PHASE-3

PROJECT TITLE: PREDICTING CUSTOMER CHURN USING MACHINE LEARNING TO UNCOVER HIDDEN PATTERNS

Student Name: Kesavan .S

Register Number: 511723205304

Institution: Pallavan college of Engineering

Department: INFORMATION TECHNOLOGY

Date of Submission: 15-05-2025

Github Link: https://github.com/kesavanIT/NAANMUDHALVAN

---

1. PROBLEM STATEMENT

Customer churn, or the loss of customers over time, is a critical issue for businesses in competitive industries. Retaining existing customers is often more cost-effective than acquiring new ones. However, identifying which customers are likely to churn can be difficult due to the complexity and volume of customer behavior data. This project aims to build a machine learning-based predictive system that identifies customers likely to churn based on their behavior, transaction history, demographics, and service usage patterns. The goal is to provide early warning indicators so that businesses can take proactive measures to improve retention and customer satisfaction.

2. ABSTRACT

Customer churn presents a significant challenge for businesses, particularly in subscription-based industries, where retaining existing customers is more cost-effective than acquiring new ones. This project aims to develop a predictive model

using machine learning techniques to identify customers at risk of churning. The approach involves collecting and preprocessing customer data, including demographics, usage patterns, and service interactions. Feature engineering is applied to extract meaningful attributes influencing churn behavior. Various machine learning algorithms, such as logistic regression, decision trees, random forests, and gradient boosting machines (e.g., XGBoost), are trained and evaluated. To address class imbalance in the dataset, techniques like Synthetic Minority Over-sampling Technique (SMOTE) are employed. Model interpretability is enhanced using tools like SHAP (SHapley Additive exPlanations), Streamlit or Flask to understand feature importance. The resulting model effectively predicts potential churners, allowing businesses to proactively engage at-risk customers with personalized offers or interventions, leading to improved customer retention and optimized marketing efforts.

3. SYSTEM REQUIREMENTS

- **HARDWARE**
  - Minimum 4 GB RAM (8 GB recommended)
  - Any standard processor (Intel i3/i5 or AMD equivalent)

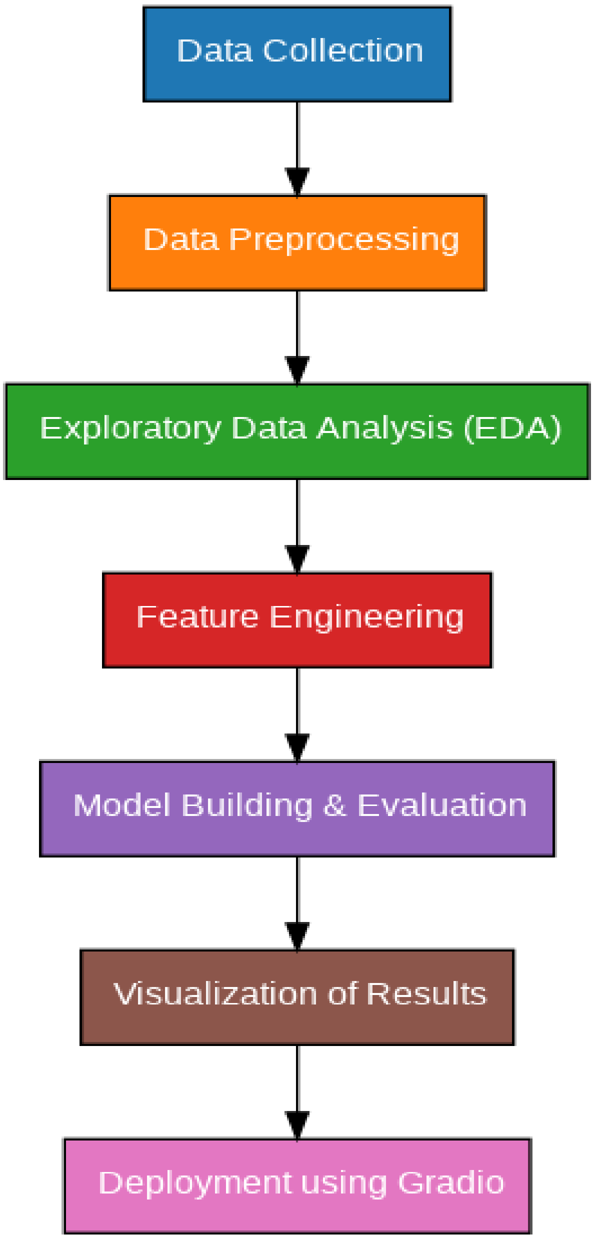
- **SOFTWARE**
  - Python 3.10+
  - Libraries: pandas, numpy, matplotlib, seaborn, scikit-learn, gradio, plotly
  - IDE: Google Colab (preferred for free GPU and easy setup)

