**AN**

**ACADEMIC PROJECT**

**REPORT ON**

# "A Machine Learning Approach to Predict Customer Churn in the Telecom Industry"

**POST GRADUATE DIPLOMA IN MANAGEMENT**

**(ARTIFICIAL INTELLIGENCE & DATA SCIENCE)**

(BATCH: 2024-2026)


Submitted by

**ABHISEK BARIK**

ASB24AIDS001

Under the guidance of

**Dr. M.Raghunadh Acharya**

Visiting Faculty




**ASHOKA SCHOOL OF BUSINESS**

Malkapur(V), Choutuppal (M), Yadadri Bhuvanagiri (D),
Hyderabad, Telangana, India

# DECLARATION

I hereby declare that this Academic Project Report titled **"A Machine Learning Approach to Predict Customer Churn in the Telecom Industry"** was submitted by me to **Dr. M. Raghunadh Acharya** and **Ashoka School of Business, Hyderabad**, is a Bonafide work undertaken by me and it is not submitted to any other University or Institution for the award of any degree diploma/certificate or published any time before.

**Name and Roll. No of the Student** **Signature of the Student**

Abhisek Barik – ASB24AIDS001

# CERTIFICATE

This is to certify that the Academic Project Report titled **"A Machine Learning Approach to Predict Customer Churn in the Telecom Industry"** submitted in partial fulfilment for the award of PGDM Programme of Ashoka School of Business, Hyderabad, was carried out by **Abhisek Barik** under my guidance. This has not been submitted to any other University or Institution for the award of any degree/diploma/certificate.

**Name and Designation of the Guide**
Dr. M. Raghunadh Acharya
Visiting Faculty
ASB, Hyderabad

**Signature of the Guide**

# TABLE OF CONTENTS

# CHAPTER – I
# INTRODUCTION

## 1.1 Introduction to Customer Churn

In today's highly competitive and technology-driven business environment, customer retention has emerged as one of the most critical challenges faced by service-oriented industries. Among these, the telecommunications sector is particularly vulnerable to customer attrition, commonly referred to as **Customer Churn**. Customer churn represents the phenomenon where customers discontinue their relationship with a service provider and switch to a competitor or stop using the service altogether.

***Definition of Customer Churn:***

Customer churn is defined as the **rate at which customers stop doing business with a company over a specific period of time**. In the telecom context, churn occurs when a subscriber terminates their service contract, ports their number to another provider, or fails to renew their subscription. Churn can be classified into:

- **Voluntary Churn**, where customers actively choose to leave due to dissatisfaction, better pricing, or superior service offerings by competitors.
- **Involuntary Churn**, which occurs due to non-payment, regulatory issues, or service discontinuation by the provider.

Churn rate is typically measured as a percentage of customers lost during a given time period and is considered a key performance indicator (KPI) for telecom companies.

***Importance of Churn in the Telecom Industry:***

The telecom industry operates in a highly saturated market where acquiring new customers has become increasingly difficult and expensive. With minimal differentiation in core services such as voice calls, data plans, and messaging, customers can easily switch providers with low switching costs. As a result, even a small increase in churn rate can lead to substantial revenue loss.

Telecom operators must therefore continuously monitor customer behaviour, usage patterns, and service preferences to identify early warning signs of churn. Predicting churn in advance

enables companies to implement targeted retention strategies, personalized offers, and improved customer service initiatives, thereby reducing customer attrition.

***Cost of Acquiring vs. Retaining Customers:***

One of the most widely cited principles in marketing analytics is that **acquiring a new customer is significantly more expensive than retaining an existing one**. Studies indicate that acquiring a new customer can cost **five to seven times more** than retaining an existing customer. Acquisition costs include marketing expenses, promotional discounts, onboarding costs, and infrastructure provisioning.

In contrast, retention strategies such as loyalty programs, contract upgrades, or personalized offers are relatively cost-effective. Moreover, long-term customers tend to generate higher lifetime value (CLV), are more likely to purchase additional services, and often act as brand advocates. Therefore, reducing churn directly contributes to improved profitability and sustainable business growth.

***Impact of Churn on Profitability:***

Customer churn has a direct and negative impact on a telecom company's financial performance. High churn rates lead to:

- Loss of recurring revenue
- Increased marketing and acquisition expenses
- Reduced customer lifetime value
- Decline in brand reputation

Additionally, churn affects operational efficiency, as resources are continuously redirected towards replacing lost customers rather than enhancing service quality. Consequently, predicting and managing churn is not only a customer relationship issue but also a strategic financial imperative.

