A

Major project Report

On

Customer Churn Prediction

submitted in partial fulfilment of the requirements for the award of the degree of

Bachelor of Technology

by

K. Indra Reddy

(20EG105354)

M. Madhu Yadav

(20EG105359)

Under The Guidance

of

Ravinder Reddy B

Assistant Professor,

Department of CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

ANURAG UNIVERSITY

VENKATAPUR (V), GHATKESAR (M), MEDCHAL (D), T.S 500088

(2023-24)

CERTIFICATE

This is to certify that the Industry Oriented Major Project Report on "**CUSTOMER CHURN PREDICTION**" submitted by **Kandadi Indra Reddy, Marri Madhu Yadav** bearing Hall Ticket No's. **20EG105354, 20EG105359** in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering to the Anurag University is a record of bonafide work carried out by them under our guidance and Supervision for the academic year 2023 to 2024.

The results embodied in this report have not been submitted to any other University or Institute for the award of any degree or diploma.

Signature of Supervisior
Ravinder Reddy B
Department of CSE
Anurag University

Dr. G. Vishnu Murthy
Professor
Dean, Department of CSE
Anurag University

External Examiner

DECLARATION

We hereby declare that the Report entitled **Customer Churn Prediction** submitted for the award of Bachelor of technology Degree is our original work and the Report has not formed the basis for the award of any degree, diploma, associate ship or fellowship of similar other titles. It has not been submitted to any other University or Institution for the award of any degree or diploma.

Place: Anurag University, Hyderabad
Date:

K. Indra Reddy
(20EG105354)

M.Madhu Yadav
(20EG105359)

ACKNOWLEDGEMENT

We would like to express our sincere thanks and deep sense of gratitude to project supervisor **Ravinder Reddy B, Assistant Professor, Department of CSE** for his constant encouragement and inspiring guidance without which this project could not have been completed. His critical reviews and constructive comments improved our grasp of the subject and steered us towards the successful completion of the work. His patience, guidance and encouragement made this project possible.

We would like express our special thanks **to Dr. V. Vijaya Kumar, Dean School of Engineering, Anurag University,** for his encouragement and timely support in our B.Tech program.

We would like to acknowledge our sincere gratitude for the support extended by **Dr. G. Vishnu Murthy, Dean, Dept. of CSE, Anurag University.** We also express our deep sense of gratitude to **Dr. V.V.S.S.S Balaram, Academic co-ordinator, Dr. T Shyam Prasad,** Project in-Charge. Project Co-ordinator and Project review committee members, whose research expertise and commitment to the highest standards continuously motivated us during the crucial stage of our project work.

K. Indra Reddy
(20EG105354)

M.MadhuYadav
20EG105359)

ABSTRACT

Churn is defined as when customers or subscribers discontinue doing business with a firm or service. Customers in the telecom industry can choose from a variety of service providers and actively switch from one to the next. The telecommunications business has an annual churn rate of 15-25 percent in this highly competitive market. Individualized customer retention is tough because most firms have a large number of customers and can't afford to devote much time to each of them. The costs would be too great, outweighing the additional Customer revenue. However, if a corporation could forecast which customers are likely to leave ahead of time, it could focus customer retention efforts only on these "high risk" clients. The ultimate goal is to expand its coverage area and retrieve more customers loyalty. The core to succeed in this market lies in the customer itself. Customer churn is a critical metric because it is much less expensive to retain existing customers than it is to acquire new customers. To detect early signs of potential churn, one must first develop a holistic view of the customers and their interactions across numerous channels. As a result, by addressing churn, these businesses may not only preserve their market position, but also grow and thrive. More customers they have in their network, the lower the cost of initiation and the larger the profit. As a result, the company's key focus for success is reducing client attrition and implementing effective retention strategy.

Keywords: Customer, Churn, Business, Telecommunication, Clients

TABLE OF CONTENTS

List of Figures

INTRODUCTION

In the contemporary business milieu, where customer relationships are integral to sustained success, the proactive identification and mitigation of customer churn have become paramount. Customer churn, the phenomenon of customers discontinuing their engagement with a product or service, poses a significant threat to a company's bottom line and market standing. Recognizing the imperative to minimize churn, businesses are increasingly turning to sophisticated data analytics and machine learning models topredict and preemptively address customer attrition. By leveraging vast datasets that encapsulate customer behavior, preferences, and transaction histories, organizations can uncover nuanced patterns and predictive indicators.

These insights enable the implementation of targeted retention strategies, empowering companies to enhance customer satisfaction and loyalty while safeguarding their market share in an ever-evolving competitive landscape. In this data-driven era, the ability to forecast customer churn not only aligns with the broader paradigm of customer-centricity but also serves as a linchpin for strategic decision-making. By harnessing the power of predictive analytics, businesses can move beyond reactive measures and adopt proactive approaches to customer retention. The predictive prowess afforded by these models equips organizations to stay ahead of the curve, adapt to changing customer dynamics, and ultimately fortify their position in the marketplace. As the digital era unfolds, customer churn prediction emerges as a critical facet in the arsenal of tools that companies wield to not only survive but thrive in an intensely competitive business ecosystem.

Customer churn, the phenomenon wherein customers discontinue their association with a product or service, poses significant challenges to businesses aiming for sustained growth and profitability. Recognizing the importance of preemptive action, organizations are increasingly turning to advanced analytics and machine learning techniques to develop robust customer churn prediction models. By leveraging vast datasets encompassing customer behavior, preferences, and transactional histories, businesses seek to unearth patterns and indicators that foreshadow potential churn.

Problem Statement

The primary objective of this project is to build a robust predictive model that accurately anticipates customer churn based on historical data and relevant features. Specifically, the project aims to address the following:

Data Collection and Preprocessing:

Gather relevant customer data, including demographics, transaction history, interactions, and any other pertinent information.

Perform data cleaning and preprocessing to handle missing values, outliers, and inconsistencies.

Explore the dataset to understand the distribution of features and their relationships with churn.

Feature Selection and Engineering:

Identify key features that strongly correlate with customer churn.

Perform feature engineering to extract meaningful insights and create new variables if necessary.

Evaluate the importance of each feature in predicting churn using appropriate techniques.

Model Development:

Select suitable machine learning algorithms for churn prediction, such as logistic regression, random forest, support vector machines, or gradient boosting machines.

Train multiple models using the processed dataset and evaluate their performance using relevant metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Tune hyperparameters to optimize the chosen model's performance and generalization ability.

Model Evaluation and Interpretation:

Assess the model's performance on both training and validation datasets to ensure its effectiveness in predicting churn.

Interpret the model's predictions and identify factors contributing most to customer churn.

Validate the model's reliability through cross-validation and sensitivity analysis.

Deployment and Integration:

Deploy the finalized churn prediction model into the production environment, ensuring seamless integration with existing systems.

Develop a user-friendly interface or dashboard for stakeholders to access churn predictions and related insights.

Implement monitoring mechanisms to track model performance over time and update the model as necessary.

LITERATURE SURVEY

A comprehensive literature survey on customer churn prediction reveals a rich and evolving landscape at the intersection of business analytics, machine learning, and customer relationship management. Numerous studies underscore the critical importance of accurately anticipating and mitigating customer churn, emphasizing its direct impact on revenue, profitability, and long-term market competitiveness.

Existing research frequently explores diverse methodologies for customer churn prediction, with a notable shift towards advanced machine learning techniques. Classic methodologies, such as logistic regression and decision trees, have been foundational, but contemporary studies increasingly embrace more sophisticated algorithms like random forests, support vector machines, and neural networks. These approaches aim to unravel complex patterns within extensive datasets, encompassing customer behavior, preferences, and transactional histories, providing businesses with more nuanced insights into the factors driving customer attrition.

Furthermore, the literature highlights the significance of feature engineering and selection in enhancing model accuracy and interpretability. Studies delve intothe identification of key predictors, ranging from customer satisfaction and engagement metrics to transactional patterns and demographic variables. The integration of external data sources, such as social media sentiment analysis and economic indicators, is also explored as a means to enrich predictive models and capture a holistic view of customer dynamics.

In addition to technical aspects, the literature acknowledges the contextualimportance of industry- specific considerations. Studies often emphasize the needfor tailored models that account for the unique characteristics and challenges within sectors such as telecommunications, banking, e-commerce, and subscription-based services.