4. OBJECTIVES

The primary objective of this project is to develop a machine learning model that accurately predicts customer churn by uncovering hidden patterns in customer behavior and interactions. By analyzing historical data—including demographics, usage patterns, and service interactions—the model aims to identify customers at high risk of leaving. The expected outputs include a ranked list of at-risk customers, insights into key factors driving churn, and actionable recommendations for targeted retention strategies. These predictions enable businesses to proactively engage with vulnerable customers through personalized interventions, thereby reducing churn rates. This approach not only helps in retaining valuable customers but also enhances customer satisfaction and loyalty. Ultimately, the project seeks to improve revenue stability and profitability by minimizing customer attrition and optimizing marketing efforts.

5. FLOWCHART OF PROJECT WORKFLOW

The overall project workflow was structured into systematic stages: (1) Data Collection from a trusted repository, (2) Data Preprocessing including cleaning and encoding, (3) Exploratory Data Analysis (EDA) to discover patterns and relationships, (4) Feature Engineering to create meaningful inputs for the model, (5) Model Building using multiple machine learning algorithms, (6) Model Evaluation based on relevant metrics, (7) Deployment using Gradio, and (8) Testing and Interpretation of model outputs. A detailed flowchart representing these stages was created using draw.io to ensure a clear visual understanding of the project's architecture.



6. Dataset Description

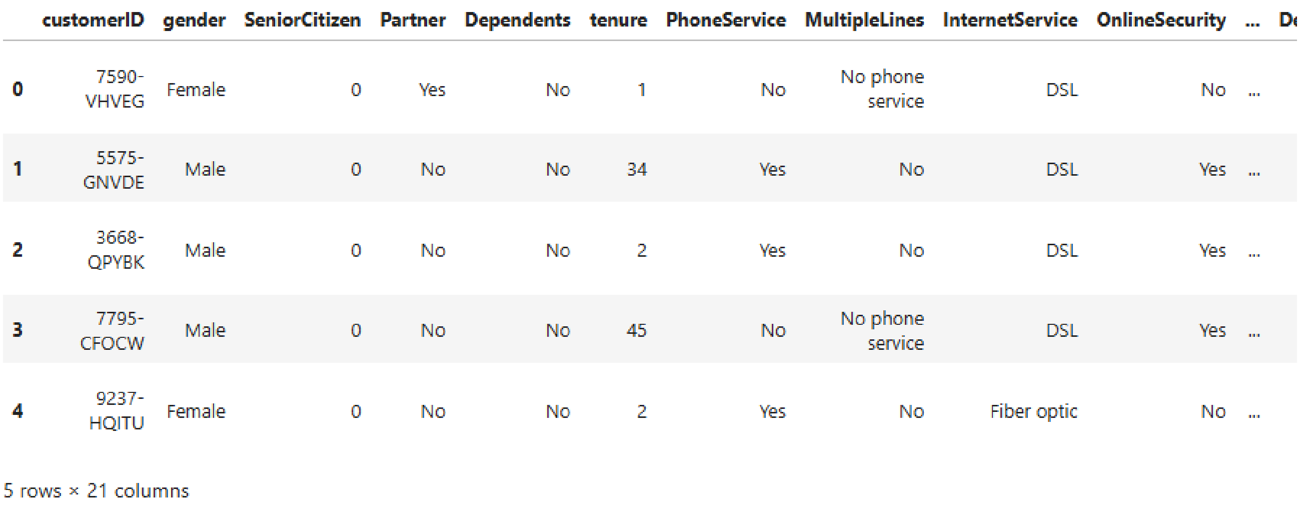
- **Source:** Kaggle / Company CRM
- **Type of Data:** Structured tabular data
- **Records and Features:** ~7,000+ records, with 20-30 features
- **Size:** 7043 rows * 21 columns
- **Attributes Covered:**

  Customer Demographics: Gender, Age, SeniorCitizen, etc.

