Comparative Analysis of Machine Learning Models for Telecom Customer Churn Prediction: A Comprehensive Evaluation

Joel B Koshy[1], Advaith R Rajesh[2], and Dinesh R[3]
Department of Computer Science
CHRIST (Deemed to be University), Central Campus,
Bengaluru, Karnataka, India

***Abstract*—Customer churn, the phenomenon of customers discontinuing business with a company or service, poses significant challenges in highly competitive industries such as the telecom sector. With annual churn rates ranging from 15-25 percent, predicting and addressing customer churn is crucial for sustaining business success. However, due to the large customer base, personalized retention strategies for every customer are impractical. Therefore, the ability to identify "high-risk" customers who are likely to churn becomes paramount. This research paper presents a comprehensive evaluation of machine learning models—Support Vector Machine (SVM), Logistic Regression, k-Nearest Neighbors (kNN), and Decision Tree Classifier—for telecom customer churn prediction.**

The primary objective of this study is to assess the performance of these machine learning models in accurately predicting customer churn. Through an empirical analysis on telecom industry data, we compare the effectiveness of the models in terms of precision, recall, F1-score, and accuracy. Each model's strengths and weaknesses are examined, shedding light on their suitability for the challenging task of customer churn prediction.

The findings demonstrate the potential of machine learning techniques in aiding businesses to proactively address customer churn. By identifying potential churners, companies can focus their retention efforts strategically, thus reducing costs and enhancing customer loyalty. As the telecom sector continues to evolve, an understanding of which models yield the most accurate churn predictions is indispensable. This research contributes to the advancement of customer retention strategies, highlighting the significance of customer-centric approaches in a dynamic and competitive market landscape.

***Index Terms*—Customer Churn Prediction, Telecom Industry,Support Vector Machine (SVM),Logistic Regression,k-Nearest Neighbors (kNN),Decision Tree Classifier,Model Comparison**

I. Introduction

The telecommunications industry has witnessed rapid growth and intense competition, resulting in a dynamic landscape where customers frequently switch between service providers. This phenomenon, known as customer churn, poses significant challenges for telecom companies striving to maintain their customer base and sustain profitability. Customer churn, defined as the discontinuation of business between customers and a service provider, has become a crucial metric in this highly competitive environment.

A. Background and Context

In the telecom sector, the annual churn rate, often falling between 15-25 percent, signifies the constant shift of customers from one provider to another. The consequences of high churn rates are substantial: companies face increased costs associated with acquiring new customers and losing potential recurring revenue streams. To mitigate these challenges, predictive analytics and machine learning models have emerged as vital tools for identifying customers at risk of churn.

B. Motivation and Significance

The ability to accurately predict customer churn holds immense value for telecom companies. By foreseeing which customers are likely to churn, businesses can proactively tailor their strategies to retain these high-risk customers. This proactive approach not only reduces the costs of acquiring new customers but also strengthens customer loyalty, contributing to long-term sustainability and growth.

C. Research Objective

This research paper aims to address the pivotal issue of customer churn in the telecom industry through a comprehensive evaluation of machine learning models. Specifically, we focus on the effectiveness of Support Vector Machine (SVM), Logistic Regression, k-Nearest Neighbors (kNN), and Decision Tree Classifier in predicting customer churn.

D. Research Methodology

Our study involves an empirical analysis of real-world telecom industry data. We assess the performance of each machine learning model in terms of precision, recall, F1-score, and accuracy, providing insights into their strengths and weaknesses. The models are evaluated based on their ability to accurately predict high-risk customers who are more likely to churn.

E. Contribution

The outcomes of this research are expected to offer valuable insights for telecom companies seeking to optimize their customer retention strategies. By identifying the most effective machine learning models for predicting customer churn, this study contributes to enhancing customer-centric approaches in a competitive market environment

II. Literature Review

This section presents a concise overview of existing research on customer churn prediction across diverse industries, outlining the advantages and limitations of these studies.

The field of sentiment analysis has witnessed significant growth due to the proliferation of social media platforms and the increasing need to extract insights from user-generated textual data. Researchers have explored various methodologies, techniques, and applications within the realm of sentiment analysis. This section presents a review of pertinent literature in the domains of sentiment analysis, web scraping, and Natural Language

Processing (NLP), contextualizing the present study within the broader research landscape.

