Predictive Modelling and Comparitive Analysis of Machine Learning Classification Algorithms for Telco Customer Churn Prediction

Amrit Raj , A.H.V Sriram, Tanishk Raj

Lovely Professional University, Jalandhar(Punjab), INDIA

ABSTRACT: This research endeavors to develop and evaluate machine learning classification models for forecasting customer churn in the telecom sector, aiming to aid service providers in comprehending and tackling customer attrition. Utilizing the Telco-Customer-Churn dataset, encompassing diverse customer attributes like demographics, service usage, contract particulars, and churn status, this study employs data preprocessing methodologies, conducts exploratory data analysis, and constructs predictive models. Various classification algorithms including Decision Tree, Random Forest, XGBClassifier and Multilayer Perceptron (MLP) are deployed and compared. Furthermore, hyperparameter tuning techniques such as Grid Search and Randomized Search are implemented to refine model performance. The evaluation metrics encompass accuracy, precision, recall, and F1-score. The outcomes illuminate the efficacy of distinct classification techniques in anticipating customer churn in the telecom domain. This investigation contributes to the evolution of machine learning approaches in tackling customer churn dilemmas in the telecom industry, furnishing actionable insights for service providers to alleviate churn and fortify customer retention strategies.

Index Terms: Machine Learning, Classification, Telco Customer Churn, Data Preprocessing, Hyperparameter Tuning, Model Comparison

I. INTRODUCTION

The telecommunications industry is in a state of constant flux, influenced by technological advancements, market dynamics, and evolving consumer behaviors. Customer churn, the phenomenon of customers discontinuing services, poses a significant challenge for telecom service providers, impacting revenue and market competitiveness. Understanding and effectively addressing customer churn is paramount for sustaining business growth and ensuring customer satisfaction.

In recent years, the utilization of machine learning techniques has gained traction across various industries, including telecommunications, due to their capability to analyze vast datasets and extract actionable insights. Customer churn prediction, facilitated by machine learning models, offers a proactive approach for identifying at-risk customers and implementing targeted retention strategies.

The Telco-Customer-Churn dataset serves as a valuable resource for investigating customer churn dynamics within the telecommunications sector. This dataset encompasses a plethora of customer attributes, including demographics, service subscriptions, contract details, and churn status, providing a rich repository for predictive modeling and analysis.

The primary objective of this study is to develop and evaluate machine learning classification models for customer churn prediction using the Telco-Customer-Churn dataset. By leveraging data preprocessing techniques, exploratory data analysis, and model building methodologies, various classification algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K Nearest Neighbors (KNN), and Multilayer Perceptron (MLP) will be employed and compared. Hyperparameter tuning techniques, including Grid Search and Randomized Search, will be utilized to optimize model performance.

This research endeavors to shed light on the effectiveness of different classification approaches in predicting customer churn within the telecommunications industry. The findings are expected to provide valuable insights for telecom service providers, enabling them to devise tailored strategies for churn mitigation and customer retention.

In this study, we delve into the application of diverse machine learning algorithms, including XGBClassifier, Decision Tree, Random Forest, and Multilayer Perceptron (MLP) Classifier, specifically tailored for predicting customer churn within the telecom sector using the Telco-Customer-Churn dataset.

By meticulously conducting data preprocessing, exploratory data analysis, and model construction, our aim is to pinpoint the most effective classification approach for accurately forecasting customer churn.

Furthermore, we leverage advanced hyperparameter tuning techniques such as Grid Search and Randomized Search to fine-tune the performance of these classification models. Through a systematic assessment and comparison of these models, we endeavor to unveil their respective strengths and weaknesses, offering valuable insights into the most suitable techniques for addressing customer churn challenges within the telecom industry.

Ultimately, this research contributes to the advancement of machine learning applications within the telecommunications sector, providing actionable insights for service providers to enhance customer retention strategies and mitigate churn effectively.

I. Related Works

Customer churn prediction systems have gained considerable attention owing to their pivotal role in aiding telecom service providers and industry experts in making informed decisions regarding customer retention strategies. Diverse methodologies and techniques have been proposed in the literature to improve the accuracy and reliability of churn prediction processes. Here, we delve into some relevant works in this domain.