Service Usage: Internet, Phone, Streaming, etc.

- **Dataset Link:** https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Sample dataset (df.head())

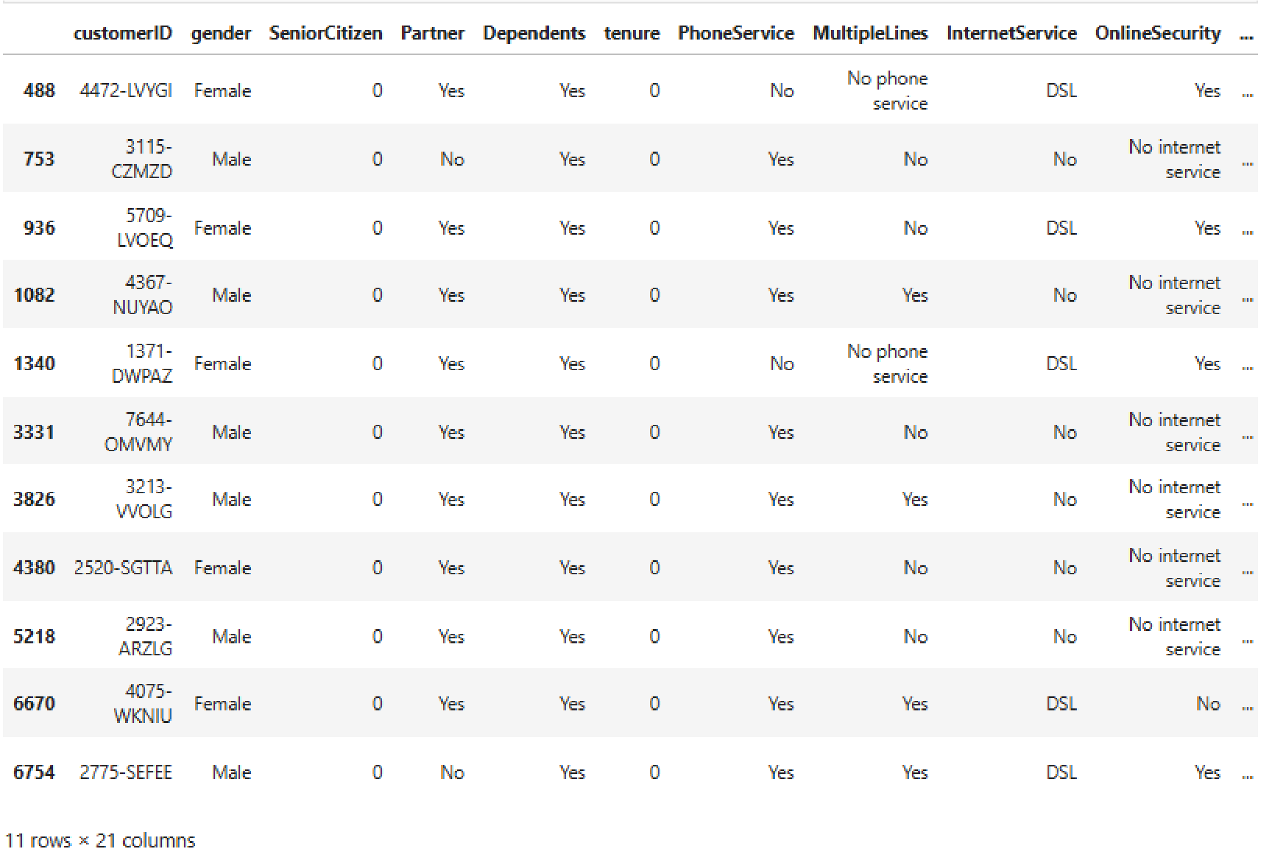


	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	De
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	

5 rows × 21 columns

7. Data Preprocessing

- **Missing values** : 11 detected in Total charges.
- **Missing value treatment :** Since the % of these records compared to total dataset is very low ie 0.15%, it is safe to ignore them from further processing.
- **Duplicates** : Checked and none found.
- **Outliers**:
  - Detected using countplots, kdeplot and lmplots.
  - Total Charges increase as Monthly Charges increase as we expected.
- **Encoding**:
  - One-Hot Encoding for multi-class categorical variables.
  - Label Encoding for binary categorical variables (e.g., yes/no features)
- **Scaling**:
  - StandardScaler applied to numeric features (e.g., gender,tenure).



