ISSN No:-2456-2165

Machine Learning Based Telecom-Customer Churn Prediction

C. Subalakshmi; G. Bhanu Praveen; C. V. Saketh; N. Reddy Samba Siva Reddy
IV Year B.Tech CSE DS(AI) Students, Dept of Computer science and Engineering,
DR. M.G.R Educational and Research Institute, Maduravoyal, Chennai-95, Tamil Nadu, India

Abstract:- In the highly competitive telecom sector, maintaining client loyalty is a critical obstacle to long-term profitability and expansion. This research uses the Random Forest and Logistic Regression algorithms to give a detailed investigation of customer attrition prediction specifically for the telecom industry. Building a strong predictive model to identify possible churners will enable telecom businesses to implement focused customer loyalty campaigns.

Our methodology incorporates a wide range of telecom-specific characteristics, such as call trends, usage information, and customer support exchanges. By utilizing the Random Forest and Logistic Regression methods, we may increase the forecasting accuracy by exploring the complex patterns that indicate customer churn. Carefully considered feature engineering techniques are used to improve the model's capacity to capture subtleties specific to the telecom . Our approach is validated using a real-world telecom dataset that includes a range of customer categories. Performance metrics such as F1 score, recall, accuracy, and precision show how well our model forecasts customer attrition in the dynamic telecom market.

Keywords:- *Customer Churn, Machine Learning, Telecom Sector, Performance Metrics*

I. INTRODUCTION

The telecommunications industry is characterized by constant change and fierce competition, making customer retention essential to long-term survival, market share, and profitability. Telecom firms confront a daunting task as consumer options grow and technology advances: not only must they attract new consumers, but more importantly, they must hold onto their current clientele. One major barrier to reaching this retention target is customer churn, the occurrence when users defect to rival service providers.

In the telecom sector, which is renowned for its quick innovation and changing consumer expectations, this study conducts a thorough investigation into the field of customer churn prediction. Our research is centered on the use of sophisticated machine learning methods, particularly the Random Forest and Logistic Regression algorithms. Our goal is to create a strong predictive model that can accurately detect probable churners by utilizing these advanced approaches.

Our approach's telecom-specific peculiarities are highlighted by the incorporation of a wide range of industry-specific elements. These include call trends, use information, and contacts with customer service representatives, among other aspects of consumer behavior. Taking into account the complex nature of customer attrition, our approach attempts to analyze and comprehend the complex dynamics that lead up to the loss of subscribers.

Because of its capacity to manage intricate, non-linear interactions within the data, the Random Forest algorithm was selected, whereas Logistic Regression offers interpretability and insights into the importance of specific features. By carefully crafting its features, we improve the model's ability to identify nuanced patterns that are exclusive to the telecom sector, hence enhancing its predictive power.

We perform a thorough analysis on a real-world telecom dataset that covers a variety of client groups in order to validate the effectiveness of our strategy. In order to provide a thorough evaluation of the Important metrics including recall, accuracy, precision, and the F1 score are utilised to assess the model's success in predicting customer attrition in the dynamic telecom industry.

In the fiercely competitive and technologically-advancing telecom industry, keeping customers loyal poses a significant issue. The importance of this issue is highlighted by recent industry studies, which show that telecom businesses face an annual turnover rate of 10% to 15%. These figures underscore the necessity of taking proactive steps to reduce customer attrition in addition to highlighting the financial impact of the problem. The strategic importance of anticipating and reducing customer attrition is becoming more and more evident as telecom carriers come to terms with the harsh fact that recruiting new customers can cost up to five times more than keeping existing ones.This study uses sophisticated machine learning methods, including the Random Forest and Logistic Regression algorithms, to explore the complex field of customer churn prediction in the telecom sector. By doing this, we hope to give telecom companies a powerful prediction model that does more for them than just identify possible churners; instead, it gives them useful information that they can use to launch targeted, successful customer loyalty efforts.

