Customer Churn Analysis and Prediction

[1]M. Sai Krishna Reddy, [2]P.Guru Vamshi,[3]N.Roja Ashritha, [4]V.Ravi Teja [1,2,3]UG Student, [4]Assistant Professor
[1,2,3,4]CSE- Artificial Intelligence and Machine Learning [1,2,3,4]Sreenidhi Institute of Science and Technology, Hyderabad, Telangana.

Abstract:*Customer churn is one of the big issues for businesses these days, especially in competitive markets such as telecommunications. This paper introduces a sentiment analysis-based model using Twitter data to determine customer churn. The dataset is obtained through the Tweepy API and comprises thousands of tweets referring to major telecom companies. These tweets undergo preprocessing through various NLP techniques, including tokenization, removal of stopwords, and lemmatization. A lexicon-based approach determines the sentiment labels. The labeled data train an ML classifier based on the XGBoost algorithm. The evaluation metrics accuracy, precision, recall, and F1-score confirm the adequacy of the proposed model. By integrating social sentiment into churn prediction models, telecom providers can gain valuable insights into factors that lead to customer dissatisfaction so they can adopt proactive retention strategies.*

Keywords:Customer Churn, Sentiment Analysis, Twitter, NLP, XGBoost, Machine Learning, Churn Prediction, Social Media Analytics

---

1. INTRODUCTION

1.1 Background

Customer churn has a significant impact on profitability in the telecom sector. Traditional churn models rely on demographic and transaction data, but social media offers further insights into customers' emotions. Customer dissatisfaction can be signaled by analyzing tweets that refer to telecom providers. This project utilizes tweets from actual customers to augment churn prediction with sentiment analysis.

Motivation

Social media is a reflection of customer sentiment. Displeased customers usually tweet their complaints first before canceling a service. This study intends to utilize such signals as early detection indicators. We thus try to incorporate the results of sentiments analysis of tweets into a churn prediction model in order to increase the accuracy and timelines of identifying at risk customers.Proactive retention is one feature through machine learning in predicting customer churn by pointing out those most likely to leave before it actually happens. It gives better accuracy and scalability because it would be possible to handle large complex data. Information obtained identifying particular drivers for different customer segments provides for tailored retention strategies that target those customers that are bringing high risk. It will then optimize resources by only focusing efforts on customers labeled as either high-risk or high-value customers.It shall definitely promise a competitive edge at re-attributing the existing customers, further enhanced branding loyalty.

1.2 Objectives

The objectives of the customer churn rate prediction are as follows:

- Collecting Twitter data about telecom providers using the Tweepy API.
- Preprocess tweet data using NLP: tokenization, stopword removal, lemmatization.
- Assign sentiment labels to tweets using rule-based sentiment scoring.
- Build and evaluate churn prediction models using XGBoost.
- Analyze feature importance and assess the impact of sentiment features.

2. RELATED WORKS

2.1. Predictive Analytics and Customer Retention

Role of prediction of customer churn for retention strategies. Role of Predictive Analytics in Business Profitability Applications of Predictive Modeling in various Industries

2.2. Machine Learning Algorithms for Churn Prediction

Comparison of logistic regression, random forests, and Rule-Based algorithms, XG Boost algorithm for churn prediction

strengths and weaknesses of the supervised learning techniques in modeling churn Hyperparameter Tuning to Enhance model performance.

2.3. Data Preprocessing for Churn Analysis

Working on categorical features such as geography and gender. Scaling of numeric features such as credit score and balance. Encoding the use of binary features, having a credit card, being an active member or otherwise.

2.4. Feature Engineering and Selection

Demographic along with transaction history plays major influences the churn.

Feature Importance - use of random forest coefficients of logistic regression. Drawing up the feature derived by balance-to-salary, age groups, etc.

2.5. Evaluation Metrics of the Churn Prediction Model

Accuracy, precision, recall, and F1-score for model evaluation.

2.6. Churn prediction Real World Applications

Bank and financial services: churn prediction from balance, tenure and credit score perspective. Telecom and retail: prediction based on transactional and behavioral patterns of churn.

2.7. Advanced Features in Churn Prediction

Real-time churn prediction with real-time analytics Deep learning models like neural networks for better accuracy Hybrid approach that combines traditional algorithms with ensemble techniques like XG Boost.

2.8. Ethical and Business Considerations

Ethical use of customer data for churn prediction Balancing model interpretability with complexity for business take-up Cost-benefit analysis of implementing churn-prediction systems.