	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...
488	4472-LVYGI	Female	0	Yes	Yes	0	No	No phone service	DSL	Yes	...
753	3115-CZMZD	Male	0	No	Yes	0	Yes	No	No	No internet service	...
936	5709-LVOEQ	Female	0	Yes	Yes	0	Yes	No	DSL	Yes	...
1082	4367-NUYAO	Male	0	Yes	Yes	0	Yes	Yes	No	No internet service	...
1340	1371-DWPAZ	Female	0	Yes	Yes	0	No	No phone service	DSL	Yes	...
3331	7644-OMVMY	Male	0	Yes	Yes	0	Yes	No	No	No internet service	...
3826	3213-VVOLG	Male	0	Yes	Yes	0	Yes	Yes	No	No internet service	...
4380	2520-SGTTA	Female	0	Yes	Yes	0	Yes	No	No	No internet service	...
5218	2923-ARZLG	Male	0	Yes	Yes	0	Yes	No	No	No internet service	...
6670	4075-WKNIU	Female	0	Yes	Yes	0	Yes	Yes	DSL	No	...
6754	2775-SEFEE	Male	0	No	Yes	0	Yes	Yes	DSL	Yes	...

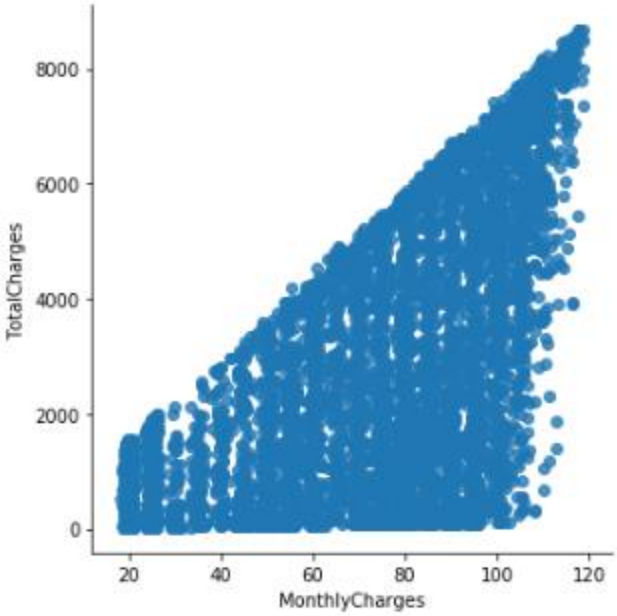
11 rows × 21 columns

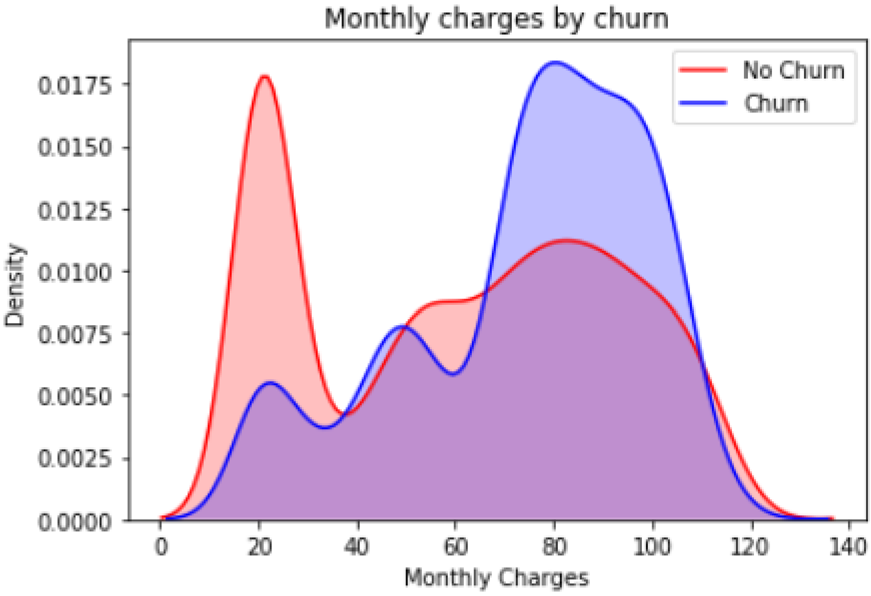
8. Exploratory Data Analysis (EDA)

