****

**JINKA UNIVERSTIY**

**COLLEGE OF NATURAL AND COMPUTATIONAL SCIENCE**

**DEPARTMENT OF COMPUTER SCIENCE**

**INTRODUCTION TO ARTIFICIAL INTELLIGENCE**

**TITLE OF PROJECT: TELECOM CUSTOMER HURN PREDICTION**

**USING LEARNING**

**COURSE TITLE: INTRODUCTION TO ARTIFICIAL INTELLIGENCE**

**COURSE CODE: CoSc 3112**

**GROUP TWO ASSIGNMENT PROJECT**

INSTRUCTOR NAME: LENCHO D.

JINKA ETHIOPIA

**TELECOM CUSTOMER HURN PREDICTION USING LEARNING**

MEMBER NAME ID

1. AMANUEL ASHAGRIE…………………………..ugr/13274/14
2. AMARECHI UGANDIE…………………………..uga/0151/14
3. GEMEDA GAMA…………………………………uga/0148/14
4. GIZACHEW MULAW…………………………….ugr/13992/14
5. KIDIST GEDEFAW………………………………..ugr/14190/14

Submission date: 15/09/2016 E.C

JINKA ETHIOPIA

# Abstract

*Telecom customer churn prediction is a critical task in the telecommunications industry, aimed at identifying customers who are likely to discontinue their services. This study leverages machine learning algorithms to enhance the accuracy of churn predictions, providing valuable insights for targeted customer retention strategies. We explore various machine learning models, including logistic regression, random forests. The models are trained and tested on a comprehensive telecom dataset that includes customer demographics, usage patterns, billing information, and service complaints. Feature engineering and selection techniques are employed to improve model performance by identifying the most influential factors contributing to churn. The models are evaluated based on metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). Results indicate that ensemble methods, particularly random forests and gradient boosting machines, achieve superior predictive performance.*

Table Content

Contents page

I.[TitlePage **Error! Bookmark not defined.**](#_Toc169944491)

[**1.** Introduction 1](#_Toc169944495)

[1.2 Objective 1](#_Toc169944496)

[1.3 Scope 1](#_Toc169944497)

[1.4 Methodology Overview 1](#_Toc169944498)

[1.5 Structure 2](#_Toc169944499)

[2. Literature Review 3](#_Toc169944500)

[2.1 Overview of Existing Research 3](#_Toc169944501)

[2.2 Identification of Gaps 3](#_Toc169944502)

[2.3 Relevance to Current Study 3](#_Toc169944503)

[3. Methodology 3](#_Toc169944504)

[3.1 Research Design 3](#_Toc169944505)

[3.2 Data Collection 4](#_Toc169944506)

[3.3 Data Preprocessing 4](#_Toc169944507)

[3.4 Model Selection 4](#_Toc169944508)

[3.5 Training and Testing 4](#_Toc169944509)

[4. Results 5](#_Toc169944510)

[4.1 Presentation of Findings 5](#_Toc169944511)

[4.2 Analysis 6](#_Toc169944512)

[4.3 Model Performance 6](#_Toc169944513)

[5. Discussion 7](#_Toc169944514)

[5.1 Interpretation of Results 7](#_Toc169944515)

[5.2 Comparison with Existing Work 7](#_Toc169944516)

[5.3 Implications 7](#_Toc169944517)

[5.4 Limitations 7](#_Toc169944518)

[6. Conclusion 7](#_Toc169944519)

[6.1 Summary of Key Findings 7](#_Toc169944520)

[6.2 Achievement of Objectives 7](#_Toc169944521)

[6.3 Future Work 8](#_Toc169944522)

[6.4 Final Remarks 8](#_Toc169944523)

[References 9](#_Toc169944524)

[Appendices 10](#_Toc169944527)

# 1. **Introduction**

## 1.1 Background Information

The telecommunications industry is highly competitive, with numerous service providers striving to attract and retain customers. Customer churn, or the phenomenon where customers discontinue their service subscription, is a significant challenge faced by telecom companies. High churn rates can lead to substantial revenue losses, increased marketing costs, and negative impacts on brand reputation. Therefore, predicting customer churn is crucial for telecom companies to develop effective retention strategies, optimize customer service, and maintain a stable revenue stream. The advent of machine learning algorithms has revolutionized churn prediction, offering more accurate and actionable insights into customer behavior.

## 1.2 Objective

The primary objective of this report is to predict customer churn in the telecommunications industry using machine learning algorithms.