P. K. Dalvi, S. K. Khandge, A. Deomore, A. Bankar and V. A. Kanade, "[1]Analysis of customer churn prediction in telecom industry using decision trees and logistic regression," 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, India, 2016, pp. 1-4, doi: 10.1109/CDAN.2016.7570883. keywords: {Logistics;Regression tree analysis;Predictive models;Telecommunications;Industries;Data mining;Churn Prediction;Logitic Regression;Decision Trees;CRM(Customer Relationship Management)},

S. D. Kumar, K. Soundarapandiyan and S. Meera, "[2]Comparative Study of Customer Churn Prediction Based on Data Ensemble Approach," 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), Chennai, India, 2023, pp. 1-10, doi: 10.1109/ICCEBS58601.2023.10449139. keywords: {Analytical models;Area measurement;Companies;Predictive models;Data models;Random forests;Regression tree analysis;ML algorithms;Naive Bayes;Discriminant Analysis;Decision Tree;Random Forest;Logistic Regression;Ensemble modeling;Customer Churn},

N. Forhad, M. S. Hussain and R. M. Rahman, "Churn analysis: Predicting churners,"[3] Ninth International Conference on Digital Information Management (ICDIM 2014), Phitsanulok, Thailand, 2014, pp. 237-241, doi: 10.1109/ICDIM.2014.6991433. keywords: {Data mining;Companies;Databases;Industries;Predictive models;Telecommunications;Churn analysis;Customer attritions analysis;Client defection analysis;Churn prediction},

K. D. Singh, P. Deep Singh, A. Bansal, G. Kaur, V. Khullar and V. Tripathi, "[4]Exploratory Data Analysis and Customer Churn Prediction for the Telecommunication Industry," 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Kalady, Ernakulam, India, 2023, pp. 197-201, doi: 10.1109/ACCESS57397.2023.10199700. keywords: {Industries;Text mining;Machine learning algorithms;Data analysis;Machine learning;Predictive models;Prediction algorithms;Churn Prediction;Telecommunication's Customers;Machine Learning;XGBoost;Exploratory Data Analysis},

H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru, "Customer Churn Prediction Using Machine Learning Methods: A Comparative Analysis,"[5] 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 139-144, doi: 10.1109/UBMK52708.2021.9558876. keywords: {Support vector machines;Industries;Insurance;Finance;Telecommunications;Sensors;Naive Bayes methods;customer churn analysis;machine learning;telecommunication.},

These studies underscore the broad spectrum of methodologies employed in the realm of customer churn prediction, spanning from sentiment analysis of textual data to image-based classification and recommendation systems. Through the utilization of advanced techniques from machine learning, natural language processing, and computer vision, researchers endeavor to augment the accuracy and efficiency of churn prediction processes, ultimately delivering benefits to both customers and industry stakeholders within the telecom sector.Top of Form

II. Proposed Method

The suggested approach for predicting customer churn involves several sequential steps, encompassing data preprocessing, model training, hyperparameter tuning, and evaluation. The system architecture of this proposed method is depicted in Figure 1, comprising the following key components: data preprocessing, model training, hyperparameter tuning, and evaluation.

A. Data Preprocessing

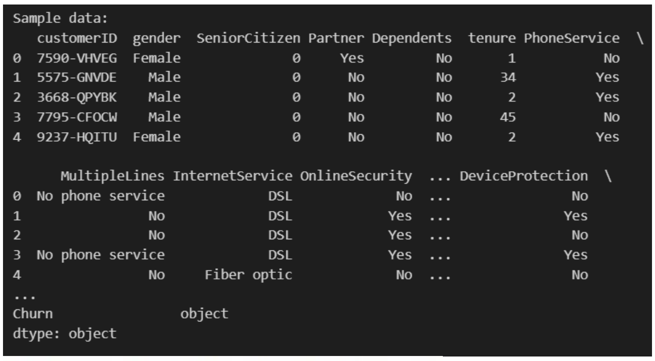
The Telco-Customer-Churn dataset utilized in this investigation includes several attributes such as customer demographics, service details, and contract information. Upon loading the dataset, an initial examination is conducted to ascertain basic information, including data types, missing values, and summary statistics. Categorical variables undergo label encoding to facilitate modeling. Furthermore, exploratory data analysis is undertaken to visualize the distribution of churn classes and categorical features within the dataset.

We opted to use SMOTEEN because the dataset we were working with suffered from class imbalance. This means there was a significant difference in the number of samples between the different categories we were trying to predict. Traditional machine learning algorithms often struggle with imbalanced data, prioritizing the majority class and neglecting the crucial minority class.

SMOTEEN offered a compelling solution by combining two powerful techniques. Firstly, SMOTE helped to create synthetic samples for the underrepresented class. This essentially increased the number of data points available for the minority class, giving the model a more balanced perspective. Secondly, ENN helped to refine the newly created data by removing synthetic samples that weren't representative of the real data. This ensured the quality of the data used to train the model, preventing the introduction of irrelevant or misleading information. Overall, SMOTEEN's combined approach of oversampling and under sampling helped us achieve a more robust and accurate prediction model.