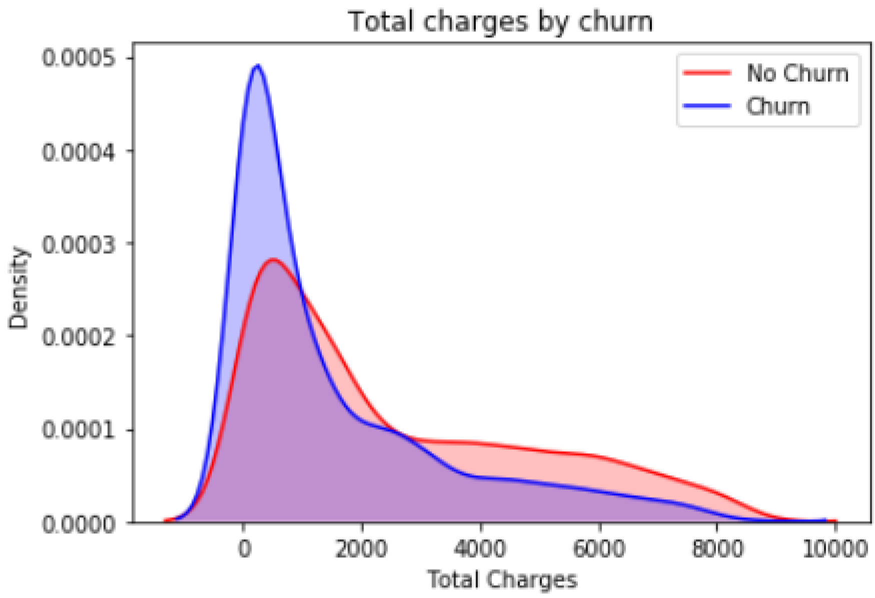
- **Univariate Analysis:**
  - Histograms for gender, senior citizen, partner, dependents, phone services, multiple lines, fiber optic internet services, online security, etc.
  - Countplots for Churns, Total charges, Monthly charges

- **Bivariate/Multivariate Analysis:**
  - Heatmap showing correlation between features.
  - Boxplots comparing churn vs. non-churn across various metrics.
- **Key Insights:**
  - Month-to-month contract customers are more likely to churn.
  - Customers with high monthly charges and short tenure show high churn rate.

- Electronic check payment users are more prone to churn.







9. FEATURE ENGINEERING

New features:

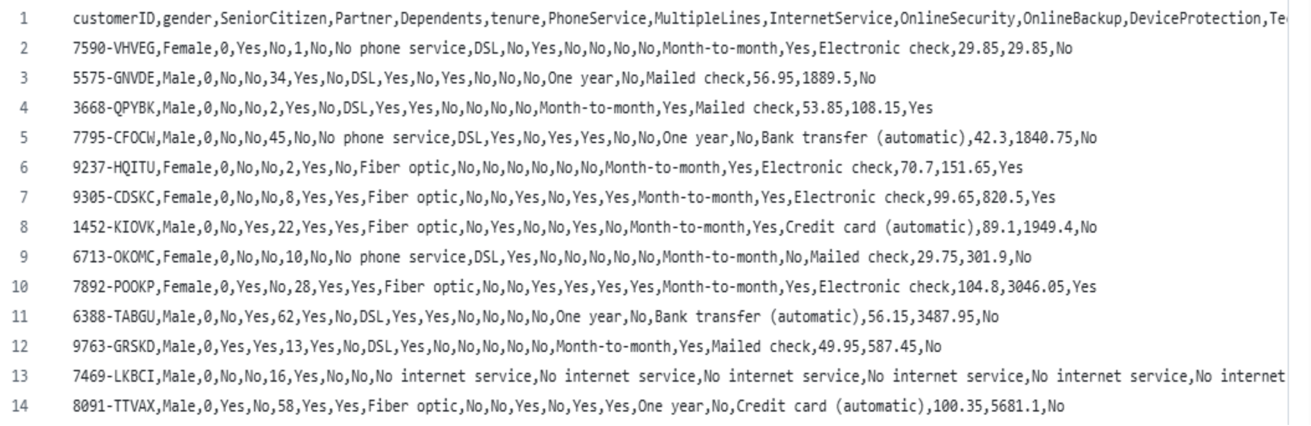
- Average charges per tenure, Service count

