Оглавление

[Введение](#_heading=h.30j0zll) 2

[1. Постановка задачи](#_heading=h.1fob9te) 3

[1.1. Описание бизнес-области, постановка задачи, обоснование ее актуальности](#_heading=h.3znysh7) 3

[1.2. Определение стейкхолдеров, их интересов](#_heading=h.2et92p0) 3

[1.3. Построение системы метрик](#_heading=h.tyjcwt) 3

[2. Методики анализа и прогнозирования целевого значения](#_heading=h.3dy6vkm) 4

[2.1. Предварительный анализ, очистка и преобразование данных](#_heading=h.1t3h5sf) 4

[2.2. Методы решения задачи и способы визуализации данных.](#_heading=h.4d34og8) 5

[3. Анализ влияния факторов на отток клиента и прогнозирование оттока](#_heading=h.2s8eyo1) 5

[3.1. Анализ влияния факторов на целевой показатель.](#_heading=h.17dp8vu) 5

[3.2. Построение и оценка моделей для прогнозирования оттока](#_heading=h.3rdcrjn) 9

[3.3. Выдвижение и проверка гипотез по снижению оттока.](#_heading=h.26in1rg) 12

[4. Представление и оценка результатов](#_heading=h.lnxbz9) 13

[4.1. Представление аналитического дашборда](#_heading=h.35nkun2) 13

[4.2. Выводы по полученным результатам](#_heading=h.1ksv4uv) 13

[4.3. Сравнение решения задачи с аналогичными решениями](#_heading=h.jdb5m6cimf1v) 14

[4.4. Способы улучшения решения задачи](#_heading=h.jgg7yuck94q3) 14

[Заключение](#_heading=h.2jxsxqh) 15

# Введение

Дипломный проект посвящен исследованию влияния ряда факторов на отток клиента в области предоставления услуг домашнего интернета, а также построению моделей машинного обучения, способных спрогнозировать отток.

Исходные данные взяты из открытого источника: https://www.kaggle.com/vpfahad/telecom-churn-data-sets

Описание датасета представлено в файле «final\_project.ipynb»

# 1. Постановка задачи

## 1.1. Описание бизнес-области, постановка задачи, обоснование ее актуальности

Заказчиком является интернет-провайдер, желающий на основании предоставленных им данных о своих клиентах (физических лицах) получить знания о том, по каким причинам от него уходят клиенты, рекомендации, следуя которым, можно уменьшить показатель оттока и прогностическую модель, которая с высокой долей вероятности могла бы предсказать, склонен ли клиент прекратить обслуживание, и на основании этого знания предпринять шаги к удержанию.

Как известно, одна из основных целей любого бизнеса – это максимизация прибыли, которая достигается, как правило, за счет увеличения числа клиентов и средней суммы чека. В свою очередь увеличение числа клиентов можно разложить на две составляющие: привлечение новых и удержание существующих клиентов. При этом, удержать клиента дешевле, чем привлечь нового. А чтобы разрабатывать эффективные меры по удержанию клиентов, необходимо понимать, почему они уходят, и кто находится в «красной зоне». Ввиду того, что отток является неотъемлемой частью бизнеса в сфере предоставления услуг, проблема его минимизации актуальна и важна.

## 1.2. Определение стейкхолдеров, их интересов

Внутренними стейкхолдерами можем считать персонал, задействованный в процессах по привлечению и удержанию клиентов, их линейных руководителей, чьей задачей будет минимизация клиентского оттока, преследуя цель по увеличению доходов.

К внешним стейкхолдерам отнесем инвесторов, поставщиков, подрядчиков, так же заинтересованных в максимизации прибыли.

