Системы искусственного интеллекта (ПОМС)

ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ ТЕЛЕКОМ-ОПЕРАТОРА

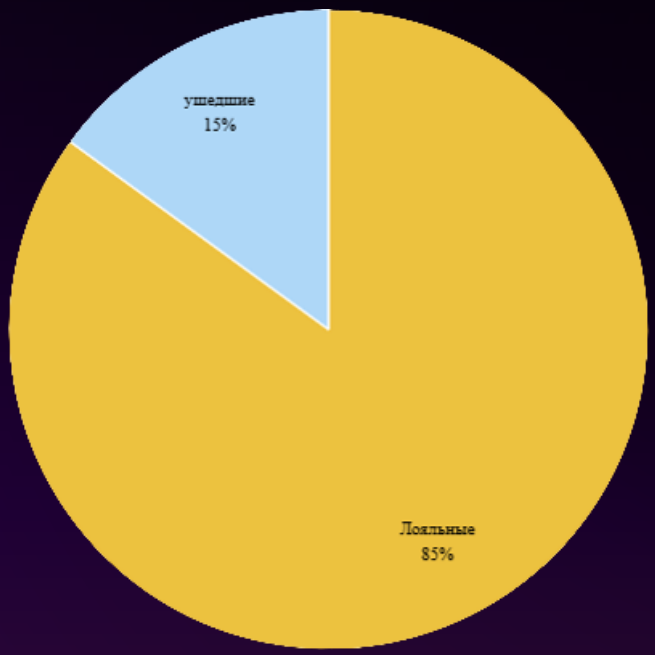
Команда 7• Курсовой проект

КОМАНДА 7

- Ошлаков Константин (Работы над кодом, создание отчета)
- Зырянов Иван (Работы над кодом, создание презентации)

АКТУАЛЬНОСТЬ ПРОБЛЕМЫ И ИСХОДНЫЕ ВЫЗОВЫ

Высокий отток клиентов – **значительные финансовые потери для телеком-оператора**



- Значительный дисбаланс классов: **~85% лояльных клиентов против ~15%** ушедших (в обучающей выборке). Это затрудняет обучение моделей и требует специальных подходов
- Исходное кодирование категориальных признаков **могло вносить ложный порядок и искажать зависимости для моделей.**
- Высокая корреляция между некоторыми признаками (например, Total day minutes и Total day charge) **приводила к избыточности данных и могла влиять на стабильность моделей.**
- Базовый набор признаков мог не полностью отражать сложные факторы, **влияющие на решение клиента уйти.**

ОБЗОР ДАТАСЕТА И FEATURE ENGINEERING

• **Набор данных телеком-оператора:** информация о клиентах, разделенная на обучающую (80%) и тестовую (20%) выборки.

• **Исходные признаки (19)**: включали демографию (штат, планы), статистику использования услуг (минуты, звонки).

• **Целевая переменная**: Churn (бинарная: True – клиент ушел, False – остался)

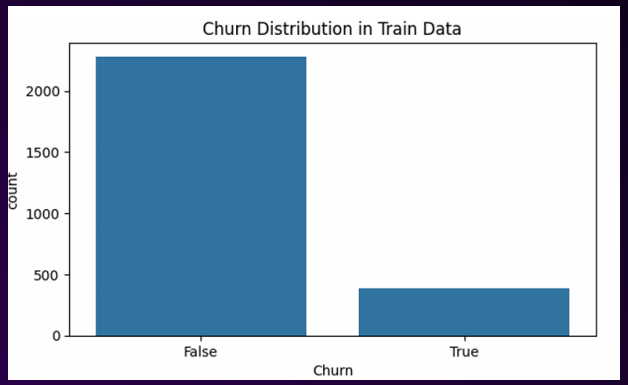
• **Инжиниринг признаков** для повышения информативности модели:

-**Агрегация штатов в Region**: уменьшение количества категорий, обобщение географического фактора

