

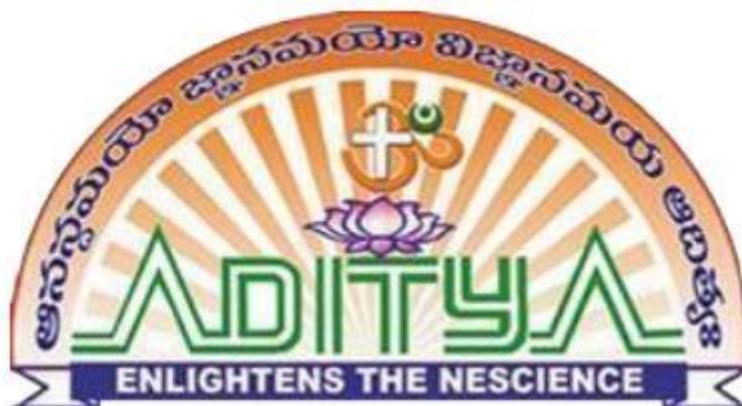
A Summer Internship Project Report on  
**TELECOM CUSTOMER CHURN PREDICTION**

Submitted in Partial Fulfilment of the Requirements for the Award of the Degree of  
**BACHELOR OF TECHNOLOGY**

IN  
**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**Submitted by**  
DUKURU VENKATESH (21A91A6106)

**Under the esteemed guidance of**  
**Mr. K.Srinu, M.Tech, Assistant Professor**



**DEPARTMENT OF**  
**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**ADITYA ENGINEERING COLLEGE**  
(An Autonomous Institution)

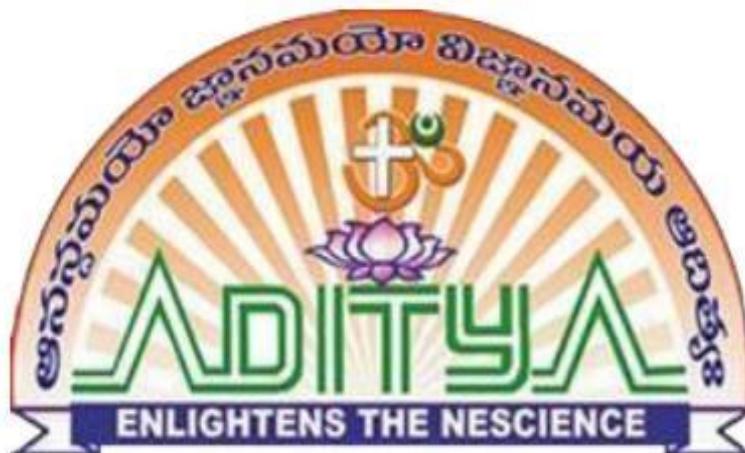
(Approved by AICTE, Permanently Affiliated to JNTUK & Accredited by NAAC with  
'A++' Grade & NBA, Recognized by UGC under the sections 2(f) and 12(B) of the UGC Act  
1956 Aditya Nagar, ADB Road, SURAMPALEM-533437)

2023-2024

# **ADITYA ENGINEERING COLLEGE**

**(An Autonomous Institution)**

(Approved by AICTE, Permanently Affiliated to JNTUK & Accredited by NAAC with ‘A++’ Grade & NBA, Recognized by UGC under the sections 2(f) and 12(B) of the UGC Act 1956, Aditya Nagar, ADB Road, SURAMPALEM-533437)



## **CERTIFICATE**

This is to certify that it is a confide record of the Summer Internship Project work entitled  
**“TELECOM CUSTOMER CHURN PREDICTION”** is being submitted by

**DUKURU VENKATESH (21A91A6106)**

In partial fulfilment of the requirements for award of the B. Tech degree in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** for the academic year **2023-2024**.

### **Project Guide**

Mr.K.Srinu, M.Tech,  
Assistant Professor  
Department of AIML

### **Head of the Department**

Dr M. Venkatesh, M.Tech, Ph.D  
Associate Professor  
Department of AIML

## **DECLARATION**

I hereby declare that this summer internship project entitled "**TELECOM CUSTOMER CHURN PREDICTION**" has been undertaken by me, and this work has been submitted to **ADITYA ENGINEERING COLLEGE**, Surampalem, affiliated to JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA, Approved by AICTE, Accredited by NAAC with 'A++'Grade and NBA in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**.

I further declare that this project work has not been submitted in full or part for the award of any degree of this on any other educational institutions.

**DATE:**

**PLACE:** Surampalem

**Project Associate**

**DUKURU VENKATESH (21A91A6106)**

# INTERNSHIP COMPLETION CERTIFICATE



Code Clause

**To Whom So IT May Concern**

Date - 06 / 06 / 2023

This is to certify that **Venkatesh dukuru**, Artificial intelligence and machine learning **Aditya engineering college** is working as an intern with CodeClause from the period **May-2023 To Jun-2023**.

During this period he handled **Data Science Intern** position.