## 1.3. Построение системы метрик

Рассматривая отток, как влияющий на доход фирмы фактор, можем выстроить иерархию метрик и посмотреть, где в системе метрик находится интересующий показатель:

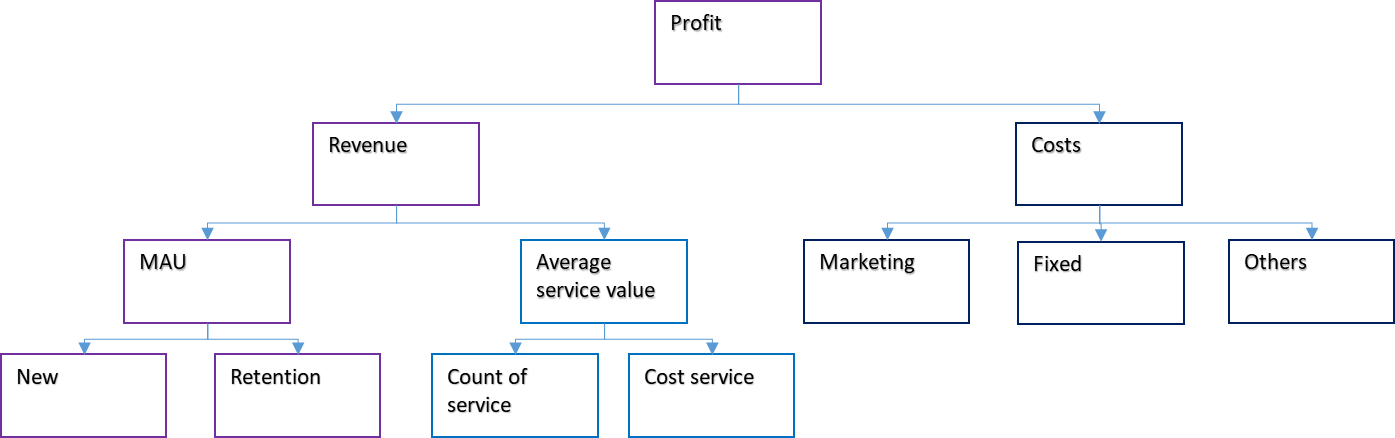


Рисунок 1. Система метрик для интернет-провайдера

Удержание (отток) – составляющая показателя количества активных пользователей, от которого напрямую зависит доход.

# 2. Методики анализа и прогнозирования целевого значения

## 2.1. Предварительный анализ, очистка и преобразование данных

Датасет представлен тремя таблицами, загружаем таблицы с помощью команды pd.read\_csv, затем объединяем в один датафрейм:

*df = churn.merge(custom, how='left', on='customerID').merge(internet, how='left', on='customerID')*

Чтобы оценить размерность и типы данных вызываем метод df.info(). Видим, что датафрейм имеет размерность 7043х21, при этом большинство переменных имеют тип object, который необходимо будет преобразовать для проведения анализа и построения модели машинного обучения.

Переменную TotalCharges (сумма денежных средств, полученных компанией с клиента) необходимо преобразовать к вещественному типу:

*# встречаются строки с пробелом, заменяем пробел на значение 'NaN', после чего преобразуем формат*

*df['TotalCharges'] = df['TotalCharges'].str.replace(' ', 'NaN').astype(float)*

*# проверяем количество сделанных замен*

*df['TotalCharges'].isnull().sum()*

*# т.к. 'NaN'-значений мало, заменяем их средним значением переменной*

*df['TotalCharges'] = df['TotalCharges'].fillna(df['TotalCharges'].mean())*

Категориальные переменные, принимающие только два значения, приводим к количественному типу, заменяя на 1 и 0 соответственно. Пишем функцию, для замены значений используем метод map(), примененный к словарю старого и нового значений:

*def categorial\_convert\_numeric(df, list\_columns):*

*dict\_yes\_no = {'Yes': 1, 'No': 0, 'Female': 0, 'Male': 1}*

*for column in list\_columns:*

*df[column] = df[column].map(dict\_yes\_no)*

*return df*

Категориальные переменные, принимающие больше двух значений, приводим к количественному типу с помощью метода подсчета "каунтов". Не используем "One hot encoding", чтобы не увеличивать в разы количество столбцов, т.к. таких переменных много. Реализуем в функции:

*def use\_counts(list\_columns\_counts, df = df, target\_column = 'Churn'):*

*dict\_category = {}*

*for column in list\_columns\_counts:*

*dict\_category[column] = (df[df[target\_column] == 1][column].value\_counts() / df[column].value\_counts()).to\_dict()*

*for column in list\_columns\_counts:*

*df[column] = df[column].apply(dict\_category[column].get)*

*return df*

Очистка и преобразование данных завершены.

