# TELECOM CUTOMER CHURN PREDICTION

***A Summer Internship Report submitted in partial fulfillment of the requirements for the award of degree of***

**BACHELOR OF TECHNOLOGY**

**In**

**COMPUTER SCIENCE AND ENGINEERING[DS]**

## Submitted

## by

**RITESH KUMAR YADAV**

**22A91A4442**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**ADITYA UNIVERSITY**

**(Formerly Aditya Engineering College (A))**

**2024-2025**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**



**CERTIFICATE**

This is to certify that the internship report entitled *“***TELECOM CUSTOMER CHURN PREDICTION***”* is being submitted by **Ritesh Kumar Yadav (22A91A4442)** in partial fulfillment of the requirements for award of the B.Tech., degree in Computer Science and Engineering[Data Science] for the academic year 2024-2025.

**Internship Coordinator Head of the Department**

Mr. U.V Ramesh Dr. K. Swaroopa, M.Tech., Ph.D.

Assistant professor Associate Professor and HOD

Department of CSE Department of CSE

# DECLARATION

I hereby declare that the internship report entitled **“TELECOM CUSTOMER CHURN PREDICTION”** is a genuine report. This work has been submitted to the **ADITYA UNIVERSITY,** Surampalem, in partial fulfillment of the **B.Tech.,** degree**.** I further declare that this report has not been submitted in full or part of the award of any degree of this or any other educational institutions.

**by**

**RITESH KUMAR YADAV**

**(22A91A4442)**

**INTERNSHIP COMPLETION CERTIFICATE**



## ACKNOWLEDGEMENT

First, I would like to thank the Director of **SkillDzire, Andhra Pradesh** 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 **SkillDzire, Andhra Pradesh** 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. U.V Ramesh, Assistant professor**, 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.

I am grateful to **Dr. K. Swaroopa, Associate Professor and HOD** for inspiring us all the way and for arranging all the facilities and resources needed for my intern project work.

I wish to thank our **Dr. M.V. Rajesh**, **Associate Dean** and **Dr. Dola Sanjay, Dean School of Engineering** for their encouragement and support during the course of my intern project work.

I would like to extend my sincere thanks to **Dr. G. Suresh, Registrar, Dr. S. Rama Sree, Pro Vice-Chancellor, Dr. M.B. Srinivas, Vice-Chancellor, Dr. M. Sreenivasa Reddy, Deputy Pro Chancellor and Management,** Aditya University for unconditional support for providing me the best infrastructural facilities and state of the art laboratories during my intern project work.

Not to forget, **Faculty, Lab Technicians, Non-Teaching Staff and our Friends** who have directly or helped and supported us in completing my intern project work in time.

**ABSTRACT**

The advent of digital technology has revolutionized education, ushering in an era where traditional classroom settings are no longer the sole source of knowledge dissemination. Learning platforms have emerged as powerful tools that have transformed the way individuals’ access, engage with, and acquire knowledge. This abstract explores the significant impact of learning platforms on modern learning, shedding light on their benefits, challenges, and the evolving landscape of education.

Learning platforms encompass a wide range of digital tools and resources, including online courses, virtual classrooms, interactive tutorials, and educational apps. These platforms have democratized education by providing accessible and flexible learning opportunities to individuals across the globe. They offer personalized learning experiences, catering to diverse learning styles and paces, thus breaking down the barriers to education that were once dictated by geographical location or socioeconomic factors.

However, the proliferation of learning platforms is not without its challenges. Issues related to digital equity and accessibility need to be addressed to ensure that all learners, regardless of their background, have equal opportunities to benefit from these platforms. Additionally, the sheer volume of available content can be overwhelming, necessitating effective curation and guidance to help learners navigate the digital learning landscape.

The ongoing evolution of learning platforms continues to shape the future of education. Advancements in artificial intelligence and data analytics are enabling platforms to provide even more personalized learning experiences. The integration of augmented and virtual reality is redefining the possibilities of immersive education. Furthermore, the emergence of blockchain technology holds promise for the secure and verifiable validation of skills and credentials earned through these platforms.