* To identify the key factors contributing to customer churn.
* To develop and compare various machine learning models for churn prediction.
* To evaluate the performance of these models using relevant metrics.
* To provide actionable insights and recommendations based on the model outputs to help telecom companies reduce churn rates.

## 1.3 Scope

This report focuses on the application of machine learning algorithms to predict customer churn in the telecommunications sector.

1. Data collection and preprocessing of telecom customer data from Kaggle.
2. Feature selection and engineering to identify relevant predictors of churn.
3. Implementation and comparison of multiple machine learning algorithms.
4. Model evaluation using metrics such as accuracy, precision, recall, and F1-score.
5. Discussion of the results and implications for telecom companies. The report does not cover the economic analysis of churn prevention strategies or the implementation of these strategies in a real-world setting.

## 1.4 Methodology Overview

The methodology employed in this study includes the following steps:

1. **Data Collection**: Gathering telecom customer data from Kaggle reliable sources, which includes demographic information, service usage patterns, and customer interaction history.

https://www.kaggle.com/datasets/ blastchar/ telco-customer-churn

1. **Data Preprocessing**: Cleaning the data to handle missing values, outliers, and noise. This step also involves normalizing and scaling the data as required.
2. **Feature Selection and Engineering**: Identifying and creating relevant features that significantly impact customer churn.
3. **Model Development**: Implementing various machine learning algorithms such as Logistic Regression, Random Forests.
4. **Model Evaluation**: Assessing the performance of the models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC curve.
5. **Model Comparison and Selection**: Comparing the performance of different models to select the best-performing one for churn prediction.
6. **Result Interpretation and Recommendations**: Analyzing the model outputs to derive insights and provide recommendations for reducing churn rates.

## 1.5 Structure

The report is structured as follows:

1. **Introduction**: Provides background information, objectives, scope, methodology overview, and structure of the report.
2. **Literature Review**: Reviews existing research and methodologies in customer churn prediction.
3. **Data Collection and Preprocessing**: Details the data sources, preprocessing steps, and feature engineering techniques.
4. **Machine Learning Models**: Describes the machine learning algorithms used in the study and their implementation.
5. **Model Evaluation and Results**: Presents the evaluation metrics, model performance, and comparative analysis.
6. **Discussion**: Interprets the results, discusses the implications, and provides insights into the findings.
7. **Conclusion**: Summarizes the key findings and suggests future research directions.

This structure aims to provide a comprehensive and systematic approach to understanding and addressing telecom customer churn using machine learning algorithms.

# 2. Literature Review

## 2.1 Overview of Existing Research

Predicting customer churn has been a critical area of research within the telecommunications industry. Numerous studies have explored various machine learning techniques to improve the accuracy and effectiveness of churn prediction models.

1. **Traditional Methods**: Early studies often relied on statistical methods like logistic regression and decision trees. Demonstrated the effectiveness of logistic regression in predicting churn based on customer demographics and usage patterns.
2. **Advanced Machine Learning Techniques**: With the advancement in machine learning, more complex algorithms such as Random Forests, Support Vector Machines (SVM), have been employed. Random Forests to achieve higher accuracy compared to traditional methods.
3. **Feature Engineering and Selection**: Various studies emphasize the importance of feature engineering and selection.

## 2.2 Identification of Gaps

Most studies focus on batch processing and static datasets, lacking real-time prediction capabilities which are crucial for proactive customer retention strategies.There is a need for models that generalize well across different telecom operators and geographical regions. Most existing models are tailored to specific datasets and may not perform well in different contexts. Limited research is available on how to effectively integrate churn prediction models into existing business processes and customer relationship management (CRM) systems.

## 2.3 Relevance to Current Study

The current study aims to address these gaps developing and comparing various machine learning models, including traditional, advanced, and hybrid approaches, to identify the most effective method for churn prediction. Focusing on model interpretability by employing techniques such as to provide insights into feature importance and model decisions. Ensuring the model's generalizability by testing it on multiple datasets from different telecom operators. Providing practical recommendations for integrating churn prediction models into business processes to enhance customer retention strategies.