Вызываем метод describe(), смотрим статистическую информацию. Переменные tenure, MonthlyCharges, TotalCharges имеют больший размах в сравнении с другими переменными, для качественной визуализации по ним строим отдельные графики boxplot. Как видно из графиков, данные выбросов не имеют.

Строим матрицу корреляций, видим, что с оттоком в сильной линейной зависимости ни одна из переменных не находится, но между некоторыми переменными есть сильная корреляция.

Отображаем гистограммы распределения признаков, по графикам можем судить о том, что ни одна из количественных переменных не распределена нормальным образом.

## 2.2. Методы решения задачи и способы визуализации данных.

В ходе предварительного анализа данных получили информацию о том, что данные не распределены нормально, не имеют выбросов, целевая переменная не имеет сильных линейных зависимостей с другими переменными. Большая часть признаков является категориальными.

На основании данных выводов определяем методику дальнейшей работы с данными. Для проверки выдвигаемых статистических гипотез будем использовать непараметрические критерии в виду ненормальности распределения. Для прогнозирования целевой переменной будем применять методы деревьев принятия решений, логистической регрессии и «Случайный лес».

Для визуализации результатов будем использовать возможности библиотек Matplotlib и Seaborn.

# 3. Анализ влияния факторов на отток клиента и прогнозирование оттока

## 3.1. Анализ влияния факторов на целевой показатель.

Сначала определим процент ушедших клиентов от всех наших клиентов – он составляет 26,54 %. Данные представлены за весь период существования интернет-провайдера (либо с момента формирования базы данных), взяты в конкретный момент времени, где для клиента указан период обслуживания в компании в количестве месяцев и классификатор – ушел/не ушел. Разделение на периоды отсутствует. Считаем, что полученный показатель оттока достаточно высокий, и он требует проведения анализа причинно-следственных связей и определения мероприятий по повышению лояльности клиентов на основании проведенного анализа.

Чтобы ответить на вопрос, какие показатели оказывают наибольшее влияние на отток клиента, в первую очередь необходимо определить наличие связи между оттоком и прочими факторами.

Как мы видим из таблицы корреляций, многие независимые переменные коррелируют между собой, но с целевой переменной сильных взаимосвязей нет. Считая, что зависимость между переменными обуславливает наличие корреляционной связи между ними (но не наоборот), выбираем для дальнейшего анализа наиболее коррелирующие с зависимой величиной переменные – это tenure, Contract, InternetService, OnlineSecurity, TechSupport.

Т.к. в таблице корреляции расчет выполняется по критерию Пирсона, а наши данные не количественные, то рассчитаем для выбранных переменных коэффициент корреляции по критерию Спирмена. Данные с таблицей корреляции совпадают.

Рассматриваем зависимость оттока от времени, в течение которого клиент пользуется услугами. Строим график (рисунок 2):

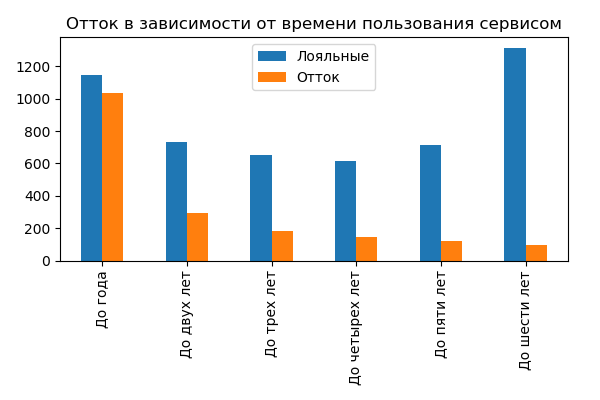


Рисунок 2. Отток в зависимости от времени пользования сервисом

По визуализации отмечаем, что почти половина клиентов, которые пользовались услугами интернет-провайдера меньше года, ушли. Чем длительнее клиент пользуется услугами, тем реже он уходит.

Для подтверждения полученных графическим способом результатов выдвигаем гипотезу о том, что отток одинаков для всех временных групп. Проверяем гипотезу с помощью критерия Хи-квадрат Пирсона, т.к. переменные качественные.

В результате получили p-value близкое к нулю, следовательно, зависимость частоты случаев оттока от времени пользования интернет-сервисом статистически значима.