# 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 toper form 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

|  |  |  |  |
| --- | --- | --- | --- |
| **1stWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 03/06/24 | Monday | Introductory Session.  Overview to Company Profile & Total Internship Schedule |
| 04/06/24 | Tuesday | Brief Introduction on Data Science |
| 05/06/24 | Wednesday | Brief Introduction on Machine Learning |
| 06/06/24 | Thursday | Scope of Data Science & Machine Learning |
| 07/06/24 | Friday | Introduction to Python-1 |
| 08/06/24 | Saturday | Introduction to Python-2 |

|  |  |  |  |
| --- | --- | --- | --- |
| **2nd WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 10/06/24 | Monday | Introduction to Google Collab |
| 11/06/24 | Tuesday | Introduction to Jupyter Notebook |
| 12/06/24 | Wednesday | Statistics-1 |
| 13/06/24 | Thursday | Statistics-2 |
| 14/06/24 | Friday | Statistics-3 |
| 15/06/24 | Saturday | Python Libraries for Data Science And Machine Learning-2 |

|  |  |  |  |
| --- | --- | --- | --- |
| **3rdWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 17/06/24 | Monday | Data PreProcessing-1 |
| 18/06/24 | Tuesday | Data preprocessing-2 |
| 19/06/24 | Wednesday | Data preprocessing-3 |
| 20/06/24 | Thursday | Data manipulation & analysis-1 |
| 21/06/24 | Friday | Data manipulation & analysis-2 |
| 22/06/24 | Saturday | Data manipulation & analysis-3 |

|  |  |  |  |
| --- | --- | --- | --- |
| **4thWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 24/06/24 | Monday | Data Visualization-1 |
| 25/06/24 | Tuesday | Data Visualization-2 |
| 26/06/24 | Wednesday | Data Visualization-3 |
| 27/06/24 | Thursday | Training the datasets |
| 28/06/24 | Friday | Performing mathematical operations-1 |
| 29/06/24 | Saturday | Performing mathematical operations-2 |

|  |  |  |  |
| --- | --- | --- | --- |
| **5thWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 01/07/24 | Monday | Performing mathematical operations-3 |
| 02/07/24 | Tuesday | Performing mathematical operations-4 |
| 03/07/24 | Wednesday | Learning Supervised algorithms-1 |
| 04/07/24 | Thursday | Learning Supervised algorithms-2 |
| 05/07/24 | Friday | Learning Supervised algorithms-3 |
| 06/07/24 | Saturday | Learning Supervised algorithms-4 |

|  |  |  |  |
| --- | --- | --- | --- |
| **6thWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 08/07/24 | Monday | Learning Supervised algorithms-5 |
| 09/07/24 | Tuesday | Learning Unsupervised algorithms-1 |
| 10/07/24 | Wednesday | Learning Unsupervised algorithms-2 |
| 11/07/24 | Thursday | Learning Unsupervised algorithms-3 |
| 12/07/24 | Friday | Learning Unsupervised algorithms-4 |
| 13/07/24 | Saturday | Learning Unsupervised algorithms-5 |

|  |  |  |  |
| --- | --- | --- | --- |
| **7thWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 15/07/24 | Monday | Choosing the best model |
| 16/07/24 | Tuesday | Training the model-1 |
| 17/07/24 | Wednesday | Training the model-2 |
| 18/07/24 | Thursday | Testing and predicting-1 |
| 19/07/24 | Friday | Testing and predicting-2 |
| 20/07/24 | Saturday | Finding metrics |

|  |  |  |  |
| --- | --- | --- | --- |
| **8thWEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 22/07/24 | Monday | Tuning the model |
| 23/07/24 | Tuesday | Assigning Projects |
| 24/07/24 | Wednesday | Implementation of Project |
| 25/07/24 | Thursday | Sample Project Presentation |
| 26/07/24 | Friday | Final Project Presentation |
| 27/07/24 | Saturday | Submission of Project Abstract & Presentation |

