**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

###### Центр непрерывного образования

###### Факультета компьютерных наук

**ИТОГОВАЯ АТТЕСТАЦИОННАЯ РАБОТА**

Прогнозирование и анализ оттока пользователей по сообщениям в соцсети

|  |  |  |
| --- | --- | --- |
|  | Выполнил: | |
|  | | Крылов Сергей Викторович  Ф.И.О. |
|  | Руководитель: | |
|  | Селезнев Артем  Руководитель направления по аналитике данных в Сберанке  степень, звание, должность Ф.И.О. | |
|  |  | |

Москва, 2021

**Введение**

Предсказание оттока является крайне важной задачей для любого бизнеса, и особенно в сфере телекома. Ключевой момент в предсказании оттока заключается в понимании причин неудовлетворенности пользователей. И именно Twitter является тем крайне полезным и эффективным источником данных для аналитиков данных, который может предоставить неплохое представление о реакции пользователей в реальном времени. Твит – это короткое сообщение, который может содержать информацию на любые топики, он может отражать как субъективную, так и фактическую информацию, а также может быть окрашен разными тональностями – от раздражения и до восхищения. Поэтому твит – это очень хороший объект для анализа и прогнозирования.

Ввиду того, что задача крайне актуальна для бизнеса, по этому вопросу написано множество работ и статей. Обобщая изученные источники, можно сказать, что задача прогнозирования оттока сводится к задаче классификации. При этом очень часто отсутствует историческая информация о таргете для того, чтобы можно было использовать данные в машинном обучении. Поэтому ставится дополнительная задача искусственного создания целевых значений – кластеризации, которая в свою очередь основывается на анализе тональности текста. Так, в работе «Sentiment Analysis: A Comparative Study On Different Approaches» авторы отмечают[[1]](#footnote-2), что существует несколько подходов к кластеризации тональности - подходы, основанные на правилах, подходы, основанные на словарях, машинное обучение с учителем, машинное обучение без учителя.

В данной работе целью является использование существующих алгоритмов по определению оттока клиентов операторов связи, а также его прогнозирования. В качестве алгоритма были изучены:

* подходы машинного обучения с учителем,
* без учителя,
* подход на правилах.

Важно отметить, что т. к. Twitter в России не является таким широко используемым, как на западе, в работе анализировались сообщения клиентов западных сотовых операторов.

**Краткое описание алгоритма**

Алгоритм начинается со сбора данных с помощью API Twitter. Далее эти данные обрабатываются, делается разведочный анализ и генерируются переменные для дальнейшей возможности использования их в моделях машинного обучения. Важно также отметить, что для целей прогнозирования оттока необходим таргет. Ввиду того, что в случае с Twitter, истинного таргета нет, необходимо создать таргет искусственно при помощи кластеризации. Для увеличения точности некоторые подходы предлагают применить на кластеризованные данные rule-based подход. И уже на этих данных можно обучить модели прогноза.

**Получение данных**

Для извлечения твитов был использован пакет Tweepy. В этих целях была создана функция, которая принимает на вход текст запроса, дату и количество сообщений. Все полученные данные складывались в один датасет. В итоге исходный датасет выглядел следующим образом:

Таблица 1. Исходные данные

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **screen\_name** | **date\_time** | **location** | **text** |
| 0 | TheSkubis | 2021-03-30 23:59:19 | Pennsylvania, USA | @VerizonSupport I have issues with closed capt... |
| 1 | VerizonSupport | 2021-03-30 23:56:55 |  | @\_carolinek This could be due to regional rest... |
| 2 | VerizonSupport | 2021-03-30 23:55:30 |  | @tvmurray We'll be happy to help with anything... |

Количество данных изначально небольшое и сводится к указанию автора твитта, времени твитта и его содержания, поэтому для дальнейшего анализа они требуют обработки. Для этой цели был проведен разведочный анализ данных (EDA), который включает:

* preprocessing;
* статистику по сегментам;
* общую статистику датафрейма;
* частотный анализ;
* динамический анализ;
* сентиментальный анализ;
* разделение корпуса на топики.

Preprocessing предполагает использование следующих методов на текстовых данных:

* исключение символов обращений «#» и «@»;
* исключение ссылок;
* токенизация – разделение на компоненты;
* лемматизация – приведение слова к исходной форме;
* составление словаря n-gram;
* фильтрация stop-слов.