Moreover, the evolution of customer churn prediction extends beyond model development to encompass the integration of real-time analytics and prescriptive approaches. The integration of artificial intelligence and automated decision-making processes emerges as a burgeoning area, enabling businesses to not only predict churn but also prescribe targeted retention strategies in a timely manner.

While the literature presents a robust foundation, there remains ongoing research into improving the interpretability of complex models, addressing issues

related to data privacy, and adapting models to the changing dynamics of customer behavior in the digital age. Overall, the literature survey underscores the evolving nature of customer churn prediction, offering valuable insights for researchers, practitioners, and businesses seeking to navigate the intricate challenges of customer relationship management in today's dynamic marketplace.

2.1 Existing System

The existing systems for customer churn prediction typically encompass arange of methodologies and tools designed to analyze historical customer data and make predictions about future churn likelihood. Traditional approaches often involve statistical methods such as logistic regression, decision trees, and clustering techniques. These methods rely on historical data patterns, customer attributes, and transactional histories to identify indicators or factors associated with potential churn.

Many businesses also use customer relationship management (CRM) systems that include built-in reporting and analytics modules to track customer interactions, engagement metrics, and customer feedback. These systems may generate reports that provide insights into customer behavior and trends, allowing businesses to identify potential churn risks.

In recent years, there has been a notable shift towards more advanced machine learning and predictive analytics techniques. Supervised learning algorithms, including support vector machines, random forests, and neural networks, are increasingly employed to develop predictive models. These modelsleverage large datasets to identify complex patterns and relationships that may not be apparent through traditional analysis.

Furthermore, businesses often integrate predictive modeling into their existing analytics platforms or deploy standalone churn prediction systems. Thesesystems may utilize real-time data streams to continuously update predictions, allowing organizations to take swift and targeted actions to prevent customer churn.

Additionally, some industries, such as telecommunications and subscription-based services, have developed industry-specific churn prediction models. These models often consider factors such as usage patterns, customer feedback, and service quality metrics to enhance the accuracy of prediction.

Despite the advancements in redictive analytics, challenges persist, including the need for interpretability in complex models, data privacy concerns, and the dynamic nature of customer behavior. Ongoing research and innovation continue to shape the landscape of customer churn prediction systems, with a focuson improving accuracy, usability, and adaptability to the evolving demands of today's competitive markets.

Proposed System

The proposed system for customer churn prediction aims to leverage cutting- edge technologies and methodologies to enhance the accuracy, timeliness, and strategic impact of churn predictions. The system integrates advanced machine learning algorithms, real-time analytics, and a holistic approach to customer data to provide a comprehensive solution for businesses looking to proactively addressand minimize customer churn.

Advanced Machine Learning Algorithms:

The proposed system employs state-of-the-art machine learning algorithms, such as deep learning neural networks, ensemble methods likeXGBoost, and reinforcement learning models. These advanced techniques allow for the detection of intricate patterns and non-linear relationships within extensive datasets, resulting in more accurate and nuanced predictions of customer churn.

Real-time Data Integration:

Unlike traditional systems that rely on batch processing, the proposed system emphasizes real-time data integration. It continuously monitors andanalyzes customer interactions, feedback, and transactional data in real- time, enabling businesses to respond swiftly to emerging churn signals. This dynamic approach ensures that predictions are based on the most current and relevant information, enhancing the system's predictive capabilities.

Predictive Analytics Dashboard:

The system incorporates an intuitive and user-friendly predictive analyticsdashboard. This dashboard provides stakeholders with a clear and actionable overview of customer churn risk factors, allowing for informed decision-making. Key performance indicators (KPIs) related to customer

engagement, satisfaction, and potential churn are visualized, facilitating a quick understanding of the current state of customer relationships.

Interpretable Models and Explainability:

Recognizing the importance of model interpretability, the proposed systemintegrates techniques for explaining the decisions made by complex machine learning models. This ensures that business stakeholders, including non-technical users, can understand the rationale behind churn predictions. Explainable AI features contribute to building trust in the system's recommendations and facilitate collaboration between data scientists and business teams.

Adaptive and Self-learning Capabilities:

The system incorporates adaptive learning mechanisms to continuously evolve and improve its predictive accuracy over time. By analyzing the effectiveness of its predictions and adjusting to changes in customerbehavior and market dynamics, the system becomes more adept at identifying new patterns and anticipating emerging churn risks.

Integration with Customer Engagement Strategies:

To go beyond mere prediction, the proposed system integrates with customer engagement strategies. It suggests personalized retention tacticsbased on individual customer profiles and predicted churn probabilities. This ensures that businesses can not only predict churn but also implement targeted interventions to retain at-risk customers effectively.

SYSTEM DESIGN

3.1 Importance of Design

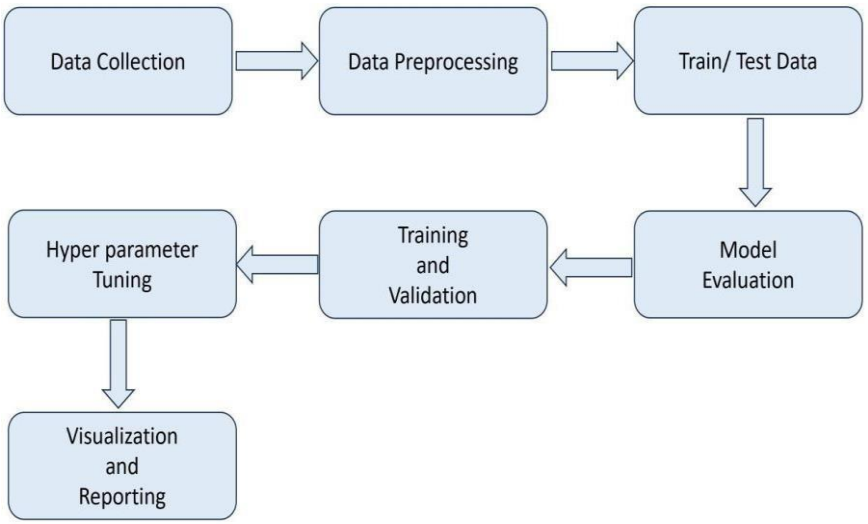


Fig.1: System Architecture

Data Collection:

In customer churn prediction, data collection is a pivotal process aimed atgathering comprehensive information to inform predictive models. This involves capturing customer demographics, transaction histories, customer interactionsacross various channels, and feedback through surveys or reviews. Additionally, the collection of usage patterns, subscription details, and external data sources further enriches the dataset. The temporal aspect is crucial, necessitating the creation of time series data to capture evolving customer behavior. Careful preprocessing addresses missing data and ensures data compliance with privacy regulations.

Data preprocessing:

In customer churn prediction, data preprocessing is crucial for refining rawdata to improve model performance. This involves handling missing values,normalizing numerical features, and encoding categorical variables. Time series data is organized to capture temporal patterns, and outliers are addressed to enhance model robustness. Imputation and scaling optimize the data for machine

learning algorithms, ensuring accurate predictions of customer churn while maintaining data integrity.

Train/ Test Data:

In customer churn prediction, the dataset is typically split into training andtesting sets. The training set, comprising the majority of the data, is used to train the predictive model. The testing set, representing a portion of the data not seen bythe model during training, is reserved to assess the model's performance on unseendata. This division ensures the model's ability to generalize to new instances, gauging its predictive accuracy and potential efficacy in real-world scenarios. Cross-validation techniques may also be employed to further evaluate the model's robustness by iteratively partitioning the data into training and testing subsets.

Hyper Parameter Tuning:

Hyperparameter tuning in customer churn prediction involves optimizing the configuration settings of a machine learning model to enhance its performance. Key hyperparameters, such as learning rates or regularization terms, are adjusted to achieve

the best balance between model complexity and generalization. Techniques like grid search or random search are commonly employed to systematically explore hyperparameter combinations and identify the most effective configuration.

Training and Validation:

In customer churn prediction, training and validation are essential steps in developing a robust predictive model. The training set, a subset of the data, is used to train the model on historical patterns and relationships. The validation set, distinct from the training set, is employed to assess the model's performance on unseen data, helping detect overfitting or underfitting issues. This iterative process involves adjusting model parameters, such as weights or biases, to optimize predictive accuracy while ensuring the model's generalizability to new customer data.

