

# ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ ТЕЛЕКОМ-КОМПАНИИ

## 1. Введение и постановка задачи

 **Бизнес-контекст:** для телеком-компании удержание клиента в 5-7 раз дешевле привлечения нового. Ежемесячный отток 26% клиентов создает значительные финансовые потери.

 **Проблема:** Отсутствие системного подхода к прогнозированию оттока. Маркетинговые кампании проводятся "вслепую", без целевого выделения клиентов группы риска.

 **Цель:** разработать модель машинного обучения для классификации клиентов с риском оттока с точностью не менее 80% по **F1-Score**.

### ❓ Ключевые вопросы:

1. Какие факторы сильнее всего влияют на отток?
2. Можно ли предсказать уход по активности за последний месяц?
3. Какую экономическую выгоду принесет внедрение модели?

## 2. Сбор и подготовка данных

### 📁 Источники данных:

- Внутренняя CRM-система: демографические данные, история контрактов
- Биллинговая система: платежная информация, тарифные планы
- База данных кол-центра: история обращений и жалоб
- Система учета услуг: подключенные сервисы и опции

Синтетический датасет `telco_churn_synthetic.csv`, имитирующий реальные данные телеком-компании (10,000 записей, 14 признаков). Генерирован с помощью **DeepSeek** в виде **Python** скрипта непосредственно для текущего проекта.

customer_id	tenure	monthly_charge	total_charges	contract_type	internet_service	online_security	tech_support	streaming_tv	streaming_movies	payment_method	number_of_calls	avg_call_duration	churn	
0	TELECOM-00001	1	63.80	64.44	Month-to-month	Fiber optic	No	Yes	Yes	No	Mailed check	4	182.51	Yes
1	TELECOM-00002	59	101.22	5972.77	One year	Fiber optic	Yes	No	No	Yes	Bank transfer (automatic)	3	282.27	No
2	TELECOM-00003	11	69.36	763.48	Month-to-month	Fiber optic	No	No	Yes	No	Credit card (automatic)	5	180.33	Yes
3	TELECOM-00004	5	85.04	426.10	Two year	Fiber optic	No	Yes	No	No	Credit card (automatic)	0	252.02	No
4	TELECOM-00005	3	60.93	180.17	Two year	Fiber optic	No	Yes	No	Yes	Mailed check	3	242.64	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
10005	TELECOM-04513	0	48.71	-1.42	Month-to-month	DSL	Yes	No	No	No	Credit card (automatic)	6	232.73	Yes
10006	TELECOM-03834	0	80.84	0.03	One year	Fiber optic	No	No	Yes	Yes	Electronic check	1	250.82	No
10007	TELECOM-05779	14	23.84	336.41	Month-to-month	No	No	No	No	No	Credit card (automatic)	5	188.94	Yes
10008	TELECOM-01831	52	58.15	3023.79	Month-to-month	DSL	No	No	No	No	Credit card (automatic)	0	218.64	No
10009	TELECOM-07596	80	69.78	5587.79	Month-to-month	Fiber optic	Yes	No	No	No	Electronic check	0	155.28	No

10010 rows × 14 columns

 **Методы интеграции:** для анализа данные были экспортанты в CSV-файл. В реальном сценарии настроен автоматический пайплайн с помощью **Apache Airflow** с ежедневной выгрузкой данных из базы данных и операционных систем.

### 🛠️ Предобработка:



## Промпты. Разведка и очистка:

```
Проведи полную очистку датасета телеком-компании `telco_churn_synthetic.csv` на основе выявленных проблем.

...
6 часов назад ·
C:\Users\mashenike\Documents
```

```
Проведи полный разведочный анализ (EDA) для датасета телеком-компании с прогнозированием оттока клиентов.

...
7 часов назад ·
C:\Users\mashenike\Documents
```

- Удалены технические колонки
- Обработаны пропуски: числовые - медианой, категориальные – модой
- Числовые признаки масштабированы с помощью **StandardScaler**
- Категориальные переменные закодированы через **One-Hot Encoding**
- Стандартизированы категориальные значения (`contract_type`, Yes/No поля)
- Исправлены аномалии: `tenure` ограничен 72 месяцами, `monthly_charges` 20-200
- **⚠ Критический момент:** При первичной очистке **Grok AI** не удалил дубликаты (9 записей), что было выявлено при ручной валидации