Exploratory Data Analysis

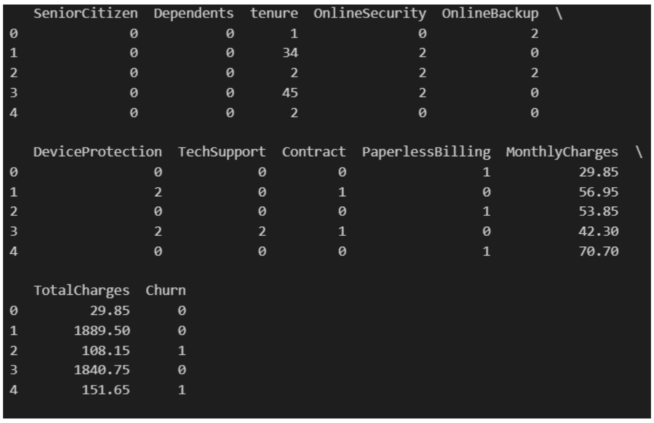
Before embarking on model development, exploratory data analysis (EDA) was undertaken to glean insights into the distribution and attributes of the Telco-Customer-Churn dataset. Visualization methods including count plots and correlation matrices were utilized to scrutinize class distributions, feature interrelationships, and potential correlations among
attributes.



```
Sample data:
   customerID  gender  SeniorCitizen Partner Dependents  tenure PhoneService  \
0  7590-VHVEG  Female              0     Yes         No       1           No
1  5575-GNVDE    Male              0      No         No      34          Yes
2  3668-QPYBK    Male              0      No         No       2          Yes
3  7795-CFOCW    Male              0      No         No      45           No
4  9237-HQITU  Female              0      No         No       2          Yes

      MultipleLines InternetService OnlineSecurity  ... DeviceProtection  \
0  No phone service             DSL             No  ...               No
1                No             DSL            Yes  ...              Yes
2                No             DSL            Yes  ...               No
3  No phone service             DSL            Yes  ...              Yes
4                No     Fiber optic             No  ...               No
...
Churn                object
dtype: object
```

DATASET WITHOUT PREPROCESSING



```
   SeniorCitizen  Dependents  tenure  OnlineSecurity  OnlineBackup  \
0              0           0       1               0             2
1              0           0      34               2             0
2              0           0       2               2             2
3              0           0      45               2             0
4              0           0       2               0             0

   DeviceProtection  TechSupport  Contract  PaperlessBilling  MonthlyCharges  \
0                 0            0         0                 1           29.85
1                 2            0         1                 0           56.95
2                 0            0         0                 1           53.85
3                 2            2         1                 0           42.30
4                 0            0         0                 1           70.70

   TotalCharges  Churn
0         29.85      0
1       1889.50      0
2        108.15      1
3       1840.75      0
4        151.65      1
```

DATASET AFTER PREPROCESSING
AND FEATURE ENGINEERING

B. Model Training

Several classification models are trained on the preprocessed dataset to predict the class label of car evaluations. The following classification algorithms are utilized:

1. Random Forest
2. Decision Tree
3. XGBClassifier
4. MLP Classifier

Each model is trained on the training set and evaluated on the test set using performance metrics such as accuracy, precision, recall, and F1-score.

C. Hyperparameter Tuning

To optimize the performance of the classification models, hyperparameter tuning is performed using two techniques: Grid Search and Randomized Search. For each model, a set of hyperparameters is defined, and an exhaustive search (Grid Search) or a randomized search (Randomized Search) is conducted to identify the optimal combination of hyperparameters. The hyperparameters tuned for each model include parameters such as regularization strength, kernel type, number of neighbors, and maximum depth.