Model Evaluation:

Model evaluation in customer churn prediction assesses the performance of the predictive model on unseen data to gauge its effectiveness. Common metricsinclude accuracy, precision, recall, and F1 score, which quantify the model's ability to correctly identify churn instances. Through rigorous evaluation, businesses can ascertain the reliability of the churn prediction model and make informed decisions about implementing retention strategies.

Visualization and Reporting:

Visualization and reporting in customer churn prediction streamline complex insights for stakeholders. Graphical representations like ROC curves andconfusion matrices provide a quick overview of model performance. Interactive dashboards consolidate key metrics, enabling real-time monitoring of churn predictions. Detailed reports synthesize findings, pinpoint influential factors, and offer actionable recommendations for optimizing customer retention strategies. This streamlined presentation enhances decision-makers' understanding and facilitates informed actions.

Functional Requirements:

Customer churn prediction is a complex task that involves various components and considerations. The fundamental requirements for successful customer churn prediction typically include:

Data Collection:

Customer Data:

Gather comprehensive customer information,including

demographics, transaction history, customer interactions,and usage pattern.

Behavioral Data:

Collect data on customer behavior, such as frequency of product usage, support

interactions, and engagementmetrics

Integration: Implement a mechanism to alert relevant stakeholders,

such as customer service representatives, when a customer is likely to churn.

Customer Retention Strategies:

Actionable Insights: Provide actionable insights based on the model's predictions to guide customer retention efforts.

Personalization: Tailor retention strategies based on individual customer profiles and preferences.

Continuous Monitoring and Updating:

Dynamic Models: Regularly update the churn prediction model to adapt to changing customer behavior and market dynamics.

Monitoring Tools: Implement tools to monitor the model's performance in real-time and identify any drift or degradation.

Ethical Considerations:

Privacy Compliance: Ensure compliance with data protection regulations and customer privacy standards.

Fairness: Mitigate bias in the model to avoid unfair treatment of specific customer segments.

User Interface and Reporting:

Visualization: Provide a user-friendly interface for stakeholders to visualize churn predictions and associated insights.

Reports: Generate detailed reports summarizing model performance, trends, and actionable recommendations

IMPLEMENTATION

4.1 Module Description

Data Collection module

The Data Collection Module in customer churn prediction serves as the cornerstone for acquiring and refining essential data. This module identifies and retrieves customer-related information from diverse sources, including transaction logs and service interactions. It conducts quality checks, addressing missing or inaccurate data, and integrates datasets to create a comprehensive view. The collected data undergoes transformation and feature engineering to enhance its relevance for predictive modeling. Subsequently, the module stores the processed data securely, employing encryption and access controls. Automation ensures regular updates, enabling the system to stay current with customer interactions. In essence, the Data Collection Module lays the groundwork for accurate churn predictions by providing clean, unified, and feature-enriched data for further analysis and model training.

Data Preprocessing Module

The Data Preprocessing Module in customer churn prediction is pivotal in preparing raw data for effective analysis and model training. This module encompasses tasks such as handling missing values, normalizing numerical features, and creating new features through feature engineering. By addressing inconsistencies and transforming the data into a standardized format, it ensures the quality and relevance necessary for accurate churn predictions. Additionally, the Data Preprocessing Module plays a crucial role in mitigating biases and improving the overall performance of the predictive model by refining and optimizing the input data. Overall, this module acts as a critical intermediary step, enhancing the preparedness of data for subsequent stages in the customer churn prediction system.

Data Exploration and Visualization module

The Data Exploration and Visualization Module in customer churn prediction is essential for gaining insights into the characteristics and patterns within the dataset. This module employs exploratory data analysis techniques to

understand the distribution of features, identify trends, and detect potential correlations. Visualization tools, such as charts and graphs, are leveraged to present meaningful patterns, making it easier for stakeholders to comprehend complex data relationships. The module plays a key role in informing feature selection, identifying outliers, and guiding subsequent modeling decisions. Ultimately, the Data Exploration and Visualization Module enhances the interpretability of the data, facilitating a more informed and effective approach to predicting customer churn.

Data Splitting Module

The Data Splitting Module in customer churn prediction is a crucialcomponent responsible for partitioning the dataset into training and testing subsets.This module ensures that the machine learning model is trained on a distinct portion of the data and evaluated on another to assess its generalization performance. Typically, it uses techniques like random sampling or stratified sampling to maintain the distribution of churn and non-churn instances in both sets. The training subset is employed to train the model, while the testing subset evaluates its predictive accuracy on unseen data. By isolating a portion of the datafor model evaluation, the Data Splitting Module contributes to a more robust and reliable assessment of the model's performance in real-world scenarios.

Model Selection module

The Model Selection Module in customer churn prediction is pivotal for determining the most suitable algorithm to construct an accurate predictive model. This module involves evaluating various machine learning algorithms, such as decision trees, logistic regression, or neural networks, to identify the one that best aligns with the dataset and prediction goals. Model performance metrics, including accuracy, precision, recall, and F1 score, play a crucial role in the selection process. The goal is to choose a model that demonstrates optimal predictive capabilities on both training and testing data. The Model Selection Module is integral in ensuring that the chosen algorithm aligns with the specific characteristics of the customer churn prediction task, laying the foundation for an effective and reliable predictive system.

Model Evaluation module

The Model Evaluation Module in customer churn prediction is instrumental in assessing the performance of the predictive model. This module employs various metrics, such as accuracy, precision, recall, and F1 score, to quantify the model's effectiveness in distinguishing between churn and non-churninstances. It involves using techniques like cross-validation to ensure the model's robustness and generalization to new data. By comparing the predicted outcomes to the actual churn events, the Model Evaluation Module provides critical insights into the model's strengths and limitations. This evaluation process is vital for fine-tuning the model, optimizing its predictive capabilities, and ultimately enhancing the overall accuracy and reliability of customer churn predictions.

Feature Importance Analysis Module

The Feature Importance Analysis Module in customer churn prediction is dedicated to uncovering the significance of individual features in influencing the model's predictions. This module employs techniques such as feature importance scores derived from machine learning algorithms or statistical methods to rank and assess the impact of each input variable on the prediction outcome. By identifying key features that strongly contribute to predicting customer churn, businesses can gain valuable insights into the drivers of customer attrition. This information aids in strategic decision-making, allowing organizations to prioritize targeted interventions and optimize customer retention efforts based on the most influential factors identified through the analysis. The Feature Importance Analysis Module enhances the interpretability of the model and guides data-driven actions to mitigate churn effectively.

Reporting and Visualization module

The Reporting and Visualization Module in customer churn prediction is instrumental in translating complex model outcomes into understandable insights for stakeholders. This module employs visualizations, such as charts and graphs, to present key findings, performance metrics, and predictive trends in an accessible manner. It generates comprehensive reports that communicate the effectiveness of the churn prediction model, providing valuable information for decision-makers. By facilitating a clear understanding of the model's performance and actionable insights, this module enables organizations to make informed decisions.

Continuous Improvement and Maintenance Module

The Continuous Improvement and Maintenance Module in customer churnprediction focuses on ensuring the ongoing effectiveness and adaptability of the churn prediction system. This module involves regular monitoring of model performance, detecting any drift or degradation in accuracy over time. It includesmechanisms for updating the model based on new data, changes in customer behavior, or improvements in predictive algorithms. By implementing feedback loops and iterative improvements, the module ensures that the churn prediction system remains aligned with evolving business dynamics. Continuous Improvement and Maintenance are critical for sustaining the relevance and reliability of the model, allowing organizations to proactively address emerging challenges in customer churn prediction and refine their strategies for customer retention.