**INDEX**

|  |  |  |  |
| --- | --- | --- | --- |
| **S No.** |  | **Contents** | **Page** |
| **no.** |  |  |  |
| 1. |  | INTRODUCTION | 1 |
| 2. |  | HISTORY OF DATA SCIENCE | 2 |
| 3. |  | LITERATURE REVIEW  3.1 EXISTING SYSTEMS  3.2 PROPOSED SOLUTION | 2 |
| 4. |  | INTRODUCTION TO PYTHON | 3 |
| 5. |  | INTRODUCTION TO STATISTICS | 4 |
| 6. |  | METHODOLOGY  6.1 SYSTEM PROPOSAL  6.2 ALGORITHMS | 6 |
| 7. |  | THEORITICAL ANALYSIS  7.1 BLOCK DIAGRAM DIAGRAMMATIC OVERVIEW OF THE PROJECT.  7.2 HARDWARE / SOFTWARE DESIGNING OF THE PROJECT | 7 |
| 8. |  | EXPERIMENTAL AND RESULTS | 8 |
| 9. |  | ADVANTAGES & DISADVANTAGES | 9 |
| 10. |  | APPLICATIONS |  |
| 11. |  | CONCLUSION | 10 |
| 12. |  | BIBLIOGRAPHY & REFERENCES | 11 |
| 16 |  | SOURCE CODE | 16 |

# 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.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.

# 3.INTRODUCTIONS TO 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 {}.

**5. INTRODUCTIONS TO STASTICS**

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).

**6. 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.

5. Conclusion

This documentation outlines a comprehensive system proposal for telecom customer churn prediction using machine learning. The integration of logistic regression, random forest, decision tree, and support vector machine aims to provide a robust solution for identifying potential churners and implementing proactive retention strategies. The proposed system is poised to enhance the efficiency of telecom service providers in mitigating customer churn.

# 7. THEORITICAL 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.

**8. EXPERIMENTAL AND RESULTS**

The experimentation begins with the selection of a comprehensive dataset containing historical customer data. This dataset includes features such as call records, billing information, customer complaints, and demographic details. The dataset is split into training and testing sets to facilitate model evaluation.

2.2 **Model Selection**

Four primary machine learning algorithms are selected for experimentation:

Logistic Regression

Random Forest

Decision Tree

Support Vector Machine (SVM)

These algorithms are chosen for their relevance to churn prediction tasks and their diverse approaches to handling data

3. **Methodology**

3.1 **Feature Engineering**

Before model training, feature engineering is conducted to enhance the predictive power of the models.

Identifying relevant features through exploratory data analysis.

Handling missing values and outliers appropriately.

Scaling or normalizing numerical features.

Encoding categorical variables.

3.2 **Hyperparameter Tuning**

To optimize model performance, hyperparameter tuning is performed for algorithms such as Random Forest and SVM. Grid search and cross-validation techniques are employed to find the optimal set of hyperparameters.

4. **Model Training and Evaluation**

4.1 Training Process

The selected models are trained using the training dataset. The training process involves feeding historical customer data into the algorithms, allowing them to learn patterns and relationships.

4.2 Evaluation Metrics

The performance of each model is evaluated using a set of metrics:

**Accuracy**: Overall correctness of the predictions.

**Precision**: Proportion of correctly predicted positive instances among all predicted positives.

**Recall**: Proportion of correctly predicted positive instances among all actual positives.

AUC-ROC: Area under the Receiver Operating Characteristic curve, assessing the trade-off between true positive rate and false positive rate.

