Project Title :Predicting customer churn using machine learning to uncover hidden patterns

PHASE-2

1. ProblemStatement

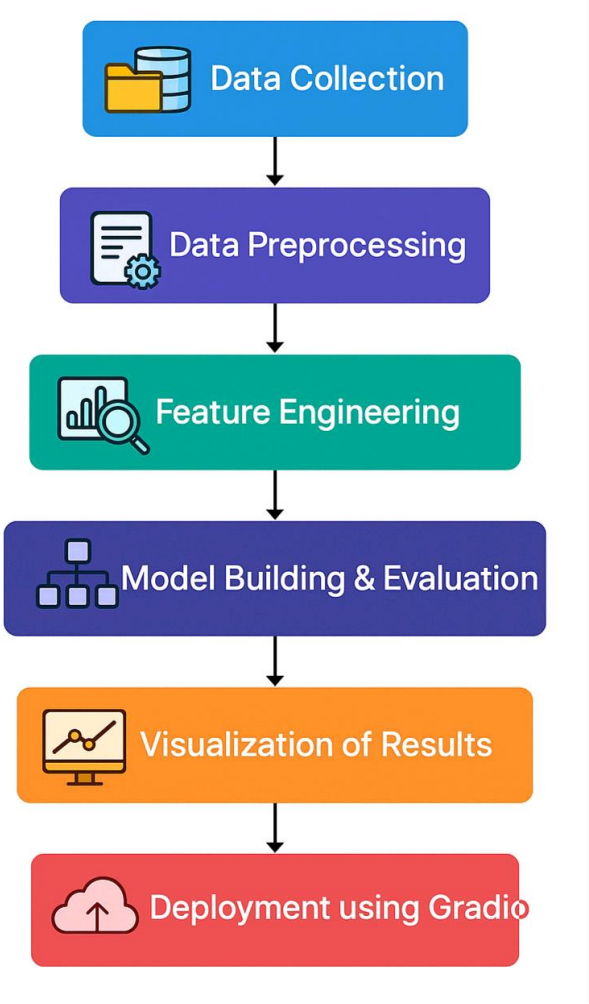
In today's highly competitive business environment, retaining existing customers is more cost-effective than acquiring new ones. However, many companies struggle with unexpected customer churn, which negatively impacts revenue and long-term growth. Traditional analysis often fails to identify subtle indicators of churn in complex and large customer datasets.

This project aims to develop a machine learning-based predictive model to accurately identify customers who are likely to churn. By uncovering hidden patterns and relationships within customer behavior, transaction history, and demographic data, the model will enable proactive retention strategies, thereby enhancing customer loyalty and reducing churn rates.

2. Project Objectives

1. **Understand and define churn behavior:** Clearly identify what constitutes customer churn based on the specific business context (e.g., service cancellation, inactivity over a period)
2. **Collect and preprocess data:** Gather relevant customer data including demographics, transaction history, usage patterns, and support interactions, followed by cleaning and preprocessing for model training.
3. **Explore and analyze data:** Perform exploratory data analysis (EDA) to uncover trends, correlations, and patterns that may influence churn behavior.
4. **Feature engineering:** Create meaningful features from raw data that enhance the predictive power of machine learning models.
5. **Build and evaluate machine learning models:** Implement various classification algorithms (e.g., Logistic Regression, Random Forest, XGBoost) to predict churn, and compare their performance using metrics such as accuracy, precision, recall, and F1-score.
6. **Uncover hidden churn patterns**: Use model insights and explainability tools (e.g., SHAP, LIME) to identify hidden patterns and key churn indicators.
7. **Deploy the best model:** Prepare the selected model for integration into a business environment for real-time churn prediction and decision-making.
8. **Recommend strategies**: Provide actionable insights and retention strategies based on the model's findings to reduce future churn

3.Flowchart of the Project Workflow



4.Data Description

customerID: Unique ID assigned to each customer

gender: Gender of the customer (Male or Female)

SeniorCitizen: Indicates if the customer is a senior citizen (1 = Yes, 0 = No)

Partner: Whether the customer has a partner (Yes/No)

Dependents: Whether the customer has dependents (Yes/No)

tenure: Number of months the customer has stayed with the company

PhoneService: Whether the customer has phone service (Yes/No)

StreamingTV: Whether the customer has streaming TV service

StreamingMovies: Whether the customer has streaming movie service

Contract: Type of contract (Month-to-month, One year, Two year)

PaperlessBilling: Whether the customer uses paperless billing

PaymentMethod: Customer's payment method (e.g., Electronic check, Credit card)

MonthlyCharges: Amount charged to the customer monthly

TotalCharges: Total amount charged to the customer

Churn: Target variable; indicates if the customer has churned (Yes/No)

5.Data Preprocessing

- **Data Cleaning:**

Removed duplicate records (if any).
Handled missing values in fields like TotalCharges by imputing or removing them.