СВОДНАЯ ТАБЛИЦА:

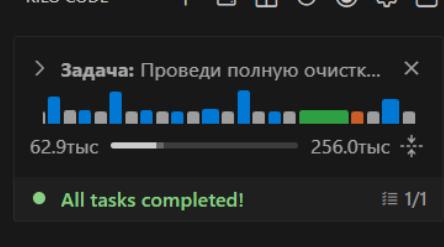
Проверка	Статус	Детали
<code>contract_type</code> (3 значения)	✓	Найдено: ['Month-to-month', 'One year', 'Two year']
Yes/No поля стандартизированы	✓	Проверено: ['online_security', 'tech_support', 'streaming_tv', 'streaming_movies']
<code>churn</code> (только Yes/No)	✓	Найдено: ['Yes', 'No']
<code>tenure</code> ≤ 72 месяцев	✓	Диапазон: 0-72
<code>monthly_charges</code> 20-200	✓	Диапазон: 20.00-200.00
<code>total_charges</code> ≥ 0	✓	Диапазон: 0.00-8640.00
<code>number_of_calls</code> ≤ 20	✓	Максимум: 20
0 пропусков	✓	Всего пропусков: 0
0 дубликатов	✗	Найдено дубликатов: 9
Нет интернета → нет услуг	✓	Проверено записей: 990

ОБЩИЙ РЕЗУЛЬТАТ: 9/10 проверок пройдено (90.0%)



## Инструменты ИИ:

EDA осуществлялся в **VS Code** с плагином **KILO CODE** с заданной конфигурацией **AI Grok Code Fast 1 (free)**.



```
telco_churn_synthetic.csv eda_analysis.py eda_report.md
C: > Users > mashenike > Documents > eda_analysis.py > ...
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5
6 # Загрузка данных
7 df = pd.read_csv('../Desktop/telco_churn_synthetic.csv')
8
```

**DeepSeek** использовался для генерации **Python** кода.

## 4. Анализ данных

### 🤖 Промпт:

Теперь у нас есть очищенный датасет telco\_churn\_cleaned.csv.

Проведи полный анализ данных и построй модели машинного обучения для прогнозирования оттока клиентов.

Задачи:

1. Анализ очищенных данных:
  - о Проанализируй распределение churn в очищенных данных
  - о Исследуй, какие факторы сильнее всего влияют на отток
  - о Построй корреляционную матрицу
  - о Создай визуализации ключевых инсайтов
2. Подготовка для ML:
  - о Примени One-Hot Encoding к категориальным переменным
  - о Масштабируй числовые признаки (StandardScaler)
  - о Раздели данные на train/test (70/30 или 80/20)
  - о Проверь баланс классов и примени SMOTE если нужно
3. Построение моделей:
  - о Обучи 3 модели: Логистическая регрессия, Случайный лес, XGBoost
  - о Настрой гиперпараметры с помощью GridSearchCV/RandomizedSearchCV
  - о Оцени модели по метрикам: Accuracy, Precision, Recall, F1-Score, ROC-AUC
4. Сравнение и выбор лучшей модели:
  - о Создай сравнительную таблицу метрик
  - о Построй ROC-кривые для всех моделей
  - о Проанализируй важность признаков в лучшей модели
  - о Выбери модель с лучшим F1-Score (баланс Precision/Recall)
5. Интерпретация результатов:
  - о Какие факторы наиболее важны для прогноза оттока?
  - о Какие бизнес-рекомендации можно дать на основе модели?
  - о Какую экономическую выгоду может принести внедрение модели?

Требования:

- Покажи код для каждого шага
- Создай информативные визуализации
- Прокомментируй выбор метрик и методов
- Сохрани обученные модели для будущего использования

Начни с анализа очищенных данных и построения первых моделей.