Далее с помощью библиотеки textblob была рассчитана средняя полярность и субъективность предложения. В рамках разведочного анализа данных была выведена статистика в разрезе полярности предложений и в разрезе операторов связи.

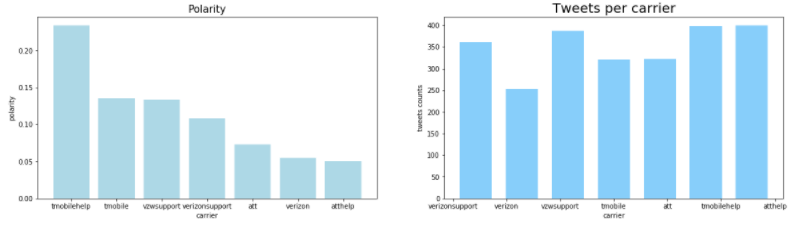
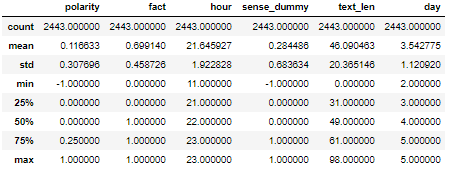


График 1. Полярность и количество твиттов

Количественная статистика датасета отражена в таблице 2.

Таблица 2. Статистическое описание исходных данных



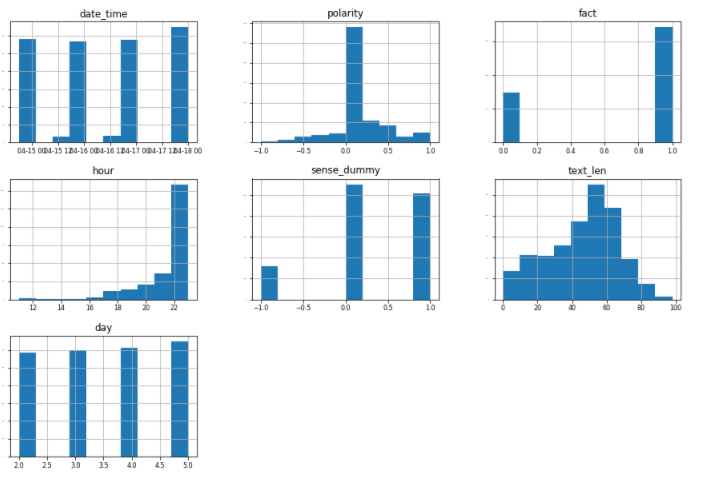
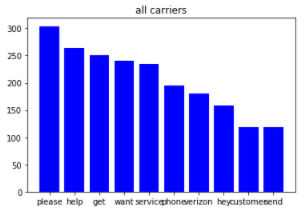
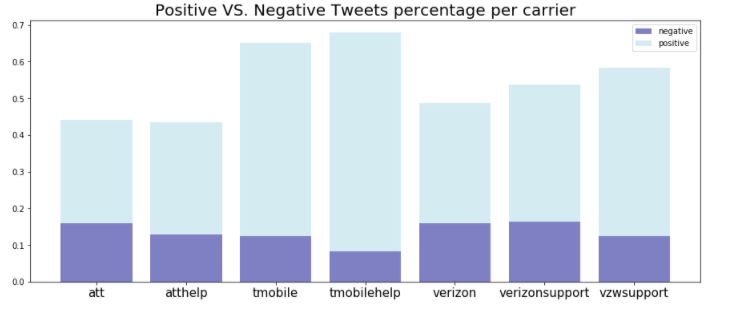


График 2. Статистическое описание исходных данных

Самая высокая полярность – у оператора TMobile, при этом у него же самое большое количество твиттов.

Ниже представлены результаты составления словаря самых часто употребляемых слов среди всех операторов, а также динамика количества твиттов по времени и распределение твиттов по сентиментам .

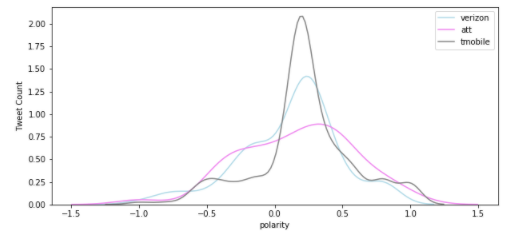


График 3. Частотный и сентиментный анализ сообщений

При помощи библиотеки gensim (LdaMulticore) все сообщения были разделены на 20 топиков – это будет началом для следующего шага по созданию фичей для нашей модели.