- **Data Type Conversion:**

Converted TotalCharges from string to numerical format.
Encoded categorical variables (like gender, Contract, PaymentMethod) using Label Encoding or -Hot Encoding.

- **OneFeature Engineering:**

Created new features if required (e.g., tenure groups).
Scaled numerical features like MonthlyCharges and TotalCharges using StandardScaler or MinMaxScaler.

- **Outlier Detection:**

Checked for and handled outliers in numerical columns using statistical methods or visualization tools.

- **Data Splitting:**

Split the dataset into training and testing sets (typically 80% training, 20% testing).
Optionally created a validation set for hyperparameter tuning.

- **Balancing the Dataset**:

Addressed class imbalance in the Churn column using techniques like SMOTE (Synthetic Minority Over-sampling Technique) or undersampling.

6.Exploratory Data Analysis (EDA)

- **Univariate Analysis**:

Analyzed the distribution of individual features using histograms and count plot(e.g., gender, tenure, Contract, Churn).
Observed class imbalance in the target variable Churn.

- **Bivariate Analysis:**

Explored relationships between features and Churn (e.g., churn rate by Contract type or MonthlyCharges).
Plotted bar charts and box plots to compare churn across different **customer** segments.

- **Correlation Analysis:**

Computed correlation matrix to identify relationships between numerical features.
Used heatmaps to visualize correlations (e.g., MonthlyCharges and TotalCharges).

- **Churn Patterns:**

Found that month-to-month contract customers had higher churn rates.
Identified that customers with fiber optic internet or no tech support were more likely to churn.

- **Tenure Insights:**

Observed that customers with shorter tenure had a higher chance of churning.
Created tenure groups to better visualize trends.

- **Multivariate Analysis:**

Combined multiple features to observe complex patterns (e.g., Churn rate among senior citizens with monthly contracts and high charges).

7.FeatureEngineering

1. Encoded categorical variables using Label Encoding and One-Hot Encoding.
2. Created new features like tenure_group and TotalServicesUsed.
3. Scaled numerical features (MonthlyCharges, TotalCharges, tenure) using MinMaxScaler.
4. Removed irrelevant columns like customerID.
5. Selected important features using correlation and model-based methods.

8.Model Building

- Tried multiple classification models: Logistic Regression, Random Forest, XGBoost, and SVM.
- Split the data into training (80%) and testing (20%) sets.
- Used cross-validation and grid search to fine-tune hyperparameters.
- Evaluated models using accuracy, precision, recall, F1-score, and ROC-AUC.
- Selected the best-performing model for final prediction.

9.VisualizationofResults&ModelInsights

- Confusion Matrix: Showed true positives, false positives, true negatives, and false negatives for model evaluation.
- ROC Curve: Visualized the trade-off between true positive rate and false positive rate; calculated AUC score.
- Feature Importance Plot: Highlighted key features influencing churn (e.g., Contract, tenure, MonthlyCharges).
- SHAP Values: Used SHAP plots to interpret individual predictions and understand model decisions.
- Churn Trends: Plotted churn rate by tenure, contract type, and service usage to uncover patterns.

10.Tools and Technologies Used

1. Python – Programming language for data analysis and model building
2. Pandas & NumPy – Data manipulation and numerical operations
3. Matplotlib & Seaborn – Data visualization
4. Scikit-learn – Machine learning models and preprocessing tools
5. XGBoost – Advanced gradient boosting algorithm
6. SHAP – Model explainability and interpretation

7. Jupyter Notebook / Google Colab – Development environmen
8. Git & GitHub – Version control and project collaboration

11.TeamMembersandContributions

- ***Data Analyst***

Collected and cleaned raw customer data
Conducted exploratory data analysis (EDA) and created visualizations

- ***Data Engineer***

Handled data preprocessing and feature engineering
Encoded variables, scaled features, and managed data pipelines
Built and trained classification models
Tuned hyperparameters and evaluated model performance using various metrics

- ***Model Interpreter & Presenter***

Analyzed model results using SHAP and feature importance
Summarized insights and prepared the final presentation/report