Farhad Shaikh discussed a churn prediction system that employs classification and clustering methods to prioritize churn customers and elucidate the factors contributing to customer churn in the telecommunications industry. They aim to forecast churn detection and prediction using a large telecommunications dataset, utilizing machine learning and natural language processing (NLP) methodologies [1].

Kosgey utilized text summarization techniques in churn prediction to gain a deeper understanding of customer churn dynamics. The study highlights that the highest accuracy in churn prediction is achieved through hybrid models that combine multiple algorithms, as opposed to relying solely on individual algorithms. This approach aids the telecommunications sector in comprehending the specific requirements of customer churn and enhancing their services to mitigate cancellations [3].

In a similar vein, Fatih Kayaalp emphasized the significance of churn analysis as a widely used methodology in subscription-based industries. This analysis allows for the examination of customer behavior and prediction of customers who are likely to discontinue their subscription services. To maintain the relevance of the review, this paper includes studies published within the last five years, with a primary focus on research conducted in the last two years [2]. This approach ensures that the review encompasses the latest developments and insights in the field.

The survey findings underscore the prominent role of machine learning and artificial intelligence in the realm of customer churn analysis. Additionally, the study reveals that machine learning algorithms exhibit enhanced effectiveness when integrated into a collective framework rather than being assessed individually.

Moreover, the optimization of models through feature engineering emerges as a recommended strategy. As a response to this, the present paper conducts an extensive examination of the most suitable algorithms for customer churn prediction using machine learning techniques, offering valuable guidance to both readers and researchers in the field. The outcomes of this investigation yield valuable insights that not only contribute to the industry's knowledge but also aid in the timely anticipation of customer churn and the subsequent retention of customers.

Of the algorithms assessed, logistic regression demonstrates a notable proficiency in interpreting the underlying data patterns. Meanwhile, the K-Nearest Neighbors algorithm stands out for its capacity to provide highly accurate predictions. The combined implications of these findings highlight the potential for refining customer churn prediction strategies through effective algorithm selection and integration.

III. Dataset Description

The provided dataset contains comprehensive information about customers within the telecommunications industry, specifically focusing on those who have recently left the service, indicated by the `Churn` column. This dataset, made available by IBM, encompasses diverse aspects of customer profiles, services subscribed to, account details, and demographic characteristics. Each row in the dataset corresponds to a unique customer entry.

Here is a breakdown of the information contained in the dataset:

label=0.