**Создание параметров**

Классически кластеризацию на текстах делается с помощью мешка слов или TF-IDF, но для уточнения модели необходимо добавить еще несколько фичей. В данных есть такой параметр, как время. Без предобработки он не может быть взят в модель, т. к. он некорректно отображает разницу в данных, то есть, если посмотреть на график секунд за день, можно отметить, что разница между 23-55 и 00-05 равна 23-50

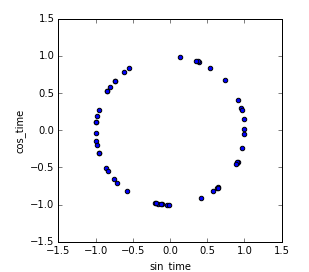
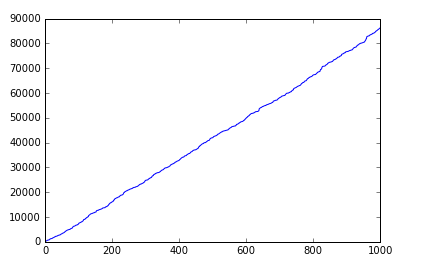


График 4. Исходный и требуемый формат периодических данных

Для того, чтобы можно было использовать временные данные в параметрах, необходимо их трансформировать в периодический формат. Это делается с помощью sin/cos-трансформации, после которой временные данные выглядят более подходящими для использования в модели.

Следующий параметр – это TF-IDF. Это простой и удобный способ оценить важность термина для какого-либо документа относительно всех остальных документов. Принцип такой — если слово встречается в каком-либо документе часто, при этом встречаясь редко во всех остальных документах — это слово имеет большую значимость для того самого документа. Таким образом, слова, неважные для вообще всех документов, например, предлоги или междометия — получат очень низкий вес TF-IDF (потому что часто встречаются во всех-всех документах), а важные — высокий. Для увеличения веса этого параметра рекомендуется убрать 20% незначимых слов с помощью параметра TfidfVectorizer (max\_features). Так же неплохой характеристикой для анализа будет использование проекции RFM-анализа на текстовые данные. Из показателей классического RFM (recency, frequency, monetary) в данных есть все для расчета recency (как разницу между текущим днем и датой твитта) и frequency – подсчетом количества твиттов каждого пользователя. Polarity в данном случае будет аналогом monetary. Таким образом группировка пользователей по этим 3 показателям позволяет получить новый параметр.

Параметр геолокации также может быть использован в модели, предварительно переведя текстовое значение в числовое. Для этого, например, можно поставить каждому значению локации в соответствие ее вероятность.

Ранее были выведены такие метрики, как номер топика и значение субъективности. Матрица параметров для кластеризации готова. Забегая вперед, надо отметить, что нормализация данных в данном случае ухудшила результаты кластеризации.

**Кластеризация**

Для кластеризации данных были использованы евклидово расстояние, квадратное евклидово расстояние и расстояние Чебышева.

Таблица 3. Формулы расстояний при кластеризации

|  |  |
| --- | --- |
| Расстояние | Формула |
| Евклидово расстояние |  |
| Квадратное евклидово расстояние |  |
| Расстояние Чебышева |  |

Эти расстояния реализованы в sklearn в K-means (евклидово расстояние) и K-medoids (квадратное евклидово расстояние и расстояние Чебышева). Качества этих методов предлагается сравнить 2 способами – метрикой Силуэт (Silhouette score), а также сравнением с «наивным» таргетом метрика accuracy, roc\_auc. В качестве наивного таргета можно использовать гипотезу о том, что человек уйдет, если полярность его твитта меньше, чем 0,5. Таким образом, «наивный» таргет получается указанием 1 для подобных твиттов.

В результате фаворитом является разделение по евклидову расстоянию (табл.4) .

Таблица 4. Результаты кластеризации

|  |  |  |  |
| --- | --- | --- | --- |
|  | **silhouette** | **roc\_auc** | **accuracy** |
| **euclidean\_churn** | 0.53 | **0.57** | **0. 63** |
| **chebyshev\_churn** | -0.36 | 0.5 | 0.97 |
| **sqeuclidean\_churn** | 0.72 | 0.57 | 0.63 |

С помощью метода главных компонент (PCA) посмотрим на качество кластеризации.

