# ПРОГНОЗИРОВАНИЕ ОТТОКА КЛИЕНТОВ ТЕЛЕКОММУНИКАЦИОННОЙ КОМПАНИИ

CUSTOMER CHURN PREDICTION IN TELECOMMUNICATIONS COMPANY

**УДК 004.89**

**Чечнев А.А.,** студент, alexey-che@mail.ru, Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана, кафедра «Системы обработки информации и управления»

**Chechnev A.A**, student, alexey-che@mail.ru, Russia, 105005, Moscow, Bauman Moscow State Technical University, Department of Information Processing and Management Systems

**Аннотация**

В статье рассматриваются вопросы применимости методов машинного обучения для прогнозирования оттока клиентов. Клиенты – это основа успеха любого бизнеса, поэтому компании заинтересованы в создании систем для более гибкой работы с ними. С проблемой оттока клиентов сталкиваются абсолютно все компании, предоставляющие товары или услуги. В связи с высокой стоимостью привлечения новых пользователей и ростом конкурентоспособности рынка увеличивается значимость удержания клиентов. Машинное обучение нашло применение во многих задачах прогнозирования и классификации и способно преждевременно выявить клиентов, склонных перейти к одному из конкурентов.

Основной целью статьи является исследование современных методов машинного обучения и построение прогнозирующей модели оттока клиентов в телекоммуникационной компании.

**Annotation**

This article examines the applicability of machine learning methods for predicting customer churn. Customers are the basis for the success of any business, so companies are interested in creating systems for more flexible work with them. All companies that provide goods or services face the problem of customer churn. Due to the high cost of attracting new users and the growing competitiveness of the market, the importance of customer retention increases. Machine learning has found application in many forecasting and classification tasks and is able to prematurely detect customers exposed to care.

The main purpose of the article is to study modern methods of machine learning and build a predictive model of customer churn in a telecommunications company.

**Ключевые слова:** отток клиентов, машинное обучение, прогнозирование, классификация, Случайный лес

**Key words:** customer churn, machine learning, prediction, classification, Random Forest

## Введение

Сегодняшний мир целиком и полностью состоит из информации. Люди веками собирали данные в книгах, летописях. Теперь огромные массивы данных хранятся на электронных носителях, которые надо анализировать, структурировать и извлекать ценную информацию. Простой человеческий мозг не способен обработать весь объем информации и найти все связи между данными, которые в большинстве случаев неявно выражены. Исследования в области построения прогнозирующих моделей ведутся специалистами разных областей, и уже существует огромное множество подходов и методов к реализации моделей анализа данных.

Удержание клиентов – одна из ключевых задач бизнеса. Ежедневно рынок пополняется все новыми конкурентами, которые готовы предпринять любые усилия, чтобы отвоевать часть клиентуры у других. Потребителю такое

положение только на руку: он может выбирать самые выгодные условия покупки, требовать скидок и дополнительных опций. А если что-то не устроит, он совершенно спокойно уйдет, и продавец лишится покупателя. Поэтому, если ваше направление бизнеса предполагает не разовые сделки, а длительные отношения с потребителями, а их потеря весьма критична, то хорошо бы заранее озаботиться удержанием существующих клиентов.

В большинстве компаний, где собирается большое количество данных, пригодных для анализа, применяются техники машинного обучения и интеллектуального анализа данных. Одним из популярных примеров использования машинного обучения в реальной жизни - это прогнозирование оттока клиентов. Прогнозированием и управлением оттоком клиентов занимаются телекоммуникационные компании, банки, страховые компании и другие. В условиях жёсткой конкуренции, прогнозирование оттока клиентов с целью удержания становится одним из самых важных направлений в современном бизнесе.