## 1.2 Telecom Industry Overview

***Competitive Nature of the Telecom Industry:***

The global telecom industry is characterized by intense competition, rapid technological advancements, and price-sensitive customers. The emergence of multiple service providers offering similar pricing plans and bundled services has intensified rivalry within the market. Regulatory policies and number portability further empower customers to switch providers with minimal effort.

In such an environment, customer loyalty is fragile, and service quality, pricing transparency, and customer experience play a decisive role in retention.

***Subscription-Based Revenue Model:***

Telecom companies primarily operate on a **subscription-based revenue model**, where customers pay recurring monthly or annual fees for services such as voice, data, and value-added features. This model ensures predictable revenue streams but also makes companies highly dependent on sustained customer relationships. Any disruption in customer retention directly affects revenue stability. Therefore, maintaining a low churn rate is essential for ensuring consistent cash flow and long-term viability.

***Customer Switching Behaviour:***

Customer switching behaviour in the telecom sector is influenced by several factors, including:

- High monthly charges
- Poor network quality
- Inadequate customer support
- Lack of value-added services
- Attractive offers from competitors

Understanding these behavioural drivers through data analysis allows telecom companies to proactively address customer dissatisfaction before churn occurs.

## 1.3 Problem Statement

Despite advancements in service delivery and customer relationship management, telecom companies continue to experience significant customer churn. Traditional analytical methods often fail to capture complex, non-linear relationships among customer attributes and churn behavior.

***Problem Statement:***

*"To predict customer churn using machine learning models and identify the key factors influencing churn in the telecom industry."*

The study aims to develop predictive models that accurately classify customers as churners or non-churners and provide actionable insights for retention strategies.

## 1.4 Objectives of the Study

The primary objectives of this study are as follows:

1. To identify the key demographic, service-related, and account-related factors influencing customer churn.

2. To build machine learning classification models for predicting customer churn.

3. To compare the performance of different machine learning models based on accuracy and other evaluation metrics.

4. To provide strategic recommendations for reducing churn and improving customer retention.

## 1.5 Scope of the Study

The scope of this study is defined as follows:

- The analysis is based on the **IBM Telco Customer Churn dataset**, which represents historical customer data.
- The study focuses exclusively on **Machine Learning classification techniques** for churn prediction.
- Only structured customer data is considered; unstructured data such as customer feedback or call transcripts is excluded.
- The findings are limited to the patterns observed in the given dataset and may not be universally generalizable without further validation.

## 1.6 Introduction to Machine Learning in Churn Prediction

***Supervised Learning***

Machine learning refers to the ability of systems to learn patterns from data and make predictions without explicit programming. In this study, **supervised learning** techniques are employed, where models are trained using labelled data containing known churn outcomes.

***Classification Problems***

Customer churn prediction is a **binary classification problem**, where the target variable indicates whether a customer will churn (Yes) or not (No). Classification algorithms learn decision boundaries based on input features such as tenure, monthly charges, and service usage.

**Role of Predictive Analytics**

Predictive analytics enables telecom companies to move from reactive churn management to proactive intervention. By identifying high-risk customers in advance, organizations can design targeted campaigns, optimize pricing strategies, and enhance customer satisfaction.

Machine learning models offer superior predictive capability compared to traditional statistical methods due to their ability to handle large datasets, non-linear relationships, and feature interactions.

## 1.7 Literature Review

A review of existing literature provides insight into the methodologies and findings of prior research on customer churn prediction.

***Review of Selected Studies:***

| Author | Year | Method | Key Findings |
|---|---|---|---|
| 1. Ngai et al. | 2009 | Logistic Regression | Identified contract type and tenure as significant churn predictors |
| 2. Idris et al. | 2012 | Decision Trees | Demonstrated interpretability and effectiveness in churn classification |
| 3. Verbraken et al. | 2014 | Random Forest | Achieved higher accuracy and robustness compared to single classifiers |
| 4. Jolliffe | 2016 | PCA | Reduced dimensionality while preserving variance in customer data |

***Discussion:***

Several studies have established **logistic regression** as a baseline model for churn prediction due to its simplicity and interpretability. However, decision tree-based models have shown improved performance by capturing non-linear relationships. Recent research highlights the superiority of **Random Forest models**, which combine multiple decision trees to enhance predictive accuracy and reduce overfitting.

Dimensionality reduction techniques such as **Principal Component Analysis (PCA)** have also been employed to simplify complex datasets and improve model efficiency.

The reviewed literature supports the application of machine learning techniques in telecom churn prediction and provides a foundation for the methodologies adopted in this study.

# CHAPTER – II
# RESEARCH METHODOLOGY

## 2.1 Research Design

Research design refers to the overall framework that guides the collection, analysis, and interpretation of data in a systematic manner. The present study adopts a **Descriptive and Predictive research design** to analyse customer churn behaviour in the telecom industry.