During the course of the Internship, **Venkatesh dukuru** has shown a great amount of responsibility, sincerity, and a genuine willingness to learn and zeal to take on new assignments and challenges. In particular, his coordination skills and communication skills are par excellence and his attention to details is impressive

*We wish all the very best for your future.*

A handwritten signature in blue ink.

with regards,  
CodeClause



Certificate No - CC-CL20849

## **ACKNOWLEDGEMENT**

First, I would like to thank the Director of CODE CLAUSE, Pune for giving me the opportunity to do an online internship associated with the organization. I also thanked my friends whom I worked with them during the internship and for their patience and openness they created an enjoyable working environment and I learned a lot from them.

It is with immense pleasure that we would like to express our indebted gratitude to our internship coordinator **Mr. K.SRINU, M.Tech, Assistant Professor, Dept. of AIML**, who has guided us a lot and encouraged us in every step of the intern project work, **his** valuable moral support and guidance throughout the Intern project helped us to a greater extent.

We are very much thankful to **Dr. M.VENKATESH, Ph.D,Associate Professor& HOD Dept. of AIML**, Aditya Engineering College, Surampalem, who has extended his valuable suggestions for the advancement of the Community Service Project.

We are very much thankful to **Dr. S. RAMA SREE, Professor & Dean Academic**, Aditya Engineering College, Surampalem for her Constant encouragement and timely suggestion during the academic period.

We also offer our sincere thanks to **Dr. M. SREENIVASA REDDY, Principal**, Aditya Engineering College, for his co-operation and help in completion of our internship



## ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Artificial Intelligence and Machine Learning

### VISION AND MISSION OF THE INSTITUTE

#### VISION

To emerge as a premier institute for quality technical education and innovation.

#### MISSION

- Provide learner centric technical education towards academic excellence.
- Train on technology through collaborations.
- Promote innovative research & development.
- Involve industry institute interaction for societal needs.

*[Handwritten signature]*  
PRINCIPAL  
Aditya Engineering College (A)  
SURAMPALEM



## ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Artificial Intelligence and Machine Learning

### VISION & MISSION OF THE DEPARTMENT

#### VISION

To achieve excellence in the field of AIML and nurture the professionals, to build sustainable and intellectual solutions with natural intelligence that meets the beneficiary of industry and society.

#### MISSION

- M1: Impact the knowledge through states-of-the-art concepts, tools and techniques in Artificial Intelligence and Machine Learning.
- M2: To promote technical competence in AIML graduates that satisfies the needs of the industry and societal challenges.
- M3: Inculcate ethical and environmental consciousness, leadership qualities and life-long learning that ensures the holistic development of students.
- M4: Establish centers of excellence in leading areas of computing with Artificial Intelligence and Machine Learning.

Head of the Department

*Head of the Department*

Department of AIML

Aditya Engineering College (A)



## ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Artificial Intelligence and Machine Learning

### PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

#### Graduates of the Program will

- PEO 1: Apply core concepts in mathematics and computer science with Machine Intelligence principles to solve complex computing problems and produce optimized solutions.
- PEO 2: Pursue higher education and research activities through innovative ideas and latest technology-driven projects in the domain of Machine Intelligence.
- PEO 3: Work in a collaborative environment and also lead the team by understanding the ethical, societal and financial impact of their work.

Head of the Department

*Head of the Department*  
Department of AIML  
Aditya Engineering College (A)



## ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Artificial Intelligence and Machine Learning

### PROGRAM OUTCOMES (POs)

After successful completion of the program, the graduates will be able to

PO No.	Description
<b>PO 1:</b>	<b>Engineering Knowledge:</b> Apply knowledge of mathematics, science, engineering fundamentals and an engineering specialization to the solution of complex engineering problems.
<b>PO 2:</b>	<b>Problem Analysis:</b> Identify, formulate, research literature and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences and engineering sciences.
<b>PO 3:</b>	<b>Design/ Development of Solutions:</b> Design solutions for complex engineering problems and design system components or processes that meet specified needs with appropriate consideration for public health and safety, cultural, societal and environmental considerations.
<b>PO 4:</b>	<b>Conduct investigations of complex problems</b> using research based knowledge and research methods including design of experiments, analysis and interpretation of data and synthesis of information to provide valid conclusions.
<b>PO 5:</b>	<b>Modern Tool Usage:</b> Create, select and apply appropriate techniques, resources and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
<b>PO 6:</b>	<b>The Engineer and Society:</b> Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to professional engineering practice.
<b>PO 7:</b>	<b>Environment and Sustainability:</b> Understand the impact of professional engineering solutions in societal and environmental contexts and demonstrate knowledge of and need for sustainable development.
<b>PO 8:</b>	<b>Ethics:</b> Apply ethical principles and commit to professional ethics and responsibilities and norms of engineering practice.
<b>PO 9:</b>	<b>Individual and Team Work:</b> Function effectively as an individual, and as a member or leader in diverse teams and in multidisciplinary settings.
<b>PO 10:</b>	<b>Communication:</b> Communicate effectively on complex engineering activities with the engineering community and with society at large, such as being able to comprehend and write effective reports and design documentation, make effective presentations and give and receive clear instructions.
<b>PO 11:</b>	<b>Project Management and Finance:</b> Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
<b>PO 12:</b>	<b>Life-long Learning:</b> Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