## Основная часть

### 1. Понятие оттока клиентов

Для любого предприятия, предоставляющего товары или услуги, важным аспектом является клиентская база. В нее поступают новые клиенты, какое-то время активно пользуются услугами и через некоторый промежуток времени перестают. Весь это промежуток называется «Жизненный цикл клиента» (англ. Customer Lifecycle) — это термин, описывающий этапы, которые проходит клиент, когда узнает о продукте, принимает решение о покупке, платит, использует и становится лояльным потребителем, и в конечном счете перестает пользоваться продуктами по тем или иным причинам. Соответственно понятие “отток” описывает завершающую стадию жизненного цикла клиента, а для

бизнеса это означает, что клиент перестал приносить прибыль и вообще какую-либо пользу.

Отток клиентов выражается в сокращении клиентской базы и снижении показателей выручки[1]. Ранее проблема оттока клиентов компенсировалась за счет привлечения новых клиентов, но в современном мире темпы роста постоянно ускоряются, рынок заполнен огромным количеством конкурентов, что приводит к постоянному повышению требований клиентов к продукции и оказываемым услугам. В связи с этим становится более затратным и трудоемким процесс привлечения новых клиентов, что в свою очередь усиливает значимость в удержании уже существующей клиентской базы.

Понятие отток клиентов является не строгим понятием и не описывает этап расторжения. Существует три основных подхода к определению оттока клиентов:

1. Оттоком является отказ клиентов приобретать продукцию и услуги компании, расторжение клиентами договоров на обслуживание и т.д.
2. Оттоком признается прекращение использования клиентом продукции компании или какой-либо из оказываемых услуг.
3. Оттоком является ситуация, в которой клиент вначале активно пользуется продукцией компании или услугой, а затем сводит пользование до минимума.

## 2. Почему важно удерживать клиентов?

Привлечение новых клиентов, для борьбы с оттоком является более сложной и ресурсоемкой процедурой, требующей огромного количества средств на рекламу, социальные сети и иные каналы поиска клиентов, нежели увеличение лояльности существующих клиентов. Также операционная эффективность работы с существующими клиентами выше, чем с новыми, по нескольким причинам:

1. Существующие клиенты более лояльны и готовы заплатить больше, если их устраивает уровень представляемого сервиса
2. Существующие клиенты приобретут ваш новый продукт или услугу с вероятностью 70-80%, для новых клиентов этот показатель не превышает 20-30%
3. Не нужно тратить маркетинговые бюджеты на привлечение клиентов, чтобы получить прибыль от существующих клиентов

## 3. Анализ оттока клиентов

Первым этапом работы с оттоком клиентов является преждевременное выявление группы людей, предрасположенных к прекращению пользования услугами. Зная заранее о возможности ухода клиента можно применить стратегические решения.

Основная цель анализа оттока клиентов состоит в создании списка контрактов (клиентов), которые с большой вероятностью в ближайшем будущем будут прерваны. Существуют разные подходы к анализу оттока клиентов. Большинство из них основано на интеллектуальном анализе данных, показывающих в современном мире высокую эффективность и производительность.

Для прогнозирования ухода клиентов используются различные математические модели, среди них — логистическая регрессия[2], деревья решений, метод ближайших соседей, SVM[3], случайный лес. Поскольку компании имеют ограниченные возможности по связям с клиентами, важной задачей также является определение апостериорной вероятности классификации. Зная список наиболее вероятных отказчиков, компания может построить оптимальную стратегию проведения акций.

## 4. Анализ оттока клиентов телекоммуникационной компании

### 4.1. Выбор программных средств

Для исследования различных методов построения моделей для предсказания оттока клиентов был выбран язык python, поскольку он является бесплатным, и имеет библиотеку машинного обучения sklearn[4], в которой реализованы современные модели машинного обучения.