Our technique takes into account a wide range of telecom-specific characteristics, such as call statistics, usage data, and customer support interactions. This strategy is based on the understanding that complex datasets specific to the telecom sector require advanced algorithms that can identify minute trends that indicate customer attrition.Utilizing advanced machine learning techniques, such as the Random Forest and Logistic Regression algorithms, this study delves into the intricate domain of telecom customer churn prediction. By doing this, we intend to provide telecom firms with a robust prediction model that helps them with more than just identifying potential churners; rather, it provides them with actionable insights that they can utilize to initiate focused, fruitful customer loyalty initiatives. Numerous telecom-specific factors, including call records, usage data, and customer support interactions, are taken into consideration by our method. This approach is predicated on the knowledge that sophisticated algorithms are needed to detect minute trends that signify customer attrition in complicated datasets unique to the telecom industry.

In conclusion, this study not only fills a critical gap in the telecom industry's understanding of customer churn prediction, but it also advances the conversation about the strategic business insights that can be obtained from advanced machine learning techniques. By giving telecom companies a customized approach and useful data, we hope to enable them to continue their unwavering focus on customer retention and long-term profitability.

II. EXISTING WORK

In the ever-changing telecom industry, the need to combat customer churn has sparked a slew of research projects targeted at applying advanced analytics and machine learning approaches. Previous research in the realm of customer churn prediction has laid the groundwork for our current study, offering light on approaches, obstacles, and outcomes that have impacted our understanding of this essential subject.

A thorough analysis of the available literature indicates a plethora of ways to predicting customer attrition in the telecommunications industry. Researchers have investigated the effectiveness of various machine learning techniques, such as decision trees, neural networks, and ensemble methods. Notably, studies have highlighted the potential of Random Forest and Logistic Regression algorithms, which are relevant to our current research.Several significant publications have delves into the complexities of telecom datasets, emphasizing the value of feature engineering and the nuanced interpretation of customer behavior patterns. For example, highlighted the efficacy of call detail records and customer contact data in predicting churn, revealing temporal and behavioral elements impacting subscriber attrition. Furthermore, the introduction of big data analytics has sparked a paradigm shift in the forecast of client attrition. Recent research has underlined the importance of utilizing huge datasets to identify hidden patterns and gain previously unreachable insights. Our study contributes to the continuing discussion on the practical application of machine learning in tackling industry-specific difficulties since it coincides with the current trend of leveraging real-world telecom information.

Despite the advances made in the field, problems remain. The intrinsic complexity of telecom data, privacy concerns, and the necessity for interpretability in predictive models remain hot topics in academic and corporate circles. By expanding on previous studies, we hope to improve our understanding of customer churn prediction in the telecom business. We hope to add nuanced insights and practical approaches geared to the unique issues presented by the telecom sector by focusing primarily on the Random Forest and Logistic Regression algorithms.

We will elaborate on our methodology in the following sections of this paper, detailing the application of these algorithms to a real-world telecom dataset. By connecting our work with and expanding on existing studies, we hope to provide a thorough and forward-thinking contribution to the subject of customer attrition prediction in telecoms.

III. ALGORITHMS

Using cutting edge machine learning algorithms is essential when it comes to telecom sector predictive modeling for customer attrition. This study uses two well-known algorithms, Random Forest and Logistic Regression, which each bring unique advantages to the difficult issue of subscriber attrition prediction.

A. Logistic Regression

A statistical technique for simulating the likelihood of a binary outcome is called logistic regression. The binary result in customer churn prediction usually indicates whether a client will churn (1) or not (0). The logistic function (sigmoid function) is used to convert a linear combination of input data into a probability value between 0 and 1. It is used to predict a categorical dependent variable from a given set of independent variables.

- With logistic regression, the result of a categorical dependent variable is predicted. As a result, a discrete or category value must be the result.
- Instead of providing the exact values, which are 0 and 1, it provides the probabilistic values, which fall between 0 and 1. It can be either Yes or No, 0 or 1, true or False, etc.
- With the exception of how they are applied, logistic regression and linear regression are very similar. While logistic regression is used to solve classification difficulties, linear regression is used to solve regression problems.
- In logistic regression, we fit a "S" shaped logistic function, which predicts two maximum values, rather than a regression line. (0 or 1).