Аналогично проверяем зависимость оттока от типа заключенного с клиентом договора на оказание телематических услуг. Строим график (рисунок 3):

*table\_contract = pd.crosstab(original\_df['Contract'], original\_df['Churn'])*

*table\_contract.plot(kind='bar', rot=0)*

Отмечаем, что максимальный отток среди тех клиентов, которые заключают договор с ежемесячной оплатой. Кто заключил договор на год или два, уходят значительно реже.

Проверим зависимость частоты оттока от наличия/отсутствия сервиса интернет-безопасности, построим график (рисунок 4).

Наблюдаем высокий отток клиентов, у которых нет сервиса онлайн-безопасности.

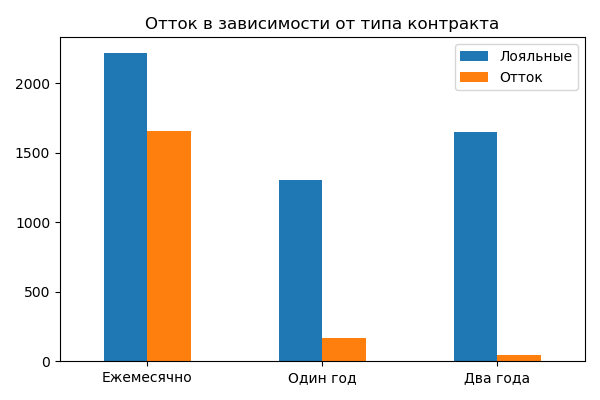


Рисунок 3. Отток в зависимости от типа контракта

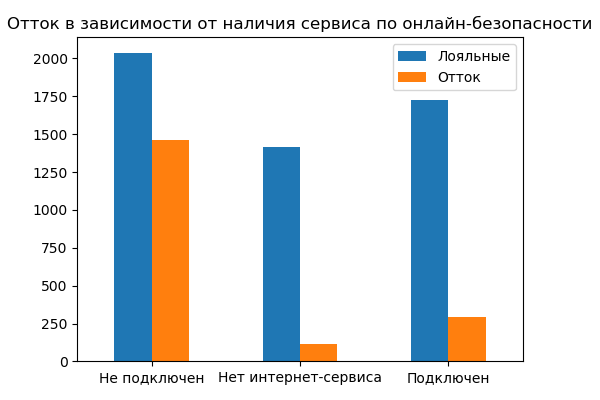


Рисунок 4. Отток в зависимости от наличия/отсутствия сервиса интернет-безопасности

Теперь рассмотрим взаимосвязь оттока с количественным признаком – ежемесячным платежом.

Переменная 'MonthlyCharges' распределена не нормально, для проверки гипотезы об отсутствии различий в платежах между лояльными и ушедшими клиентами используем непараметрический критерий Манна-Уитни:

*# выдвигаем гипотезу о том, что ежемесячные платежи у лояльных и ушедших клиентов не отличаются*

*churn\_monthly = df[df['Churn'] == 1]['MonthlyCharges']*

*nochurn\_monthly = df[df['Churn'] == 0]['MonthlyCharges']*

*u, p\_value = stats.mannwhitneyu(churn\_monthly, nochurn\_monthly)*

*print('p-value =', p\_value)*

Получили p-value, стремящееся к нулю. Считаем различия в ежемесячных платежах лояльных и ушедших клиентов статистически значимыми.

Строим график (рисунок 5):

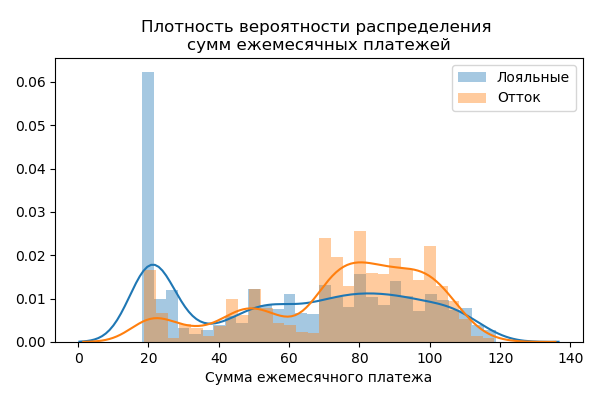


Рисунок 5. Плотность вероятности распределения сумм ежемесячных платежей

Из графика можем сделать вывод о том, что те клиенты, у которых высокие ежемесячные платежи (от 70 до 110 денежных единиц), наиболее часто склонны к оттоку. И наоборот, самые лояльные клиенты - это те, у которых минимальные ежемесячные платежи (в районе 20 единиц). Для уточнения параметров, от которых зависит сумма платежа и отток, разобьем абонентов на две группы: с высокими ежемесячными платежами (от 70 до 110 единиц), для которых видим максимальный отток, и с платежами меньше 70 и больше 110 единиц измерения и меньшим оттоком.