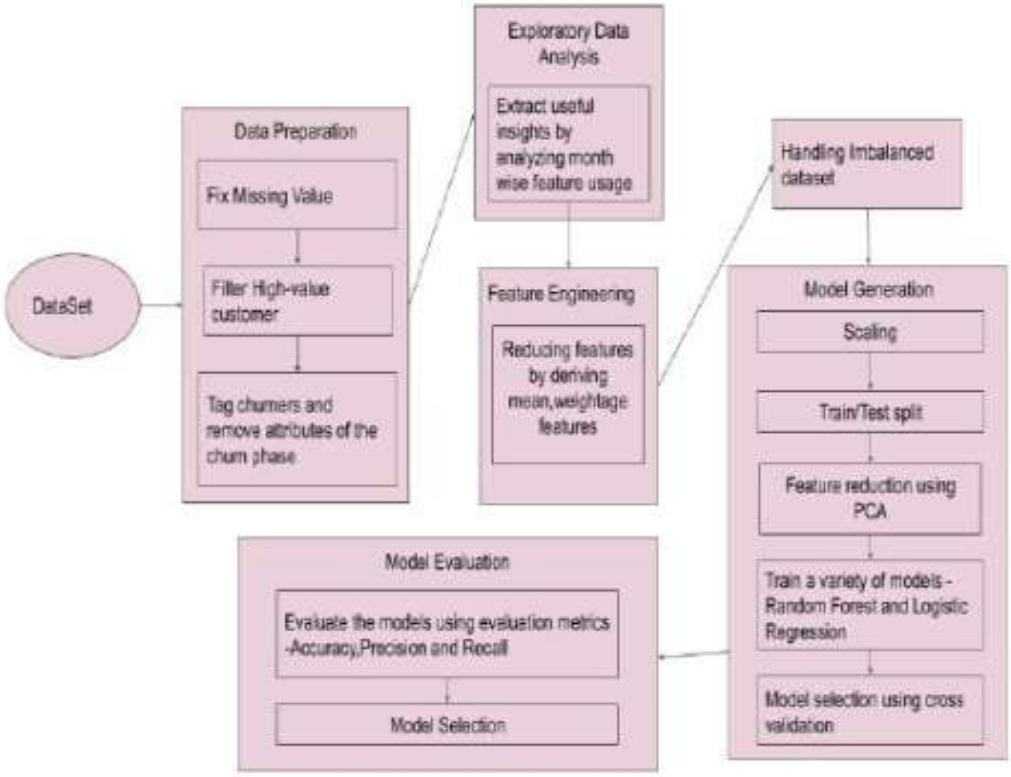
2.9. Visualization and Insights

Graphical representation of feature distribution to understand the churn patterns Dashboard design for real-time churn monitoring The Use of Data Storytelling: Presenting Churn Insights in an Effective Manner.

3. SYSTEM ARCHITECTURE

It is very crucial to make the data useful because unwanted or null values can cause unsatisfactory results or may lead to producing less accurate results. In the data set, there are a lot of incorrect values and missing values. We analyzed the

whole dataset and listed out only the useful features. The listing of features can result in better accuracy and contains only
valuable features as to come up with the specific information like the owner, place of registration, address. pictorial representation of system architecture is shown in which includes various phases, namely, Data pre-processing and feature selection, Splitting of Pre-processed Data into train and test set, training and testing of models respectively.



4. PROPOSED SYSTEM

The proposed system is a holistic pipeline that integrates real-time social media monitoring with machine learning to forecast customer churn. The architecture is modular and comprises the following primary components:

Tweet Extraction: Based on relevant keywords, tweets about telecom service providers are extracted using the Tweepy API. This ensures that the data collected is domain-specific and represents real customer feedback.

Text Preprocessing and NLP Module: Every tweet is subjected to a strong cleaning process with the Natural Language Toolkit (NLTK). This involves tokenization by TweetTokenizer, stopword removal, removal of URLs, hashtags, user names, and lemmatizing tokens to their root forms. This process cleans and standardizes the input text for downstream processing.

Sentiment Classification: Sentiment analyzer like VADER is used to classify cleaned tweets on the basis of a rule-based sentiment analysis. Tweets are given a sentiment polarity score and classified into positive, neutral, or negative classes. These classes are used as proxy measures of customer satisfaction.