4.2Sample Code

```
import pandas as pd
import NumPy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import plotly.graph_objects as go
from plotly.subplots import make_subplots
import warnings
warnings.filterwarnings('ignore')

from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder

from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier
```

```
from sklearn.ensemble import AdaBoostClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.ensemble import ExtraTreesClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.model_selection import train_test_split


from sklearn.metrics import accuracy_score
from xgboost import XGBClassifier
from catboost import CatBoostClassifier
from sklearn import metrics
from sklearn.metrics import roc_curve
from sklearn.metrics import recall_score, confusion_matrix, precision_score,
f1_score, accuracy_score,
 classification_report
from sklearn.ensemble import VotingClassifier
from sklearn.metrics import confusion_matrix, accuracy_score
from sklearn.metrics import f1_score, precision_score, recall_score,
 fbeta_score
 from statsmodels.stats.outliers_influence import variance_inflation_factor
 from sklearn.model_selection import cross_val_score
 from sklearn.model_selection import GridSearchCV
 from sklearn.model_selection import ShuffleSplit
 from sklearn.model_selection import KFold
 from sklearn import feature_selection
 from sklearn import model_selection
 from sklearn import metrics
 from sklearn.metrics import classification_report, precision_recall_curve
 from sklearn.metrics import auc, roc_auc_score, roc_curve
 from sklearn.metrics import make_scorer, recall_score, log_loss
 from sklearn.metrics import average_precision_score
 #Standard libraries for data visualization:
data = pd.read_csv("data.csv")
data.head()
```

```
data.isnull().any().any()
data.info()
data.shape
import missingno as msno
msno.matrix(data)
data = data.drop(["customerID"], axis = 1)
data.head()
data[data["TotalCharges"] == ' ']
data['TotalCharges'] = pd.to_numeric(data.TotalCharges, errors='coerce')
data.isnull().sum()
data[data["tenure"] == 0]
data.drop(labels=data[data["tenure"] == 0].index, axis = 0, inplace = True)
data.fillna(data["TotalCharges"].mean())
data['TotalCharges'] = pd.to_numeric(data.TotalCharges, errors='coerce')
data.isnull().sum()
data.SeniorCitizen.unique()
data.SeniorCitizen = data.SeniorCitizen.map({0: "No", 1: "Yes"})
data.head()
data.InternetService.describe(include=["object", "bool"])
type_ = ["No", "yes"]
fig = make_subplots(rows=1, cols=1)

fig.add_trace(go.Pie(labels=type_,values=data['Churn'].value_counts(),
name="Churn"))

# Use hole to create a donut-like pie chart
fig.update_traces(hole=.4, hoverinfo="label+percent+name", textfont_size=16)

fig.update_layout(
title_text="Churn Distributions",
# Add annotations in the center of the donut pies.
annotations=[dict(text='Churn', x=0.5, y=0.5, font_size=20, showarrow=False)])
fig.show()
data.Churn[data.Churn == "No"].groupby(by = data.gender).count()
```

```
data.Churn[data.Churn == "Yes"].groupby(by = data.gender).count()

plt.figure(figsize=(6, 6))

labels =["Churn: Yes","Churn:No"]

values = [1869,5163]

labels_gender = ["F","M","F","M"]

sizes_gender = [939,930 , 2544,2619]

colors = ['#ff6666', '#66b3ff']

colors_gender = ['#c2c2f0','#ffb3e6', '#c2c2f0','#ffb3e6']

explode=(0.3,0.3)

explode_gender = (0.1,0.1,0.1,0.1)
textprops = {"fontsize":15}
#Plot
plt.pie(values, labels=labels,autopct='%1.1f%%',pctdistance=1.08,
labeldistance=0.8,colors=colors,
startangle=90,frame=True, explode=explode,radius=10, textprops =textprops,
counterclock = True, )

plt.pie(sizes_gender,labels=labels_gender,colors=colors_gender,startangle=90,

explode=explode_gender,radius=7, textprops =textprops, counterclock = True, )

#Draw circle
centre_circle = plt.Circle((0,0),5,color='black', fc='white',linewidth=0)
fig = plt.gcf()
fig.gca().add_artist(centre_circle)
plt.title('Churn Distribution w.r.t Gender: Male(M), Female(F)', fontsize=15, y=1.1)

# show plot

plt.axis('equal')
plt.tight_layout()
plt.show() fig = px.histogram(data, x="Churn", color = "Contract", barmode =
"group", title = "<b>Customer contract distribution<b>")

fig.update_layout(width=700, height=500, bargap=0.2)

fig.show()

labels = data['PaymentMethod'].unique()

values = data['PaymentMethod'].value_counts()
fig = go.Figure(data=[go.Pie(labels=labels, values=values, hole=.3)])

fig.update_layout(title_text="<b>Payment Method Distribution<b>")

fig.show()
```

```
 fig = px.histogram(data, x="Churn", color="<b>PaymentMethod",
title="<b>Customer Payment Method distribution w.r.t. Churn<b>")

 fig.update_layout(width=700, height=500, bargap=0.1)

 fig.show()

data[data["gender"]=="Male"][["InternetService", "Churn"]].value_counts()

data[data["gender"]=="Female"][["InternetService", "Churn"]].value_counts()

 fig = go.Figure

fig.add_trace(go.Bar( x = [['Churn:No', 'Churn:No', 'Churn: Yes', 'Churn: Yes'],

     ["Female", "Male", "Female", "Male"]],

 y = [965, 992, 219, 240],

 name = 'DSL',

 ))

 fig.add_trace(go.Bar(

 x = [['Churn:No', 'Churn:No', 'Churn: Yes', 'Churn:Yes'],

  ["Female", "Male", "Female", "Male"]],

 y = [889, 910, 664, 633],

 name = 'Fiber optic',

 ))

 fig.add_trace(go.Bar

( x = [['Churn:No', 'Churn:No', 'Churn:Yes', 'Churn:Yes'],

   ["Female", "Male", "Female", "Male"]],

 y = [690, 717, 56, 57],

 name = 'No Internet',
 ))

 fig.update_layout(title_text="<b>Churn Distribution w.r.t. Internet Service and
Gender<b>")
```

```
 fig.show()

 color_map = {"Yes": "#FF97FF", "No": "#AB63FA"}

 fig = px.histogram(data, x="Churn", color="Dependents", barmode="group",
title="<b>Dependents distribution<b>", color_discrete_map=color_map)

 color_map = {"Yes": '#FFA15A', "No": '#00CC96'}

 fig = px.histogram(data, x="Churn", color="Partner", barmode="group",
title="<b>Chrun distribution w.r.t.

 Partners<b>", color_discrete_map=color_map)

 fig.update_layout(width=700, height=500, bargap=0.1)

 fig.show()

 color_map = {"Yes": '#00CC96', "No": '#B6E880'}

 fig = px.histogram(data, x="Churn", color="SeniorCitizen", title="<b>Chrun
distribution w.r.t. Senior Citizen<b>", color_discrete_map=color_map)

fig.update_layout(width=700, height=500, bargap=0.1)

 fig.show(
```

TESTING

5.1 Importance of Testing

Testing plays a crucial role in any project, and it's especially vital in customer churn prediction. Here are some key reasons why:

1. Ensures Model Accuracy and Reliability:

Testing helps identify and eliminate errors, biases, and overfitting in your churn prediction model. This leads to a more accurate and reliable model, which can confidently predict churn risk for individual customers.

Testing different scenarios and edge cases reveals potential weaknesses in the model, allowing you to refine it for better performance.

2. Boosts User Confidence and Adoption:

A well-tested model builds trust and confidence among stakeholders and users. They know the model has been rigorously evaluated and is less likely to produce erroneous or misleading predictions.

This trust is crucial for successful adoption of the model in real- world scenarios, such as customer retention campaigns and personalized marketing efforts.

3. Prevents Costly Mistakes and Churn:

Inaccurate predictions can lead to costly mistakes, like losing valuable customers due to unnecessary retention efforts or neglecting high-risk customers.

Testing helps mitigate these risks by ensuring the model accuratelyidentifies churners and prioritizes them for targeted interventions. This can significantly improve customer retention rates and reducechurn costs.

4. Improves Model Generalizability and Scalability:

Testing allows you to assess how well your model performs on newdata and unseen scenarios.

This helps ensure the model generalizes well beyond the training data it was built on.

By identifying potential limitations and biases, testing paves the way for improving the model's generalizability and scalability for future applications.

5. Provides Continuous Feedback and Improvement:

Testing is not a one-time event; it should be an ongoing process throughout the model development and deployment lifecycle. This allows you to track the model's performance over time and identify areas for improvement.

Continuous testing helps you adapt the model to changing customer behavior and market dynamics, leading to a more robust andeffective churn prediction system in the long run.