Для каждой группы построили гистограмму по трем параметрам: тип контракта, тип подключения и наличие/отсутствие онлайн-безопасности (рисунки 6 и 7):

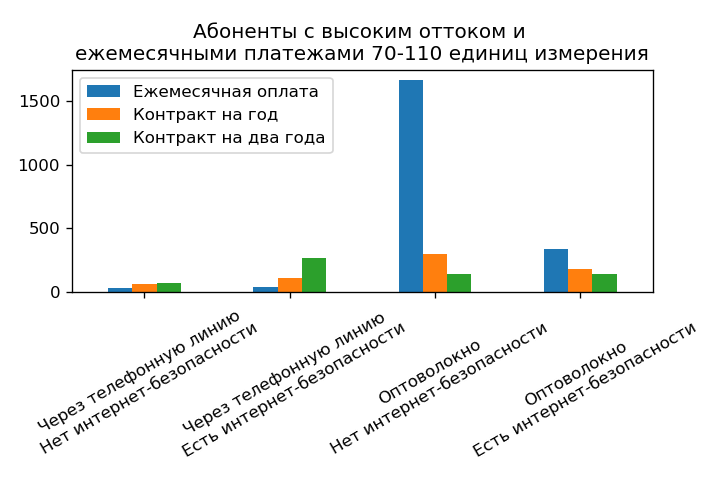


Рисунок 6. Гистограмма распределения параметров для абонентов с высоким оттоком

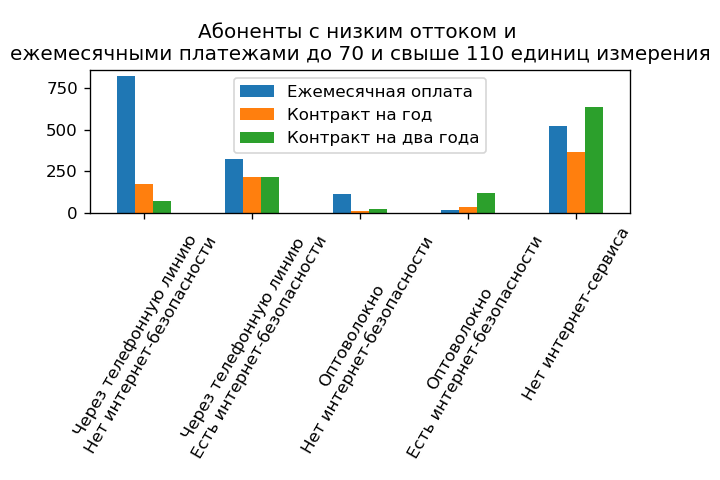


Рисунок 7. Гистограмма распределения параметров для абонентов с низким оттоком

По визуализации параметров понимаем, что абоненты с платежами в пределах 70-110 единиц, наиболее склонные к оттоку, в большинстве подключены через оптоволоконный кабель, отсюда, вероятно, более высокие тарифы. Вторая группа абонентов (рисунок 7) преимущественно подключены через линии телефонной связи либо совсем не имеют интернет-сервиса, что объясняет более низкие платежи.

## 3.2. Построение и оценка моделей для прогнозирования оттока

Прогнозирование оттока является задачей классификации, поэтому будем использовать модели деревьев принятия решений, логистической регрессии, «Случайного леса». При оценке качества моделей ввиду несбалансированности классов будем анализировать показатели точности и полноты модели (рrecision и recall), их агрегированный показатель F-меру и значение AUC-ROC.