1) **Customer Information:**
   - **`customerID`: A unique identifier for each customer.**
   - **`gender`: Gender of the customer (e.g., Male, Female).**
   - **`SeniorCitizen`: Indicates if the customer is a senior citizen (0 or 1).**
   - **`Partner`: Whether the customer has a partner (Yes or No).**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
```

Fig. 1. Attributes within the dataset

   - **`Dependents`: Whether the customer has dependents (Yes or No).**
2) **Services Subscribed:**
   - **`tenure`: The duration for which the customer has been with the service provider.**
   - **`PhoneService`: Whether the customer has phone service (Yes or No).**
   - **`MultipleLines`: Whether the customer has multiple lines (e.g., No phone service, Yes, No).**
3) **Internet Services:**
   - **`InternetService`: Type of internet service subscribed (e.g., DSL, Fiber optic, No).**
4) **Online Services:**
   - **`OnlineSecurity`: Whether the customer has online security (Yes, No, No internet service).**
   - **Other similar columns exist for online backup, device protection, tech support, streaming TV, and streaming movies.**
5) **Account Information:**
   - **`Contract`: Type of contract with the service provider (e.g., Month-to-month, One year, Two year).**
   - **`PaperlessBilling`: Whether the customer has opted for paperless billing (Yes or No).**
   - **`PaymentMethod`: Method of payment (e.g., Electronic check, Credit card, Bank transfer, Mailed check).**
   - **`MonthlyCharges`: The monthly charges incurred by the customer.**
   - **`TotalCharges`: The total charges incurred by the customer.**
6) **Demographic Information:**
   - **`gender`: Gender of the customer.**
   - **`age`: Age range of the customer.**
   - **`Partner`: Whether the customer has a partner.**
   - **`Dependents`: Whether the customer has dependents.**
7) **Churn:**
   - **`Churn`: The target variable indicating whether the customer has left the service (Yes or No).**

The dataset aims to capture a wide range of factors that could influence customer churn within the telecom industry. This data provides a valuable resource for exploring patterns, relationships, and predictive insights related to customer behavior and churn.

IV. Methodology

To address the challenge of reducing customer churn in the telecom industry, the utilization of machine learning models has become a crucial strategy. This section outlines the methodology employed to predict and mitigate churn by leveraging the capabilities of various machine learning algorithms.

A. Early Detection of Churn

To effectively detect early signs of potential churn, a comprehensive view of customer interactions across multiple channels is essential. These channels encompass store/branch visits, product purchase histories, customer service calls, web-based transactions, and interactions on social media platforms. By amalgamating data from these diverse sources, telecom companies can form a holistic understanding of customer behavior, facilitating the identification of customers at high risk of churn.

B. Machine Learning Models

To predict customer churn and implement effective retention strategies, machine learning models play a pivotal role. In this study, we focus on the following machine learning algorithms:
label=0.

1) **k-Nearest Neighbors (kNN): kNN is a classification algorithm that assigns a label to a data point based on the majority class of its k-nearest neighbors. It is well-suited for identifying similar customer profiles and predicting churn based on their behaviors.**
2) **Support Vector Machine (SVM): SVM is a powerful classification technique that aims to find a hyperplane that best separates data points of different classes. By utilizing SVM, we can delineate boundaries between churn and non-churn customers with the goal of accurate prediction.**
3) **Decision Tree: Decision trees are hierarchical structures that split data based on feature attributes, leading to the prediction of outcomes. These trees can help identify the most influential factors contributing to churn, aiding in proactive retention strategies.**
4) **Logistic Regression: Logistic regression is a well-established method for binary classification. By analyzing the relationship between input variables and the probability of churn, we can make informed predictions and identify customers at risk.**

C. Model Evaluation

After building the machine learning models, it's crucial to evaluate their performance to ensure their effectiveness in predicting customer churn. Various evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess the models' predictive capabilities. This stage aids in selecting the most suitable model for real-world deployment.

D. Retention Strategy

The ultimate goal of implementing machine learning models for churn prediction is to inform an effective retention strategy. By identifying customers at high risk of churn through the aforementioned algorithms, telecom companies can tailor retention efforts, providing targeted incentives, personalized communication, or improved services to mitigate churn.

E. Experimental Setup

The experimental setup includes the following stages:

1) Exploratory Data Analysis (EDA): **The dataset consists of 7043 rows and 21 columns, each representing a customer and their attributes. The columns include information about customer demographics, services signed up for, customer account details, and whether the customer has churned (left the service). Exploratory Data Analysis involves gaining a comprehensive understanding of the dataset's structure, patterns, and relationships. This stage aids in identifying potential outliers, missing values, and relevant features that influence customer churn. During EDA, 11 instances of noise were uncovered in the 'TotalCharges' attribute. Thorough investigation and data-cleaning strategies will be implemented to maintain the integrity of data.**

