1. Executive Summary

Customer retention is one of the most critical factors for long-term business success. Acquiring a new customer costs significantly more than retaining an existing one. Therefore, predicting customer churn in advance enables companies to take proactive retention measures.

This project focuses on analyzing telecom customer data to identify churn patterns and build a machine learning model capable of predicting whether a customer is likely to leave the service. Using data preprocessing techniques and a Random Forest classification model, the project successfully identifies high-risk customers and provides actionable business insights.

The developed solution can help organizations reduce revenue loss, improve customer engagement strategies, and enhance overall profitability.

---

2. Introduction

In subscription-based industries such as telecom, banking, SaaS, and streaming platforms, customer churn is a major challenge. Customer churn occurs when a customer discontinues the use of a company's product or service.

High churn rates can:

- Decrease company revenue
- Increase customer acquisition costs
- Reduce market competitiveness
- Impact brand loyalty

To address this issue, companies use predictive analytics and machine learning to detect early warning signs of churn. This project applies classification algorithms to predict churn based on customer attributes.

---

3. Problem Statement

The main problem addressed in this project is:

Given customer demographic, account, service, and payment data, predict whether the customer will churn or not.

This is a Binary Classification Problem where:

- 1 → Customer will churn
- 0 → Customer will not churn

The objective is to build a predictive model that can accurately classify customers into churn and non-churn categories.

---

4. Project Objectives

4.1 Primary Objective

- To develop a machine learning model that predicts customer churn accurately
- To identify key factors influencing churn
- To help businesses design customer retention strategies

4.2 Secondary Objectives

- Perform data cleaning and preprocessing
- Convert categorical data into numerical form
- Split dataset into training and testing data
- Evaluate model performance using standard metrics
- Generate business insights from model output

---

5. Dataset Description

The dataset contains telecom customer information including:

5.1 Demographic Features

- Gender
- Senior Citizen
- Partner
- Dependents

5.2 Account Information

- Tenure (number of months customer stayed)
- Contract Type
- Monthly Charges
- Total Charges

5.3 Service Information

- Internet Service

- Online Security
- Tech Support
- Streaming Services
- Device Protection

5.4 Payment Information

- Payment Method

Target Variable

- Churn (Yes/No)

The dataset consists of both numerical and categorical variables, making preprocessing essential before model training.

---

6. Data Preprocessing

Data preprocessing is one of the most crucial steps in any machine learning project.

6.1 Handling Missing Values

- Checked dataset for null or inconsistent values
- Cleaned or removed missing data
- Ensured dataset consistency

6.2 Label Encoding

Since machine learning algorithms require numerical input, categorical columns were converted using Label Encoding.

Example:

```
LE = LabelEncoder()
for col in df.columns:
    if df[col].dtype == 'object':
        df[col] = LE.fit_transform(df[col])
```

This converted categories like "Yes/No" and "Male/Female" into numerical values.

---

7. Feature and Target Separation

The dataset was divided into:

- X (Independent Variables)

- Y (Target Variable – Churn)

x = df.drop('Churn', axis=1)

y = df['Churn']

---

8. Train-Test Split

To evaluate the model fairly:

x_train , x_test , y_train , y_test = train_test_split(x,y,test_size=0.25)

- 75% used for training
- 25% used for testing

This ensures unbiased evaluation.

---

9. Model Selection

Random Forest Classifier

The Random Forest algorithm was chosen for this project.

model = RandomForestClassifier(n_estimators=50 , max_depth=3)

Why Random Forest?

- Ensemble technique combining multiple decision trees
- Reduces overfitting
- Handles non-linear relationships
- Provides high accuracy
- Works well on structured datasets

Parameters Used

- n_estimators = 50
- max_depth = 3

These parameters control model complexity and stability.

---

10. Model Training

The model was trained using training data:

model.fit(x_train, y_train)

The algorithm learned patterns between customer attributes and churn behavior.

---

11. Model Evaluation

Model performance was evaluated using test data.

accuracy = model.score(x_test, y_test)

11.1 Evaluation Metrics

Accuracy

Measures overall correct predictions.

Precision

Measures how many predicted churn customers were actually churned.

Recall

Measures how many actual churn customers were correctly predicted.

F1-Score

Harmonic mean of precision and recall.

These metrics ensure balanced model evaluation.

---

12. Key Insights from Analysis

High Churn Risk Observed In:

- Month-to-month contract customers
- Short tenure customers
- High monthly charge customers
- Customers without security services

Low Churn Risk Observed In:

- Long-term contract customers
- High tenure customers
- Customers using automatic payment
- Customers subscribed to multiple services

These insights show that contract type and tenure significantly influence churn behavior.

13. Business Impact

The model helps businesses:

- Identify high-risk customers early
- Design targeted retention campaigns
- Reduce revenue loss
- Optimize marketing cost
- Improve customer satisfaction

Predictive churn analysis increases strategic decision-making efficiency.

14. Limitations

- Limited dataset features
- No real-time behavioral data
- Accuracy depends on data quality

15. Future Scope

- Use advanced algorithms (XGBoost, Gradient Boosting)
- Perform feature importance analysis
- Deploy model as a web application
- Integrate with CRM systems
- Use real-time prediction pipeline

16. Tools and Technologies Used

- Python
- Pandas
- NumPy
- Scikit-Learn
- Random Forest
- Jupyter Notebook

17. Skills Demonstrated

- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Model Building
- Model Evaluation
- Business Interpretation
- Critical Thinking

18. Conclusion

This project successfully demonstrates how machine learning can be used to predict customer churn in telecom businesses. By analyzing customer attributes and applying the Random Forest classification algorithm, high-risk customers can be identified effectively.

The insights derived from this analysis can help businesses reduce churn, improve retention strategies, and increase profitability.

This project reflects the practical application of data science techniques to solve real-world business problems.