ISSN No:-2456-2165

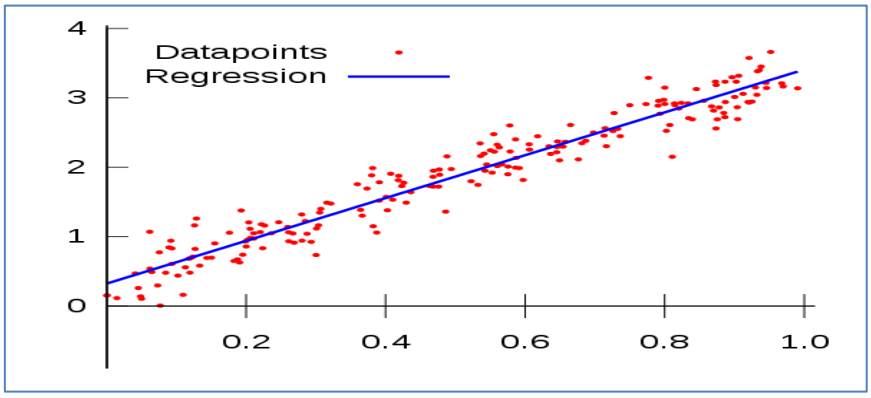

Fig. 1: Linear Regression Model for Predicting

Mathematics:-The logistic regression formula is:
$P(y = 1 /x) = \sigma(b0 + b1*x)$

Where:

- $P(y = 1 \mid x)$ is the probability of the event y=1, given x
- $\sigma(z)$ is the standard logistic function, also known as the sigmoid function, which maps any input to a value between 0 and 1.
- b0 and b1 are the parameters to be estimated, where b0 is the intercept and b1 is the slope.

B. Random forest

An ensemble learning technique called a random forest generates a large number of decision trees during the training phase and outputs a forecast that is the mean of the predictions made by each individual tree. Each tree in the forest is built by selecting a random subset of features at each node and finding the best split among those features to minimize the impurity measureAn ensemble algorithm known as random forest is made up of decision trees, it is also a supervised machine algorithm. The many individual decision trees that make up a random forest are unrelated to one another. The main steps in the operation of random forest are:

- For an N-sample-size-minimum sample, N samples are taken for N times, and getting an N sample is created by taking a single sample each time;
- When there are M attributes per sample and each node of the decision tree needs to be partitioned, then m attributes will be chosen at random (meeting the criterion of $m \leq M$). Then, one of these m attributes is chosen as the node's splitting attribute using some method;
- During the decision tree-building process, each node is divided in accordance with the precious step;

Follow the above procedures to construct a huge number of decision trees. Random forest can be applied to highly dimensional data, judge the significance of features and the interaction between different features, balance the error for unbalanced data sets, and can maintain accuracy for missing features. Compared to decision trees, it is less prone to overfitting. However, it has been shown to overfit on certain noisy classification or regression problems

IV. TOOLS AND LIBRARIES

- **Python:** Python is a popular high-level, interpreted, dynamically typed programming language that is easy to understand and use. It is versatile and widely used for applications such as web development, data analysis, artificial intelligence, machine learning, and automation. Python emphasizes code readability with a clean and straightforward syntax, reducing the cost of program maintenance.
- **Pandas:** Pandas is a Python-based open-source data analysis and manipulation package. The Data Frame, a two-dimensional table for holding structured data, is at the heart of Pandas.It provides key data structures like Series and DataFrame, enabling efficient data manipulation and analysis. pandas facilitates reading data from various file formats, such as CSV and Excel, making it a versatile tool for data loading
- **Numpy:** Numpy is an open-source numerical computation package for Python. It includes fundamental tools for working with huge, multidimensional arrays and matrices, as well as mathematical functions that may be applied to them. Numpy is a core Python module for scientific computing and data analysis that prioritises efficiency and performance.
- **Streamlit:** Streamlit is an open-source Python library designed for creating web applications for data science and machine learning with minimal effort. It simplifies the process of turning data scripts into interactive web applications. With a straightforward and intuitive API, streamlit enables users to build web apps quickly and efficiently. Key features include automatic widget generation, real-time updates, and the ability to integrate charts, tables, and interactive components seamlessly.
- **Plotly:** An open-source Python package called Plotly is used to create engaging and eye-catching data visualizations. Because of its adaptability to different chart types, it can be used for a wide range of data visualization applications. Plotly is renowned for its capacity to create dynamic dashboards that let people engage and examine data in real time.