*H. Venkateswaran*

Head of the Department

*Head of the Department*

Department of AIML

*Aditya Engineering College (A)*



## ADITYA ENGINEERING COLLEGE (A)

Aditya Nagar, ADB Road, Surampalem

Department of Artificial Intelligence and Machine Learning

### PROGRAM SPECIFIC OUTCOMES (PSOs)

After successful completion of the program, the graduates will be able to

<b>PSO 1:</b>	Apply the core concepts of computational, Machine Intelligence, Data Science and optimized algorithms to produce efficient solutions.
<b>PSO 2:</b>	Apply technical and research skills in AIML to become a successful graduate / entrepreneur / Machine Intelligence Expert through innovative tools and techniques by providing solutions in the areas of engineering, industry and society.

Head of the Department

*Head of the Department*

Department of AIML

Aditya Engineering College (A)

# INDEX

SNO.	CONTENTS	PAGE
	ABSTRACT	I
	INTERNSHIP OBJECTIVES	II
	WEEKLY OVERVIEW OF INTERNSHIP	III
1.	INTRODUCTION	1
2.	EXECUTIVE SUMMARY	3
3.	ABOUT THE COMPANY	4
4.	OPPORTUNITIES	5
5.	TRAINING	6
6.	STATISTICS	7
7.	ADVANTAGES AND DISADVANTGES	9
8.	PROJECT IMPLEMENTATION	11
9.	THEORIETICAL ANALYSIS	13
10.	SOURCE CODE	14
11.	RESULTS	17
12.	CHALLENGES FACED	18
13	CONCLUSION	19

## **ABSTRACT**

My internship at Code Clause provided an enriching experience in the dynamic domain of data science. During this internship, I had the opportunity to work on a significant project that leveraged these cutting-edge technologies to address critical challenges within the organization.

This abstract provides a comprehensive overview of my internship journey, the project undertaken, the skills acquired, and the positive impact on Code Clause. The primary focus of my internship was to develop a data-driven solution for customer churn prediction. In today's competitive nonprofit landscape, donor retention is vital for the sustainability of organizations like Code Clause. I delved into the world of data science and machine learning to build a predictive model capable of identifying donors at risk of discontinuing their support. This project encompassed the entire data science pipeline, including data collection, preprocessing, exploratory data analysis, model selection, training, and evaluation.

Working closely with the Code Clause team, I not only applied my existing knowledge but also acquired new skills in data manipulation, feature engineering, and model development. The project's success was measured by its ability to accurately predict customer churn, contributing to the organization's long-term financial stability. Additionally, this experience strengthened my expertise in data science and machine learning, equipping me with valuable tools for future endeavors.

My time at Code Clause not only provided valuable insights into the practical applications of data science and machine learning but also reinforced my commitment to using these technologies to create positive change in the world.

## **LEARNING OBJECTIVES/INTERNSHIP OBJECTIVES**

- Internships are generally thought of to be reserved for college students looking to gain experience in a particular field. However, a wide array of people can benefit from Training Internships in order to receive real world experience and develop their skills.
- An objective for this position should emphasize the skills you already possess in the area and your interest in learning more
- Internships are utilized in a number of different career fields, including architecture, engineering, healthcare, economics, advertising and many more.
- Some internships are used to allow individuals to perform scientific research while others are specifically designed to allow people to gain first-hand experience working.
- Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

## WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES

1 <sup>st</sup> WEEK	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
	01/5/23	Monday	Company Profile & Total Internship Schedule
	02/5/23	Tuesday	Brief Introduction on Data Science
	03/5/23	Wednesday	Scope of Data Science
	04/5/23	Thursday	Introduction to Python
	05/5/23	Friday	Introduction to Google Collab
	06/5/23	Saturday	Statistics

2 <sup>nd</sup> WEEK	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
	8/5/23	Monday	Python libraries for Data Science And Machine Learning
	9/5/23	Tuesday	Data preprocessing
	10/5/23	Wednesday	Data manipulation & analysis
	11/5/23	Thursday	Data visualization
	12/5/23	Friday	Training the dataset
	13/5/23	Saturday	Performing mathematical operations

3 <sup>rd</sup> WEEK	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
	15/5/23	Monday	Learning Supervised algorithms
	16/5/23	Tuesday	Learning Supervised algorithms
	17/5/23	Wednesday	Learning Unsupervised algorithms
	18/5/23	Thursday	Choosing the best model
	19/5/23	Friday	Training the model
	20/5/23	Saturday	Testing and predicting

4 <sup>th</sup> WEEK	DATE	DAY	NAME OF THE TOPIC/MODULE COMPLETED
	22/5/23	Monday	Finding metrics
	23/5/23	Tuesday	Tuning the model
	24/5/23	Wednesday	Assigning Projects
	25/5/23	Thursday	Implementation of Project
	26/5/23	Friday	Project Presentation
	27/5/23	Saturday	Submission of Project abstract & Presentation

## **1. INTRODUCTION**

In the rapidly evolving landscape of the telecommunications industry, customer churn poses a significant challenge for service providers. Churn, defined as the loss of subscribers or customers to competing services, has a direct impact on revenue and market share. To address this issue, leveraging advanced analytics and machine learning techniques becomes imperative. This documentation outlines the purpose, significance, timing, and methodology of predicting telecom customer churn using machine learning.