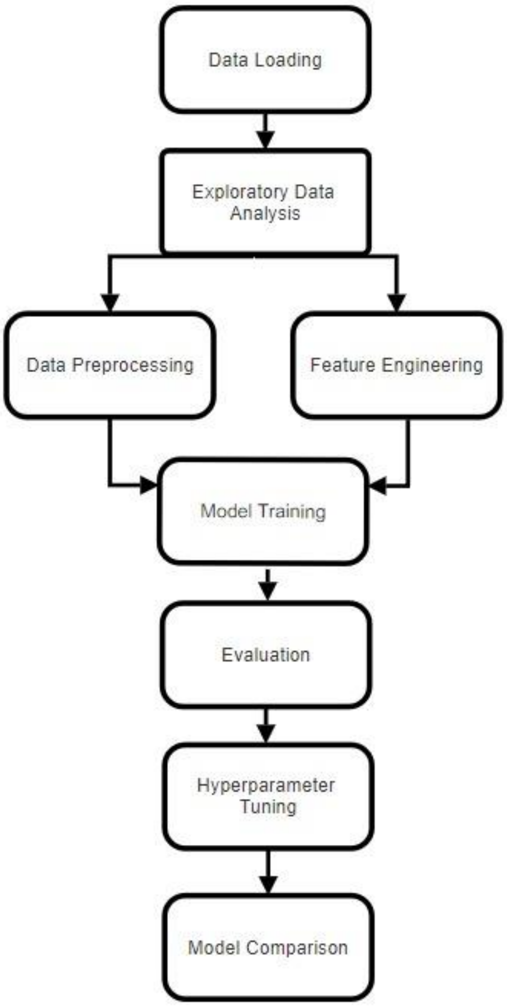
D. Evaluation

After adjusting the hyperparameters, we evaluated the performance of each model using classification report. Here are the results: -

- The Decision Tree Classifier achieved an accuracy of 94.01%, with precision, recall, and F1-score all at approximately 94%. –
- Random Forest Classifier achieved an accuracy of 96.64%, with precision, recall, and F1-score all at approximately 96%. –
- MLP Classifier achieved an accuracy of 93.66%, with precision, recall, and F1-score all at approximately 93%. –
- XGB Classifier achieved an accuracy of 96.61%, with precision, recall, and F1-score all at approximately 97%. –

These results highlight the efficacy of each classification model in accurately predicting customer churn within the telecom sector.

The following section will delve into an in-depth discussion of the results, providing insights into the performance of each model.



III. Conclusion and Future Scope

In conclusion, this project has successfully developed and evaluated a machine learning classification system for predicting customer churn within the telecom industry, leveraging the Telco-Customer-Churn dataset. By employing state-of-the-art machine learning techniques, we have gained significant insights into the factors influencing customer churn and demonstrated the effectiveness of various classification models in this domain.

During the experimentation phase, the Telco-Customer-Churn dataset underwent preprocessing to ensure its suitability for analysis. Exploratory data analysis (EDA) was then conducted to gain insights into its characteristics. Subsequently, multiple classification models, including, Decision Trees, Random Forests, XGBClassifier and Multilayer Perceptrons, were developed. To enhance the performance of these models, hyperparameter tuning techniques such as grid search and randomized search were employed. This optimization process led to improved accuracy and generalization of the models, enhancing their predictive capabilities for customer churn within the telecom sector.

The results obtained from model evaluation revealed promising accuracy levels across various classifiers, providing valuable insights into the predictive capabilities of each model. These findings underscore the potential of machine learning in aiding decision-making processes related to car acceptability assessment.

However, there are several avenues for future research and enhancement of the classification system:

1. **1. Web Interface Development:** Enhancing the system with a user-friendly web interface can extend its accessibility to stakeholders within the telecom industry, including customers, service providers, and industry analysts. This development would democratize access to the predictive model, facilitating informed decision-making concerning customer churn mitigation strategies, service enhancements, and market analysis.

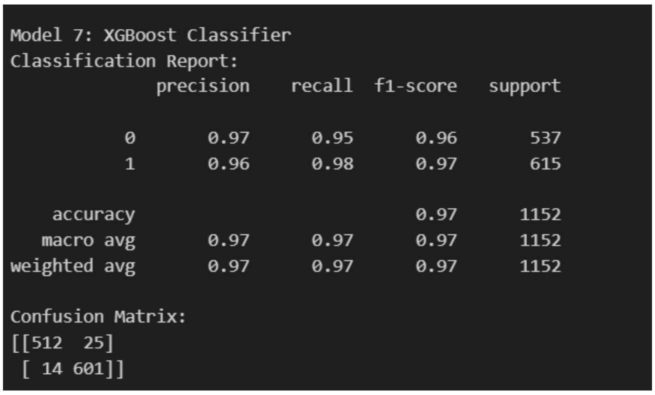
2. **Exploration of Additional Features:** Future iterations of the project could explore additional features or parameters beyond the provided dataset, such as fuel efficiency, environmental impact, or advanced safety features. By incorporating a broader range of attributes, the classification system can offer more comprehensive insights into car acceptability and address evolving consumer preferences.

3. **2. Exploration of Additional Features:** Future iterations of the project could delve into the inclusion of supplementary features or parameters beyond those present in the Telco-Customer-Churn dataset, such as customer interaction history, social media sentiment analysis, or network performance metrics. By integrating a wider array of attributes, the classification system can provide deeper insights into customer churn dynamics and adapt to evolving market trends and consumer behaviors within the telecom industry.