### 4.2.Предобработка данных

Проанализируем набор данных по оттоку клиентов телекоммуникационной компании. Структура данных представлена на рис. 2.

| | state | account length | area code | phone number | international plan | voice mail plan | number vmail messages | total day minutes | total day calls | total day charge | total eve minutes | total eve calls | total eve charge | total night minutes | total night calls | total night charge | total intl minutes | total intl calls | total intl charge | customer service calls | churn |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | KS | 128 | 415 | 382-4657 | no | yes | 25 | 265.1 | 110 | 45.07 | 197.4 | 99 | 16.78 | 244.7 | 91 | 11.01 | 10.0 | 3 | 2.70 | 1 | False |
| 1 | OH | 107 | 415 | 371-7191 | no | yes | 26 | 161.6 | 123 | 27.47 | 195.5 | 103 | 16.62 | 254.4 | 103 | 11.45 | 13.7 | 3 | 3.70 | 1 | False |
| 2 | NJ | 137 | 415 | 358-1921 | no | no | 0 | 243.4 | 114 | 41.38 | 121.2 | 110 | 10.30 | 162.6 | 104 | 7.32 | 12.2 | 5 | 3.29 | 0 | False |

Рисунок 2 – структура набора данных

Некоторые колонки представлены в виде текстовых значений, поэтому преобразуем их в числовой формат, подставив на каждые уникальные значения столбца числовое значение. Также избавимся от колонки с номером телефона, поскольку она никак не влияет на отток клиента.

| | state | account_length | area_code | international_plan | voice_mail_plan | number_vmail_messages | total_day_minutes | total_day_calls | total_day_charge | total_eve_minutes | t |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 16 | 128 | 415 | 0 | 1 | 25 | 265.1 | 110 | 45.07 | 197.4 | |
| 1 | 35 | 107 | 415 | 0 | 1 | 26 | 161.6 | 123 | 27.47 | 195.5 | |
| 2 | 31 | 137 | 415 | 0 | 0 | 0 | 243.4 | 114 | 41.38 | 121.2 | |

| total_eve_minutes | total_eve_calls | total_eve_charge | total_night_minutes | total_night_calls | total_night_charge | total_intl_minutes | total_intl_calls | total_intl_charge | customer_service_calls | churn |
|---|---|---|---|---|---|---|---|---|---|---|
| 197.4 | 99 | 16.78 | 244.7 | 91 | 11.01 | 10.0 | 3 | 2.70 | 1 | False |
| 195.5 | 103 | 16.62 | 254.4 | 103 | 11.45 | 13.7 | 3 | 3.70 | 1 | False |
| 121.2 | 110 | 10.30 | 162.6 | 104 | 7.32 | 12.2 | 5 | 3.29 | 0 | False |

Рисунок 3 – Данные после преобразования в числовой формат

### 4.3.Описательные статистики

Посмотрим на сбалансированность классов целевой переменной на рис. 4.

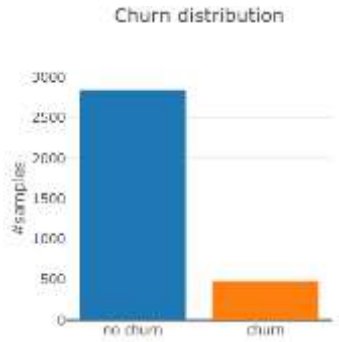


Рисунок 4 – распределение целевой переменной

Мы видим, что у нас явно больше образцов для клиентов без оттока, чем для клиентов с оттоком. Таким образом, мы имеем классовый дисбаланс для целевой переменной, который может привести к прогностическим моделям, которые смещены в сторону большинства (т.е. отсутствие оттока). Для решения этой проблемы, при построении и оценке моделей воспользуемся метрикой f1-score.