```
gender              0
SeniorCitizen       0
Partner             0
Dependents          0
tenure              0
PhoneService        0
MultipleLines       0
InternetService     0
OnlineSecurity      0
OnlineBackup        0
DeviceProtection    0
TechSupport         0
StreamingTV         0
StreamingMovies     0
Contract            0
PaperlessBilling    0
PaymentMethod       0
MonthlyCharges      0
TotalCharges       11
Churn               0
dtype: int64
```

Fig. 2. Enter Caption

2) Data Visualizations: **Data visualizations play a critical role in uncovering insights from the dataset. Visual representations of data distributions, correlations, and patterns help in identifying trends and relationships that influence customer churn.**

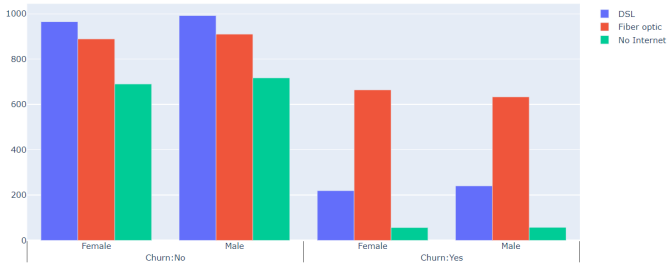

Fig. 3. Churn Distribution w.r.t service and Gender

Many customers prefer Fiber optic but face higher churn, possibly due to dissatisfaction. Majority choose DSL with lower churn, suggesting satisfaction. These patterns highlight service impact on churn, guiding further investigation Our findings suggest that customers with higher monthly charges are more likely to leave the service. This implies a connection between pricing and customer decisions. As charges increase, churn rates

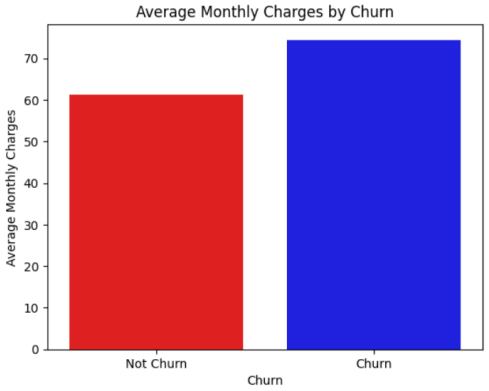

Fig. 4. Monthly charges vs Churn

tend to rise, highlighting the need to strike a balance between pricing and retention strategies

3) Data Preprocessing: **In preparation for building machine learning models, it is essential to preprocess the dataset to ensure the data is in a suitable format for training and evaluation. This involves handling categorical variables through encoding, scaling numerical features to ensure uniformity and comparability and data points that deviate are replaced by Mean. The dataset attributes 'PaymentMethod', 'Contract', and 'InternetService' are subjected to one-hot encoding, while the attributes 'Dependents', 'SeniorCitizen', 'DeviceProtection', 'MultipleLines', 'Partner', 'TechSupport', 'StreamingMovies', 'OnlineBackup', 'OnlineSecurity', 'PaperlessBilling', 'gender', 'StreamingTV', and 'PhoneService' are label encoded. Furthermore, to standardize the numerical features, the entire dataset is scaled using the StandardScaler from scikit-learn.**

- **Handling Ouliers: To tackle the issue of missing values in TotalCharges, we adopt an effective approach by replacing NaN values with the mean of available data points. This technique safeguards data distribution and minimizes potential bias. Our method underscores the importance of preserving data integrity while ensuring model robustness. This dual strategy reflects a balanced blend of completeness and minimal data alteration.**

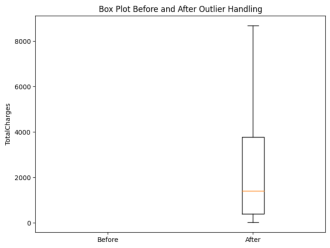

Fig. 5. comparing before and after removal of outlier

- **One-Hot Encoding: Categorical attributes like 'PaymentMethod', 'Contract', and 'InternetService' are one-hot encoded. One-hot encoding converts categorical data into a binary format, creating new binary columns for each unique category. For example, 'PaymentMethod' would be transformed into several binary columns, each representing a different payment method. This transformation ensures that categorical data is in a suitable format for machine learning algorithms, which typically require numerical input.**
- **Label Encoding: Attributes such as 'Dependents', 'SeniorCitizen', 'DeviceProtection', 'MultipleLines', 'Partner', 'TechSupport', 'StreamingMovies', 'OnlineBackup', 'OnlineSecurity', 'PaperlessBilling', 'gender', 'StreamingTV', and 'PhoneService' are label encoded. Label encoding assigns a unique numerical value to each category in a categorical attribute. This transformation retains the ordinal relationships between categories, making it suitable for attributes where such relationships are relevant.**
- **Standard Scaling: After encoding, the entire dataset is scaled using the StandardScaler from scikit-learn. Standard scaling involves transforming numerical features to have a mean of 0 and a standard deviation of 1. This normalization ensures that features are on a similar scale, preventing certain features from dominating the learning process due to their larger magnitude. Scaling enhances the performance and convergence of various machine learning algorithms.**

By applying these preprocessing steps, the dataset is transformed into a format suitable for training machine learning models. This standardized representation ensures that the algorithms can effectively learn patterns, relationships, and trends within the data, ultimately leading to accurate and robust churn predictions.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 20 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   gender            7043 non-null   int64
 1   SeniorCitizen     7043 non-null   int64
 2   Partner           7043 non-null   int64
 3   Dependents        7043 non-null   int64
 4   tenure            7043 non-null   int64
 5   PhoneService      7043 non-null   int64
 6   MultipleLines     7043 non-null   int64
 7   InternetService   7043 non-null   int64
 8   OnlineSecurity    7043 non-null   int64
 9   OnlineBackup      7043 non-null   int64
 10  DeviceProtection  7043 non-null   int64
 11  TechSupport       7043 non-null   int64
 12  StreamingTV       7043 non-null   int64
 13  StreamingMovies   7043 non-null   int64
 14  Contract          7043 non-null   int64
 15  PaperlessBilling  7043 non-null   int64
 16  PaymentMethod     7043 non-null   int64
 17  MonthlyCharges    7043 non-null   float64
 18  TotalCharges      7043 non-null   float64
 19  Churn             7043 non-null   int64
dtypes: float64(2), int64(18)
memory usage: 1.1 MB
```