Проводим подготовку данных: избавляемся от мультиколлинеарности, разбиваем данные на тренировочные и тестовые в соотношении 70/30, приводим данные к одному масштабу:

*from sklearn.model\_selection import train\_test\_split*

*X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3)*

*from sklearn.preprocessing import StandardScaler*

*sc = StandardScaler()*

*X\_train = sc.fit\_transform(X\_train)*

*X\_test = sc.transform(X\_test)*

Обучаем модель деревьев принятия решений:

*from sklearn.tree import DecisionTreeClassifier*

*tree = DecisionTreeClassifier(criterion='gini')*

Оцениваем качество:

*tree\_predict = tree.predict(X\_test)*

*from sklearn.metrics import classification\_report*

*print(classification\_report(y\_test, tree\_predict))*

*tree.score(X\_test, y\_test)*

Модель деревьев принятия решений неплохо справляется с классификацией лояльных клиентов, точность, полнота и F-мера в районе 80%, но достаточно плохо определяет ушедших клиентов, показатели в районе 50%. Значение AUC-ROC равно 65%, т.е. модель примерно с вероятностью 65% способна различать классы.

Обучаем модель логистической регрессии и оцениваем ее качество, результат в таблице 1:

Таблица 1. Показатели качества модели логистической регрессии

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.85 | 0.89 | 0.87 | 1552 |
| 1 | 0.63 | 0.56 | 0.60 | 561 |
| accuracy |  |  | 0.80 | 2113 |

Показатели качества модели логистической регрессии выше, чем модели деревьев принятия решений, F-мера для лояльных клиентов стремится к 90%, а для ушедших - около 60%, соответственно повысилась способность модели различать классы и составила 70%, что по-прежнему не очень хорошо.

Обучаем модель «Случайный лес»:

*from sklearn.ensemble import RandomForestClassifier*

*forest=RandomForestClassifier(criterion='entropy', n\_estimators=50, max\_depth=5, min\_samples\_leaf=5)*

*forest.fit(X\_train,y\_train)*

Параметры модели подобраны эмпирическим путем.

Показатели качества модели приведены в таблице 2:

Таблица 2. Показатели качества модели "Случайный лес"

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.83 | 0.92 | 0.87 | 1552 |
| 1 | 0.69 | 0.46 | 0.55 | 561 |
| accuracy |  |  | 0.80 | 2113 |

Качество модели «Случайный лес» примерно такое же, как и у модели логистической регрессии. Попробуем поработать ещё с данной моделью, определим наиболее важные атрибуты для модели (рисунок 8) и отсечем те, чей вес минимальный, и обучим модель заново.

Попробуем улучшить модель, исключив 5 наименее важных атрибутов.

Повторили процесс обучения, но качество модели осталось приблизительно на том же уровне: агрегированный показатель точности и полноты определения класса лояльных клиентов стремится к 90%, для ушедших клиентов этот показатель стремится к 60%, в целом способность модели разделять классы составляет порядка 70%.

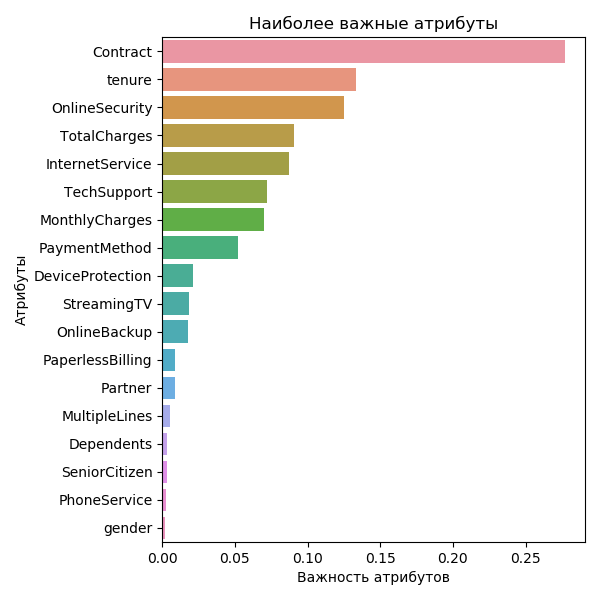


Рисунок 8. Наиболее важные атрибуты для модели «Случайный лес»

С целью повышения качества моделей уравновесим классы в тренировочных данных:

*from imblearn.over\_sampling import RandomOverSampler*

*ros = RandomOverSampler(random\_state=0)*

*ros.fit(X\_train, y\_train)*

*X\_resampled, y\_resampled = ros.fit\_sample(X\_train, y\_train)*

И обучим заново модели логистической регрессии и «Случайного леса». В результате у обеих моделей получили более уравновешенные для классов показатели: (таблица 3):

Таблица 3. Показатели качества модели «Случайный лес» после уравновешивания классов

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.90 | 0.73 | 0.81 | 1556 |
| 1 | 0.51 | 0.78 | 0.62 | 557 |
| accuracy |  |  | 0.74 | 2113 |

Значительно увеличился показатель полноты определения recall (с 50% до 80%), т.е. 80% клиентов, которые действительно ушли, мы им это предсказали, при этом для класса лояльных этот показатель несколько уменьшился, но в целом способность модели различать классы (показатель AUC-ROC) повысилась до 75%. Ниже на рисунке 9 представлена ROC-кривая для модели «Случайный лес» после уравновешивания классов.

Данную модель возьмем для проверки гипотез, направленных на снижение оттока. Исходя из тезиса о том, что удержание клиента, как правило, дешевле, чем его привлечение, полагаем, что механизмы по удержанию лучше применить к тем, кто, возможно, не планирует прекращать обслуживание в компании, чем пропустить того клиента, кто действительно собирается уйти, поэтому отдаем предпочтение той модели, у которой выше показатель полноты охвата, а не точности.

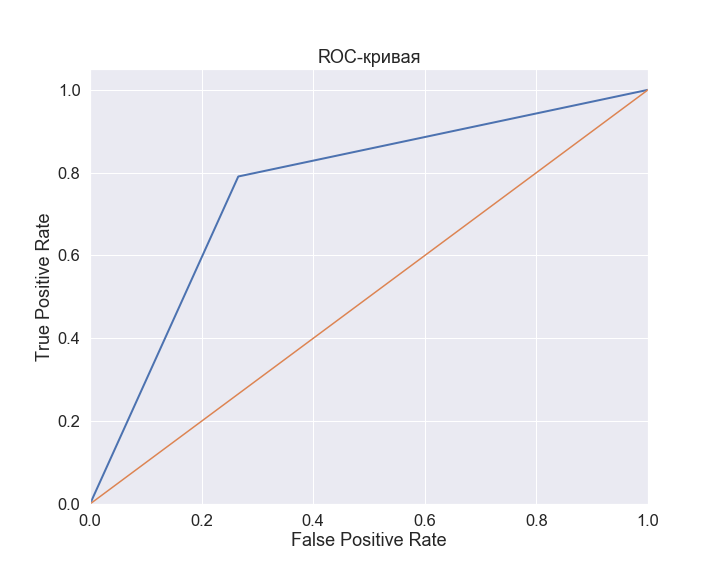


Рисунок 9. ROC-кривая для модели «Случайный лес» после уравновешивания классов

## 3.3. Выдвижение и проверка гипотез по снижению оттока.

Проверим гипотезы, основанные на изменении двух факторов: типа заключенного договора на оказание услуг и использования сервиса интернет-безопасности (таблица 4).

Ввиду того, что нет возможности проверить гипотезы на реальных данных, уберем из датасета целевую переменную и с помощью выбранной в п.3.2 модели спрогнозируем отток, далее после корректировки данных в соответствии с гипотезой изменение оттока будем рассчитывать, сравнивая с полученным результатом и используя для предсказания ту же модель.

Таблица 4. Гипотезы по снижению оттока

|  |  |  |  |
| --- | --- | --- | --- |
| Гипотеза | Действия | Аналитика | Выводы |
| 1. Перевод 10% клиентов на годовой договор на оказание услуг вместо договора с ежемесячной оплатой снизит отток на 5%. | Изменяем в десяти процентах записей значение контракта с ежемесячного на годовой. | Используя модель машинного обучения, прогнозируем отток. | Отток снизился на 3,12%, что меньше ожидаемого. Выручка при этом увеличилась на 2,36% или на 379 156 ден. единиц. Гипотезу можно рассматривать, как один из механизмов по снижению оттока. |
| 2. Подключение 50% клиентов сервиса по онлайн-безопасности приведет к снижению оттока на 5% | Изменяем в пятидесяти процентах записей значение онлайн-безопасности с отсутствия на наличие. | Используя модель машинного обучения, прогнозируем отток. | Снижение оттока составило 2,09%, что более чем вдвое меньше ожидаемого, но выручка при этом увеличилась на 2,63%, или 423 101 ден. единиц. Если затраты на подключение сервиса будут ниже, то принимаем гипотезу как жизнеспособную. |

В результате проверки гипотез получили показатель снижения оттока на 2-3 %, аналогичный показатель увеличения выручки. При низких затратах на реализацию предложенных мер их можно рекомендовать к использованию.