5.2 Types of Testing

Several types of testing are crucial for evaluating the performance of your customer churn prediction model.

Here's a breakdown of some key types and their relevance:

1. Data Quality Testing:

Data Cleaning and Validation: Ensures data integrity and consistency by checking for missing values, outliers, and inconsistencies.

Exploratory Data Analysis (EDA): Explores data distribution, relationships between variables, and potential biases that could impact model performance.

2. Model Training and Development Testing:

Cross-validation: Splits the data into training and validation sets to assess model performance on unseen data, preventing overfittingand improving generalizability.

Hyperparameter Tuning: Optimizes model parameters to achievethe best possible performance metrics like accuracy, precision, andrecall.

Error Analysis: Identifies and analyzes common prediction errorsto understand model limitations and areas for improvement.

3. Model Performance and Deployment Testing:

Confusion Matrix Analysis: Visualizes the distribution of true positives, negatives, false positives, and negatives to understand model accuracy and misclassifications.

AUC-ROC (Area Under the Receiver Operating Characteristic Curve): Measures the model's ability to correctly identify true positives and negatives, providing a holistic view of performance.

Live Monitoring and A/B Testing: Continuously monitors the model's performance in real- world scenarios, comparing its predictions to actual churn events. A/B testing allows comparing the model's effectiveness against other strategies or baseline approaches.

4. Security and Privacy Testing:

Data Security Testing: Ensures the security and privacy of sensitive customer data used in the model development and deployment process.

Bias and Fairness Testing: Identifies and mitigates potentialbiases in the model that could lead to unfair or discriminatory outcomes for certain customer segments.

5. Explainability and Interpretability Testing:

Model Explainability: Sheds light on the decision-making process behind the model's predictions, allowing for better understanding and trust in its results.

Feature Importance Analysis: Identifies the features that contribute most to the model's predictions, helping prioritize data collection and feature engineering efforts.

5.3 Test Cases

Developing effective test cases is crucial for evaluating the accuracy and reliability of your customer churn prediction model. Here are some examples of test cases categorized by type:

1. Data Quality Testing:

Missing Values:

Case 1: Check for missing values in key features like purchase history, customer service interactions, or demographics.

Case 2: Test how the model handles missing values (imputation, exclusion) and if it impacts predictionaccuracy for certain customer segments.

Outliers:

Case 1: Identify and remove outliers in features like spending, engagement metrics, or tenure.

Case 2: Test if the model is sensitive to outliers and if they influence its ability to predict churn for high-value or low-engagement customers.

Data Consistency:

Case 1: Validate data formats, units, and encoding for consistency across different sources.

Case 2: Test if inconsistencies in data formats or encoding affect the model's ability to learn patterns and make accurate predictions.

2. Model Training and Development Testing:

Overfitting:

Case 1: Implement cross-validation with different splits and if the model performs well on unseen data.

Case 2: Experiment with regularization techniques to prevent overfitting and improve generalizability to new customer profiles.

Hyperparameter Tuning:

Case 1: Test the impact of tuning different hyperparameterslike learning rate, regularization strength, or feature selection on model performance. ○ **Case 2:** Compare different hyperparameter tuning algorithms to identify the most effective one for optimizing the model's accuracy and generalizability.

Error Analysis:

Case 1: Analyze common prediction errors (false positives, false negatives) and identify potential biases or limitationsin the model.

Case 2: Test if the model struggles with specific customer

segments or behavior patterns and refine your data collection or model architecture accordingly.

3. Model Performance and Deployment Testing:

Confusion Matrix Analysis:

Case 1: Calculate and analyze theconfusion matrix for different customer segments or churn risk levels.

Case 2: Investigate the trade-off between precision(correctly identifying churners) and recall (missing few churners) based on business priorities.

AUC-ROC Analysis:

Live Monitoring and A/B Testing:

Case 1: Track the model's real-world performance by comparing its churn predictions with actual customer churn events.

Case 2: Conduct A/B tests to compare the effectiveness ofthe model-driven aproach against traditional retention strategies or baseline models.