Feature selection :

- Reduced multicollinearity by dropping highly correlated features

Impact :

Performed feature selection using Recursive Feature Elimination (RFE) and feature importance rankings.



```
customerID,gender,SeniorCitizen,Partner,Dependents,tenure,PhoneService,MultipleLines,InternetService,OnlineSecurity,OnlineBackup,DeviceProtection,Te
7590-VHVEG,Female,0,Yes,No,1,No,No phone service,DSL,No,Yes,No,No,No,No,Month-to-month,Yes,Electronic check,29.85,29.85,No
5575-GNVDE,Male,0,No,No,34,Yes,No,DSL,Yes,No,Yes,No,No,No,One year,No,Mailed check,56.95,1889.5,No
3668-QPYBK,Male,0,No,No,2,Yes,No,DSL,Yes,Yes,No,No,No,No,Month-to-month,Yes,Mailed check,53.85,108.15,Yes
7795-CFOCW,Male,0,No,No,45,No,No phone service,DSL,Yes,No,Yes,Yes,No,No,One year,No,Bank transfer (automatic),42.3,1840.75,No
9237-HQITU,Female,0,No,No,2,Yes,No,Fiber optic,No,No,No,No,No,No,Month-to-month,Yes,Electronic check,70.7,151.65,Yes
9305-CDSKC,Female,0,No,No,8,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,Month-to-month,Yes,Electronic check,99.65,820.5,Yes
1452-KIOVK,Male,0,No,Yes,22,Yes,Yes,Fiber optic,No,Yes,No,No,Yes,No,Month-to-month,Yes,Credit card (automatic),89.1,1949.4,No
6713-OKOMC,Female,0,No,No,10,No,No phone service,DSL,Yes,No,No,No,No,No,Month-to-month,No,Mailed check,29.75,301.9,No
7892-POOKP,Female,0,Yes,No,28,Yes,Yes,Fiber optic,No,No,Yes,Yes,Yes,Yes,Month-to-month,Yes,Electronic check,104.8,3046.05,Yes
6388-TABGU,Male,0,No,Yes,62,Yes,No,DSL,Yes,Yes,No,No,No,No,One year,No,Bank transfer (automatic),56.15,3487.95,No
9763-GRSKD,Male,0,Yes,Yes,13,Yes,No,DSL,Yes,No,No,No,No,No,Month-to-month,Yes,Mailed check,49.95,587.45,No
7469-LKBCI,Male,0,No,No,16,Yes,No,No,No internet service,No internet service,No internet service,No internet service,No internet service,No internet
8091-TTVAX,Male,0,Yes,No,58,Yes,Yes,Fiber optic,No,No,Yes,No,Yes,Yes,One year,No,Credit card (automatic),100.35,5681.1,No
```

10. MODEL BUILDING

- **Algorithms Used:**
    - Logistic Regression
    - Random Forest Regressor (Advanced)
    - XGBoost
    - Support Vector Machine (SVM)
- **Evaluation Metrics:**
    - Accuracy
    - Precision, Recall, F1-Score
    - Confusion Matrix
    - ROC-AUC Score
- **Logistic Regression:**
    - Easy interpretability and baseline.
- **Random Forest & XGBoost:**
    - High accuracy, handles feature interactions well
- **SVM:**
    - Effective with high-dimensional space and limited samples