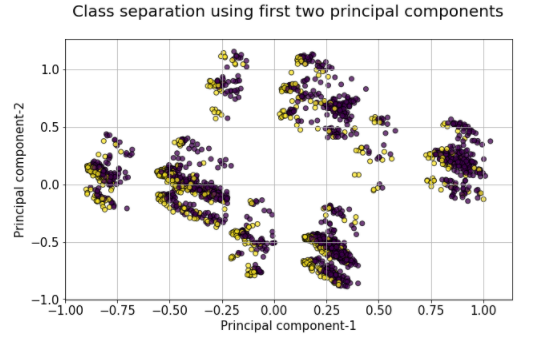


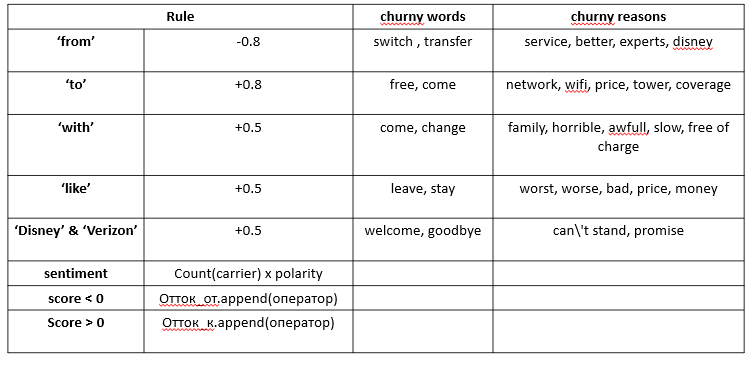
График 5. Визуализация кластеризации при помощи PCA

Далее для улучшения качества кластеризации некоторые методологии предлагают применить к нему так называемый rule-based подход. Он предполагает использование скоринга за выполнение указанных условий. Для этого создается функция, которая:

* определяет субъект, то есть обращение к определенному сотовому оператору,
* причину оттока по наличию таких слов как ‘slow','free of charge','worst','worse','bad','price','money', ’tower’
* правила скоринга:
  1. если в твитте имеются слова, говорящие об уходе (‘leave’,’come’, ‘switch’, ‘change’, ‘transfer’ и т.д.), то при наличии слова ‘to’ модель добавляет 0,8 баллов, тому оператору, о котором идет речь в тексте и наоборот, если в предложении используется название оператора и слова ‘from’, баллы в словаре скоринга у данного оператора вычитаются;
  2. если в тексте имеется слово ‘like’ и один из операторов, этому оператору добавляется 0,5 балла;
  3. при наличии слов ‘verizon’ и ‘disney’ одновременно, оператору Verizon добавляется 0,5 баллов;
  4. последний параметр скоринга – это произведение количества упоминаний оператора и полярности сообщения;
  5. если итоговое количество очков у оператора отрицательно, добавляем данному сообщению и оператора в список оттока и выводим статистику по каждому сообщению.

Таким образом, если результатом данной функции является непустой список с указанием входящего или исходящего оттока, данное сообщение может быть определено как отток (с лейблом 1). Схему данного подхода можно формализовать в таблице (табл.5).

Таблица 5. Применение rule-based подхода для кластеризации



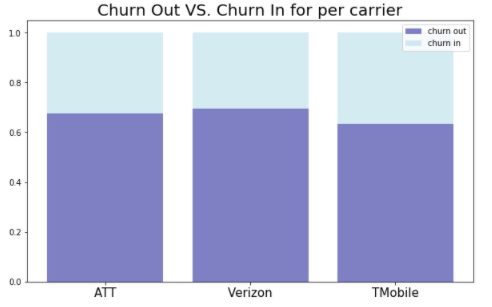
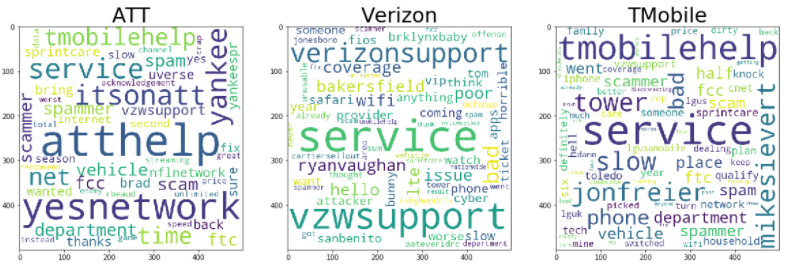
 

График 6. Статистика причин оттока по операторам

Как видим, очень часто характеристикой оттока являются такие слова, как “service”, “coverage”, “network”, “slow” и “price”.