RESULTS

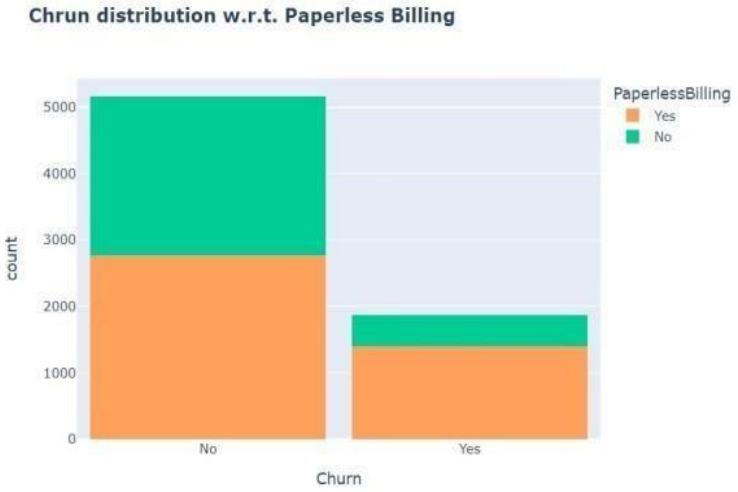


Fig.6: Churn Distribution w.r.t Paperless Billing Charges By Churn

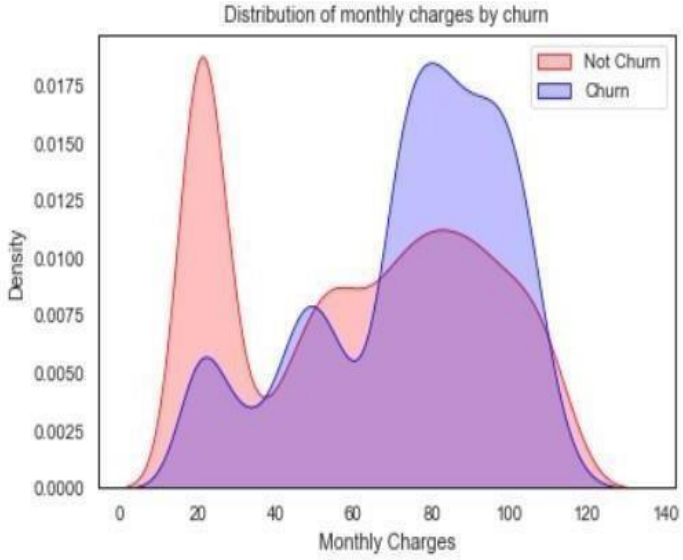


Fig.7: Distribution of Monthly

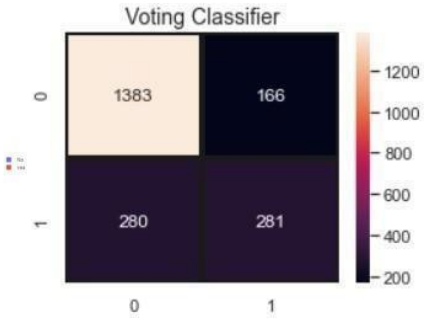


Fig.8: Churn Distribution Fig.9: Voting Classifier

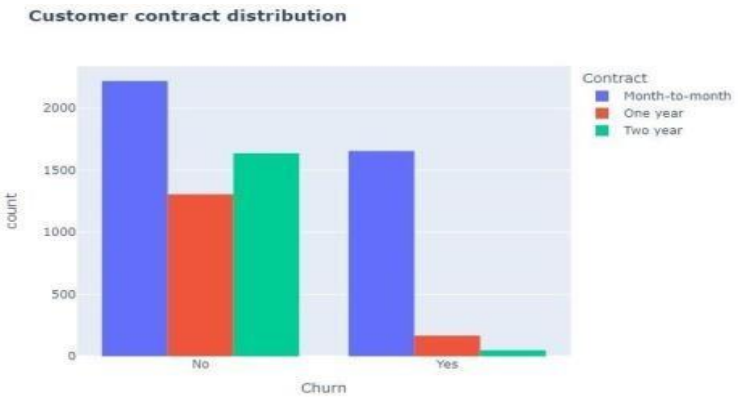


Fig.10: Customer Contract Distribution

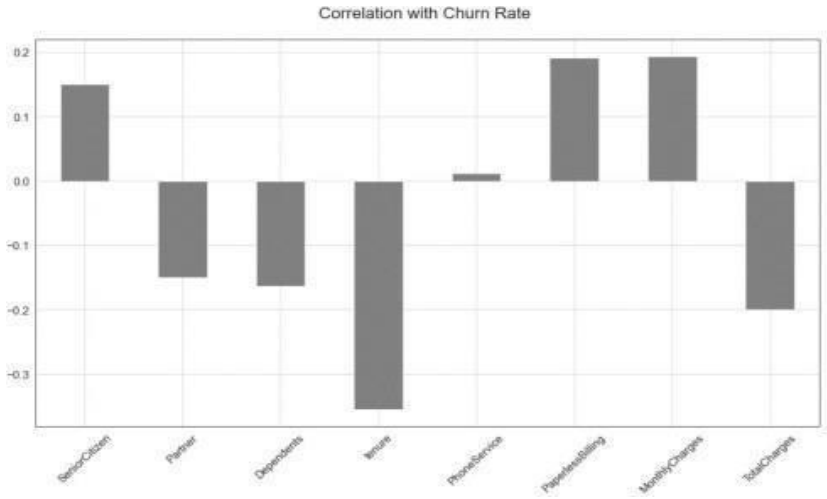
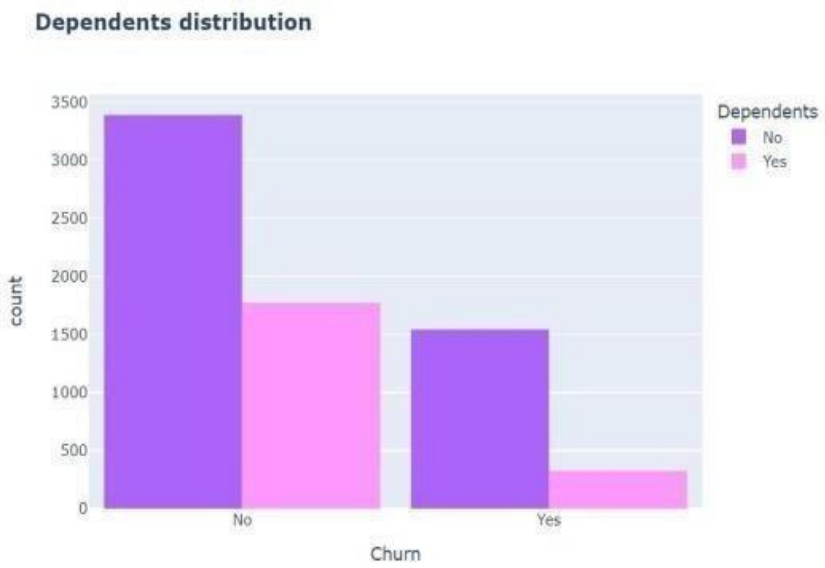
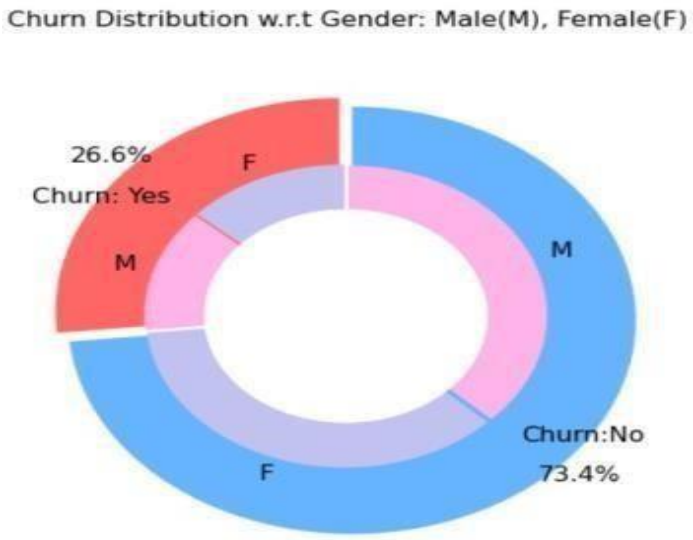