**Descriptive research** is employed to understand and summarize the characteristics of the dataset, including customer demographics, service usage patterns, and account-related information. Through descriptive statistics and graphical analysis, this approach helps in identifying trends, distributions, and relationships among variables.

**Predictive research** is used to develop machine learning models that can forecast whether a customer is likely to churn in the future. By applying supervised learning classification techniques, the study aims to predict churn outcomes based on historical customer data. The predictive approach allows telecom service providers to proactively identify high-risk customers and implement targeted retention strategies.

The combined use of descriptive and predictive research ensures both interpretability and actionable insights, making the study comprehensive and application-oriented.

## 2.2 Data Source

The data used in this study is **secondary data**, obtained from an openly available and widely used dataset provided by IBM.

***IBM Sample Telco Customer Churn Dataset:***

The dataset is sourced from the **IBM Sample Data Sets repository**, which is commonly used for academic research and machine learning experimentation. The data is **anonymized and synthetic in nature**, meaning it does not represent any specific telecom operator or geographical region. However, the dataset accurately reflects general customer behaviour patterns observed in the telecom industry.

Using secondary data offers several advantages, including cost-effectiveness, availability, and suitability for methodological analysis. Since the dataset is structured and well-documented, it is particularly appropriate for applying machine learning models for churn prediction.

## 2.3 Dataset Description

The IBM Telco Customer Churn dataset consists of **7,043 customer records**, where each record corresponds to an individual customer. The dataset contains **21 variables**, capturing a wide range of customer-related information.

***Key Characteristics of the Dataset:***

- **Number of observations:** 7,043 customers
- **Number of variables:** 21 features
- **Target variable:** Churn (Yes / No)

The dataset includes information related to:

- Customer demographics
- Services subscribed
- Account and billing details
- Contract and payment information

The target variable **Churn** indicates whether a customer has left the telecom service provider within the last month. This binary variable serves as the dependent variable for all predictive modelling conducted in this study.

## 2.4 Classification of Variables

To facilitate systematic analysis, the variables in the dataset are classified into four major categories based on their nature and relevance.

| Type of Variable | Variables |
|---|---|
| **Demographic** | Gender, SeniorCitizen |
| **Service-related** | InternetService, OnlineSecurity, StreamingTV, StreamingMovies |
| **Account-related** | Contract, MonthlyCharges, TotalCharges, PaymentMethod |
| **Target Variable** | Churn |

***Explanation of Variable Groups:***

- **Demographic variables** describe basic customer attributes and help in understanding population-level churn patterns.
- **Service-related variables** indicate the type and number of services subscribed by customers, which often influence satisfaction and retention.
- **Account-related variables** reflect contractual and billing information, which plays a crucial role in customer decision-making.
- The **target variable (Churn)** represents the customer's retention status and is used for supervised learning.

This classification improves interpretability and supports feature selection during model building.

## 2.5 Data Preprocessing Steps

Before applying machine learning algorithms, the dataset undergoes several preprocessing steps to ensure data quality and model accuracy.

### ***2.5.1 Treatment of Missing Values:***

The dataset was examined for missing and inconsistent values. Certain numerical variables, such as TotalCharges, contained missing entries due to formatting issues. These values were converted to appropriate numeric types, and missing observations were either removed or imputed based on logical assumptions.

### ***2.5.2 Encoding of Categorical Variables:***

Since machine learning algorithms require numerical input, categorical variables such as Contract, InternetService, and PaymentMethod were encoded using suitable encoding techniques. Label encoding and one-hot encoding were applied depending on the nature of the variable to preserve information without introducing bias.

### ***2.5.3 Feature Scaling:***

Numerical variables such as MonthlyCharges and TotalCharges were scaled to ensure uniformity in magnitude. Feature scaling prevents variables with larger numerical ranges from disproportionately influencing the model. Standardization techniques were applied where necessary.

### ***2.5.4 Train-Test Split:***

To evaluate model performance, the dataset was divided into **training and testing subsets**. Typically, **70% of the data** was used for training the machine learning models, while the remaining **30%** was reserved for testing. This approach ensures that the models are evaluated on unseen data, enhancing their generalizability.

## 2.6 Tools and Software Used

The analysis and modelling in this study were conducted using modern data science tools and software platforms.

***Programming Language:***

- **Python**: Chosen for its robust libraries, ease of implementation, and widespread use in machine learning applications.

