSalary (0.146), CreditScore (0.142), Balance (0.141): Клиенты с различными уровнями предполагаемой зарплаты и баланса на счете имеют разные уровни склонности к оттоку. Кредитный рейтинг клиента также важен, что может указывать на связь с финансовой стабильностью клиента. Это может свидетельствовать о том, что финансовое состояние клиента связано с его лояльностью.

Количество продуктов, которыми пользуется клиент (NumOfProducts) (0.133): Клиенты, использующие больше продуктов, менее склонны к уходу.

Стаж клиента в банке (Tenure) (0.082): Клиенты, долгое время обслуживающиеся в банке, могут быть более лояльными.

**6. Обобщающие выводы анализа оттока клиентов банка**

**6.1. Корреляции признаков:**

Возраст (Age): Имеет положительную корреляцию с оттоком клиентов (0.285), что указывает на то, что клиенты старшего возраста могут чаще покидать банк.

Географическое расположение (Geography\_Germany): Корреляция с оттоком составляет 0.173, что свидетельствует о более высоком уровне оттока среди клиентов из Германии по сравнению с другими регионами.

Баланс на счету (Balance): Слабая положительная корреляция (0.119) указывает на то, что клиенты с высоким балансом менее склонны к уходу, хотя эта связь слабая. Активность клиента (IsActiveMember): Отрицательная корреляция (-0.156) указывает на то, что активные клиенты реже покидают банк.

**6.2. Анализ модели:**

Точность модели (Accuracy): Модель правильно классифицирует около 79% тренировочных и 80% тестовых примеров.

Качество предсказаний: Модель хорошо предсказывает класс 0 (не ушедшие клиенты), но плохо справляется с предсказанием класса 1 (ушедшие клиенты), что подтверждается низкими значениями Recall и F1-score для класса 1.

ROC-AUC: Значение 0.67 указывает на то, что модель имеет умеренную способность различать между классами, но далеко от идеальной.

**6.3. Важность признаков:**

Наиболее важные признаки: Возраст (0.236), зарплата (0.146), кредитный рейтинг (0.142) и баланс на счете (0.141).

Признаки с наименьшим влиянием: Пол клиента (0.019), наличие кредитной карты (0.018), географическое расположение в Испании (0.014).

**7. Общие рекомендации**

**7.1. Целевая работа с возрастными группами:**

Могут помочь специализированные программы для удержания клиентов в возрастных группах, которые наиболее склонны к оттоку.

**7.2. Финансовые стимулы и программы:**

Внимательное наблюдение за клиентами с низким кредитным рейтингом или балансом на счете и предоставление им персонализированных предложений могут изменить их лояльность.

Внедрение программ лояльности или специальных предложений для клиентов с высоким балансом, а также для клиентов с хорошим кредитным рейтингом и высокой зарплатой может помочь удерживать финансово стабильных клиентов.

**7.3. Развитие программ лояльности и увеличение числа продуктов:**

Поощрения клиентов к использованию большего количества продуктов банка (кросс-продажи и пакеты услуг) может уменьшить их склонность к уходу.

**7.4. Исследование географических факторов:**

Уделение большего внимания регионам, таким как Германия и Испания, и улучшение качества обслуживания клиентов в этих областях, проведение дополнительных исследований в регионах, могут помочь выявить конкретные причины оттока и облегчит разработку стратегии удержания на локальном уровне.

**7.5. Повышение активности клиентов:**

Внедрение программ, стимулирующих активность клиентов, например бонусы за регулярное использование услуг банка, могут оказать благоприятное влияние на активность клиентов.

**8. Заключение**

Проект продемонстрировал значимость анализа данных для понимания факторов оттока клиентов банка и построения предсказательной модели. Модель логистической регрессии, несмотря на некоторые ограничения в предсказании оттока, предоставляет ценные инсайты и может быть использована для разработки стратегий удержания клиентов. Внедрение предложенных рекомендаций позволит банку улучшить удержание клиентов, повысить их лояльность и, как следствие, увеличить доходы.