Посмотрим на распределение оттока по штатам, чтобы увидеть, насколько это состояние влияет на нашу цель на рис.5:

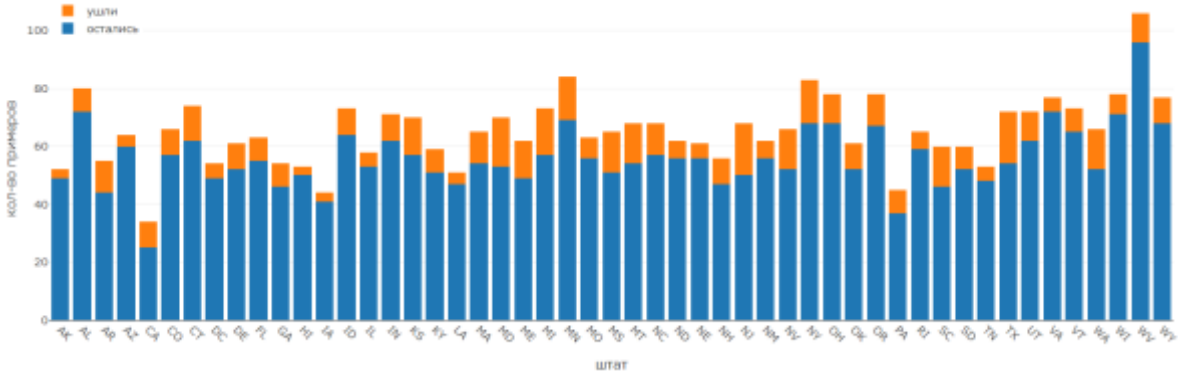


Рисунок 5 – Распределение оттока клиентов по штатам

Мы можем видеть, что некоторые штаты имеют меньшую долю клиентов с оттоком, как AK, HI, IA, а некоторые имеют более высокую, такие как WA, MD и TX. Это показывает, что мы должны включить состояние в наш дальнейший анализ, потому что это может помочь предсказать, если клиент собирается уйти.

Построим графики Ящик с усами, для остальных переменных на рис.6. Анализируя распределение количества сообщений голосовой почты ("number_vmail_messages"), мы видим, что присутствуют выбросы для клиентов с оттоком, однако большинство из них не отправляют сообщения голосовой почты. Клиенты, которые не оставались, как правило, отправляют больше сообщений голосовой почты.