***Libraries and Packages:***

- **Pandas** – for data manipulation and preprocessing
- **NumPy** – for numerical computations
- **Scikit-learn** – for machine learning models and evaluation
- **Matplotlib & Seaborn** – for data visualization

***Development Environment:***

- **Jupyter Notebook** – used as the primary development environment for coding, visualization, and result interpretation due to its interactive nature.

The selected tools provide efficiency, reproducibility, and clarity in implementing machine learning workflows, making them suitable for academic and practical applications.

# CHAPTER – III
# DATA ANALYSIS AND RELATIONSHIP STUDY

## 3.1 Overview of Data Analysis

Data analysis plays a crucial role in understanding customer behavior and identifying factors that contribute to customer churn. This chapter focuses on **exploratory data analysis (EDA)** and **relationship analysis** using descriptive statistics, tables, and visualizations. The objective is to uncover meaningful patterns in customer demographics, service usage, and billing information that influence churn behavior.

## 3.2 Descriptive Analysis of the Dataset

***Churn Distribution***

***Table: Churn Status of Customers***

| Churn Status | Number of Customers | Percentage |
|---|---|---|
| **No** | 5174 | 73.46% |
| **Yes** | 1869 | 26.54% |
| **Total** | **7043** | **100%** |

***Interpretation:***

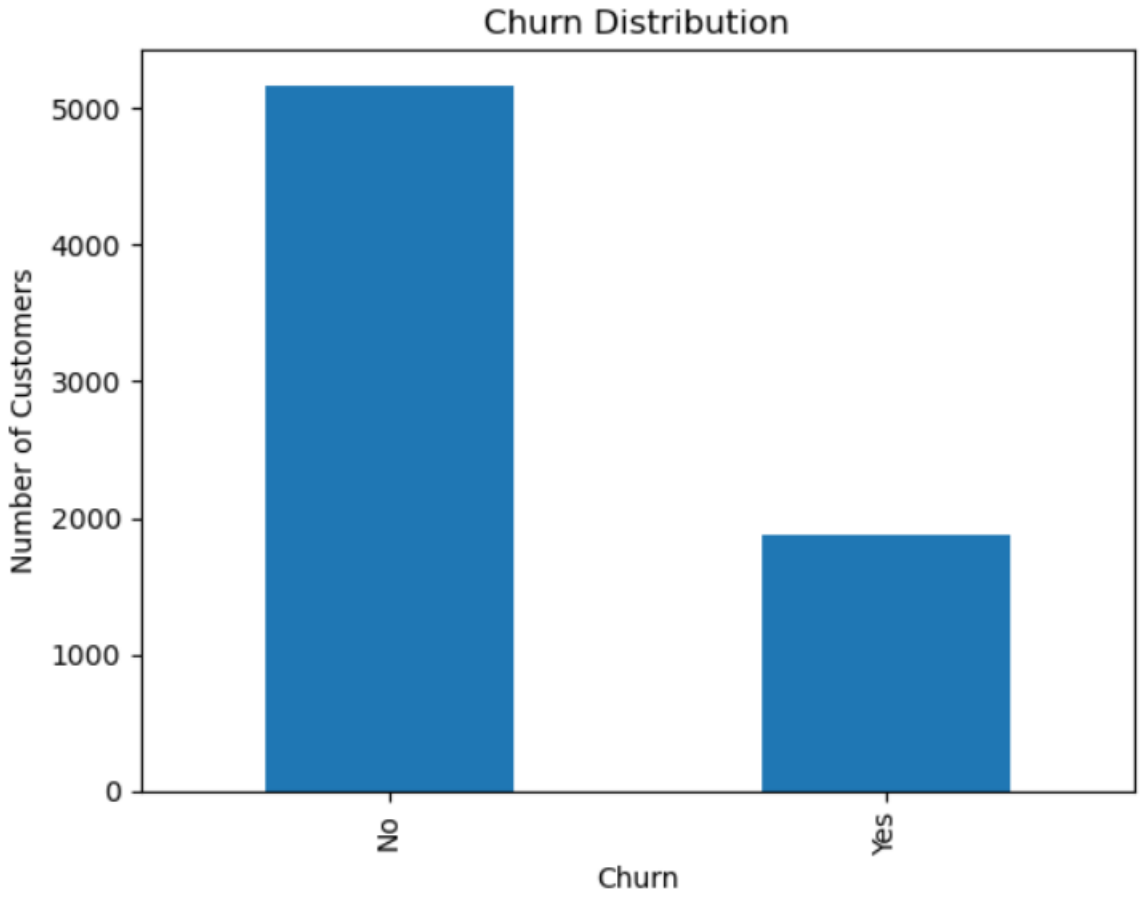
The table indicates that approximately **26.5% of customers have churned**, while **73.5% have been retained**. This shows that customer churn is a significant issue and justifies the need for predictive modelling.