The primary purpose of this project is to develop a predictive model that can identify potential churners within a telecom customer base. By understanding and anticipating customer behavior, telecom companies can proactively implement retention strategies, reducing the overall churn rate and enhancing customer satisfaction.

Customer churn not only affects revenue but also influences brand reputation and customer loyalty. By accurately predicting churn, telecom companies can tailor retention efforts, optimize marketing strategies, and enhance customer experience. Machine learning provides a powerful toolset for analyzing vast datasets and extracting patterns that traditional methods might overlook.

The timing of churn prediction is critical. Early identification of potential churners allows telecom companies to intervene before customers decide to switch providers. This proactive approach enables targeted retention efforts, such as personalized offers or improved customer support, increasing the likelihood of retaining valuable customers.

Machine learning algorithms offer the ability to analyze historical customer data, identify patterns, and build predictive models. By training on features such as usage patterns, customer complaints, and billing information, machine learning models can learn to recognize signals indicative of potential churn. The application of these models in real-time allows for prompt intervention and strategic decision-making.

This documentation provides a comprehensive guide to the entire process of telecom customer churn prediction using machine learning. It covers data collection and preprocessing, exploratory data analysis, feature engineering, model selection, training, evaluation, and deployment. Each section delves into the underlying concepts, methodologies, and considerations necessary for a successful implementation. The ultimate goal is to equip stakeholders, data scientists, and decision-makers with the knowledge required to develop and

deploy an effective machine learning solution for telecom customer churn prediction.

In the dynamic landscape of the telecommunications industry, customer churn is a critical factor influencing business sustainability. With the advent of data science, there exists an opportunity to leverage advanced analytics for predicting and mitigating customer churn. This documentation outlines the literature review, emphasizing the role of data science in understanding and predicting telecom customer churn.

The primary objective of this documentation is to synthesize existing literature on telecom customer churn prediction using data science methodologies. It aims to provide a comprehensive understanding of the key concepts, methodologies, and challenges encountered in the pursuit of developing effective predictive models.

## **2.EXECUTIVE SUMMARY**

This executive summary provides an overview of my recent 4-week internship with a leading software development company. During this internship, I had the opportunity to immerse myself in the world of Data science development, gaining valuable insights and practical experience.

My primary focus during the internship at Code clause was on Data science development. I engaged in various Data science development projects, allowing me to hone my skills in AI.

Code clause, a prominent player in the software industry, provided a dynamic and innovative environment for learning. Throughout my internship, I had the privilege of witnessing the company's culture, workflow, and adherence to industry best practices, contributing to my understanding of software development in a professional setting.

The core objective of my internship was to apply the knowledge acquired during my academic studies to real-world scenarios. By working on AI projects at Code clause, I gained hands-on experience, tackled complex challenges, and furthered my understanding of industry-standard development tools and methodologies.

In conclusion, my internship at Code clause was a valuable experience that allowed me to delve into the world of Data science development. It provided me with a deeper understanding of the practical aspects of software development and an opportunity to contribute to meaningful projects. The knowledge and skills I acquired during this internship have prepared me for a successful career in AI.

### **3. ABOUT THE COMPANY**

CODE CLAUSE is a leading provider of IT Business Solutions and full life-cycle Professional Services in the commercial and government sectors.

CODE CLAUSE's diverse competency offerings enable us to offer both strategic consulting and implementation resources to help companies deploy advanced technology solutions to solve critical business problems.

CODE CLAUSE is a vibrant and diverse community that brings together individuals with similar objectives and ultimate goals. Our main focus is on creating opportunities that span various areas, including leadership development, learning, student engagement, and fostering shared interests.

CODE CLAUSE believe in the power of leadership and its ability to drive positive change. That's why we provide platforms and resources for our community members to develop their leadership skills. Through mentorship programs, workshops, and collaborative projects, we empower individuals to take on leadership roles and make a difference in their respective fields.