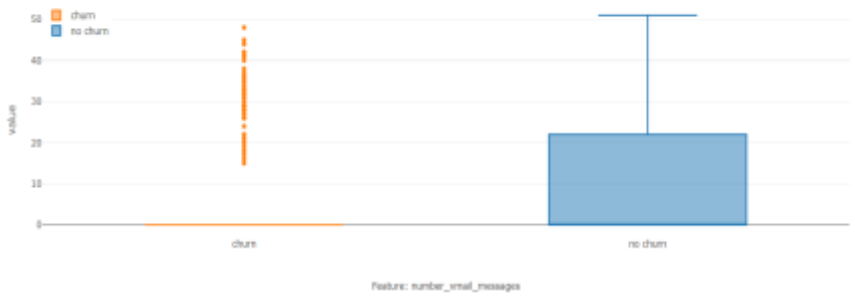


Рисунок 6 – Распределение числа голосовых сообщений

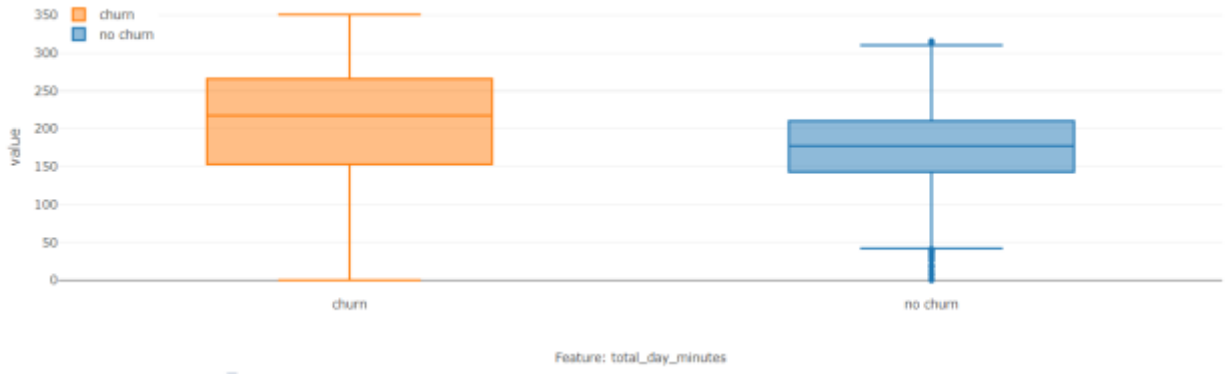


Рисунок 7 – Среднее время разговора в день

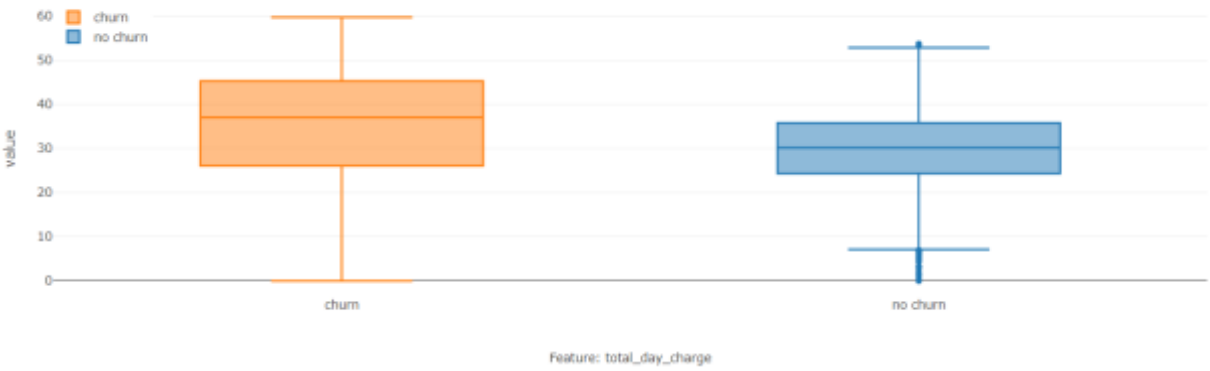


Рисунок 8 – Стоимость общей дневной платы

Из графиков на рис. 7 и рис. 8 мы видим, что медиана длительности разговора в день и общая дневная плата для ушедших клиентов выше, чем у остальных.




Рисунок 8 – Общее количество международных звонков

Если посмотреть на общее количество международных звонков ("total_intl_calls") на рис.8, то график показывает, что обе группы клиентов делают одинаковое количество международных звонков, ушедшие клиенты, как правило, делают более длинные звонки, поскольку медиана оттока клиентов для общего количества международных минут выше, чем для оставшихся клиентов.

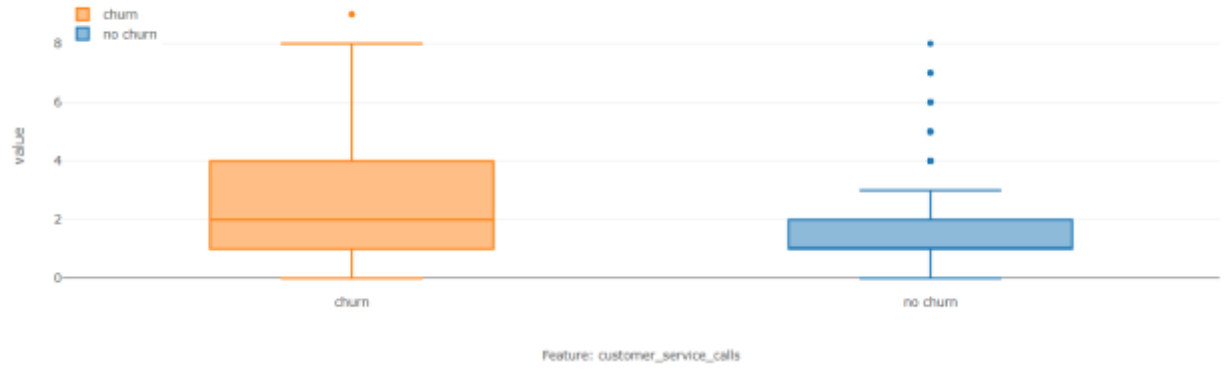


Рисунок 9 – количество обращений в службу поддержки

Наконец, график количества обращений в службу поддержки клиентов на рис.9 показывает, что ушедшие клиенты имеют более высокую медиану и более высокую дисперсию для обращений в службу поддержки клиентов.