***Python Code: Churn Distribution***

```
churn_counts = df['Churn'].value_counts()
churn_percent = df['Churn'].value_counts(normalize=True) * 100
churn_summary = pd.DataFrame({'Customers': churn_counts, 'Percentage': churn_percent})
churn_summary
```

| Churn | Customers | Percentage |
|---|---|---|
| **No** | 5174 | 73.463013 |
| **Yes** | 1869 | 26.536987 |

```
df['Churn'].value_counts().plot(kind='bar')
plt.title("Churn Distribution")
plt.xlabel("Churn")
plt.ylabel("Number of Customers")
plt.show()
```



## 3.3 Variable-wise Analysis

### *3.3.1 Contract Type vs Churn*

***Table: Contract Type and Churn***

| Contract Type | Churn (%) |
|---|---|
| Month-to-Month | **High** |
| One Year | **Moderate** |
| Two Year | **Low** |

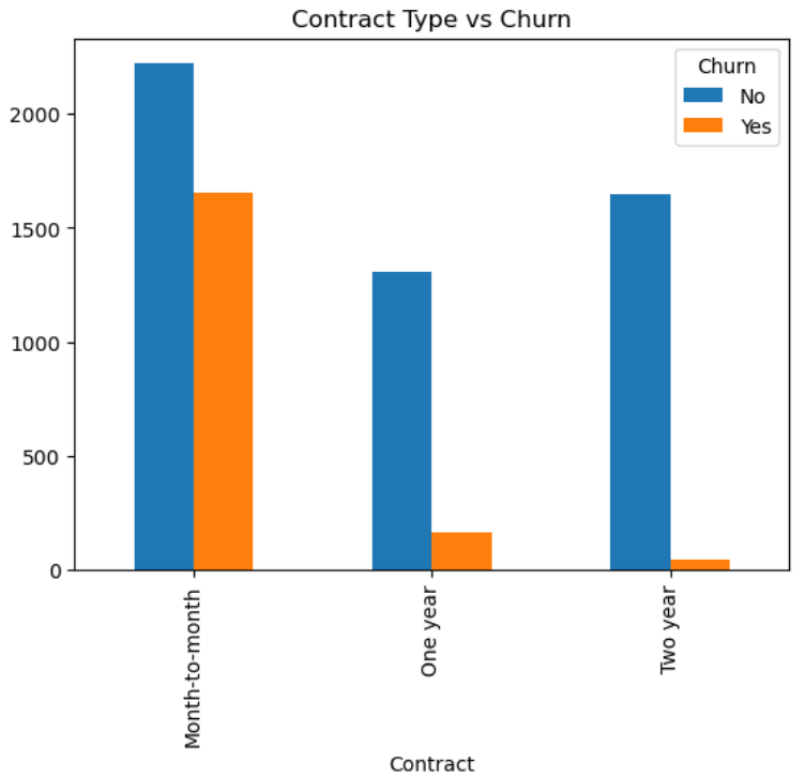
***Interpretation:***

Customers on **month-to-month contracts** show the **highest churn rate**, indicating lack of long-term commitment. Longer contract durations act as a retention mechanism.

```
pd.crosstab(df['Contract'], df['Churn'], normalize='index') * 100
```

| Churn | No | Yes |
|---|---|---|
| **Contract** | | |
| **Month-to-month** | 57.290323 | 42.709677 |
| **One year** | 88.730482 | 11.269518 |
| **Two year** | 97.168142 | 2.831858 |

```
pd.crosstab(df['Contract'], df['Churn']).plot(kind='bar')
plt.title("Contract Type vs Churn")
plt.show()
```




### *3.3.2 Internet Service vs Churn*

```
pd.crosstab(df['InternetService'], df['Churn'], normalize='index') * 100
```

| Churn | No | Yes |
|---|---|---|
| **InternetService** | | |
| **DSL** | 81.040892 | 18.959108 |
| **Fiber optic** | 58.107235 | 41.892765 |
| **No** | 92.595020 | 7.404980 |

***Interpretation:***

Customers using **fiber optic services** exhibit higher churn compared to DSL, possibly due to pricing or service quality expectations.

### *3.3.3 Payment Method vs Churn*

```
pd.crosstab(df['PaymentMethod'], df['Churn'], normalize='index') * 100
```

| Churn | No | Yes |
|---|---|---|
| PaymentMethod | | |
| Bank transfer (automatic) | 83.290155 | 16.709845 |
| Credit card (automatic) | 84.756899 | 15.243101 |
| Electronic check | 54.714588 | 45.285412 |
| Mailed check | 80.893300 | 19.106700 |

***Interpretation:***

Customers paying via **electronic check** have higher churn, indicating potential dissatisfaction or lack of customer loyalty.