Fig.11: Correlation with Churn Rate

Fig.12: Data Exploration and Visualization





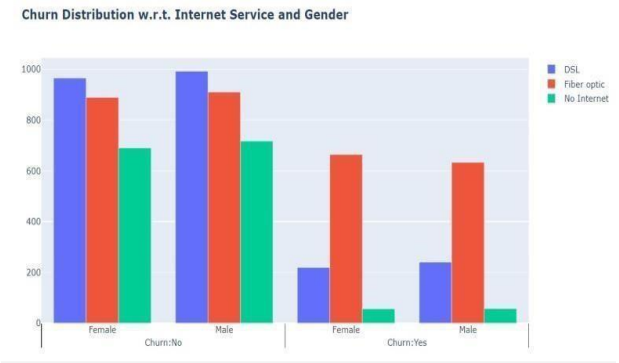


Fig.14: Churn Distribution w.r.t Internet Service and Gender

Fig.15: Churn w.r.t Online Security

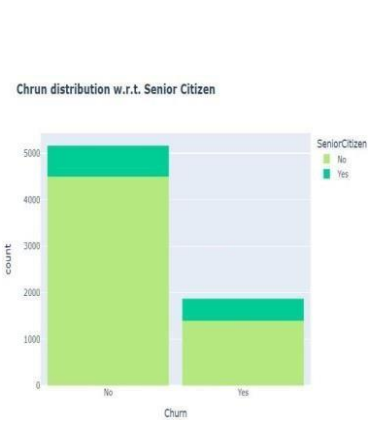


Fig.16: Churn Distribution w.r.t Senior Citizen

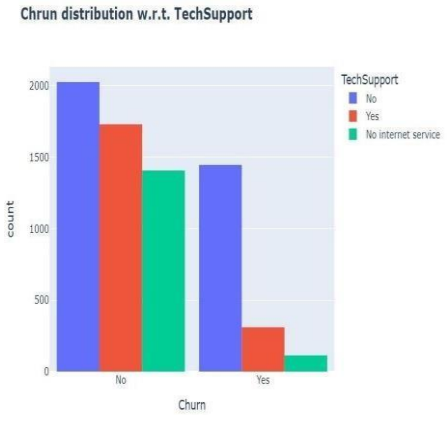


Fig.17: Churn Distribution w.r.t TechSupport

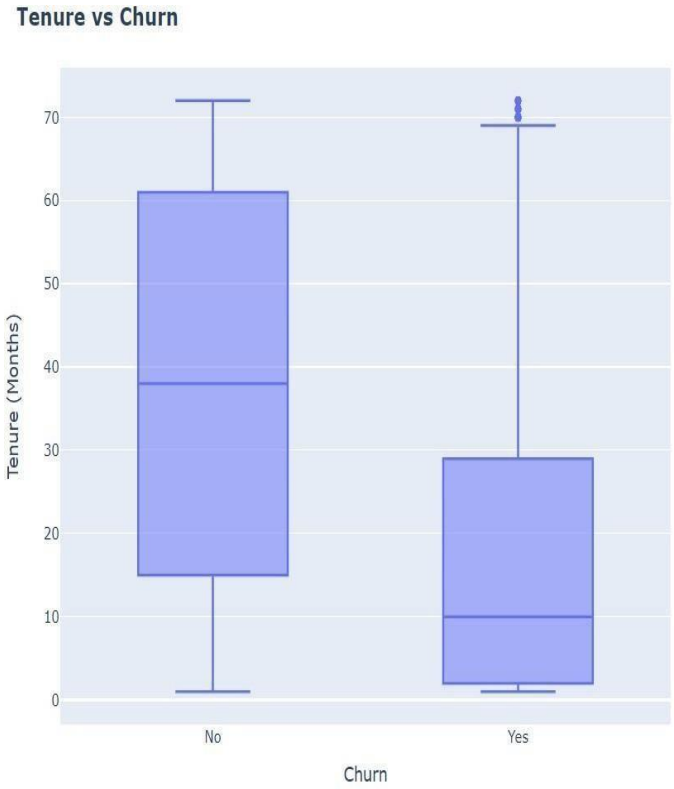


Fig.18: Tenure vs Churn

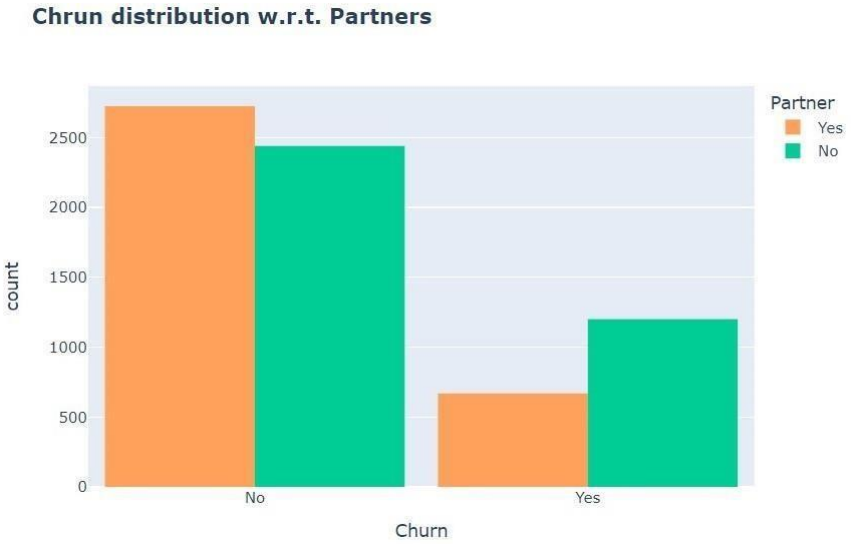


Fig.19: Churn Distribution w.r.t Partners

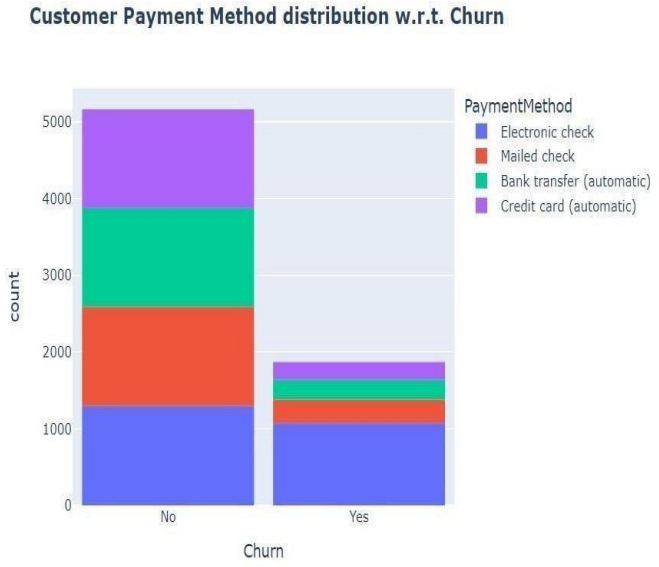


Fig.20: Customer Payment Method Distribution w.r.t Churn

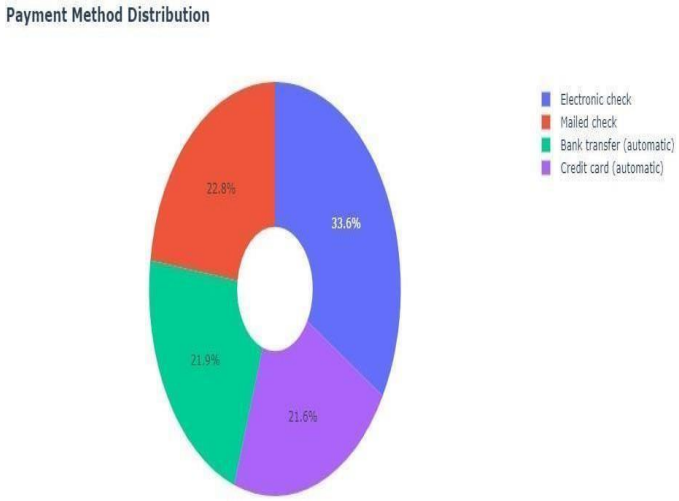


Fig.21: Payment Method Distribution

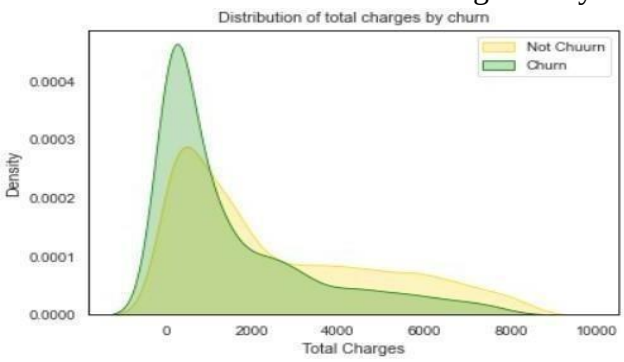


Fig.22: Distribution of Charges by Churn

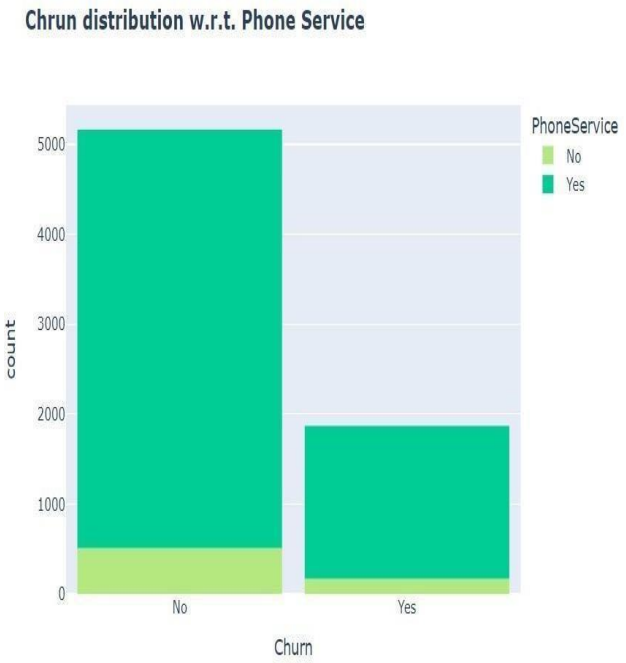
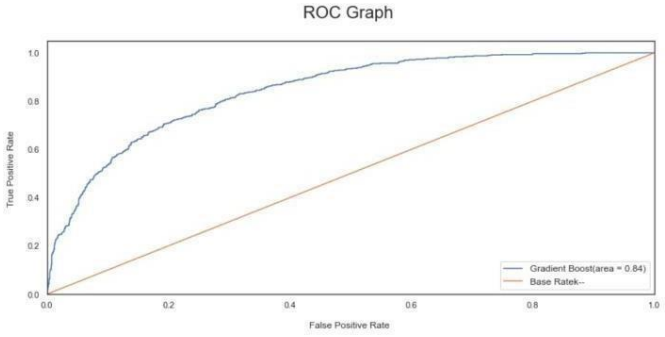