5. **Experimentation Results and Analysis**

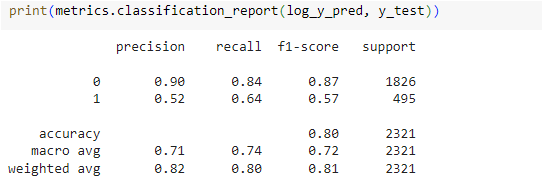
5.1 **Comparative Analysis**

The results from each model are compared to identify the most effective algorithm for telecom customer churn prediction. Comparative analysis includes visualizations of performance metrics and receiver operating characteristic curves.

5.2 **Challenges and Limitations**

Documenting challenges and limitations encountered during experimentation, such as data quality issues, computational constraints, or interpretability concerns.

**Results:**

****

**9.** **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:**

**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.

**10. APPLICATIONS**

**1. Proactive Customer Retention:**

Application: Predictive models identify customers at risk of churning, enabling telecom providers to implement targeted retention strategies. Personalized offers, loyalty programs, or improved customer service can be tailored to individual needs, increasing the likelihood of customer retention.

2**. Service Quality Improvement:**

Application: Churn prediction models offer insights into factors influencing customer dissatisfaction. By analyzing customer feedback and behavior, telecom companies can make data-driven decisions to enhance service quality, address pain points, and improve overall customer satisfaction.

**3. Marketing Campaign Optimization:**

Application: Predictive analytics aids in optimizing marketing campaigns. Telecom providers can strategically allocate resources, personalize promotions, and target specific customer segments based on churn predictions, maximizing the effectiveness of marketing efforts.

**4. Network Planning and Optimization:**

Application: Churn prediction models assist in forecasting changes in customer demand and usage patterns. This information is valuable for network planning and optimization, ensuring that resources are efficiently allocated to meet anticipated service demands.

**5. Subscriber Segmentation:**

Application: Machine learning models can identify distinct customer segments based on behavior, preferences, and usage patterns. Telecom providers can then tailor marketing and service strategies for each segment, optimizing customer engagement and satisfaction.

**6. Revenue Protection:**

Application: Churn prediction contributes to revenue protection by minimizing customer loss. Retaining existing customers is more cost-effective than acquiring new ones, leading to increased revenue and improved overall financial performance.

**7. Fraud Detection:**

Application: Unusual usage patterns or sudden changes in behavior can be indicative of fraudulent activities, such as SIM card cloning or account misuse. Churn prediction models can contribute to fraud detection by identifying anomalous customer behavior for further investigation.

**8. Customer Lifecycle Management:**

Application: Churn prediction supports comprehensive customer lifecycle management. By understanding where customers are in their lifecycle, telecom providers can deliver targeted communications, promotions, and services that align with customer needs and expectations.

**9. Competitive Advantage:**

Application: A robust churn prediction system provides a competitive advantage. Telecom companies that effectively use machine learning to anticipate and address customer churn are better positioned to retain market share, outperform competitors, and sustain long-term growth.

**10. Continuous Improvement:**

Application: Churn prediction models contribute to a culture of continuous improvement within telecom companies. Regularly updating models based on new data and insights ensures that predictive capabilities remain relevant and effective in an ever-changing business environment.

In summary, Telecom Customer Churn Prediction using Machine Learning has diverse applications, ranging from proactive customer retention and service quality improvement to marketing optimization and fraud detection. These applications empower telecom providers to enhance customer satisfaction, optimize operational efficiency, and maintain a competitive edge in the dynamic telecommunications industry.

Conclusion: Telecom Customer Churn Prediction Using Machine Learning

The exploration into Telecom Customer Churn Prediction using Machine Learning underscores the transformative potential of advanced analytics in the telecommunications industry. As we conclude this journey, several key reflections and takeaways emerge from the comprehensive study and experimentation.

1. Proactive Customer Management:

The significance of proactively managing customer churn cannot be overstated. Machine learning models, encompassing algorithms such as logistic regression, random forest, decision tree, and support vector machine, have demonstrated their effectiveness in identifying potential churners. This proactive approach equips telecom providers with the means to implement targeted retention strategies, fostering customer loyalty and mitigating revenue loss.