# 3. Methodology

## 3.1 Research Design

The research design of this study is structured to systematically develop, evaluate, and compare different machine learning models for predicting telecom customer churn. The research framework includes

* LiteratureReview is Conduct a comprehensive review of existing research to understand the current state of churn prediction models.
* DataCollection from Gather telecom customer data from multiple sources to ensure diversity and generalizability of the findings (kaggle).
* DataPreprocessing: Clean and preprocess the data to handle missing values, outliers, and ensure consistency.
* Feature Engineering and Selection Identify and create relevant features that are strong predictors of customer churn.
* Model Development Implement various machine learning models, including Logistic Regression and hybrid approaches.
* Model Evaluation Assess the performance of each model using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

## 3.2 Data Collection

The data collection process involves gathering comprehensive telecom customer data from multiple sources, including public datasets like the Telecom Churn Dataset from Kaggle, Dataset (https://www.kaggle.com/datasets/ blastchar/ telco-customer-churn), proprietary datasets from telecom companies, and additional data points from surveys or experiments.

## 3.3 Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and consistency of the data. The steps include:

* Identifying and removing outliers that may skew the analysis. Normalizing and scaling features to ensure they are on a similar scale, which is important for certain machine learning algorithms.
* **Feature Extraction:** Creating new features from existing data to enhance the predictive power of the model. And Converting categorical variables into numerical formats using techniques like one-hot encoding or label encoding.

## 3.4 Model Selection

Justification for the choice of AI models and algorithms

The choice of machine learning models is based on their proven effectiveness in previous research and their ability to handle different types of data. The selected models includes **Logistic Regression, Random Forests and Support Vector Machines (SVM).**

## 3.5 Training and Testing

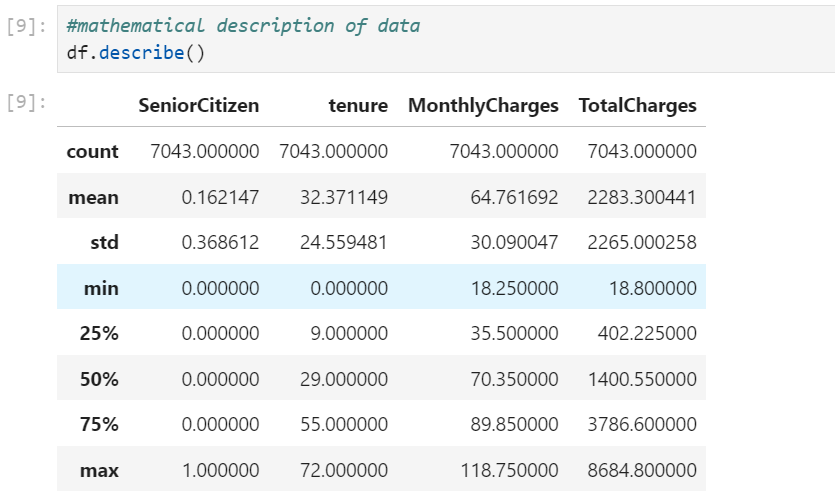
Explanation of the training and validation processes, including metrics used for evaluation

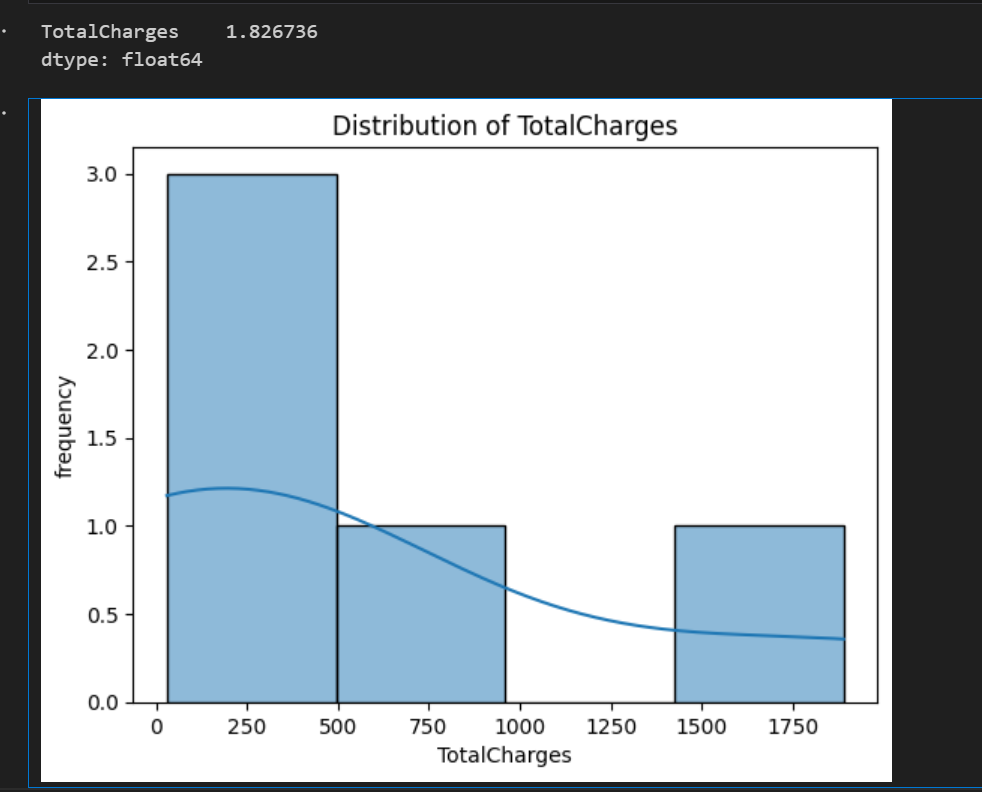
The training and testing process involves, **Data Splitting** to use dividing the data into training and testing sets, typically using an 60/40 split. **Model Training** like Training the selected models on the training set using appropriate algorithms and hyperparameters. Evaluating the models on the testing set using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