ISSN No:-2456-2165

V. FLOW CHART

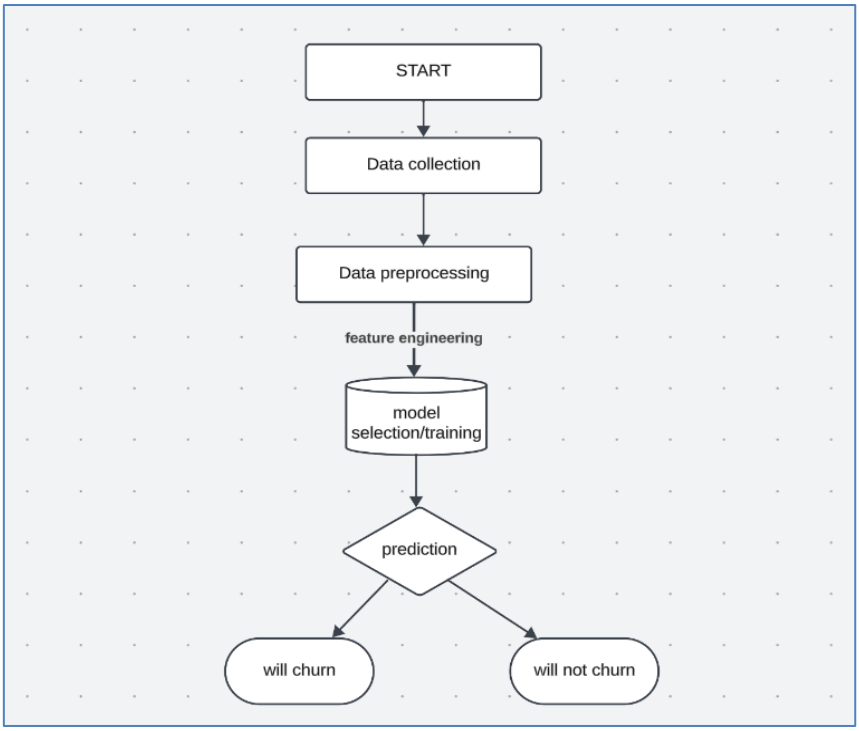

Fig. 2: Flow chart for churn process

Financial institutions are heavily dependent on client happiness for their operations, thus predicting customer attrition has a significant impact on them. These establishments operate in a very competitive market and keep clients by meeting their needs within the limitations of their resources. By fitting the model onto the existing historical data, data mining techniques are utilized to find interesting patterns and relationships in the data and anticipate the behavior of the consumers, whether they will be churning or not.

A. Data Collection

The dataset was extracted from the online of one of the leading site on internet Kaggle. The contained about 1000 customers' data with 30 attributes.

B. Data Preparation

The missing values with 30% null were removed from the dataset with the aid of Python programming language libraries. Numerical data was replaced with the 'mean' of the variables while the 'mode' was used for the categorical data. To achieve better performance, the categorical data was transformed to numerical format using the Label Encoder function in Python.

C. Data Pre-Processing

In order to prepare raw data for analysis and model training, a number of crucial tasks are involved in data preprocessing, which is an essential step in the data analysis and machine learning process. One of the main tasks is to deal with missing information by using sophisticated imputation algorithms or statistical measurements to either remove or impute them. To keep outliers from distorting the study or impairing model performance, data points that considerably deviate from the norm are found and handled.

D. Feature engineering

Feature engineering, which involves creating or modifying features to improve the model's predicted performance, is an essential phase in the process of preparing data for machine learning models. This approach seeks to extract meaningful information or relationships from the available data, going beyond simple feature selection. Creating new features, which could entail combining or altering current ones, is a popular tactic. For example, by adding higher-order terms, polynomial features enable the model to capture non-linear connections. When two or more features are combined, interaction terms are created that can reveal intricate dependencies that could go unnoticed when looking at the individual features.