### 4.4.Корреляционный анализ

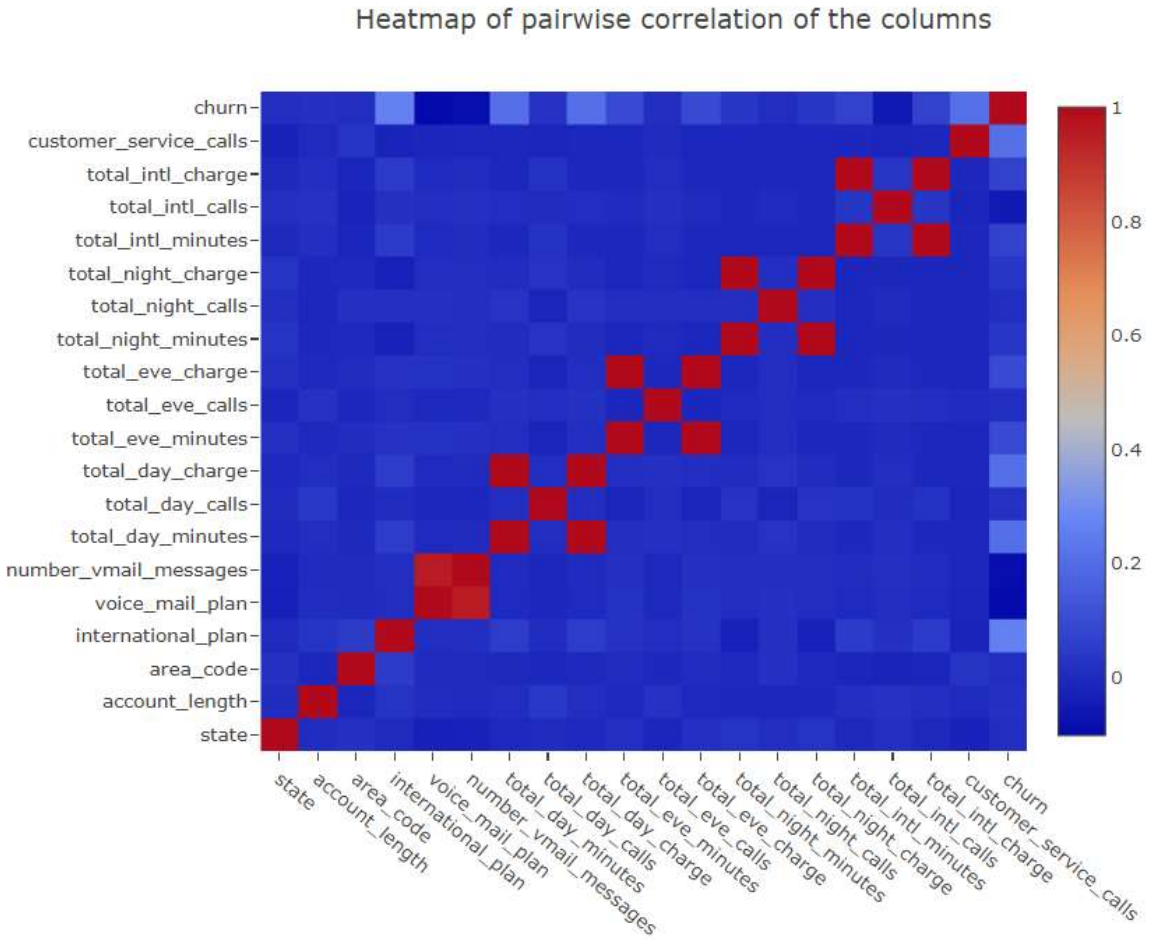


Рисунок 10 – Тепловая карта попарных корреляций

Из рис. 10 мы видим высокую корреляцию между планом голосовой почты и количеством сообщений голосовой почты. Вполне логично, что клиенты с планом голосовой почты также отправляют больше сообщений голосовой почты.