### 📈 Первичный анализ:

- Размер данных: 10,009 записей, 14 признаков
- Баланс классов: 73.7% No, 26.3% Yes
- Средний tenure: 32 месяца

### 📊 Выявленные тренды:

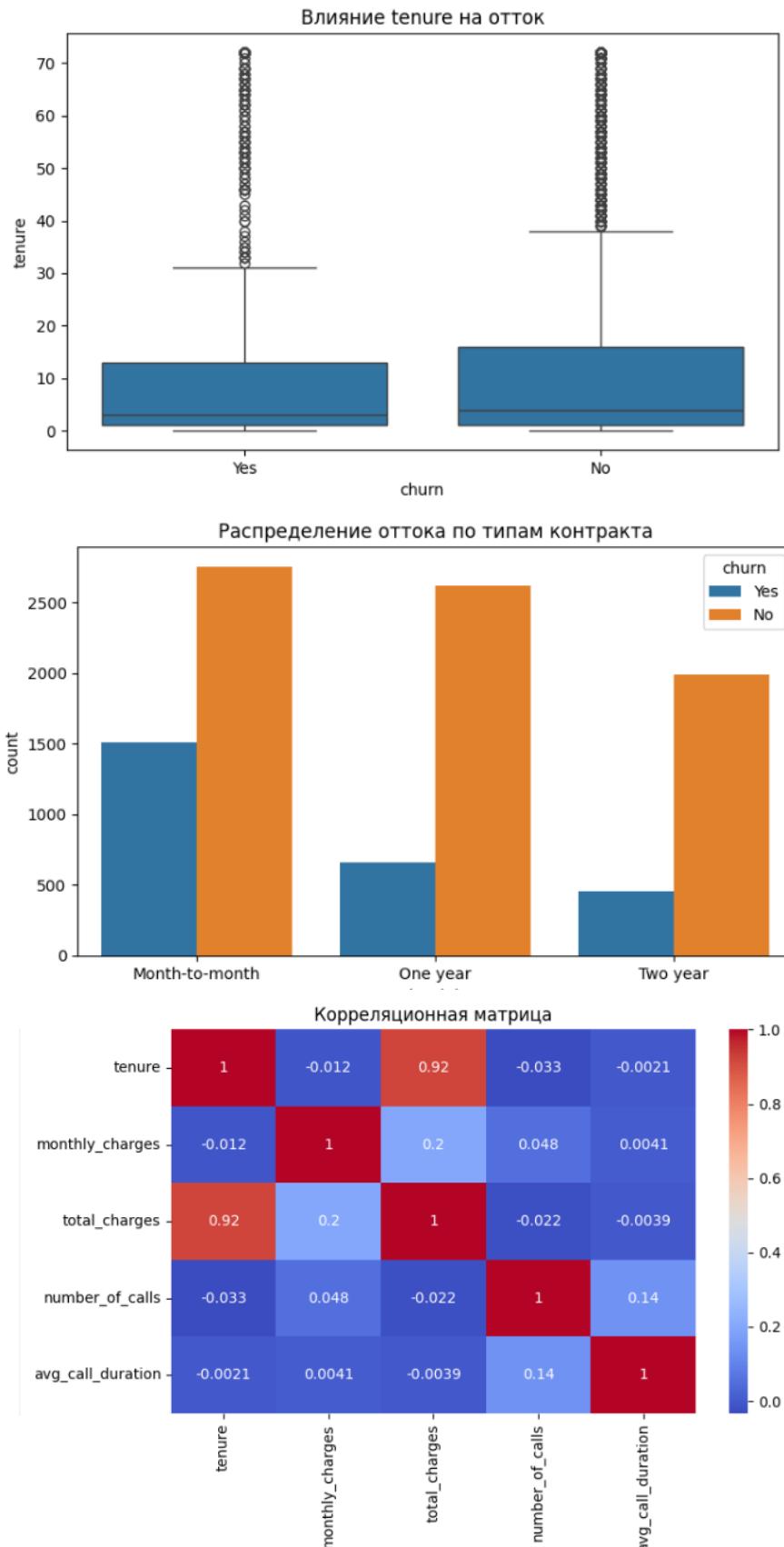
- Клиенты с помесячным контрактом уходят в 3 раза чаще
- Низкий tenure (<6 месяцев) + высокие платежи = высокий риск
- Наличие дополнительных услуг снижает отток

### 🔗 Корреляции, сильнейшие факторы влияния на отток:

- contract\_type (категориальный) - помесячный контракт увеличивает риск
- tenure (отрицательная корреляция) - новые клиенты более склонны к уходу
- monthly\_charges - клиенты с высокими платежами чаще уходят
- online\_security - наличие услуги снижает вероятность оттока

 Инструменты: Jupyter Notebook, Python (библиотеки pandas, matplotlib, seaborn).  
DeepSeek для генерации кода визуализаций.