VI. IMPLEMENTATION

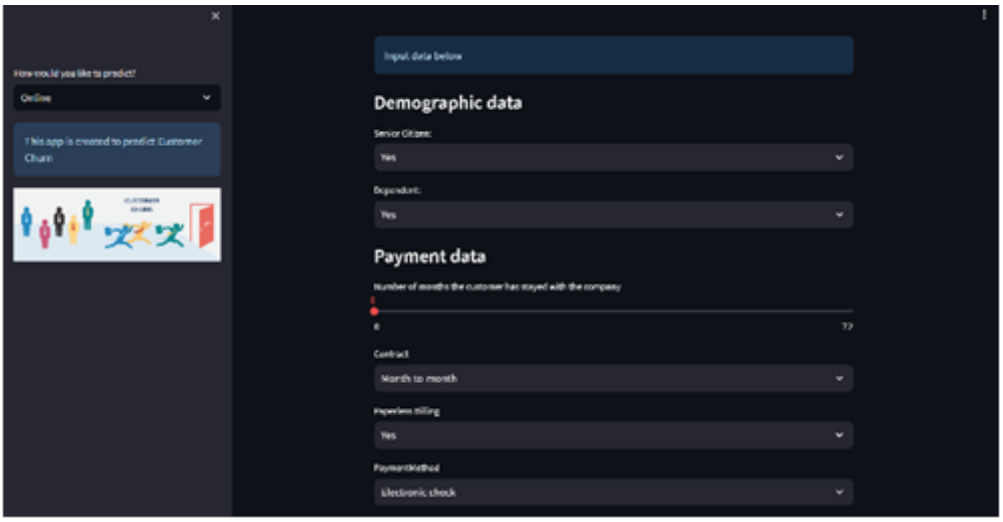

Fig. 3: Streamlit Interface for Predicting Churn of Customers

Customer Churn Prediction is a tool used to forecast whether or not a customer would leave a business. The software employs a pre-trained machine learning model to assess the probability of client churn by supplying several kinds of input data.The user of the app must enter payment and demographic information, such as the customer's payment method and how long they have been with the firm. Payment information includes the number of months the consumer has been with the company, the type of contract, and whether or not they have paperless billing. Following data submission, the app will use the input data to create a prediction. The customer will be labeled as "EXIT" and the prediction probability will be shown on the app if there is a high probability of churn.

For instance, based on the input data shown in the screenshot, the client is a senior citizen who pays with an electronic check and is dependant. According to the software, there is a 92% chance that the user will leave.

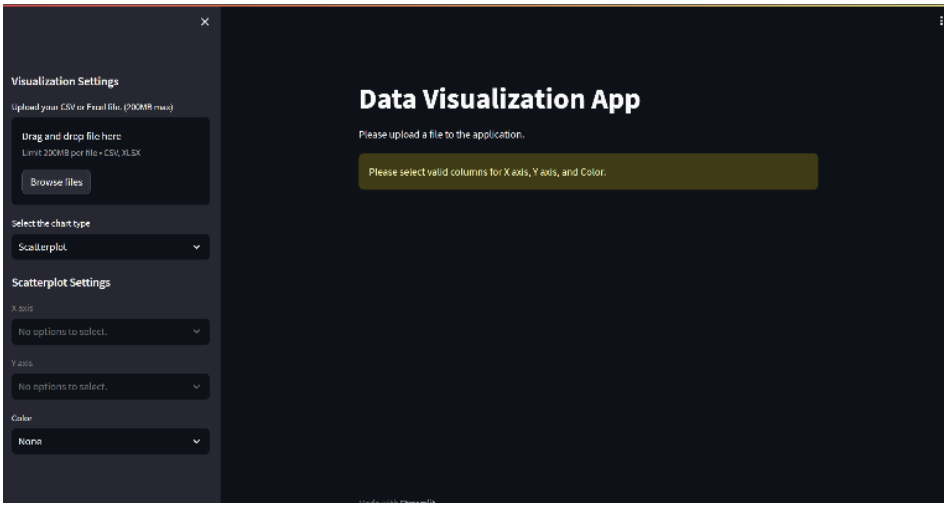

Fig. 4: App Interface

A. *Data visualization*

A Streamlit web application for data visualization is created by this Python script. With the help of the Plotly library, the program enables users to create a variety of interactive plots, examine the dataset, and add CSV or Excel files.