**Классификация**

Теперь для классификации в датасете имеется таргет. В качестве базовых моделей для классификации использовались:

1. Naive Bayes
2. Logistic Regression
3. Decision Tree
4. LightGBM

В целях повышения точности прогноза были проделаны min-max преобразование и поиск параметров регуляризации с помощью grid-search. Метрикой качества был выбран roc-auc.

Таблица 6. Качество моделей классификации оттока.

|  |  |
| --- | --- |
| **model** | **roc\_auc** |
| Naive Bayes | 0.53 |
| Logistic regression | 0.75 |
| LogRegression + grid\_search | 0.79 |
| LogRegression + MinMaxScaler | 0.79 |
| LogRegression + grid\_search + MinMaxScaler | 0.81 |
| Decision Tree | 0.82 |
| LightGBM | 0.86 |

Лучшей моделью стала LightGBM с достаточно хорошим показателем точности прогноза. Далее была сделана кросс-валидация, в результате которой можно говорить о довольно хорошей обобщающей способности модели.

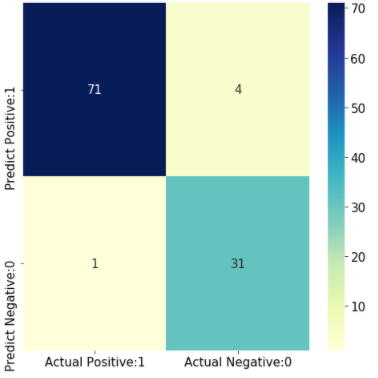
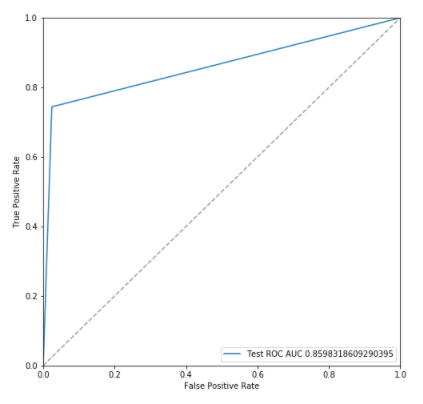


График 7. Roc-auc и confusion matrix для модели

Таким образом, в данной работе был реализован алгоритм прогнозирования оттока пользователей сотовых операторов по их сообщениям в социальной сети.

Список литературы:

1. Кластеризация и визуализация текстовой информации – статья на habr.com, 2018г.
2. Обучаем компьютер чувствам (sentiment analysis по-русски) – статья на habr.com, 2012
3. Анализ тональности в русскоязычных текстах – статья на habr.com, 2020
4. Открытый курс машинного обучения. Тема 7. Обучение без учителя: PCA и кластеризация – статья на habr.com, 2017
5. Как решить 90% задач NLP: пошаговое руководство по обработке естественного языка – статья на habr.com, 2018
6. Основы Natural Language Processing для текста– статья на habr.com, 2019
7. Encoding cyclical continuous features – 24-hours time – статья на ianlondon.github.io, 2016
8. Amiri, H., and Daume III, H. 2015. Short Text Representation for Detecting Churn in Microblogs. In Thirtieth AAAI Conference on Artificial Intelligence.
9. Devika M D, Sunitha C, Amal Ganesh.2016. Sentiment Analysis:A Comparative Study On Different Approaches
10. Силен Д., Мейсман А. - Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер,2017. – 336с.: ил. – (Серия «Библиотека программиста»).
11. Mastering Machine Learning with Python in Six Steps. Manohar Swamynathan, Bangalore, Karnataka, India, 2017

1. Devika M D, Sunitha C, Amal Ganesh.2016. Sentiment Analysis:A Comparative Study On Different Approaches [↑](#footnote-ref-2)