Churn Labeling Logic: Tweets that show negative sentiment are marked as prospective churn-indicators. This assumption is made based on the hypothesis that dissatisfaction from customers occurs before cancellation of the service.

Feature Engineering and Data Preparation: Metadata such as tweet length, sentiment score, and complaint frequency are utilized to construct enriched feature vectors. These features are then combined into a dataset appropriate for training

a machine learning model.

Model Training and Development: An Extreme Gradient Boosting (XGBoost) classifier is employed because it can efficiently work with sparse data and handle classification problems with imbalanced datasets. Hyperparameter tuning is achieved through cross-validation methods.

Evaluation and Reporting: The model is tested using typical classification metrics such as precision, recall, accuracy, F1-score, and confusion matrix analysis. Model insights can be presented through feature importance graphs.

5. METHODOLOGY

The methodology section describes the sequential process adopted for data extraction, processing, modeling, and evaluation:

5.1 Data Acquisition:

Twitter data was gathered using the Tweepy API by searching tweets that referenced mobile network operators. Tweets were restricted to those tweeted in English and geotagged or referencing Indian telecom operators. Every tweet contained metadata like user handle, creation time, location, and tweet text.

5.2 Preprocessing:

The raw tweets were preprocessed with the following steps:

i) Stopword removal, URLs, special characters, and numerics.

ii)Tokenizing using NLTK's TweetTokenizer.

iii)Lemmatization to bring words to their base form.

iv) Removed short tweets and non-informative text.

It gave us a dataset of clean, tokenized, and lemmatized tweets that were ready for sentiment analysis.

5.3 Sentiment Labeling:

Every cleaned tweet was run through the VADER sentiment analyzer that yields a compound score ranging from -1 (most negative) to +1 (most positive). Thresholds were used:

Compound score > 0.05: Positive

Compound score < -0.05: Negative

Otherwise: Neutral

These sentiment labels were used as the target variable for churn likelihood.

5.4 Feature Engineering and Dataset Construction:

New features were extracted like:

i)Tweet length (character count)

ii) Number of hashtags or mentions

iii)Sentiment polarity score

iv)Time of posting of the tweet (to study sentiment trends)

v)Features and labels were aggregated into a DataFrame for modeling.

5.5 Building and Training the Model:

The dataset was divided into a training set and a testing set. The XGBoost model was used due to its gradient boosting feature, regularization capacity, and multicollinearity robustness.

5.6 Assessment:

Model performance was assessed on:

Accuracy

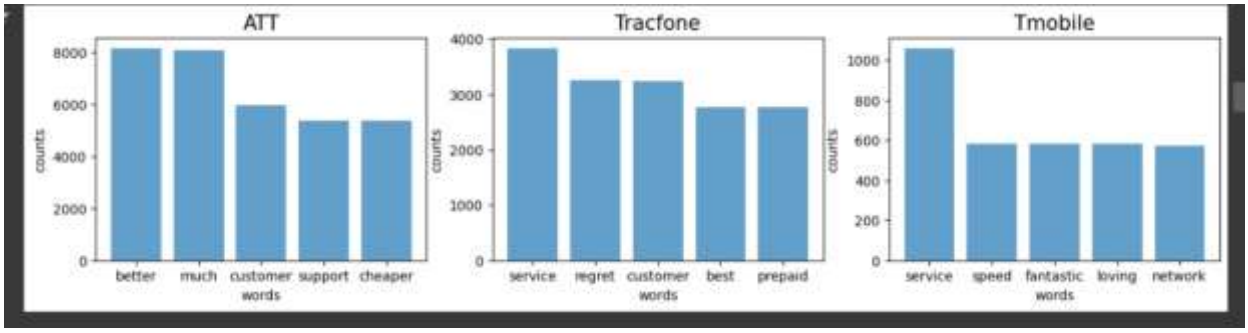
Precision

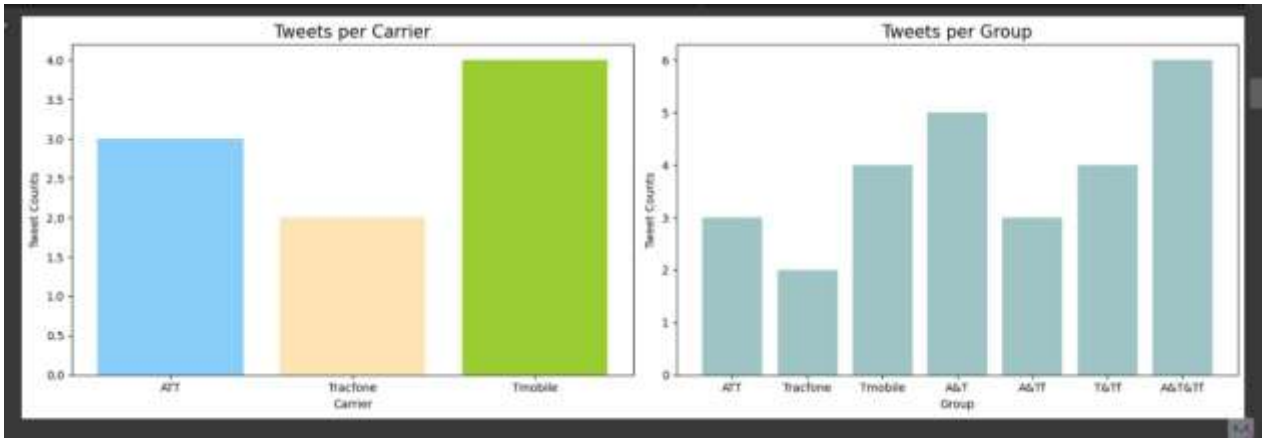
Recall

F1-score

Performance was compared between models that were trained with and without sentiment features to confirm the value added by opinion mining.

6. EXPERIMENTAL RESULTS





```
  warnings.warn(smsg, UserWarning)
XGBoost Accuracy: 84.81 %
XGBoost Precision: 0.5031111111111111
XGBoost Recall: 0.5398187887458273
Confusion Matrix:
 [[10500  1118]
 [  965  1132]]
```

7. CONCLUSION

This paper demonstrates that sentiment on social media can be a powerful predictor in models for customer churn. By incorporating the sentiments of customers as expressed on platforms like Twitter into predictive models, companies are able to obtain a better and fuller picture of customer behavior. The approach proposed here, which includes real-time data acquisition, preprocessing with natural language processing tools, sentiment classification, and modeling based on XGBoost, has evidenced encouraging results.The results show clearly that sentiment indicators are indeed precursors of possible churn, enabling firms to take proactive measures through the implementation of retention strategies. Furthermore, the framework that this project has developed is applicable in other industries beyond telecommunications, such as banking, insurance, and e-commerce, where customer comments on digital platforms are rampant.Future research can involve the use of more sophisticated sentiment analysis techniques like transformer-based models (e.g., BERT, RoBERTa) to enhance the accuracy of classification. Real-time implementation with dashboards for monitoring churn and automated alerting systems can also assist customer service teams in proactively resolving churn threats as they arise. Overall, this sentiment-informed strategy offers a scalable and efficient solution to an age-old business problem.

8. REFERENCE

[1] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning and social network analysis in big data platform,” arXiv preprint arXiv:1904.00690, 2019.
[2] C.-P. Wei and I.-T. Chiu, “Turning telecommunications call details to churn prediction: a data mining approach,” Expert systems with applications, vol. 23, no. 2, pp. 103–112, 2002.
[3] H. Faris, “A hybrid swarm intelligent neural network model for customer churn prediction and identifying the influencing factors,” Information, vol. 9, no. 11, p. 288, 2018.
[4] V. Mahajan, M. Richa, and M. Renuka, “Review on factors affecting customer churn in telecom sector,” International Journal of Data Analysis Techniques and Strategies, vol. 9, no. 2, pp. 122–144, 2017.
[5] A. Rodan, A. Fayyoumi, H. Faris, J. Alsakran, and O. Al-Kadi, “Negative correlation learning for customer churn prediction: A comparison study,” The Scientific World Journal, vol. 2015, 2015.
[6] K. Coussement, D. F. Benoit, and D. Van den Poel, “Improved marketing decision making in a customer churn prediction context using generalized additive models,” Expert Systems with Applications, vol. 37, no. 3, pp. 2132–2143, 2010.
[7] A. Sharma and P. Prabin, “A neural network based approach for predicting customer churn in cellular network services,” International Journal of Computer Applications, vol. 27, no. 11, pp. 26–31, 2011.
[8] J. Burez and D. Van den Poel, “Handling class imbalance in customer churn prediction,” Expert Systems with Applications, vol. 36, no. 3, pp. 4626–4636, 2009.
[9] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” Simulation Modelling Practice and Theory, vol. 55, pp. 1–9, 2015.