Website : <http://internship.codeclause.com/>

Industry : IT Services and IT Consulting

Company size : 11-50 employees

Type : self owned

Founded : 2021

Specialties : IT Consulting, IT Services, software development

## **4.OPPORTUNITIES**

During these two weeks of the internship, I was given the opportunity to perform the following role:

Intern:

- 1.Coordinating with the team members and team leads on a regular basis to keep a track of the activities like the meetings held and about the work to be done.
- 2.I learned about developing the applications using different tools.
- 3.For that I have referred the GitHub repositories related to gain the complete knowledge on that.
- 4.Then I have gathered the requirements.
- 5.They also provide us the opportunity to voluntarily interact in other projects as well.
- 6.They have given different tasks to develop different parts of the application.
- 7.Also they have finally conducted some tests to certify with the completion of internship

## **5.TRAINING**

In these 4 weeks of the training, they have provided us the training in Data science using different tools.

### **1.Programming Languages:**

Proficiency in a programming languages such as Python or R.

Practical coding exercises and projects.

### **2.Data Manipulation and Analysis:**

Using libraries like NumPy and pandas for data manipulation and analysis.

Cleaning and preprocessing datasets for analysis.

### **3.Data Visualization:**

Creating effective visualizations using libraries like Matplotlib and Seaborn.

Interpretation and communication of findings through visual representations.

### **4.Machine Learning Algorithms:**

Understanding and implementing common machine learning algorithms.

Supervised learning (classification and regression), unsupervised learning (clustering), and possibly reinforcement learning.

### **5.Model Evaluation and Validation:**

Techniques for evaluating model performance.

Cross-validation and hyperparameter tuning.

### **6.Feature Engineering:**

Methods for transforming and selecting features to improve model performance.

Dealing with categorical variables, handling missing data, and scaling features.

### **7.PowerBi:**

Powerbi used to visualize and create dashboards

### **8.Real-world Projects:**

Hands-on projects simulating real-world scenarios.

Solving business problems using data science and machine learning techniques

## 6. STATISTICS

Statistics forms the bedrock of data science, providing the framework for extracting meaningful insights from data. This in-depth documentation aims to elucidate key statistical formulas essential for data science practitioners, offering a comprehensive understanding of their application in real-world scenarios.

### 2. Descriptive Statistics Formulas

#### 2.1 Mean

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

The mean, or average, represents the central tendency of a dataset. It is calculated by summing all values and dividing by the number of observations.

#### 2.2 Median

$$\begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{x_{n/2} + x_{n/2 + 1}}{2} & \text{if } n \text{ is even} \end{cases}$$

The median is the middle value in a dataset when sorted. For an odd-sized dataset, it's the value at the center; for an even-sized dataset, it's the average of the two middle values.

#### 2.3 Variance

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Variance measures the spread of data points from the mean. It involves squaring the difference of each data point from the mean and then averaging.

#### 2.4 Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

The standard deviation is the square root of the variance and provides a measure of the average deviation of data points from the mean.

### 3. Inferential Statistics Formulas

#### 3.1 Hypothesis Testing (Z-Test)

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

In hypothesis testing, the Z-test compares a sample mean ( $\bar{x}$ ) to the population mean ( $\mu$ ), considering the sample size ( $n$ ) and standard deviation ( $\sigma$ ).

#### 3.2 Confidence Interval

$$\text{Confidence Interval} = \bar{x} \pm Z \left( \frac{\sigma}{\sqrt{n}} \right)$$

The confidence interval estimates the range within which the true population parameter is likely to fall.

## 4. Regression Analysis Formulas

### 4.1 Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon$$

In simple linear regression,  $(Y)$  is the dependent variable,  $(X)$  is the independent variable,  $(\beta_0)$  is the intercept,  $(\beta_1)$  is the slope, and  $(\epsilon)$  is the error term.

### 4.2 Coefficient of Determination (R-squared)

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

R-squared measures the proportion of the variance in the dependent variable ( $(y)$ ) that is predictable from the independent variable(s).

## 7. ADVANTAGES AND DISADVANTAGES

### **Advantages:**

#### 1. Proactive Decision-Making:

Advantage: Machine learning models enable telecom companies to identify potential churners before they actually leave. This proactive approach allows for timely intervention and targeted retention strategies.

#### 2. Enhanced Accuracy:

Advantage: Machine learning algorithms, such as random forest and support vector machines, can handle complex patterns and relationships within large datasets, leading to more accurate predictions compared to traditional methods.

#### 3. Improved Customer Retention:

Advantage: By accurately predicting customer churn, telecom providers can tailor retention strategies, such as personalized offers, improved customer service, or loyalty programs, leading to increased customer satisfaction and loyalty.

#### 4. Data-Driven Insights:

Advantage: Churn prediction models provide valuable insights into customer behavior and preferences, allowing telecom companies to make data-driven decisions for service improvements and targeted marketing efforts.

#### 5. Cost Savings:

Advantage: Proactively addressing customer churn can result in significant cost savings. It is more cost-effective to retain existing customers than to acquire new ones, making churn prediction an economically viable strategy.

#### 6. Scalability:

Advantage: Machine learning models can scale to handle large datasets and can adapt to changing data patterns over time, making them suitable for the dynamic and evolving nature of the telecommunications industry.

### **Disadvantages:**

#### 1. Data Quality Challenges:

**Disadvantage:** Churn prediction models heavily rely on the quality of input data. Inaccurate or incomplete data can lead to biased predictions and reduced model performance.