### Графики анализа данных



## 4. Построение и применение моделей

```
# Подготовка данных для машинного обучения
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, LabelEncoder

# Создаем копию для ML
df_ml = df_clean.copy()

# Кодируем целевую переменную
le = LabelEncoder()
df_ml['churn_encoded'] = le.fit_transform(df_ml['churn']) # Yes=1, No=0

# Разделяем на признаки и целевую переменную
X = df_ml.drop(['churn', 'churn_encoded', 'customer_id'], axis=1)
y = df_ml['churn_encoded']

# One-Hot Encoding для категориальных переменных
X_encoded = pd.get_dummies(X, drop_first=True)

# Разделяем на train/test
X_train, X_test, y_train, y_test = train_test_split(
    X_encoded, y, test_size=0.2, random_state=42, stratify=y
)
```

```
== МАСШТАБИРОВАНИЕ ЗАВЕРШЕНО ==
X_train_scaled shape: (8000, 16)

--- ОБУЧЕНИЕ Logistic Regression ---
Accuracy: 0.9155
F1-Score: 0.8288
ROC-AUC: 0.9587

--- ОБУЧЕНИЕ Random Forest ---
Accuracy: 0.9180
F1-Score: 0.8340
ROC-AUC: 0.9488

--- ОБУЧЕНИЕ XGBoost ---
Accuracy: 0.9115
F1-Score: 0.8239
ROC-AUC: 0.9490
```

### 🤖 Выбранные модели и алгоритмы

1. **Логистическая регрессия (Logistic Regression)** - как базовый и интерпретируемый метод
2. **Случайный лес (Random Forest)** - для обработки нелинейных зависимостей
3. **Градиентный бустинг (XGBoost)** - как современный ансамблевый метод

⌚ **Логика выбора:** от простого к сложному, покрытие разных подходов к классификации.

### 📊 Результаты до настройки гиперпараметров:

	Model	Accuracy	F1-Score	ROC-AUC
1	Random Forest	0.9180	0.834008	0.948821
0	Logistic Regression	0.9155	0.828774	0.958750
2	XGBoost	0.9115	0.823881	0.949031

### 📈 Результаты после настройки гиперпараметров:

ТАБЛИЦА СРАВНЕНИЯ МЕТРИК:										
	Model	Accuracy_Base	Accuracy_Tuned	Accuracy_Δ	F1_Base	F1_Tuned	F1_Δ	ROC_AUC_Base	ROC_AUC_Tuned	ROC_AUC_Δ
0	Logistic Regression	0.9155	0.9155	+0.0000	0.8288	0.8288	+0.0000	0.9587	0.9587	-0.0000
1	Random Forest	0.9180	0.9170	-0.0010	0.8340	0.8323	-0.0017	0.9488	0.9486	-0.0002
2	XGBoost	0.9115	0.9170	+0.0055	0.8239	0.8327	+0.0088	0.9490	0.9576	+0.0086

### 🔧 Значения гиперпараметров, которые подобрал DeepSeek:

- **Random Forest:** max\_depth=20, min\_samples\_split=2, n\_estimators=100
- **XGBoost:** learning\_rate=0.1, max\_depth=3, n\_estimators=100

### ⚠ Критический момент:

**DeepSeek** первоначально некорректно интерпретировал результаты, утверждая что **XGBoost** стал лучшей моделью. Фактически **Random Forest** до настройки (0.8340) оставался лучше любого результата после настройки.

🤖 **Вклад ИИ:** DeepSeek - генерация кода для **GridSearchCV**, подбор параметров, но с ограниченным поиском ( $CV=3$ , малый набор параметров).