-**Total minutes:** общая активность клиента по длительности разговоров

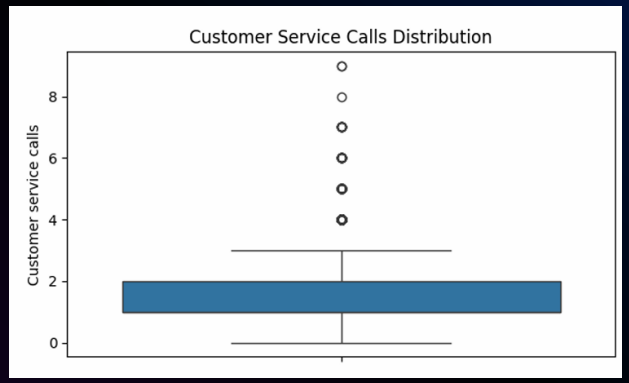
-**Avg call duration**: средняя вовлеченность клиента в один разговор

ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ (EDA) - КЛЮЧЕВЫЕ НАХОДКИ

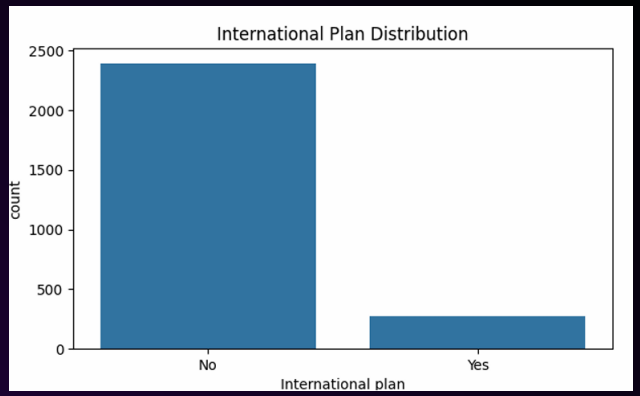


Выраженный дисбаланс классов (~15% отток)

Клиенты с оттоком в среднем чаще обращаются в поддержку (>3 раз)



Большинство без межд. плана, но наличие плана сильно связано с оттоком (показал дальнейший анализ)



Ключевые этапы построения моделей

1. Устранение дисбаланса классов с помощью SMOTE:

- **Проблема:** Существенный перевес лояльных клиентов (~85%) мешал моделям эффективно обнаруживать класс оттока (~15%).
- **Решение:** Применен SMOTE для генерации синтетических данных класса оттока, **сбалансировав обучающую выборку** и повысив значимость редкого класса.

2. Оптимальное кодирование категориальных признаков (One-Hot Encoding):

- **Проблема:** Исходный LabelEncoder мог вносить ложный числовой порядок в категориальные признаки (Region, International plan, Voice mail plan).
- **Решение:** Использован One-Hot Encoding, **преобразующий категории в независимые бинарные признаки**, что более корректно для моделей и улучшает интерпретируемость.

3. Инжиниринг признаков (Feature Engineering):

- **Цель:** Выявить скрытые закономерности и обогатить данные для моделей.
- **Решение:** Созданы новые признаки (Total minutes, Avg call duration), **отражающие общую активность клиента и среднюю вовлеченность в разговор**, что потенциально повышает предсказательную силу.

4. Оптимизация гиперпараметров моделей (GridSearchCV):

- **Цель:** Найти наилучшую конфигурацию для каждой модели (RF, XGBoost) и снизить риск переобучения.
- **Решение:** GridSearchCV **автоматически протестировал различные комбинации ключевых параметров** (глубина деревьев, скорость обучения и др.), выбрав **оптимальные на основе F1-score** при кросс-валидации.

5. Выбор моделей:

- **Обоснование:** Выбраны **мощные ансамблевые методы**, хорошо зарекомендовавшие себя на табличных данных.
- **Преимущества:** Random Forest обеспечивает **стабильность**, XGBoost – часто **лидирует по точности и эффективности** работы с бустингом градиента.

РЕЗУЛЬТАТЫ МОДЕЛЕЙ: ЭВОЛЮЦИЯ МЕТРИК

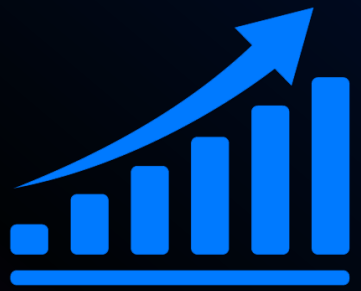
- **SMOTE – решающий шаг для Random Forest:** Учет дисбаланса кардинально повысил Recall (с 0.38 до ~0.85), сделав модель способной находить уходящих клиентов.
- **One-Hot Encoding + Новые признаки – буст для XGBoost:** Корректное кодирование и новые данные позволили XGBoost значительно улучшить Precision (с 0.68 до 0.84) и итоговый F1-score.
- **GridSearchCV стабилизировал XGBoost:** Оптимизация параметров помогла закрепить высокие метрики XGBoost (F1=0.84) и повысить общую точность до 0.95.