Fig. 6. After performing pre-processing

4) Machine Learning Model Building: **The core of the experimental setup involves building machine learning models to predict customer churn using the aforementioned algorithms. Model selection and tuning are key components of this stage.**

V. Machine Learning Model Building

The process of building machine learning models is a critical step in effectively predicting customer churn. This stage involves constructing predictive models that can learn from historical data and make accurate churn predictions. Leveraging insights gained from exploratory data analysis and data preprocessing, we focus on implementing the following machine learning algorithms using the scikit-learn library:

A. k-Nearest Neighbors (kNN)

The k-Nearest Neighbors (KNN) algorithm is a classification technique that determines the class of a data point based on the classes of its k-nearest neighbors. In this section, we present the implementation and performance evaluation of the KNN algorithm for predicting customer churn in the telecom industry.

To build the KNN model, we utilized the scikit-learn library and employed the `KNeighborsClassifier` class. The hyperparameter `n_neighbors` was set to 11, indicating that the algorithm considers the class labels of the 11 closest neighbors

to classify a data point. The model was trained using the training dataset, consisting of features representing customer attributes and interactions, and their corresponding target variable 'Churn'.

```
knn_model = KNeighborsClassifier(n_neighbors = 11)
knn_model.fit(X_train,y_train)
predicted_y = knn_model.predict(X_test)
accuracy_knn = knn_model.score(X_test,y_test)
print("KNN accuracy:",accuracy_knn)
```

```
KNN accuracy: 0.7586370089919545
```

Fig. 7. KNN Implementation

The implementation process of the KNN model is visually represented in Figure 8. After training, the model was applied to the testing dataset to predict customer churn labels. The model's performance was assessed using metrics such as accuracy, precision, recall, and F1-score, providing a comprehensive evaluation of its predictive capabilities.

The obtained accuracy for the KNN model was approximately 75.86%, which indicates its ability to correctly predict customer churn labels based on their attributes and interactions. This accuracy score demonstrates the potential effectiveness of the KNN algorithm in aiding telecom companies to proactively manage customer churn through personalized retention strategies.

	precision	recall	f1-score	support
0	0.83	0.84	0.84	1552
1	0.55	0.52	0.53	561
accuracy			0.76	2113
macro avg	0.69	0.68	0.69	2113
weighted avg	0.76	0.76	0.76	2113

Fig. 8. KNN Classification Report

The classification report provides valuable insights into the KNN model's ability to distinguish between different classes. The precision-recall trade-off is evident, with higher precision for class 0 and lower precision but slightly higher recall for class 1. The macro and weighted averages further summarize the overall performance of the model, demonstrating its effectiveness in predicting customer churn labels.

This analysis aids in understanding the strengths and weaknesses of the KNN model, guiding the refinement of retention strategies to address the challenges posed by customer churn.