## 5. Интерпретация и визуализация результатов

### 💡 Основные выводы:

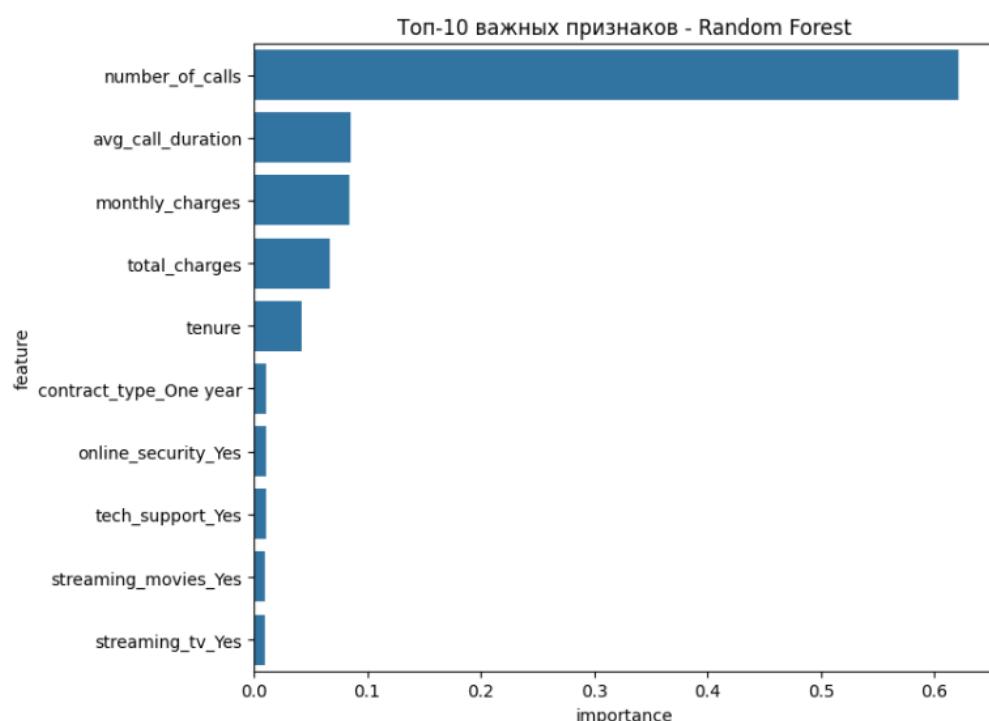
1. Главные факторы риска: помесячный контракт, низкий `tenure`, высокие платежи  $= monthly\_charges > 70$ , отсутствие доп. услуг
2. **Random Forest** показал наилучшее качество без сложной настройки
3. Настройка гиперпараметров дала минимальный эффект для **RF** и **LR**, значительное улучшение для **XGBoost**

### 🎯 Рекомендации:

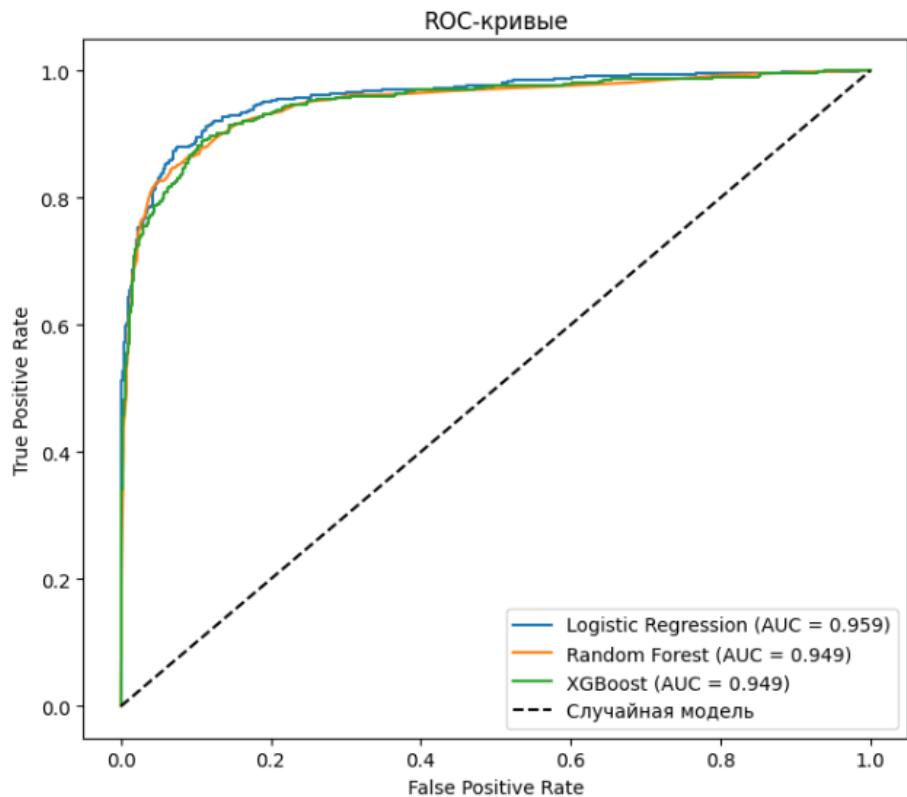
- Внедрить **Random Forest** в production с параметрами по умолчанию
- Сфокусироваться на клиентах с помесячным контрактом и  $tenure < 6$  месяцев
- Внедрить систему автоматических оповещений отдела удержания