Однако международный план лишь незначительно коррелирует с общим международным протоколом и международным сбором.

Как видно также из нашего предыдущего анализа, общий дневной заряд и общее количество дневных минут очень сильно коррелируют. Вероятно, эта

телекоммуникационная компания взимает плату за минуту. То же самое поведение можно наблюдать и для вечерних, ночных и международных звонков.

Наибольшую корреляцию с целевой переменной имеют международный план, общая стоимость в день, общее время разговоров в день и количество звонков в службу поддержки клиентов

Высокие попарные корреляции будем считать дубликатами и уберем из рассмотрения.

### 4.5.Построение предиктивных моделей

Построим и сравним результаты работы классификационных моделей, таких как Logistic Regression, K-Nearest Neighbors, Машину опорных векторов, Дерево решений, Случайный лес[5]. Для сравнения качества работы каждой модели будем использовать метрики accuracy, precision, recall, f1-score[6].

Изобразим результаты работы моделей на ROC-кривой на рис. 11.

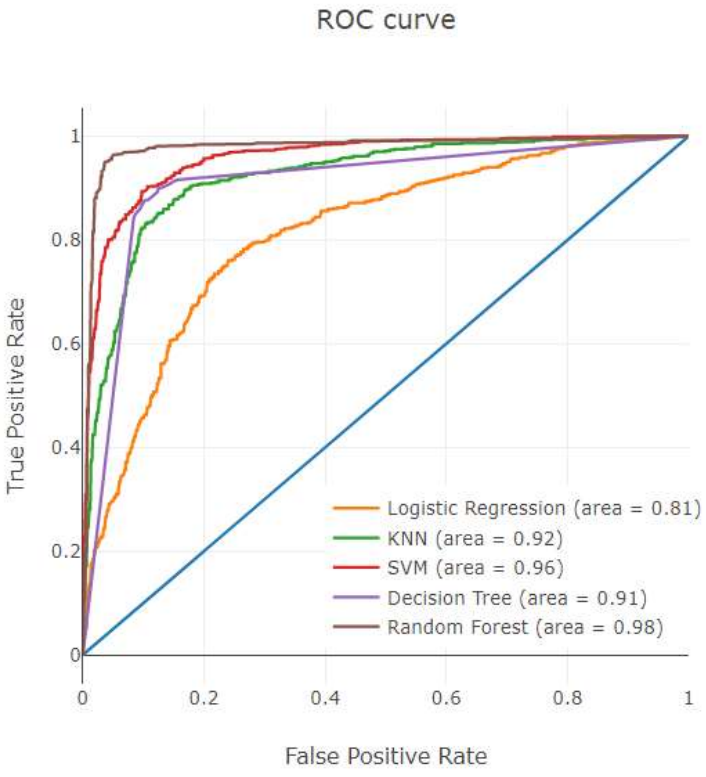


Рисунок 11 – ROC кривая для различных моделей

Мы видим, что классификатор Случайный лес имеет самое высокое значение AUC (0.98), за которым следуют машина опорных векторов (0.96), KNN и Дерево решений с, 0.92 и 0.91 соответственно. Логистическая регрессия работает хуже с AUC всего 0,81

На рис.12 изображены различные метрики для детального сравнения полученных моделей.