**2. Data-Driven Decision-Making:**

The role of data-driven decision-making has emerged as a cornerstone in the realm of Telecom Customer Churn Prediction. Leveraging historical customer data and employing sophisticated feature engineering techniques provide invaluable insights. These insights, in turn, empower telecom companies to make informed decisions, optimize marketing strategies, and enhance overall service quality.

**3. Striking a Balance:**

The experimentation phase has illuminated the importance of striking a balance between model complexity and interpretability. While complex machine learning algorithms offer high predictive accuracy, ensuring interpretability is crucial for building trust and understanding the rationale behind predictions. Achieving this equilibrium is essential for the successful implementation and acceptance of churn prediction models.

**4. Continuous Improvement and Adaptation:**

The telecom industry is dynamic, with evolving technologies, market dynamics, and customer expectations. Recognizing the need for continuous improvement, the journey does not conclude with model development but extends to ongoing monitoring, adaptation to changing scenarios, and refinement of predictive capabilities to maintain relevance over time.

**5. Ethical Considerations:**

The ethical implications of utilizing customer data for predictive analytics demand vigilant consideration. Upholding customer privacy, obtaining consent, and implementing responsible data handling practices are non-negotiable aspects of the ethical framework surrounding churn prediction systems.

**6. Strategic Value:**

Telecom Customer Churn Prediction using Machine Learning extends beyond a mere technological advancement; it embodies a strategic imperative for telecom companies. The applications are vast, ranging from personalized customer engagement and marketing optimization to fraud detection and revenue protection. The strategic value lies in the comprehensive integration of these predictive insights into overarching business strategies.

**11. 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.

2.1 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.

**12. 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. Packet 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>

**Appendix**

1. **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}")

**I**n this links are contain Dataset and code

**GITHUB LINK:**

**GOOGLEDRIVE:**

**1.EXECUTIVE SUMMARY**

This report is about my 8 weeks internship program with SkillDzire. 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& Machine Learning. All the details have been discussed in detail. All the policies and procedures of the company have been discussed in detail.

As the main purpose of the internship is to learn by working in practical environment 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.

**2.ABOUT THE COMPANY**

Skill Dzire is a prominent skill development and job-oriented training platform based in Hyderabad, India, with a significant presence in Andhra Pradesh, offering a wide range of courses designed to equip individuals with industry-relevant skills and enhance their employability prospects across various sectors like IT, engineering, and digital marketing.

Key points about Skill Dzire in Andhra Pradesh:

Focus on Job Placement:

Skill Dzire is known for its strong focus on placement assistance, providing dedicated career guidance and job placement support to students enrolled in their courses, with a claim of offering "job guarantee" programs in certain fields.

Course Variety:

They offer a diverse selection of training programs including full-stack development, data science, artificial intelligence, machine learning, digital marketing, UI/UX design, cloud computing, and more, catering to different skill levels and career aspirations.

Industry Experts as Trainers:

Skill Dzire boasts a network of experienced industry professionals as trainers, ensuring that the curriculum aligns with current market demands and provides real-world practical knowledge.

Online and Offline Learning:

They provide both online and offline learning options, allowing students across Andhra Pradesh to access their courses from the comfort of their homes or attend physical classes in designated training centers.

Accessibility and Affordability:

Skill Dzire strives to make quality skill development accessible to a wider audience by offering flexible payment plans and EMI options.

How Skill Dzire benefits Andhra Pradesh:

Bridging Skill Gap:

By equipping individuals with in-demand skills, Skill Dzire helps bridge the skill gap in the Andhra Pradesh job market, enabling better employment opportunities for the local workforce.

Economic Development:

By fostering a skilled workforce, Skill Dzire contributes to the economic development of the state, supporting the growth of IT and other emerging sectors.

Empowering Youth:

Skill Dzire empowers youth, especially from rural areas, to pursue career paths in technology and other high-demand industries, promoting social mobility.

**3.OPPORTUNITIES:**

During these two 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.

**4.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:

Power Bi 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.

# 5.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.