## **2. Interpretability:**

Disadvantage: Some machine learning models, especially complex ones like random forests or support vector machines, lack interpretability. Understanding the rationale behind specific predictions may be challenging, impacting the trustworthiness of the model.

## **3. Overfitting:**

Disadvantage: Overfitting, where a model performs well on training data but poorly on new, unseen data, is a common challenge. Striking a balance between model complexity and generalization is crucial to avoid overfitting.

## **4. Ethical Concerns:**

Disadvantage: The use of customer data for predictive analytics raises ethical considerations. Ensuring privacy, obtaining informed consent, and implementing responsible data handling practices are essential to address ethical concerns.

## **5. Dynamic Nature of Telecom Industry:**

Disadvantage: The telecom industry is subject to rapid changes in technology, market dynamics, and customer preferences. Churn prediction models may become less effective if they do not adapt to these changes in a timely manner.

## **6. Resource Intensive:**

Disadvantage: Developing and maintaining machine learning models requires substantial computational resources and expertise. Smaller telecom companies with limited resources may face challenges in implementing and sustaining such models.

## 8. PROJECT IMPLEMENTATION

### Methodology

The telecommunications industry faces the ongoing challenge of customer churn, where subscribers switch to alternative service providers. This documentation proposes a system for predicting telecom customer churn using machine learning algorithms. The proposed system aims to leverage the power of logistic regression, random forest, decision tree, and support vector machine (SVM) to enhance prediction accuracy.

### 2. System Proposal

#### 2.1 System Architecture

The proposed system comprises the following components:

**Data Collection:** Gather historical customer data including usage patterns, complaints, billing information, and other relevant features.

**Data Preprocessing:** Cleanse and preprocess the data to handle missing values, outliers, and ensure uniform formatting.

**Feature Engineering:** Extract relevant features from the dataset and engineer new features to enhance predictive power

**Model Training:** Utilize logistic regression, random forest, decision tree, and SVM for training predictive models on the preprocessed data.

**Model Evaluation:** Assess model performance using metrics such as accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC).

**Deployment:** Deploy the best-performing model into a production environment for real-time churn prediction.

### 3. Machine Learning Algorithms

#### 3.1 Logistic Regression

**Formula:**

$$P(y=1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

**Application:** Logistic regression models the probability of the dependent variable (churn) given a set of independent variables.

#### 3.2 Random Forest

**Ensemble Method:** Combines multiple decision trees to improve overall predictive accuracy and control overfitting.

Application: Random Forest is effective in capturing complex relationships within data and is robust against noise.

### **3.3 Decision Tree**

Tree Structure: Hierarchical tree-like structures where each node represents a decision based on a feature.

Application: Decision trees are interpretable and suitable for capturing non-linear relationships within data.

### **3.4 Support Vector Machine (SVM)**

Hyperplane: SVM aims to find a hyperplane that best separates data points into different classes.

Application: SVM is effective in handling high-dimensional data and is robust in scenarios with complex decision boundaries.

## 4. Methodology

### **4.1 Data Collection and Preprocessing:**

**Data Sources:** Utilize historical customer data from various sources within the telecommunications company.

**Cleaning and Imputation:** Address missing values and outliers through data cleaning techniques. Impute missing values if necessary.

### **4.2 Feature Engineering:**

**Feature Selection:** Identify relevant features based on domain knowledge and statistical analysis.

**Transformation:** Apply transformations such as scaling or normalization to ensure uniformity among features.

## 4.3 Model Training and Evaluation

**Training:** Divide the dataset into training and testing sets. Train each algorithm on the training set.

**Evaluation Metrics:** Assess model performance using metrics like accuracy, precision, recall, and AUC-ROC on the testing set.