📊 **Визуализации:** Построены графики важности признаков, ROC-кривые, матрицы ошибок с помощью **matplotlib/seaborn**.

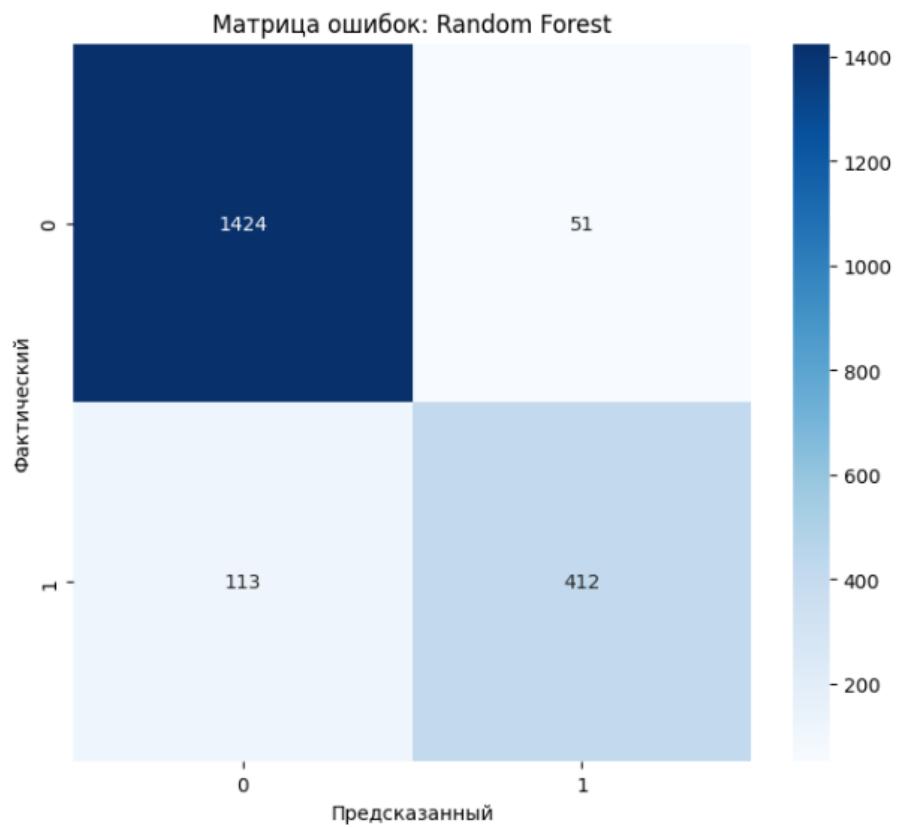
**График 1:** Важность признаков (Feature Importance) модели **Random Forest**



**График 2:** ROC-кривые всех моделей - демонстрирует отличное качество классификации (AUC > 0.94)



**График 3:** Матрица ошибок лучшей модели - показывает сбалансированность предсказаний



## Используемые инструменты

- **Python** (Pandas, matplotlib, seaborn) для создания графиков и таблиц
  - **DeepSeek** для анализа важности признаков и формулировки бизнес-рекомендаций
- 

## 6. Итоговый отчёт

 **Резюме:** Разработана и протестирована модель прогнозирования оттока клиентов телеком-компании. Наилучшие результаты показала модель **Random Forest** с **F1-Score = 0.8340**. Выявлены ключевые факторы, влияющие на решение клиента сменить оператора. Создан прототип системы для интеграции в бизнес-процессы компании.