11. MODEL EVALUATION

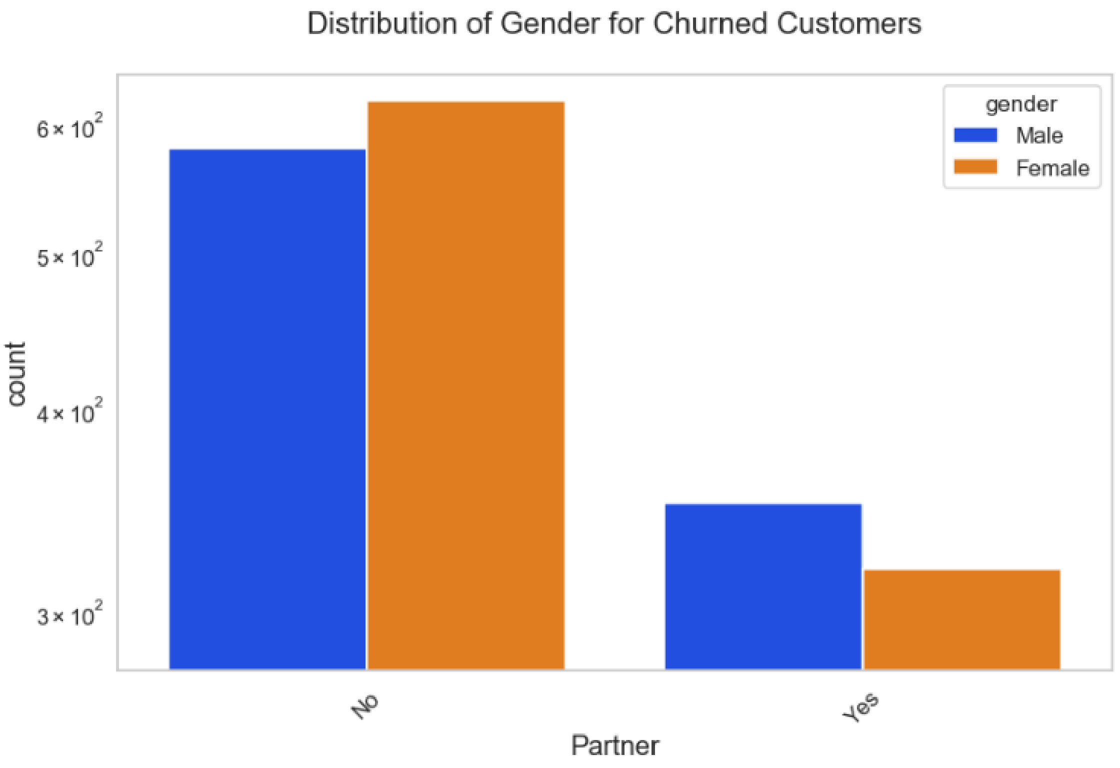
Accuracy: Overall correctness of the model.

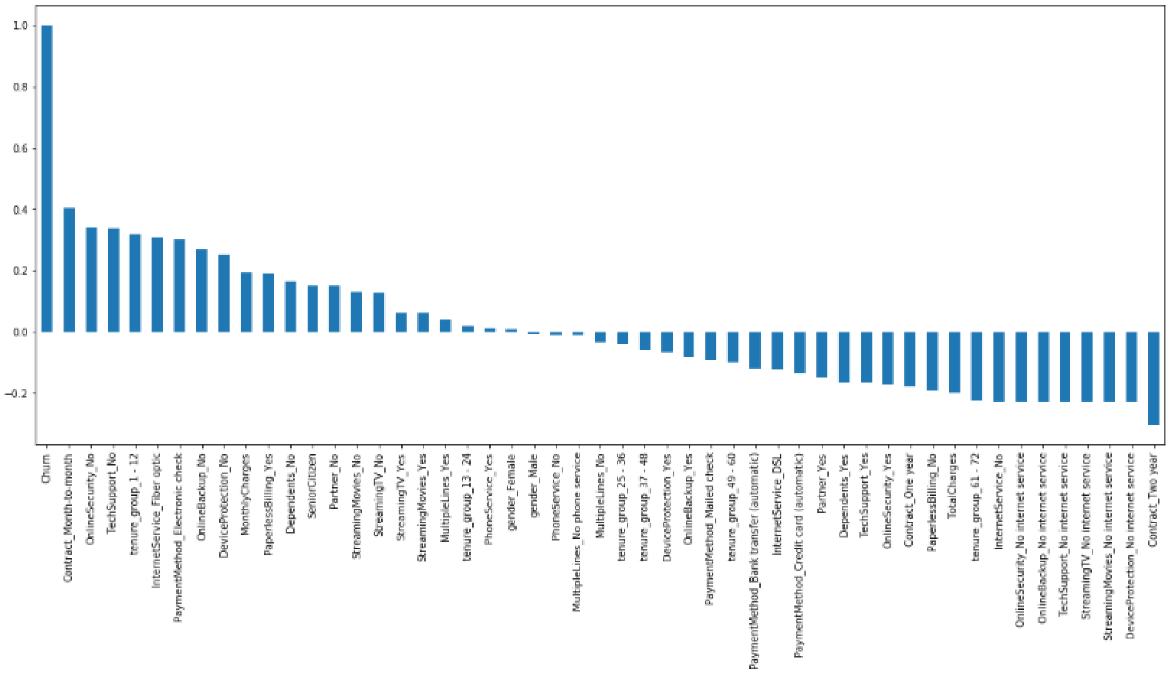
Precision: Proportion of correctly predicted churned customers among all predicted churns.