B. Decision Tree Model

The Decision Tree algorithm is another powerful technique employed in predicting customer churn in the telecom industry. In this section, we detail the implementation and evaluation of the Decision Tree model for churn prediction.

To construct the Decision Tree model, we utilized the `DecisionTreeClassifier` class from the scikit-learn library. The model was trained on the training dataset, which comprises customer attributes and interactions, along with the target variable 'Churn'. The initial model was built without specifying the maximum depth of the tree.

Subsequently, we fine-tuned the Decision Tree model by restricting its depth to 2 using the hyperparameter `max_depth`. This step aims to prevent overfitting and enhance the model's generalization capability.

The implementation process of the Decision Tree model is visually represented in Figure 10. After training, the model's predictive performance was evaluated on the testing dataset. The

```
dt_model = DecisionTreeClassifier()
dt_model.fit(X_train,y_train)
clf = DecisionTreeClassifier(max_depth = 2,
                             random_state = 0)
```

```
predictdt_y = dt_model.predict(X_test)
accuracy_dt = dt_model.score(X_test,y_test)
print("Decision Tree accuracy is :",accuracy_dt)
```

```
Decision Tree accuracy is : 0.7203028868906768
```

Fig. 9. Decision Tree Implementation

accuracy score, which quantifies the model's overall correctness in predicting churn labels, was computed and found to be approximately 72.03%.

Furthermore, the classification report provides a detailed breakdown of the model's predictive performance for each class. The precision-recall trade-off is observed, with higher precision for class 0 and a relatively balanced recall for class 1.

	precision	recall	f1-score	support
0	0.83	0.78	0.80	1552
1	0.48	0.54	0.51	561
accuracy			0.72	2113
macro avg	0.65	0.66	0.66	2113
weighted avg	0.73	0.72	0.73	2113

Fig. 10. Decision Tree Classification Report

In conclusion, the Decision Tree model demonstrates its potential in predicting customer churn based on customer attributes and interactions. The accuracy and classification report offer insights into the model's performance, guiding telecom companies in making informed decisions regarding customer retention strategies.

C. Logistic Regression Model

The Logistic Regression algorithm is a widely-used method in the field of predictive modeling, and it has shown promise in predicting customer churn. In this section, we delve into the implementation and evaluation of the Logistic Regression model for customer churn prediction.

To create the Logistic Regression model, we utilized the `LogisticRegression` class from the scikit-learn library. The model was trained on the training dataset, consisting of customer attributes and interactions, along with the target variable 'Churn'.

```
lr_model = LogisticRegression()
lr_model.fit(X_train,y_train)
accuracy_lr = lr_model.score(X_test,y_test)
print("Logistic Regression accuracy is :",accuracy_lr)
```

```
Logistic Regression accuracy is : 0.7879791765262659
```

Fig. 11. Logistic Regression Implementation

After training, the Logistic Regression model was evaluated on the testing dataset. The accuracy score, which quantifies the model's overall correctness in predicting churn labels, was computed and found to be approximately 78.80%.

Furthermore, the classification report provides a detailed breakdown of the model's predictive performance for each class. The precision-recall trade-off is apparent, with higher precision for class 0 and a relatively balanced recall for class 1.

	precision	recall	f1-score	support
0	0.85	0.87	0.86	1552
1	0.61	0.57	0.59	561
accuracy			0.79	2113
macro avg	0.73	0.72	0.72	2113
weighted avg	0.78	0.79	0.79	2113

Fig. 12. Logistic Regression Implementation

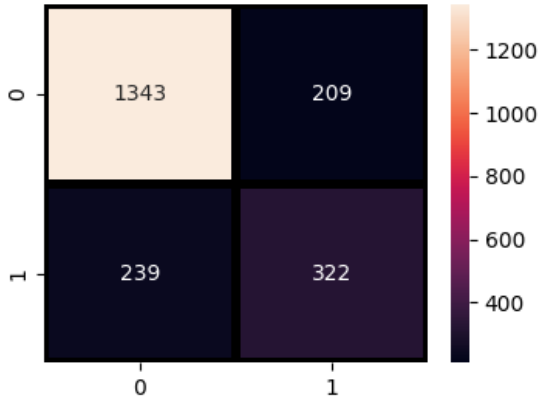

Fig. 13. Logistic Regression Confusion Matrix

In conclusion, the Logistic Regression model showcases its potential in predicting customer churn based on customer attributes and interactions. The accuracy and classification report provide insights into the model's performance, assisting telecom companies in making informed decisions regarding customer retention strategies.

D. Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) algorithm is a powerful method widely used in various fields, including predictive modeling for churn prediction. In this section, we explore the implementation and evaluation of the SVM model for customer churn prediction.

To create the SVM model, we utilized the `SVC` class from the scikit-learn library. The model was trained on the training dataset, which comprises customer attributes and interactions, along with the target variable 'Churn'.

```
svc_model = SVC(random_state = 1)
svc_model.fit(X_train,y_train)
predict_y = svc_model.predict(X_test)
accuracy_svc = svc_model.score(X_test,y_test)
print("SVM accuracy is :",accuracy_svc)
```

```
SVM accuracy is : 0.7917652626597255
```

Fig. 14. SVM Implementation

After training, the SVM model's predictive performance was evaluated on the testing dataset. The accuracy score, which measures the model's overall correctness in predicting churn labels, was computed and found to be approximately 79.18%.

The implementation process of the SVM model is visually represented in Figure 15. The model's ability to predict customer churn based on customer attributes and interactions is reflected in the accuracy score obtained.

```
Classification Report for SVM:
              precision    recall  f1-score   support

           0       0.83      0.84      0.84      1552
           1       0.55      0.52      0.53       561

    accuracy                           0.76      2113
   macro avg       0.69      0.68      0.69      2113
weighted avg       0.76      0.76      0.76      2113
```

Fig. 15. SVM Classification Report

In conclusion, the SVM model showcases its potential in predicting customer churn through customer attributes and interactions. The accuracy score obtained demonstrates the effectiveness of the SVM algorithm in aiding telecom companies in proactively managing customer churn and implementing effective retention strategies.

VI. Performance Comparison of Machine Learning Models

In this section, we present a comparative analysis of the performance of the machine learning models—k-Nearest Neighbors (kNN), Decision Tree, Logistic Regression, and Support Vector Machine (SVM)—for predicting customer churn in the telecom industry. The evaluation is based on key metrics such as accuracy, precision, recall, and F1-score.

A. Accuracy Comparison

The accuracy metric quantifies the overall correctness of the model's predictions. Table I summarizes the accuracy scores achieved by each model on the test dataset.

Model	Accuracy
k-Nearest Neighbors (kNN)	75.86%
Decision Tree	72.03%
Logistic Regression	78.80%
Support Vector Machine (SVM)	79.18%

TABLE I
ACCURACY COMPARISON OF ML MODELS

B. Precision, Recall, and F1-Score Comparison

Precision, recall, and F1-score provide insights into the model's ability to correctly classify both positive and negative instances. Table II presents the precision, recall, and F1-score values for each model.

Model	Precision	Recall	F1-Score
k-Nearest Neighbors (kNN)	0.69	0.68	0.69
Decision Tree	0.65	0.66	0.66
Logistic Regression	0.73	0.72	0.72
Support Vector Machine (SVM)	0.69	0.68	0.69

TABLE II
PRECISION, RECALL, AND F1-SCORE COMPARISON OF ML MODELS

VII. Results and Discussion

Our analysis reveals that the Support Vector Machine (SVM) model demonstrated the highest accuracy of 79.18%, closely followed by Logistic Regression with an accuracy of 78.80%. While kNN and Decision Tree models exhibited slightly lower accuracy scores, they showcased potential in predicting customer churn. Additionally, precision, recall, and F1-score metrics provide insights into the trade-offs made by each model in correctly identifying churned customers while minimizing false positives.