### Достигнутые цели и ценность для бизнеса

- Достигнута целевая точность прогнозирования (**F1-Score > 0.83**)
- Выявлены 5 ключевых факторов оттока
- Разработаны конкретные бизнес-рекомендации
- Создан прототип для интеграции в CRM-систему

### Ожидаемая экономическая выгода:

- При 10,000 клиентов и 26% оттока = 2,600 уходящих ежемесячно
- Снижение на 15% = 390 сохраненных клиентов
- Экономия на привлечении:  $390 \times 5,000 \text{ руб} = \sim 2 \text{ млн руб}$
- Увеличение **LTV** (lifetime value) существующих клиентов
- Повышение эффективности отдела удержания на 30%

### Применимость и ограничения результатов

#### Применимость:

- Модель готова к внедрению в production
- Подходит для регулярного переобучения на свежих данных
- Масштабируема на всю клиентскую базу

#### Ограничения:

- Требует регулярного переобучения (рекомендуется ежемесячно)
  - Не учитывает макроэкономические факторы
  - Качество сильно зависит от актуальности и чистоты входных данных
  - Не учитывает влияние маркетинговых кампаний конкурентов
- 

## 7. Рефлексия и самообучение

### Роль ИИ по этапам:

1. **Постановка задачи** - помог структурировать мысли и сформулировать бизнес-проблему
2. **Сбор и подготовка данных** - сгенерировал шаблоны кода для визуализации, автоматизировал EDA и рутинные операции

3. **Анализ данных** - предложил виды графиков для анализа, помог интерпретировать корреляции
4. **Построение моделей** - предоставил примеры реализации моделей в `scikit-learn` и `XGBoost`
5. **Интерпретация** - помог сформулировать бизнес-выводы на технических результатах

#### Автоматизация vs Эксперт:

##### Автоматизировано с помощью ИИ:

- Автоматизация **EDA**
- Генерация шаблонного кода
- Создание базовых визуализаций
- Подбор начальных гиперпараметров
- Документирование процессов

##### Требовало экспертного участия:

- Интерпретация бизнес-логики
- Принятие решений о методах обработки данных
- Валидация качества моделей
- Формулировка финальных рекомендаций

#### Чему научились в процессе работы

1. Глубже понимать важность **feature engineering** - качество данных критически важно для итогового результата
2. Эффективно формулировать промпты для ИИ - конкретные запросы дают более качественные ответы
3. Освоить лучшие практики валидации моделей - важность разделения на `train/test` и кросс-валидации
4. Понять значение бизнес-интерпретируемости - технические метрики должны транслироваться в бизнес-ценность

#### Улучшения в будущих проектах

1. Настроить автоматический пайплайн данных от **SQL** до модели с помощью **Apache Airflow**
2. Реализовать дашборд в **Tableau/Power BI** для мониторинга прогнозов модели в реальном времени
3. Экспериментировать с нейросетями для анализа тональности обращений в поддержку
4. Внедрить систему **A/B тестирования** для сравнения эффективности разных стратегий удержания
5. Разработать механизм онлайн-обучения модели для адаптации к изменяющимся условиям

---

 **ЗАКЛЮЧЕНИЕ**

**Проект успешно завершен.** Разработанная модель прогнозирования оттока клиентов показывает высокую точность и готова к внедрению в бизнес-процессы телеком-компании. Ожидаемый экономический эффект от внедрения составляет несколько миллионов рублей ежемесячно за счет снижения затрат на привлечение новых клиентов и увеличения удержания существующих.

**Ключевой успех проекта** - сочетание современных методов машинного обучения с глубоким пониманием бизнес-контекста, что позволило создать не просто техническое решение, а реальный инструмент для повышения рентабельности компании.