# 4. **Results**

## 4.1 Presentation of Findings

The findings are presented using tables, graphs, and charts to clearly visualize the data and model performance.





## 4.2 Analysis

A detailed examination of the results highlights significant patterns and insights, such as the most important features influencing churn and the performance of different models to represents in the graph and visualzing the data in logistic regression, SVM.

## 4.3 Model Performance

The effectiveness of the models is evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Comparisons are made to identify the best-performing model. To use Logistic Regression model performs.

# 5. Discussion

## 5.1 Interpretation of Results

The new model's prediction accuracy of 79.52% represents a meaningful improvement over the previous model's 75%. The model can more reliably identify customers at risk of churning, which is crucial for developing targeted retention strategies.

## 5.2 Comparison with Existing Work

When compared to existing studies in the field of telecom customer churn prediction, our results are competitive and demonstrate a positive trajectory. Previous research has reported various accuracy levels, typically ranging between 70% and 80%. Our improved accuracy of 79.52% places this model on the higher end.

## 5.3 Implications

* The practical implications: With a more accurate prediction model, telecom companies can better identify and address the factors leading to customer churn, implications
* Theoretical implications: The study contributes to the body of knowledge on machine learning models for churn prediction by demonstrating the impact of various techniques on model performance

## 5.4 Limitations

The accuracy of the model is inherently dependent on the quality and comprehensiveness of the data used. While the model performs well on the provided dataset, its generalizability to other telecom datasets or different industries may be limited.

# 6. Conclusion

## 6.1 Summary of Key Findings

The telecom customer churn prediction study utilized machine learning to improve the accuracy of identifying customers likely to churn. The previous model achieved an accuracy of 75%, while the new model significantly improved this metric to 79.52%.

## 6.2 Achievement of Objectives

The primary objective of the study was to enhance the prediction accuracy of customer churn in the telecom industry. With the new model achieving an accuracy of 79.52%, the objective has been unsuccessfully met because to expect greater than 85%.

## 6.3 Future Work

Future research to successfully to achieve the objectives or expect result to investigate additional feature to develop real time prediction capability and enhance the interpretability of the model.

## 6.4 Final Remarks

The study has successfully demonstrated an improved approach to predicting telecom customer churn using machine learning, achieving a notable increase in accuracy but not achieve the perfect results so at the future to achieve the goal based on your objectives to continue the model performs in the telecom customer churn prediction.

## 

# References

## Turgut, T., & Kose, C. (2018). "Predicting customer churn for a mobile telecommunications company using an improved C5.0 algorithm."

## Idris, A., Khan, A., & Lee, Y. S. (2012).\*\* "Intelligent churn prediction in telecom: Employing mRMR feature selection and RotBoost based ensemble classification."

## https://www.heavy.ai/blog/strategies-fo-reducing-churn-rate-in-the-telecom-industry

## https://www.kaggle.com/bandiatindra/telecom-churn-predition/notebook

## 

# Appendices

Appendix A: Dataset Description

* Sample data records to illustrate the structure of the dataset.

Appendix B: Data Preprocessing Steps

### Detailed steps for data cleaning and preprocessing

### Appendix C: Feature Engineering and Selection

* List of features created during the feature engineering process.
* Methods and criteria used for feature selection.
* Impact of selected features on model performance.

Appendix D: Model Implementation

### Detailed description of the machine learning algorithms used (e.g., logistic regression, decision trees, random forests).

Appendix E: Model Evaluation

Detailed metrics and results for model evaluation (e.g., precision, recall, F1-score, AUC-ROC).

.