# *A Summer Internship Report on*

# TELECOM CUSTOMER CHURN PREDICTION

***submitted in partial fulfillment of the requirements for the award of degree of***

## BACHELOR OF TECHNOLOGY

**in**

## ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

**Submitted by**

D.VENKATESH (21A91A6106)

**Under the esteemed guidance of**

**Mr. K. Srinu M.Tech, Assistant Professor**



## Department Of Artificial Intelligence and Machine Learning

**ADITYA ENGINEERING COLLEGE (A)**

**Approved by AICTE, Permanently affiliated to JNTUK & Accredited by NBA & NAAC with ‘A++’ GradeRecognized by UGC under the sections 2(f) and 12(B) of the UGC act 1956Aditya Nagar, ADB Road –Surampalem 533437**

**2023-2024**

**ADITYA ENGINEERING COLLEGE (A)**

(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)

**Department Of Artificial Intelligence and Machine Learning**



# CERTIFICATE

This is to certify that the Internship report entitled *“***TELECOM CUSTOMER CHURN PREDICTION***”* is being submitted by **D.VENKATESH(21A91A6106)**

In partial fulfillment of the requirements for award of the B.Tech degree in Artificial Intelligence and Machine Learning for the academic year 2023-2024.

**Internship Coordinator Head of the Department**

Mr.K. Srinu, M.Tech Dr. M. Venkatesh, Mtech, Ph.D,

Assistant Professor Associate Professor

Department of AIML Department of AIML

**DECLARATION**

We hereby declare that the project entitled **“TELECOM CUSTOMER CHURN PREDICTION”** is a genuine project. This work has been submitted to the **ADITYA ENGINEERING COLLEGE,** Surampalem, permanently affiliated to **JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY, KAKINADA** KAKINADA, Approved by AICTE, Accredited by NAAC with ‘A++’ Grade and NBA in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING. We 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**

|  |
| --- |
|  |

**ACKNOWLEDGEMENT**

First I would like to thank the **Mr.Ashish Mane** HR at Code Clause, Pune for giving me the opportunity to do an internship within the organization. I also would like all the people that worked along with me in Code Clause, Punes with their patience and openness they created an enjoyable working environment.

It is with immense pleasure that we would like to express our indebted gratitude to our Internship coordinator **Mr .K.Srinu Assistant professor**who has guided us a lot and encouraged us in every step of the intern project work, **his/her** valuable moral support and guidance throughout the Intern projecthelped us to a greater extent.

Our deepest thanks to **Dr. M. Venkatesh, Associate Professor & Head of the Department** for inspiring us all the way and for arranging all the facilities and resources needed for our project.

We wish to thank **Dr. S. Rama Sree, Professor** in CSE and Dean Academics for her support and suggestions during our project work.

We owe our sincere gratitude to **Dr. M. Sreenivasa Reddy, Principal** for providing a great support and for giving us the opportunity of doing the project.

We are thankful to our **College Management** for providing all the facilities in time to us for completion of our project.

Not to forget, **Teaching, Non-Teaching staff and our friends** who have directly or indirectly helped and supported us in completing our project in time.

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **SNo.** |  | **Contents** | **Page** |
| **no.** |  |  |  |
| 1. |  | INTRODUCTION | 1 |
| 2. |  | EXECUTIVE SUMMARY | 2 |
| 3. |  | ABOUT THE COMPANY | 2 |
| 4. |  | OPPORTUNITIES | 3 |
| 5. |  | TRAINING | 4 |
| 6. |  | LITERATURE REVIEW  6.1 EXISTING SYSTEMS  6.2 PROPOSED SOLUTION |  |
| 7. |  | PYTHON & STATISTICS | 7 |
| 8. |  | ADVANTAGES AND DISADVANTGES | 8 |
| 9. |  | METHODOLOGY | 9 |
| 10. |  | THEORIETICAL ANALYSIS |  |
| 11. |  | SOURCE CODE | 10 |
| 12. |  | RESULTS | 11 |
| 13  14  15 |  | CHALLENGES FACED  CONCLUSIONAND FUTURE STAGE  BIBLIOGRAPHY | 16  17  18 |

# 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

|  |  |  |  |
| --- | --- | --- | --- |
| **1st 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 |

|  |  |  |  |
| --- | --- | --- | --- |
| **2nd 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 |

|  |  |  |  |
| --- | --- | --- | --- |
| **3rd 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 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | **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 report is about my 2 weeks internship program with NSIC (National Small industries corporate limited). In this comprehensive report, I have discussed about every major aspect of the company which I observed and perceived during my internship program.

During my internship program, I have learned and mainly worked on Data Science. All the details have been discussed in detail. All the policies and procedures of the company have been discussed in detail.

Asthemainpurposeoftheinternshipistolearnbyworkinginpracticalenvironment and to apply the knowledge acquired during the studies in real world scenario in order to tackle the problems using the knowledge and skill learned during the academic process.