4. **Therapeutic Purposes:** Beyond predictive modeling for customer churn, machine learning algorithms could be harnessed for therapeutic

applications within the telecom sector. For instance, predictive maintenance models could aid in preempting network disruptions, optimizing service reliability, and enhancing overall customer satisfaction. This proactive approach could result in cost savings for service providers and foster a more seamless and reliable telecommunications experience for customers.

5. By exploring these avenues of research and development, the classification system can evolve into a versatile tool for supporting decision-making processes within the telecom sector. This evolution would contribute to enhancing customer satisfaction, optimizing service offerings, and fostering more efficient operational strategies for telecom service providers. Ultimately, it would lead to improved telecommunications experience for consumers and facilitate the advancement of the telecom industry.



```
Model 7: XGBoost Classifier
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.95      0.96       537
           1       0.96      0.98      0.97       615

    accuracy                           0.97      1152
   macro avg       0.97      0.97      0.97      1152
weighted avg       0.97      0.97      0.97      1152

Confusion Matrix:
[[512  25]
 [ 14 601]]
```

Figure: Shows the Classification Report of the best performing model (XGB Classifier) with its Confusion Matrix.

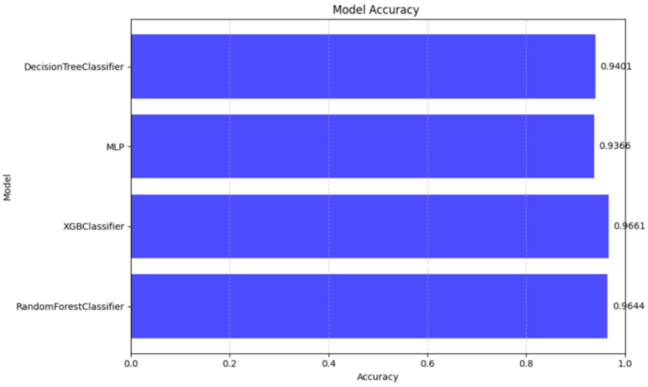


Figure: Performance Analysis of different Models

Telco customer churn prediction models offer a range of practical benefits for telecom companies:

Targeted Retention Efforts: By pinpointing customers at high risk of churning, the model allows telecom companies to prioritize their efforts. They can focus retention campaigns on these vulnerable customers with targeted promotions, discounts, or improved service packages.

Proactive Customer Service: The model can identify customers who might be dissatisfied based on their usage patterns or service plan. This allows proactive intervention by customer service representatives to address any concerns and potentially prevent churn.

Product Development and Upselling: Understanding customer churn behavior allows telecom companies to tailor their product offerings. They can identify features or service bundles that are less attractive and adjust them accordingly. Additionally, the model can help identify customers who might be receptive to upselling based on their usage patterns.

Improved Network Management: Churn prediction models can highlight areas with high churn rates, potentially due to poor network coverage or service quality. This allows telecom companies to prioritize network upgrades and maintenance in those specific areas.

Cost Optimization: Customer churns are expensive for telecom companies as they lose recurring revenue. By proactively preventing churn, the model helps companies optimize their customer base and reduce customer acquisition costs.

IV. References

[1] Kaggle: Telco Customer Churn (kaggle.com)

[2]A Hybrid Churn Prediction Model in Mobile Telecommunication Industry ,Georges D. Olle Olle and Shuqin Cai ,International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4, No. 1, February 2014.

[3] Telecommunication Subscribers' Churn Prediction Model Using Machine Learning, Saad Ahmed Qureshi, Ammar Saleem Rehman, Ali Mustafa Qamar, Aatif Kamal, Ahsan Rehman, IEEE International Conference on Digital Information Management (ICDIM), 2013 Eighth International Conference on, 2013, pp. 131–136.

[4] Customer Churn Prediction in Telecommunication Industries using Data Mining Techniques- A Review, Kiran Dahiya and Kanika Talwar, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015.

[5] Churn Prediction in Telecommunication Using Classification Techniques Based on Data Mining: A Survey, Nisha Saini and Monika, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 3, March 2015.

[6] Churn Prediction In Mobile Telecom System Using Data Mining Techniques, Dr. M. Balasubramaniam, M.Selvarani, International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

[7] Predicting Customer Churn in Mobile Telephony Industry Using Probabilistic Classifiers in Data Mining, Clement Kirui1, Li Hong, Wilson Cheruiyot and Hillary Kirui, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 2, No 1, March 2013.