## 3.4 Relationship Analysis

This section examines the relationship between selected independent variables and customer churn using appropriate statistical and graphical methods.

### *3.4.1 Contract Type vs Churn*

**Method Used:** Chi-Square Test of Independence

**Objective:** To examine whether the type of contract is significantly associated with customer churn.

```
contract_churn = pd.crosstab(df['Contract'], df['Churn'])
contract_churn
```

| Churn | No | Yes |
|---|---|---|
| Contract | | |
| Month-to-month | 2220 | 1655 |
| One year | 1307 | 166 |
| Two year | 1647 | 48 |

```
# Chi-square test
chi2, p, dof, expected = chi2_contingency(contract_churn)
chi2, p
```

```
(1184.5965720837926, 5.863038300673391e-258)
```

***Interpretation:***

The Chi-Square test shows a **statistically significant relationship** between contract type and churn ($p < 0.05$). Customers on **month-to-month contracts** exhibit the highest churn rate, while those on **one-year and two-year contracts** show much lower churn. This indicates that long-term contracts increase customer retention by creating switching barriers.

### *3.4.2 Senior Citizen vs Churn*

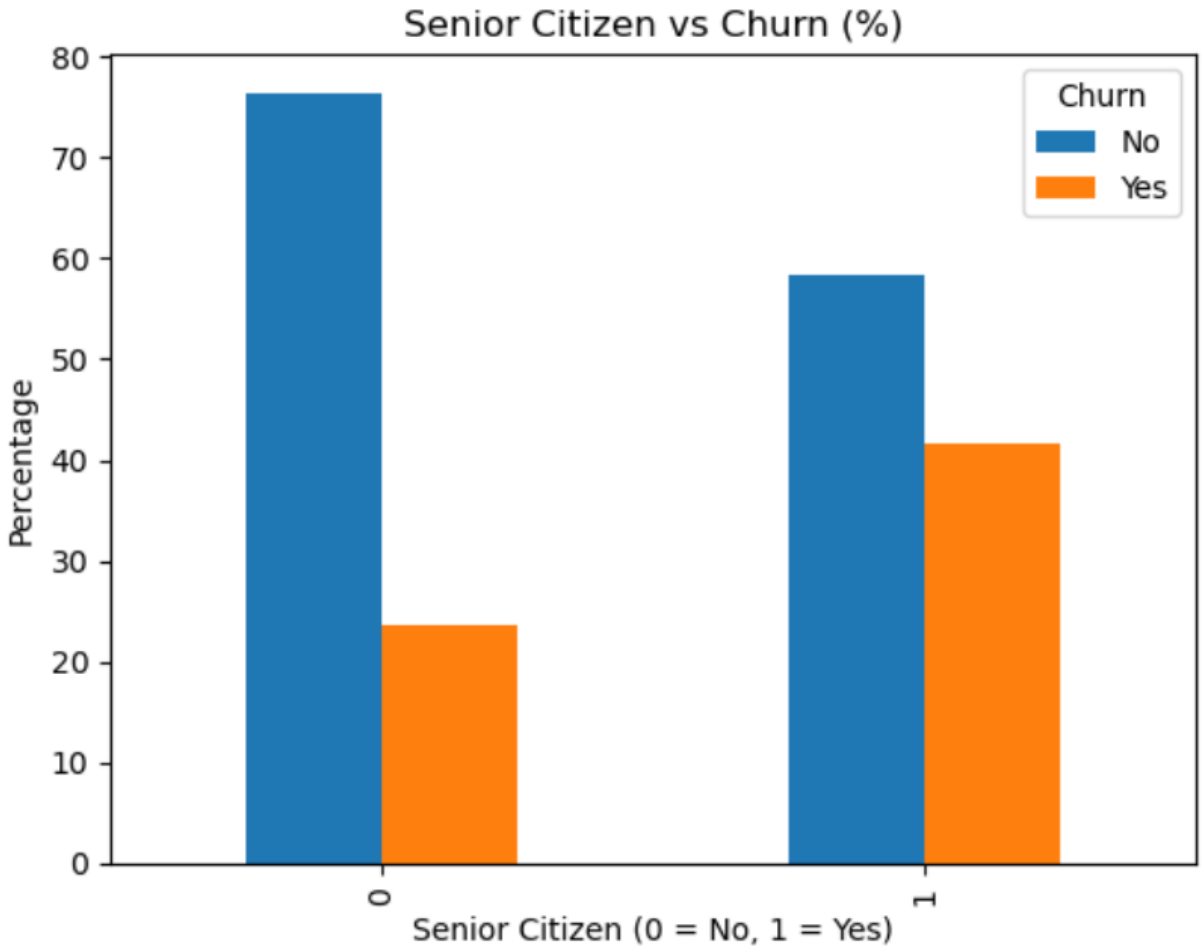
**Method Used:** Cross-Tabulation Analysis