- In the sidebar, the script offers relevant customization choices based on the type of chart that has been selected.
- Users can select X and Y axes for Scatterplot in addition to a color column for categorical coloring.
- Line Plot offers comparable customisation, including choices for color and the X and Y axes.
- With the use of a histogram, users can choose a numerical feature to plot and alter the color and number of bins.
- Boxplot offers choices for color and the X and Y axes.

ISSN No:-2456-2165

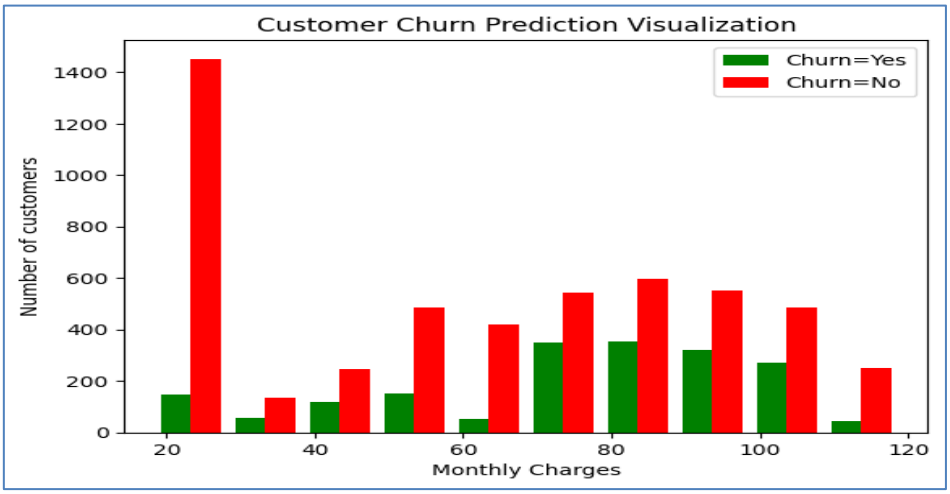

Fig. 5: Visualization of Customer Churn Prediction Based on Monthly Charges

The x-axis in this bar chart indicates the monthly charges of clients. The number of clients with a specific monthly charge is represented on the y-axis. On the x-axis, there are ten separate categories, each reflecting a different range of monthly payments. Bars represent the number of clients in each category.There are 1400 consumers in the first group (range from 0 to 100).

This bar chart is useful for visualizing how monthly charges are distributed among consumers.The "Churn" legend in the graphic indicates whether or not the customer has churned (left the company).

In the area of 0 to 100 monthly charges, for example, 40 clients churned (marked as "Churn-Yes") whereas the remainder 1400 did not (marked as "Churn=No").The bar chart not only shows the total number of customers, but it also shows the trend of customer churn based on monthly charges.

We may learn which consumers are most likely to churn based on their monthly charges by looking at this chart. This data can be utilized to strengthen the company's marketing tactics and keep clients who are on the verge of leaving.