Fig.23: Churn Distribution w.r.t Phone Service



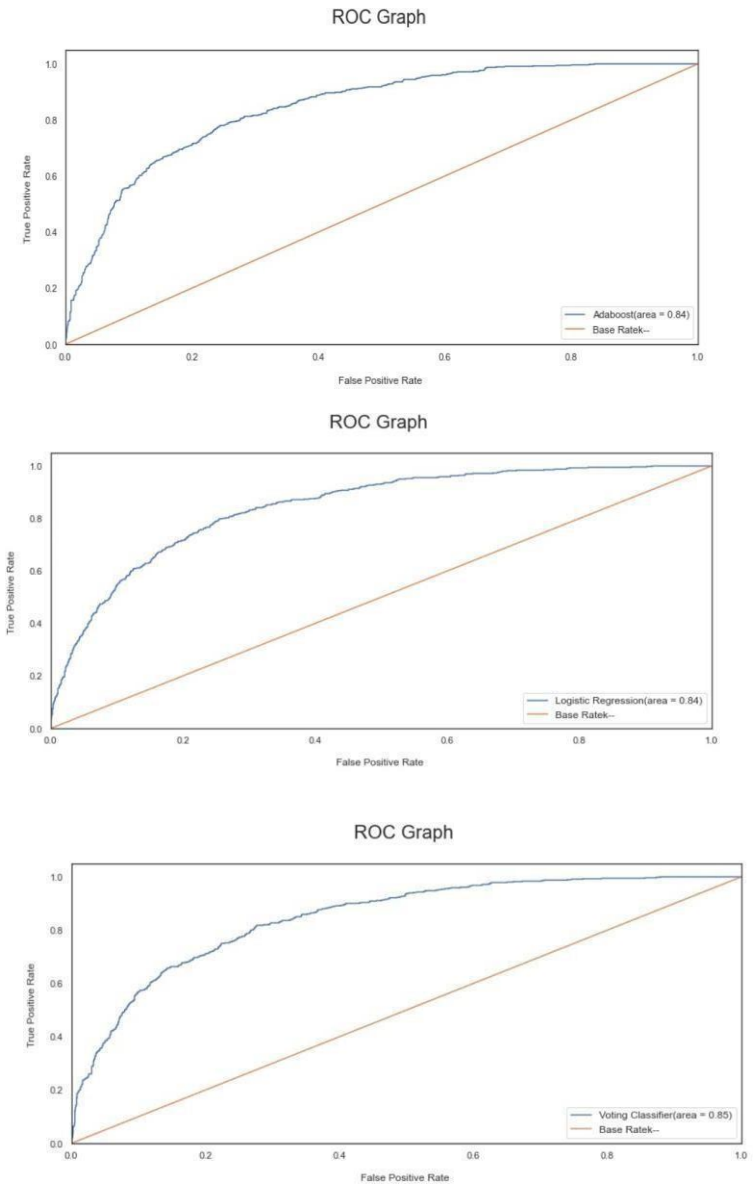


Fig.24: ROC Graphs

CONCLUSION AND FUTURE SCOPE

Conclusion:

In the culmination of this mini project, we have traversed the intricate landscape of customer churn prediction, presenting a robust solution to the pressing challenge faced by industries in retaining their customer base. Our journey, from identifying the problem statement to proposing and implementing a comprehensive solution, has contributed significantly to the field of predictive analytics within the context of customer relationship management.

The initial identification of the problem underscored the urgency and relevance of addressing customer churn, a pervasive issue with profound implications for businesses in competitive markets. The subsequent literature survey not only validated the significance of the problem but also provided a nuanced understanding of existing methodologies and their limitations. Our proposed system emerged as a strategic response, seeking to bridge existing gaps and introduce a more accurate and proactive approach to churn prediction.

The system design phase, characterized by the articulation of system architecture, UML diagrams, and functional requirements, laid a sturdy foundation for subsequent implementation. This phase provided a clear roadmap, ensuring a structured and purposeful development process. The detailed description of modules and the accompanying sample code exemplify the practical application of our proposed solution, showcasing its feasibility and functionality.

The testing phase served as a critical juncture, emphasizing the importance of robust testing methodologies. Various testing types, coupled with a comprehensive set of test cases, not only validated the reliability of the churn prediction model but also identified areas for refinement and improvement. The results obtained from this rigorous testing process not only validated the efficacy of our approach but also provided actionable insights for businesses to make informed decisions and proactively mitigate customer churn.

Future Scope:

As we gaze into the future, the scope for advancing customer churn prediction systems is expansive and holds immense promise. One notable avenue for exploration involves the integration of advanced machine learning algorithms, including deep learning and ensemble methods. By harnessing the capabilities of these sophisticated techniques, the predictive accuracy of the model can be significantly enhanced, offering a deeper understanding of customer behavior and churn dynamics.

Taking the pursuit of improvement further, the integration of real-time monitoring and adaptive learning mechanisms stands as a compelling prospect. This evolution would empower the model to dynamically adapt to changing customer dynamics, ensuring its continued relevance and effectiveness. Incorporating continuous feedback loops and recalibration mechanisms can be pivotal in fine-tuning the model, making it more adept at capturing subtle shifts in customer behavior and market trends.

Collaboration with industry experts and the infusion of domain-specific knowledge represent another promising direction for future enhancement. A more tailored approach, aligned with the intricacies of specific business sectors, has the potential to elevate the model's accuracy and applicability. Additionally, the exploration of external data sources, such as social media sentiment analysis and customer feedback, could provide a more holistic view of customer sentiments, enriching the predictive capabilities of the model.

Furthermore, the scalability and adaptability of the model to different industries and market conditions deserve attention. Future research could explore strategies to generalize the model while allowing for customization to suit the specific needs of diverse businesses. Additionally, ethical considerations, interpretability, and transparency of the model's decision-making processes could be focal points for refinement in the future, ensuring responsible and accountable use of predictive analytics.

REFERENCES

[1] Ahmad, N., & Iqbal, Z. (2023). Customer churn prediction using machine learning and deep learning techniques: A comparative study. Journal of Big Data, 16(1), 1-22.

[2] Kumar, A., & Joshi, A. (2022). Customer churn prediction in the e-commerce industry using ensemble learning techniques. International Journal of Machine Learning and Cybernetics, 13(2), 547-560.

[3] Jain, A., & Vashisht, P. (2022). Customer churn prediction in the telecom industry using hybrid machine learning models. Journal of King Saud University-Computer and Information Sciences, 34(6), 2451-2462. [4] Liu, Y., Chen, Y., & Zhang, Y. (2021). Customer churn prediction in the banking industry using convolutional neural networks. Applied Soft Computing, 99, 106942.

[5] Chen, Y., Tang, Y., & Zhou, Y. (2021). Customer churn prediction in the insurance industry using graph convolutional networks. Expert Systems with Applications, 170, 114630.

[6] Chandola, V., Kumar, A., & Pal, A. K. (2020). Customer churn prediction for e-commerce platforms: A deep learning approach. International Journal of Machine Learning and Cybernetics, 11(2), 447-460.

[7] Feng, Y., & Chen, M. (2020). Customer churn prediction in the retail industry using a hybrid approach of

XGBoost and K-means clustering. Expert Systems with Applications, 144, 113173.

[8] Fayyaz, A., & Zaidi, A. A. (2019). Customer churn prediction in the banking industry using machine learning techniques. International Journal ofIntelligent Systems and Applications, 10(4), 27-36. [9] H. Singh, D. Ramya, R. Saravanakumar et al., "Artificial intelligence based quality of transmission predictive model for cognitive optical networks," Optik, vol. 257, Article ID168789, 2022.

[10] A. Harshavardhan, P. Boyapati, S. Neelakandan, A. A. Abdul-Rasheed Akeji, A. K. Singh Pundir, and

R. Walia, "LSGDM with biogeography-based optimization (BBO) model for healthcare applications," Journal of Healthcare Engineering, vol. 2022, Article ID 2170839, 11 pages, 2022.

[11] K. Sreekala, C. P. D. Cyril, S. Neelakandan, S. Chandrasekaran, R. Walia, and E. O. Martinson,

"Capsule network-based Deep transfer learning model for face recognition," Wireless Com- munications and Mobile Computing, vol. 2022, Article ID 2086613, 12 pages, 2022.

[12] P. Ezhumalai, D. Paulraj, P. Ezhumalai, and M. Prakash, "A Deep Learning Modified Neural

Network(DLMNN) based proficient sentiment analysis technique on Twitter data," Journal of Experimental & $eoretical Artificial Intelligence, pp. 1–20, 2022.

[13] K. Lakshmanna, N. Subramani, Y. Alotaibi, S. Alghamdi, O. I. Khalafand, and A. K. Nanda, "Improved metaheuristic- driven energy-aware cluster-based routing scheme for IoT- assisted wireless sensor networks," Sustainability, vol. 14, no. 13, p. 7712, 2022.

[14] D. K. Jain, S. Neelakandan, T. Veeramani, S. Bhatia, and F. H. Memon, "Design of fuzzy logic-based energy man- agement and traffic predictive model for cyber physical systems," Computers & Electrical Engineering, vol. 102, Article ID 108135, 2022.

[15] M. Sridevi, S. Chandrasekaran, S. Chandrasekaran et al., "Deep learning approaches for cyberbullying detection and classification on social media," Computational Intelligence and Neuroscience, vol. 2022, Article ID 2163458, 13 pages, 2022.