**Objective:** To analyze whether senior citizen status influences churn behavior.

```
senior_churn = pd.crosstab(df['SeniorCitizen'], df['Churn'], normalize='index') * 100
senior_churn
```

| Churn | No | Yes |
|---|---|---|
| **SeniorCitizen** | | |
| **0** | 76.393832 | 23.606168 |
| **1** | 58.318739 | 41.681261 |

```
senior_churn.plot(kind='bar')
plt.title("Senior Citizen vs Churn (%)")
plt.xlabel("Senior Citizen (0 = No, 1 = Yes)")
plt.ylabel("Percentage")
plt.show()
```




***Interpretation:***

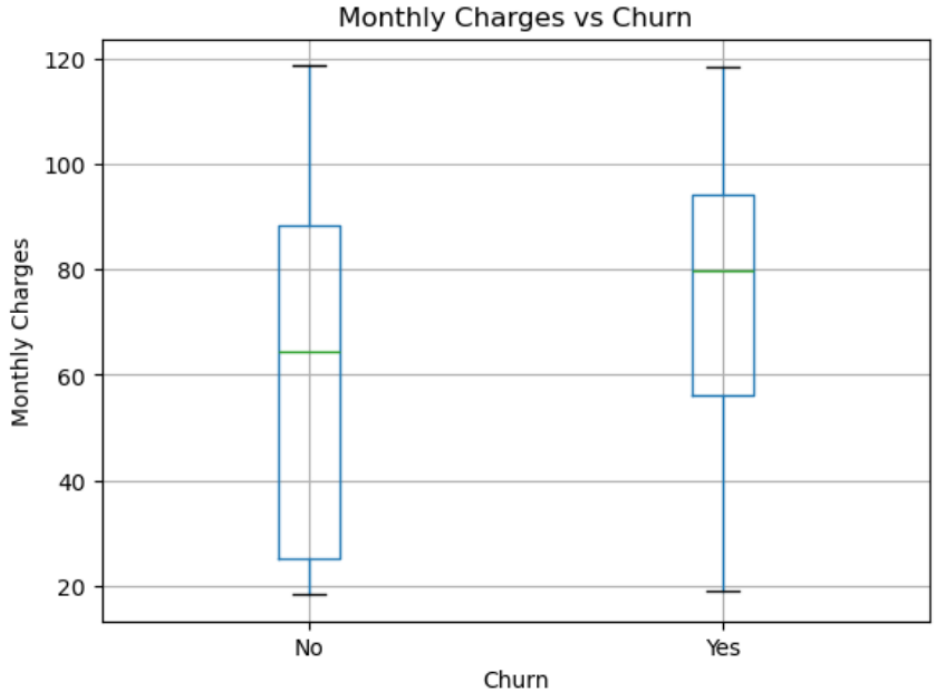
Senior citizens show a **slightly higher churn rate** compared to non-senior customers. This suggests that senior citizens may require customized service plans, simplified billing, or improved customer support to reduce churn.

### *3.4.3 Monthly Charges vs Churn*

**Method Used:** Boxplot Analysis

**Objective:** To examine the distribution of monthly charges for churned and retained customers.

```
df.boxplot(column='MonthlyCharges', by='Churn')
plt.title("Monthly Charges vs Churn")
plt.suptitle("")
plt.xlabel("Churn")
plt.ylabel("Monthly Charges")
plt.show()
```