| | accuracy | precision | recall | f1 | auc |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.760 | 0.321 | 0.788 | 0.456 | 0.812 |
| **KNN** | 0.859 | 0.470 | 0.824 | 0.598 | 0.922 |
| **SVM** | 0.901 | 0.583 | 0.788 | 0.670 | 0.958 |
| **Decision Tree** | 0.889 | 0.549 | 0.729 | 0.626 | 0.908 |
| **Random Forest** | 0.955 | 0.831 | 0.812 | 0.821 | 0.978 |

Рисунок 12 – Таблица метрик для каждой модели

Результаты показывают, что логистическая регрессия имеет очень низкое значение точности (около 0.32). Это означает, что если модель выдаст уход клиента, то она будет верна в 32% случаев.

Кроме того, по точности дерево решений работает лучше (0.89), чем KNN, который достигает только 0,86. Но в измерении AUC KNN превосходит дерево решений.

KNN даже имеет самое высокое значение полноты, но достигает плохих результатов по точности, что также делает его F1-score вторым худшим из всех классификаторов.

SVM имеет второй самый высокий балл точности (0.58), но далеко отстает от лучшего балла точности 0.83 случайного леса.

Случайный лес показывает налучшие результаты по всем метрикам, кроме полноты (0.81), однако незначительно отстает от лучшего (0.82).

Таким образом наилучшей моделью является случайный лес.

## Заключение

Таким образом в данной статье было описано понятие оттока клиентов и показана важность преждевременного выявления критичных клиентов. Также были выполнен анализ данных телекоммуникационной компании по оттоку клиентов, проведены описательные статистики и проведен корреляционный анализ. Были построены модели классификации оттока клиентов и приведен сравнительный анализ результатов функционирования моделей.

## Список использованных источников

1. Нестеров А.К. Отток клиентов // Энциклопедия Нестеровых [Электронный ресурс]. Режим доступа: http://odiplom.ru/lab/ottok-klientov.html Дата обращения: 08.04.2021
2. Толстихин И. О. // Применение логических алгоритмов классификации в задаче прогнозирования оттока клиентов – 2008 Дата обращения: 10.04.2021
3. Обзор самых популярных алгоритмов машинного обучения. [Электронный ресурс]. Режим доступа: https://tproger.ru/translations/top-machine-learning-algorithms Дата обращения: 28.03.2021
4. Sklearn документация. [Электронный ресурс]. Режим доступа: https://www.sklearn.org/documentation.html Дата обращения: 27.04.2021
5. Пол, Махеш. (2005). Классификатор случайных лесов для классификации дистанционного зондирования. Международный журнал дистанционного зондирования Дата обращения: 27.04.2021
6. Мухамедиев Р. И., Мухамедиева Е. Л., Кучин Я. И.: Таксономия методов машинного обучения и оценка качества классификации и обучаемости. [Электронный ресурс]. Режим доступа: https://cloudofscience.ru/sites/default/files/pdf/CoS_2_359.pdf. Дата обращения: 27.04.2021

**Literature**


1. Nesterov A.K. Client churn // Encyclopedia of the Nesterovs [Electronic resource]. Access mode: http://odiplom.ru/lab/ottok-klientov.html Date of request: 08.04.2021
2. Tolstikhin I. O. / / Application of logical classification algorithms in the problem of forecasting customer outflow-2008 Date of request: 10.04.2021
3. Overview of the most popular machine learning algorithms. [Electronic resource]. Access mode: https://tproger.ru/translations/top-machine-learning-algorithms Date of request: 28.03.2021
4. Sklearn documentation. [Electronic resource]. Access mode: https://www.sklearn.org/documentation.html Дата обращения: 27.04.2021
5. Pal, Mahesh. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing Date of request: 27.04.2021
6. Mukhamediev R. I., Mukhamediev E. L., Kuchin Ya. I.: Taxonomy of machine learning methods and assessment of the quality of classification and learning ability. [Electronic resource]. Access mode: https://cloudofscience.ru/sites/default/files/pdf/CoS_2_359.pdf. Date of request: 27.04.2021