# 4. Представление и оценка результатов

## 4.1. Представление аналитического дашборда

Дашборд представлен в Приложении. В нем отразили следующую информацию:

- процентное соотношение лояльных и ушедших клиентов для понимания величины текущего оттока;

- наиболее сильно влияющие на отток показатели и их детализацию: соотношение оттока в зависимости от времени пользования услугами, от договора на оказание услуг связи, распределение оттока в зависимости от ежемесячных платежей, чтобы визуально продемонстрировать, что максимальный отток наблюдается в первый год пользования услугами интернет-провайдера у клиентов с помесячным договором и с высокими платежами.

- мероприятия по снижению оттока с количественным отражением: гипотезы, которые были проверены, и показатель, на который они предположительно позволят снизить отток и повысить выручку.

## 4.2. Выводы по полученным результатам

В ходе исследования было выполнено:

- определен текущий уровень оттока клиентов – 26,54%, это высокий уровень оттока, который требует проведения мер по снижению;

- определены факторы, оказывающие наибольшее влияние на показатель оттока – это время, в течение которого клиент пользуется услугой (наибольший отток у клиентов, пользующихся услугами меньше года, с увеличением времени отток снижается), это тип договора на оказание услуг (высокий отток с типом «от месяца к месяцу» и низкий на годовой и двухгодовой договор), это наличие/отсутствие сервиса онлайн-безопасности (в случае его наличия отток значительно ниже), это суммы ежемесячных платежей (чем эти суммы больше, тем выше отток);

- по совокупности указанных факторов можно сказать, что клиенты чаще всего покидают компанию в первый год пользования услугами, в большинстве случаев это клиенты, у которых договор на ежемесячной основе, у которых высокие платежи и отсутствует сервис онлайн-безопасности, а также те, у кого трафик подается через оптоволокно, а не кабели телефонной связи, что взаимосвязано с более высокими тарифами;

- были построены несколько моделей машинного обучения по прогнозированию оттока клиентов, наилучшей была выбрана модель «Случайный лес» с уравновешенными классами, ее способность разделять классы составила 75%;

- выдвинуты и проверены с помощью выбранной модели несколько гипотез по снижению оттока на основании изменения некоторых факторов, выявлено, что при использовании предусмотренных мер показатель оттока снизится предположительно на 2-3%, и выручка увеличится так же на 2-3%.

## 4.3. Сравнение решения задачи с аналогичными решениями

Взятый датасет менее полугода назад загружен на ресурс, с которого был взят, решение к нему было предложено лишь дважды. Оба решения не включают в себя построение моделей по предсказанию оттока, содержат код по загрузке данных, объединению датафреймов, вывод информации о признаках, общих статистиках, гистограммы распределений признаков, корреляцию и точечные графики взаимодействия нескольких признаков. Данные решения можно считать низко информативными.

## 4.4. Способы улучшения решения задачи

Для получения более достоверных результатов можно детальнее проработать финансовую составляющую, выяснить, от каких факторов зависит стоимость услуги. Понять, по каким причинам клиент уходит в первые месяцы (может потребоваться дополнительная информация такая как уровень сервиса, качество услуг, в т.ч. скорость интернета и прочее).

Можно попробовать построить новые обучающие модели, например, метод опорных векторов. Или повысить качество текущих моделей. Работа по настройке качества была проведена только для модели «Случайного леса», которая была выбрана наиболее оптимальной в виду большого количества категориальных переменных.

# Заключение

В рамках выполнения дипломного проекта проведен анализ данных, влияющих на отток клиента, построены модели машинного обучения, способные предсказывать отток, выдвинуты и проверены гипотезы, направленные на снижение оттока, построен аналитический дашборд, по результатам работы сделаны выводы.