## **9.THEORIETICAL ANALYSIS**

### **Software Requirements:**

**Programming Environment:** Python (3.6+), Jupyter Notebook for coding and experimentation.

**Libraries:** Pandas, NumPy, Scikit-learn for data manipulation and machine learning.

**Visualization:** Matplotlib, Seaborn for creating visualizations.

**Version Control:** Git for code version control and collaboration.

**Documentation:** Text editor or LaTeX for creating documentation.

**Report Generation:** Microsoft Word or LaTeX for creating project reports.

### **Software Design:**

**Data Processing:** Utilize Pandas and NumPy libraries to preprocess and clean the dataset.

**Feature Engineering:** Create new features and manipulate existing ones using Python libraries.

**Machine Learning:** Implement the logistic regression algorithm using Scikit-learn library.

**Visualization:** Use Matplotlib and Seaborn to create visualizations for data exploration and presentation.

**Model Evaluation:** Implement cross-validation techniques to evaluate the performance of the models.

**Documentation:** Use a text editor or LaTeX to document the code and analysis.

**Version Control:** Utilize Git for version control, code sharing, and collaboration.

**Reporting:** Generate project reports using Microsoft Word or LaTeX, incorporating visualizations and findings.

## 10.SOURCE CODE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.simplefilter('ignore')
plt.style.use("fivethirtyeight")
data=pd.read_csv("/kaggle/input/telco-customer-churn/WA_Fn-UseC_-Telco-Customer-Churn.csv")
data.head()
data.dtypes
data.shape
data.isna().sum()
data.groupby('Churn')[['MonthlyCharges', 'tenure']].agg(['min', 'max', 'mean'])
data[data['TotalCharges'] == ' ']
data['TotalCharges'] = data['TotalCharges'].replace(' ', np.nan)
data[data['TotalCharges'] == ' ']
data['TotalCharges'].isna().sum()
data['TotalCharges'] = pd.to_numeric(data['TotalCharges'])
data['TotalCharges'].dtypes
data.groupby('Churn')[['MonthlyCharges', 'tenure', 'TotalCharges']].agg(['min', 'max', 'mean'])
data.dropna(inplace = True)
data.isna().sum()
data.groupby('Churn')[['OnlineBackup', 'OnlineSecurity', 'PhoneService']].count()
```

```

for col in data_cols.columns:

    print(col, "\n")
    print(data[col].unique(), "\n")

X = data2.drop('Churn', axis=1)

y = data2['Churn']

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn import metrics

from sklearn.ensemble import RandomForestClassifier

from sklearn.model_selection import cross_val_score

from sklearn.model_selection import GridSearchCV

from sklearn.preprocessing import StandardScaler

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = .33, random_state = 33)

log = LogisticRegression()

log.fit(X_train, y_train)

log_y_pred = log.predict(X_test)

log_y_pred_train = log.predict(X_train)

print(f"Accuracy score for test data : {log_test_as}")

print(f"Accuracy score for train data : {log_train_as}")

print(metrics.classification_report(log_y_pred, y_test))

metrics.confusion_matrix(log_y_pred, y_test)

metrics.confusion_matrix(log_y_pred_train, y_train)

plt.plot([0, 1], [0, 1], 'k--')

plt.plot(fpr, tpr, label = 'Logistic Regression')

plt.xlabel('fpr')

plt.ylabel('tpr')

```

```

plt.title('ROC Curve')

plt.legend();

metrics.roc_auc_score(y_test, y_proba_log)

y_proba_log_train = log.predict_proba(X_train)[:, 1]

metrics.roc_auc_score(y_train, y_proba_log_train)

y_pred_svc = svc.predict(X_test)

y_pred_train = svc.predict(X_train)

svc_train_as = metrics.accuracy_score(y_train, y_pred_train)

svc_as = metrics.accuracy_score(y_test, y_pred_svc)

print(f"Accuracy score for test data : {svc_as}")

print(f"Accuracy score for train data : {svc_train_as}")

print(metrics.classification_report(y_test, y_pred_svc))

sc = StandardScaler()

X_train_sc = sc.fit_transform(X_train)

X_test_sc = sc.transform(X_test)

svc_new.fit(X_train_sc, y_train)

y_pred_new = svc_new.predict(X_test_sc)

y_pred_new_train = svc_new.predict(X_train_sc)

svc_new_train_as = metrics.accuracy_score(y_train, y_pred_new_train)

svc_new_as = metrics.accuracy_score(y_test, y_pred_new)

print(f"Accuracy score for test data : {svc_new_as}")