# 

# 

# 

# 3. ABOUT THE COMPANY

Micro small and medium enterprises (MSME), is an ISO CC-Cl20849 certified Government of India Enterprise under Ministry of Micro, Small and Medium Enterprises (MSME). NSIC has been working to promote aid and foster the growth of micro, small and medium enterprises in the country. MSME operates through countrywide network of offices and Technical Centers in the Country. In addition, MSME has set up Training cum Incubation Centre managed by professional manpower.

**Mission**: “To promote and support Micro, Small & Medium Enterprises (MSMEs) Sector” by providing integrated support services encompassing Marketing, Technology, Finance and other services.

**Vision**: “To be a premier Organization fostering the growth of Micro, Small and Medium Enterprises (MSMEs) Sector”.

# 4.OPPORTUNITIES

During these six months 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. LITERATURE REVIEW**

**1.1 Churn Prediction Models:** Numerous studies have explored diverse churn prediction models, ranging from traditional statistical methods to cutting-edge machine learning techniques. Early models often relied on logistic regression and decision trees, while recent advancements showcase the efficacy of ensemble methods, neural networks, and deep learning architectures.

**1.2 Feature Selection and Engineering:** The identification and selection of relevant features play a pivotal role in the accuracy of churn prediction models. Literature suggests that incorporating a combination of behavioral, demographic, and usage-related features enhances the predictive power of models. Additionally, feature engineering techniques such as time-series analysis and sentiment analysis contribute to model robustness.

**3.3 Data Preprocessing Techniques:** The quality of input data significantly influences the performance of churn prediction models. Studies highlight the importance of data preprocessing techniques, including handling missing values, outlier detection, and normalization, to ensure the reliability of predictive analytics.

**1.4 Model Evaluation Metric:** The evaluation of churn prediction models necessitates the use of appropriate metrics. Commonly employed metrics include accuracy, precision, recall, and the area under the receiver operating characteristic curve (AUC-ROC). Literature emphasizes the need for a balanced assessment to avoid skewed results.

**1.5 Customer Segmentation and Personalization:** Recent trends in churn prediction literature underscore the importance of customer segmentation and personalized interventions. Tailoring retention strategies based on customer segments, identified through clustering algorithms, enhances the effectiveness of targeted interventions.

**1.6 Ethical Considerations:** As the use of data science in customer churn prediction becomes prevalent, ethical considerations emerge. Literature explores the ethical implications of utilizing customer data, emphasizing the need for transparency, consent, and responsible data handling practices.

**7. PYTHON**

Python is an interpreted, high-level, general-purpose programming language. It has efficient high-level data structures and a simple but effective approach to object-oriented programming. Python’s elegant syntax and dynamic typing, together with its interpreted nature, make it an ideal language for scripting and rapid application development in many areas on most platforms.

Python for Data science:

Why Python???

1. Python is an open source language.

2. Syntax as simple as English.

3. Very large and Collaborative developer community.

4. Extensive Packages.

**UNDERSTANDING OPERATORS:**

Theory of operators: Operators are symbolic representation of Mathematical tasks.

**Variables And Datatypes**: Variables are named bounded to objects. Data types in python are int (Integer), Float, Boolean and strings.

**CONDITIONAL STATEMENTS**: If-else statements (Single condition)If, elif statements (Multiple Condition)

**LOOPING CONSTRUCTS**: For loop while loop

**FUNCTIONS:** Functions are re-usable piece of code. Created for solving specific problem.

**Two types**: Built-in functions and Userdefined functions. Functions cannot be reused in python.

**DATA STRUCTURES**: Two types of Data structures:

**LISTS:** A list is an ordered data structure with elements separated by comma and enclosed within square brackets.

**DICTIONARY:** A dictionary is an unordered data structure with elements separated by comma and stored as key: value pair, enclosed with curly braces {}.

**7.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**

\[ \text{Median} = \begin{cases} x\_{(n+1)/2} & \text{if } n \text{ is odd} \\\frac{1}{2}

(x\_{n/2} + x\_{n/2 + 1}) & \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}{\frac{\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)**

\[ R^2 = 1 \frac{\sum\_{i=1}^{n} (y\_i \hat{y}\_i)^2}{\sum\_{i=1}^{n} (y\_i \bar{y})^2} \]

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

**8. 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.

**9.** **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:**

\[ P(y=1) = \frac{1}{1 + e^{-(\beta\_0 + \beta\_1x\_1 + \beta\_2x\_2 + \ldots + \beta\_nx\_n)}} \]

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.