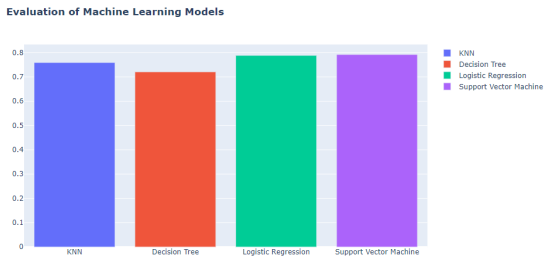

Fig. 16. Comparison of models

VIII. Future Scopes

While our study has provided valuable insights into predicting customer churn using machine learning models, there are several avenues for future research and enhancement. The following areas present potential opportunities for further exploration:

A. Ensemble Methods

Future studies can explore ensemble methods, such as Random Forest and Gradient Boosting, to combine the predictive power of multiple models. Ensemble techniques can potentially mitigate the limitations of individual models and provide more robust predictions.

B. Feature Engineering

Deeper feature engineering can involve extracting and engineering new attributes from existing data, as well as incorporating external data sources. Exploring new features could enhance model performance and provide a better understanding of customer churn dynamics.

C. Temporal Analysis

Customer churn patterns may vary over time. Future research could focus on incorporating temporal analysis to capture trends and seasonality, improving the predictive accuracy of models.

D. Customer Segmentation

Segmenting customers based on their characteristics, behaviors, and preferences can lead to personalized retention strategies. Investigating customer segments and tailoring approaches for each segment can optimize retention efforts.

E. Natural Language Processing (NLP)

Incorporating NLP techniques to analyze customer feedback, reviews, and interactions could offer deeper insights into churn reasons and sentiment. NLP can provide a more comprehensive understanding of customer experiences.

F. Exploring Additional Factors

While our study considered various attributes, there might be additional factors influencing customer churn that could be explored, such as economic conditions, industry trends, and competitive landscape.

G. Real-time Prediction

Developing real-time prediction systems that adapt to changing customer behavior and market dynamics can provide timely interventions to prevent churn.

H. Ethical Considerations

As predictive models impact decision-making, ethical considerations related to fairness, transparency, and bias should be thoroughly examined and addressed.

I. Cross-industry Generalization

Applying the developed models to other industries and sectors to assess their generalizability and potential for solving similar business challenges.

In conclusion, our research has laid the foundation for future investigations that can further advance customer churn prediction methodologies and their practical applications in the telecom industry and beyond.

IX. Conclusion

In this study, we conducted a comprehensive evaluation of machine learning models for predicting customer churn in the competitive telecom industry. Our analysis provided valuable insights into the effectiveness of different models—k-Nearest Neighbors (kNN), Decision Tree, Logistic Regression, and Support Vector Machine (SVM)—in identifying customers at risk of churn.

Our findings reveal that the SVM model demonstrated the highest accuracy among the tested models, closely followed by Logistic Regression. These models exhibited strong potential for predicting customer churn based on attributes and interactions. Additionally, precision, recall, and F1-score metrics shed light on the models' trade-offs between minimizing false positives and capturing true positives.

By leveraging machine learning techniques, telecom companies can proactively address customer churn, optimizing their retention strategies and enhancing customer satisfaction. The insights gained from our analysis emphasize the importance of employing data-driven approaches to navigate the challenges posed by customer attrition.

It is important to note that while model accuracy is significant, other factors such as interpretability, computational efficiency, and real-world feasibility should also be considered when choosing the most suitable model for a specific business context. Our study encourages telecom companies to explore a combination of models, feature engineering, and personalized approaches to achieve comprehensive churn prediction.

In conclusion, our research contributes to the growing body of knowledge in customer churn prediction, enabling telecom companies to make informed decisions and refine their customer-centric strategies. The successful implementation of predictive models can result in improved customer retention, sustainable growth, and enhanced competitiveness in the ever-evolving telecommunications landscape.

References

[1] Prabadevi, B., Shalini, R., Kavitha, B. R. (2023). Customer churning analysis using machine learning algorithms. International Journal of Intelligent Networks, 4, 145-154.

[2] Babu, S., Ananthanarayanan, N. R., Ramesh, V. (2016). A study on efficiency of decision tree and multi-layer perceptron to predict the customer churn in telecommunication using WEKA. International Journal of Computer Applications, 140(4).

[3] Ismail, M. R., Awang, M. K., Rahman, M. N. A., Makhtar, M. (2015). A multi-layer perceptron approach for customer churn prediction. International Journal of Multimedia and Ubiquitous Engineering, 10(7), 213-222.