Random Forest: Изменение метрик

Метрика	До улучшений (test)	После SMOTE (test)	После One-Hot (test)	После новых признаков и GridSearch (test)
Precision (False)	0.91	0.97	0.97	0.97
Recall (False)	1.00	0.92	0.93	0.93
F1-score (False)	0.95	0.95	0.95	0.95
Precision (True)	0.95	0.63	0.66	0.66
Recall (True)	0.38	0.85	0.82	0.82
F1-score (True)	0.54	0.73	0.73	0.73
Accuracy (общее)	0.91	0.91	0.91	0.91
Macro avg F1-score	0.75	0.84	0.84	0.84

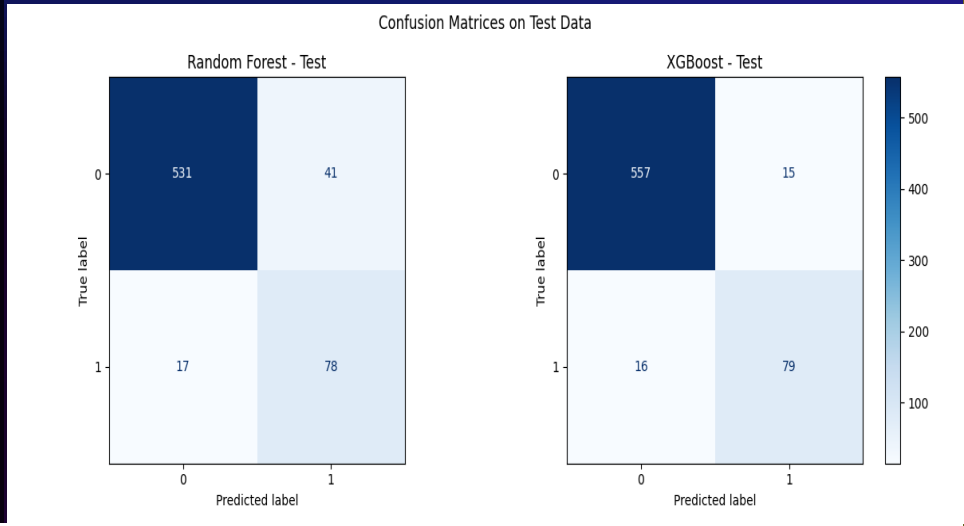
XGBoost: Изменение метрик

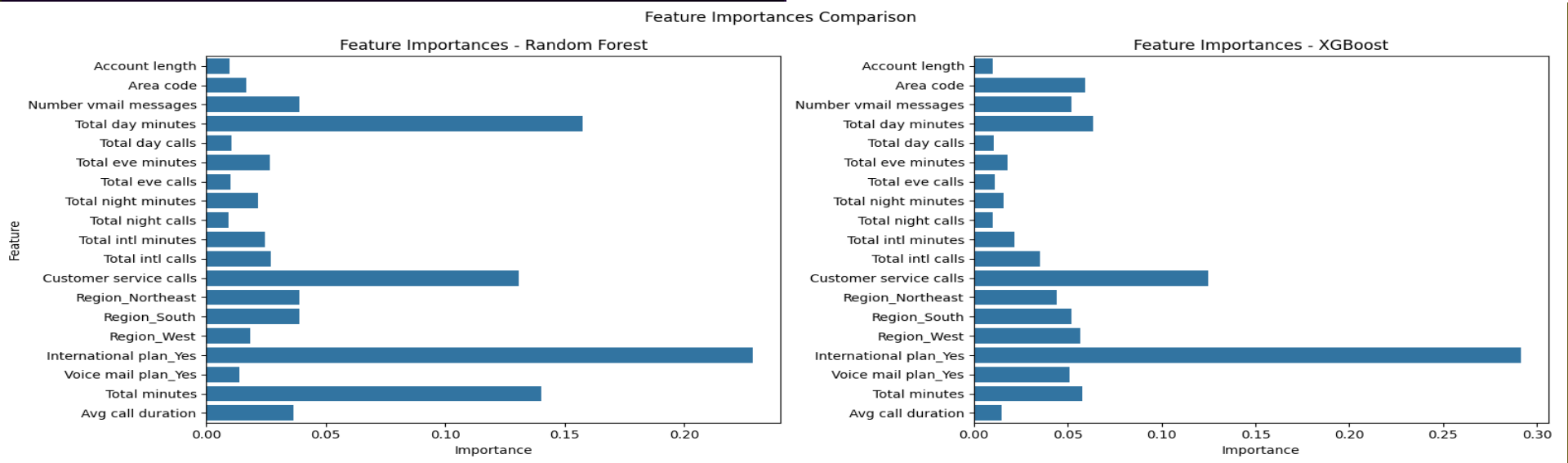
Метрика	До улучшений (test)	После SMOTE (test)	После One-Hot (test)	После новых признаков и GridSearch (test)
Precision (False)	0.97	0.98	0.98	0.97
Recall (False)	0.93	0.93	0.93	0.97
F1-score (False)	0.95	0.95	0.95	0.97
Precision (True)	0.68	0.68	0.84	0.84
Recall (True)	0.84	0.86	0.83	0.83
F1-score (True)	0.75	0.76	0.84	0.84
Accuracy (общее)	0.92	0.92	0.92	0.95
Macro avg F1-score	0.85	0.86	0.90	0.90



ВИЗУАЛЬНЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ ФИНАЛЬНЫХ МОДЕЛЕЙ (ТЕСТ)

- **XGBoost:** Эффективно находит уходящих (79 из 95), **минимизируя упущенных клиентов (FN=16)**. При этом количество ложных срабатываний (FP=15) невелико.
- **Random Forest:** Также хорошо находит уходящих (78 из 95), но **генерирует больше ложных срабатываний (FP=41)** по сравнению с XGBoost.
- **Общие ТОП-факторы:** Обе модели согласны, что Total day minutes, Customer service calls и International plan_Yes – **ключевые драйверы оттока**.
- **Уникальные акценты:** XGBoost также выделяет важность *Area code*, в то время как RF обращает внимание на *Total minutes*.





КЛЮЧЕВЫЕ ДОСТИЖЕНИЯ ПРОЕКТА

- ✓ **Достигнут высокий Recall (порядка 0.83) для XGBoost**, что позволяет эффективно выявлять клиентов, склонных к оттоку.
- ✓ **Повышен F1-score для XGBoost до 0.84**, демонстрируя хороший баланс между точностью и полнотой обнаружения оттока.
- ✓ **Модель Random Forest значительно улучшена** (F1-score вырос с 0.54 до 0.73) и показывает **стабильные результаты** с низким переобучением.
- ✓ Применение **SMOTE, One-Hot Encoding и инжиниринга признаков** доказало свою эффективность в улучшении качества моделей.
- ✓ Определены **ключевые факторы, влияющие на отток клиентов** - *International plan, Customer service calls, Total day minutes.*
- ✓ Получены **перспективные модели** для дальнейшего тестирования и возможного внедрения

ВЫВОДЫ И НАПРАВЛЕНИЯ ДЛЯ РАЗВИТИЯ

Выводы

- **XGBoost продемонстрировал наилучшие результаты** по F1-score (0.84%) и Recall (0.86%) для прогнозирования оттока, эффективно используя улучшения в предобработке и оптимизацию гиперпараметров.
- **Random Forest** после всех улучшений также достиг **хорошей стабильности и сравнимых метрик** (F1-score 0.73%).
- Методы устранения дисбаланса (SMOTE), корректное кодирование категорий (One-Hot) и создание новых признаков **существенно повысили качество прогнозирования**.
- Оценка переобучения показала, что **модели демонстрируют приемлемую обобщающую способность** на тестовых данных.

Направления для развития

- **Дальнейший инжиниринг признаков**: исследование и создание более сложных признаков.
- **Тестирование других моделей**: попробовать другие алгоритмы (например, LightGBM, CatBoost, нейронные сети).
- **Более глубокая оптимизация гиперпараметров**: использование более продвинутых техник (например, Bayesian Optimization).
- **Разработка системы мониторинга модели**: отслеживание качества модели во времени при ее реальном использовании.

СПАСИБО ЗА ВНИМАНИЕ!

- Qr на Git репозиторий