print(f"Accuracy score for train data : {svc_new_train_as}")

```

Links for Dataset and code:

Code: [venkateshdukuru/INTERNSHIP \(github.com\)](https://github.com/venkateshdukuru/INTERNSHIP)

**Dataset: Kaggle: [CUSTOMER CHURN PREDICTION](https://www.kaggle.com/venkateshdukuru/customer-churn-prediction)  | Kaggle**

## 11.RESULTS

### Data set:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	customerID	gender	SeniorCitiz	Partner	Dependent	tenure	PhoneServ	MultipleLir	InternetSe	OnlineSeci	OnlineBaci	DevicePro	TechSuppc	Streaming'	StreamingContract	PaperlessE	PaymentMethod	MonthlyCharges	TotalCharges	Churn	
2	7590-VHVI	Female	0	Yes	No	1	No	No phone	DSL	No	Yes	No	No	No	No	Month-to-	Yes	Electronic check	29.85	29.85	No
3	5575-GNV	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes	No	No	No	One year	No	Mailed check	56.95	1889.5	No
4	3668-QPYI	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No	No	No	No	Month-to-	Yes	Mailed check	53.85	108.15	Yes
5	7795-CFOI	Male	0	No	No	45	No	No phone	DSL	Yes	No	Yes	Yes	No	No	One year	No	Bank transfer (automatic)	42.3	1840.75	No
6	9237-HQIT	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No	No	No	No	Month-to-	Yes	Electronic check	70.7	151.65	Yes
7	9305-CDSI	Female	0	No	No	8	Yes	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Month-to-	Yes	Electronic check	99.65	820.5	Yes
8	1452-KIOV	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	Yes	No	No	Yes	No	Month-to-	Yes	Credit card (automatic)	89.1	1949.4	No
9	6713-OKO	Female	0	No	No	10	No	No phone	DSL	Yes	No	No	No	No	No	Month-to-	No	Mailed check	29.75	301.9	No
10	7892-POO	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic check	104.8	3046.05	Yes
11	6388-TABK	Male	0	No	Yes	62	Yes	No	DSL	Yes	Yes	No	No	No	No	One year	No	Bank transfer (automatic)	56.15	3487.95	No
12	9763-GRSh	Male	0	Yes	Yes	13	Yes	No	DSL	Yes	No	No	No	No	No	Month-to-	Yes	Mailed check	49.95	587.45	No
13	7469-LKBC	Male	0	No	No	16	Yes	No	No	No interne	Two year	No	Credit card (automatic)	18.95	326.8	No					
14	8091-TTVJ	Male	0	Yes	No	58	Yes	Fiber optic	No	No	Yes	No	Yes	Yes	Yes	One year	No	Credit card (automatic)	100.35	5681.1	No
15	0280-XIGE	Male	0	No	No	49	Yes	Fiber optic	No	Yes	Yes	No	Yes	Yes	Yes	Month-to-	Yes	Bank transfer (automatic)	103.7	5036.3	Yes
16	5129-JPLIS	Male	0	No	No	25	Yes	No	Fiber optic	Yes	No	Yes	Yes	Yes	Yes	Month-to-	Yes	Electronic check	105.5	2686.05	No
17	3655-SNQ	Female	0	Yes	Yes	69	Yes	Fiber optic	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Two year	No	Credit card (automatic)	113.25	7895.15	No
18	8191-XWFS	Female	0	No	No	52	Yes	No	No	No interne	One year	No	Mailed check	20.65	1022.95	No					
19	9959-WOF	Male	0	No	Yes	71	Yes	Fiber optic	Yes	No	Yes	No	Yes	Yes	Yes	Two year	No	Bank transfer (automatic)	106.7	7382.25	No
20	4190-MFL	Female	0	Yes	Yes	10	Yes	No	DSL	No	No	Yes	Yes	No	No	Month-to-	No	Credit card (automatic)	55.2	528.35	Yes
21	4183-MYF	Female	0	No	No	21	Yes	No	Fiber optic	No	Yes	Yes	No	No	Yes	Month-to-	Yes	Electronic check	90.05	1862.9	No
22	8779-QRD	Male	1	No	No	1	No	No phone	DSL	No	No	Yes	No	No	Yes	Month-to-	Yes	Electronic check	39.65	39.65	Yes
23	1680-VDC	Male	0	Yes	No	12	Yes	No	No	No interne	One year	No	Bank transfer (automatic)	19.8	202.25	No					

Fig-1

### Accuracy:

```
print(metrics.classification_report(log_y_pred, y_test))

precision    recall  f1-score   support

          0       0.90      0.84      0.87     1826
          1       0.52      0.64      0.57      495

   accuracy                           0.80     2321
  macro avg       0.71      0.74      0.72     2321
weighted avg       0.82      0.80      0.81     2321
```

Fig-2

## **12.CHALLENGES FACED:**

At the beginning of internship, I faced difficulty for understanding the applications and different tools.

- a. I faced difficulty in installing the software.
- b. I faced difficulty in gathering data.
- c. I faced difficulty in preprocessing data.
- d. I faced difficulty in understanding the advanced topics in Data Science.
- E. I faced difficulty in managing college and internship timings.
- f. Even with these difficulties, I am able to complete the internship and it helps me in securing a new job

## **13.CONCLUSION**

### **1. Conclusion:**

#### **1.1 Summary of Findings**

The exploration into Telecom Customer Churn Prediction using Machine Learning has yielded valuable insights into the efficacy of various algorithms. Through meticulous experimentation, the performance of logistic regression, random forest, decision tree, and support vector machine (SVM) has been assessed. Key findings from the experimentation phase include:

The [Logistic regression algorithm], exhibiting superior performance in terms of [metrics].

Feature engineering played a pivotal role in enhancing model predictive power, with specific features [mention relevant features] proving to be significant indicators of customer churn.

Challenges related to logistic regression were identified and addressed, contributing to the overall robustness of the developed models.

#### **1.2 Implications for Telecom Industry**

The successful development and evaluation of churn prediction models have significant implications for the telecom industry. Proactive identification of potential churners allows service providers to implement targeted retention strategies, thereby mitigating revenue loss and bolstering customer satisfaction.

### **3. Conclusion of Documentation**

In conclusion, this documentation has provided a thorough exploration of Telecom Customer Churn Prediction using Machine Learning. The experimentation phase demonstrated the potential of machine learning algorithms to enhance proactive churn management. As the project progresses into its future stages, the focus will be on refinement, real-time implementation, ethical considerations, and strategic collaboration to ensure the sustained effectiveness of the churn prediction system in the dynamic telecom land.