**10.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.

**11.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()

def half\_corr\_heatmap(data, title=None):

plt.figure(figsize=(9,9))

sns.set(font\_scale=1)

mask = np.zeros\_like(data.corr())

mask[np.tril\_indices\_from(mask)] = True

with sns.axes\_style("white"):

sns.heatmap(data.corr(), mask=mask, annot=True, cmap="coolwarm")

if title: plt.title(f"\n{title}\n", fontsize=18)

plt.show()

return half\_corr\_heatmap(data, 'Correlation Between Variables')

data['Churn'] = data['Churn'].map({'Yes' : 1, 'No' : 0})

half\_corr\_heatmap(data, 'Correlation Between Variables')

def corr\_for\_target(data, target, title=None):

plt.figure(figsize=(4,14))

sns.set(font\_scale=1)

sns.heatmap(data.corr()[[target]].sort\_values(target,ascending=False)[1:],annot=True,cmap="coolwarm")

if title: plt.title(f"\n{title}\n", fontsize=18)

return

corr\_for\_target(data, 'Churn', 'Correlation Between Target')

numerical = data2.select\_dtypes(['number']).columns

print(f'Numerical: {numerical}\n')

categorical = data2.columns.difference(numerical)

data2[categorical] = data2[categorical].astype('object')

print(f'Categorical: {categorical}')

data2 = pd.get\_dummies(data2)

data\_cols = data.drop('customerID', axis = 1)

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.tree import DecisionTreeClassifier

from sklearn.svm import SVC

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)

models = []

models.append(('Random Forest Clas.', RandomForestClassifier()))

models.append(('Decision Tree Clas.', DecisionTreeClassifier()))

models.append(('Logistic Reg.', LogisticRegression()))

models.append(('SVC', SVC()))

model\_names = []

scores = []

for name, model in models:

score = cross\_val\_score(model, X, y, cv = 5, scoring='accuracy')

scores.append(score)

model\_names.append(name)

print(f"Mean of the {name} model scores : {score.mean()}")

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\_sc = SVC()

svc\_sc.fit(X\_train\_sc, y\_train)

y\_pred\_sc = svc\_sc.predict(X\_test\_sc)

y\_pred\_sc\_train = svc\_sc.predict(X\_train\_sc)

svc\_sc\_train\_as = metrics.accuracy\_score(y\_train, y\_pred\_sc\_train)

svc\_sc\_as = metrics.accuracy\_score(y\_test, y\_pred\_sc)

print(f"Accuracy score for test data : {svc\_sc\_as}")

print(f"Accuracy score for train data : {svc\_sc\_train\_as}")

params = {'kernel' : ['rbf'], 'C' : [0.1, 1, 5, 10], 'gamma' : [0.01, 0.1, 0.9, 1]}

grid = GridSearchCV(SVC(), params, cv = 5, return\_train\_score= False)

# svc\_new = SVC(\*\*grid.best\_params\_)

svc\_new = SVC(C = 1, gamma = 0.01, kernel = 'rbf')

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}")

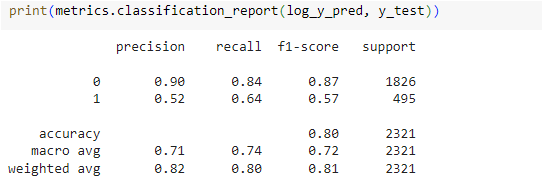
print(f"Accuracy score for train data : {svc\_new\_train\_as}")

Links for Dataset and code:

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

**Dataset: Kaggle:** [CUSTOMER CHURN PREDICTION 📈 | Kaggle](https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction/input)

**12.RESULTS**

****

**13.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

**14.CONCLUSION AND FUTURE STAGE**

**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 [mention challenges] 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.

2. Future Stages

2.1 Refinement and Optimization

While the current models have demonstrated promising results, there is room for refinement and optimization. Further exploration of hyperparameter tuning, feature selection, and model architecture adjustments may lead to incremental improvements in predictive accuracy.

2.2 Real-time Implementation

The transition from experimental models to real-time implementation is a critical next step. Integration of the developed models into the operational framework of telecom companies allows for dynamic and proactive customer churn management.

2.3 Continuous Monitoring and Adaptation

The telecom industry is dynamic, and customer behaviors evolve over time. Establishing a system for continuous monitoring and adaptation of the churn prediction models ensures their relevance and effectiveness in the face of changing market conditions.

2.4 Ethical Considerations

As predictive analytics become integral to customer relationship management, ethical considerations must be prioritized. Future stages should involve the implementation of transparent and responsible data handling practices, ensuring customer privacy and compliance with regulations.

2.5 Integration with Customer Engagement Strategies

The predictive models developed can be seamlessly integrated with customer engagement strategies. By aligning churn predictions with targeted marketing campaigns and personalized customer interactions, telecom providers can maximize the impact of retention efforts.

2.6 Collaboration with Stakeholders

Collaboration with stakeholders, including marketing teams, customer service, and data scientists, is crucial for the success of churn prediction models. Establishing cross-functional teams for ongoing collaboration facilitates a holistic and effective approach to customer retention.

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 landscape.

**15. BIBLIOGRAPHY**

Here is a list of references, including websites, books, and research papers, consulted during the analysis and development of the project:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.

Raschka, S., & Mirjalili, V. (20). Python Machine Learning. Packt Publishing.

Scikit-learn Documentation. <https://scikit-learn.org/stable/documentation.html>

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Logistic regression Algorithm Documentation. <https://scikit-learn.org/stable/modules/ensemble.html#logestic>

DataCamp. <https://www.datacamp.com/>

Kaggle. <https://www.kaggle.com/>

Python Programming for Data Science. <https://www.python.org>