Recall (Sensitivity): Proportion of actual churned customers correctly identified.

F1-Score: Harmonic mean of precision and recall.

ROC-AUC: Measures the model's ability to distinguish between classes.


Distribution of Gender for Churned Customers
count
6×10^2
5×10^2
4×10^2
3×10^2
gender
Male
Female
No
Yes
Partner


1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Churn
Contract_Month-to-month
OnlineSecurity_No
TechSupport_No
tenure_group_1 - 12
InternetService_Fiber optic
PaymentMethod_Electronic check
OnlineBackup_No
DeviceProtection_No
MonthlyCharges
PaperlessBilling_Yes
Dependents_No
SeniorCitizen
Partner_No
StreamingMovies_No
StreamingTV_No
StreamingTV_Yes
StreamingMovies_Yes
MultipleLines_Yes
tenure_group_13 - 24
PhoneService_Yes
gender_Female
gender_Male
PhoneService_No
MultipleLines_No phone service
MultipleLines_No
tenure_group_25 - 36
tenure_group_37 - 48
DeviceProtection_Yes
OnlineBackup_Yes
PaymentMethod_Mailed check
tenure_group_49 - 60
PaymentMethod_Bank transfer (automatic)
InternetService_DSL
PaymentMethod_Credit card (automatic)
Partner_Yes
Dependents_Yes
TechSupport_Yes
OnlineSecurity_Yes
Contract_One year
PaperlessBilling_No
TotalCharges
tenure_group_61 - 72
InternetService_No
OnlineSecurity_No internet service
OnlineBackup_No internet service
TechSupport_No internet service
StreamingTV_No internet service
StreamingMovies_No internet service
DeviceProtection_No internet service
Contract_Two year

12. DEPLOYMENT

- Deployment Method: Gradio Interface

13. SOURCE CODE

All source code including preprocessing, modeling, evaluation, and deployment is available in the GitHub repository:

https://github.com/kesavanIT/NAANMUDHALVAN

14. FUTURE SCOPE

Predicting customer churn using machine learning offers significant future benefits for businesses aiming to enhance customer retention and drive growth. By analyzing complex datasets, machine learning models can uncover hidden patterns and subtle indicators of potential churn, enabling proactive interventions before customers decide to leave . This proactive approach not only improves customer satisfaction but also reduces the costs associated with acquiring new customers. Furthermore, machine learning models can continuously learn and adapt from new data, ensuring that predictions remain accurate over time . By leveraging these insights, businesses can develop personalized retention strategies, optimize resource allocation, and ultimately foster long-term customer loyalty. As technology continues to advance, integrating machine learning into customer relationship management will become increasingly essential for maintaining a competitive edge in the market.

15. TEAM MEMBERS AND ROLES

Team members	Contribution
KESAVAN .S	Data Collection and Preprocessing
MANOGARAN .T	Exploratory Data Analysis
BOOPATHI .M.K	Feature Engineering & Model Development
SATHISH KUMAR .S	Evaluation and Deployment
HEMANATHAESWARAN .S	Documentation and Reporting