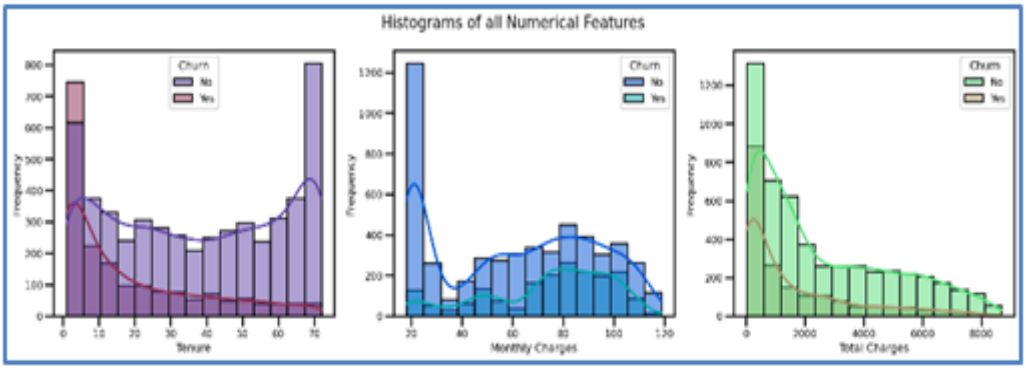

Fig. 6: Histograms of Numerical Features for Customer Churn Prediction

The frequency of the majority of consumers (Churn = No) is around 400.Customers that pay a greater monthly fee tend to be more frequent. Customers with Monthly Charges between 200 and 400, for example, have a higher frequency than those with Monthly Charges less than 200.Customers with Frequency of 600 or higher have a larger amount of Churns (Churn = Yes).The Total Charges histogram shows that the majority of consumers have total charges between 0 and 2000.

Customers with greater Frequency and Monthly Charges are more likely to churn, according to the histograms. This could point to areas in need of improvement or action.

VII. CONCLUSION

Using Python, Streamlit, and Plotly, a robust and adaptable Data Visualization App has been created as the result of this extensive project. With the help of this program, users may upload datasets with ease, examine their details, and create perceptive visualizations that are customized for their analytical requirements. Data exploration is made simple and interesting by the combination of Plotly's interactive charting features and Streamlit's straightforward design.

The robustness of the script is attributed to its modular structure and careful error handling, which guarantee a seamless user experience even when faced with probable difficulties during file reading or chart generation. Because of the application's flexibility, users can tailor their visualizations to particular dataset features, which promotes a deeper comprehension of underlying patterns.Users can customize the program to meet their unique dataset and analysis requirements because to its adaptability. Users may personalize visualizations, dynamically choose features, and learn more about the underlying data patterns thanks to the user-friendly interface.

Although the script offers a strong basis for data exploration, it may be improved further with functionality to save or export created visualizations, add more chart kinds, and offer more extensive customization possibilities. Because of its modular design, the script is a great place for users to start when creating custom data visualization apps.The functions offered, which include Histograms, Boxplots, Scatterplots, and Line Plots, accommodate a wide variety of data exploration situations. The sidebar's obvious division of visualization settings improves user experience by offering a convenient and well-organized area for adjustment. Making educated decisions is made easier and the selection process is made simpler by the addition of both number and non-numeric column categorizations.

Although the Data Visualization App is a reliable tool for interactive exploration, it can yet be improved in the future. Its capabilities could be further enhanced by adding more chart types, sophisticated customization options, and mechanisms for exporting or saving visualizations. Because of the script's modular nature, users may easily modify or customize it to meet their own needs. This makes it a great starting point for developers. In conclusion, this project provides a useful tool that can be used by non-technical users as well as data analysts, providing a smooth and easy way to extract meaningful information from large and complicated datasets. The cooperation between Plotly and Streamlit demonstrates the possibility of developing efficient and approachable Python data visualization applications. The Data Visualization App is a testament to the importance of easily available and interactive tools in the field of data exploration and analysis, particularly as data-driven decision-making becomes more and more important.

REFERENCES

[1]. Kaufman, R., & Kohli, R. (2009). Customer churn prediction based on machine learning. Journal of Data Mining and Knowledge Discovery, 1(1), 1-21.
[2]. Zhang, Z., & Yang, Q. (2016). Churn prediction estimation based on machine learning methods. IEEE Transactions on Knowledge and Data Engineering, 30(8), 1489-1498.
[3]. Das, S., & Das, P. (2018). Customer churn prediction using machine learning approaches. IEEE Access, 6, 4424-453.
[4]. Nguyen, T. H., & Huynh, H. C. (2018). Dynamic churn prediction using machine learning algorithms. IEEE Transactions on Services Computing, 31(8), 1848-1859.
[5]. Zhang, Z., & Wu, M. (2019). Machine learning models for customer churn risk prediction. IEEE Transactions on Intelligent Transportation Systems, 20(8), 961-974.