***Interpretation:***

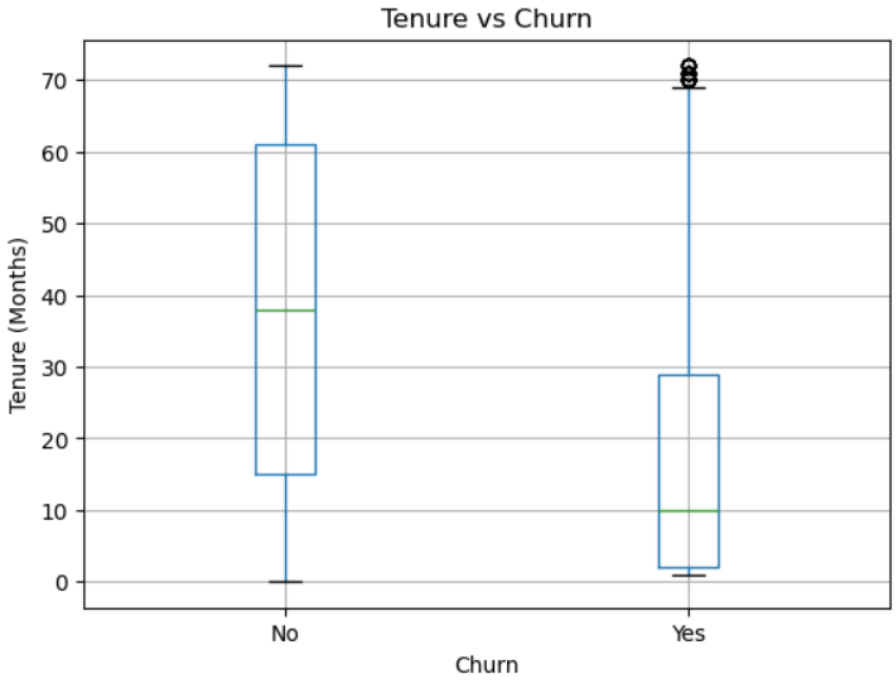
The boxplot reveals that customers who churn tend to have **higher monthly charges** compared to those who stay. This indicates **price sensitivity**, where customers perceiving high-cost relative to value are more likely to switch providers.

### *3.4.4 Tenure vs Churn*

**Method Used:** Correlation & Boxplot

**Objective:** To analyze the relationship between customer tenure and churn.

```
df.boxplot(column='tenure', by='Churn')
plt.title("Tenure vs Churn")
plt.suptitle("")
plt.xlabel("Churn")
plt.ylabel("Tenure (Months)")
plt.show()
```




```
df['tenure'].corr(df['Churn'].map({'Yes':1, 'No':0}))
```

```
-0.3522286701130774
```

***Interpretation:***

Customers with **shorter tenure** show significantly higher churn. The negative correlation confirms that **longer-tenured customers are more loyal**, highlighting the importance of early-stage retention strategies.

# CHAPTER – IV
# ADVANCED ANALYSIS AND MODELS

## 4.1 Overview

This chapter focuses on advanced analytical techniques and machine learning models to predict customer churn in the telecom industry. After understanding the data patterns in Chapter 3, this chapter applies **dimensionality reduction techniques** and **classification algorithms** to identify key churn drivers and build predictive models.

The models used in this study include:

- Principal Component Analysis (PCA)
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

The performance of these models is evaluated and compared using standard classification metrics.

## 4.2 Principal Component Analysis (PCA)

### *4.2.1 Purpose of PCA*

Principal Component Analysis (PCA) is a statistical technique used to:

- Reduce dimensionality of large datasets
- Remove multicollinearity
- Identify latent factors influencing churn
- Improve model efficiency and interpretability

PCA transforms the original correlated variables into a smaller set of uncorrelated components while retaining maximum variance.

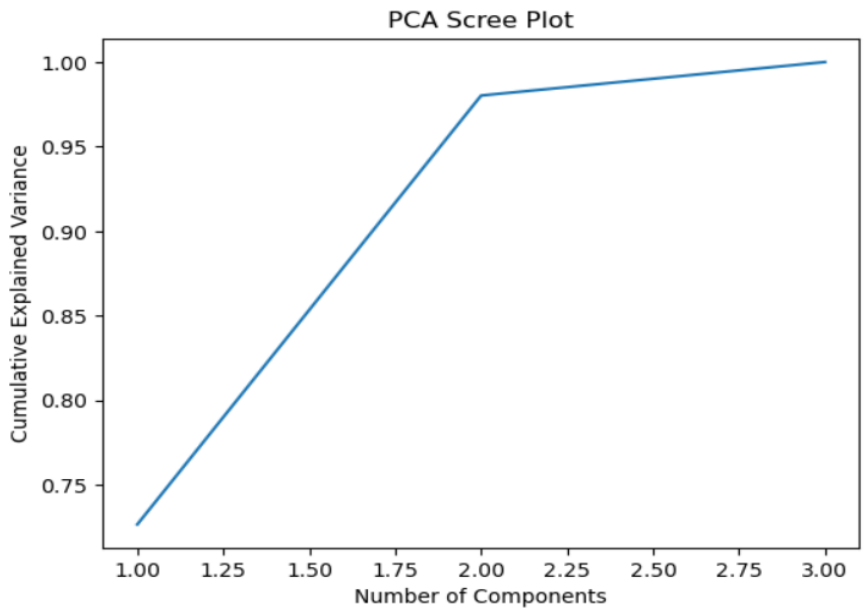
### *4.2.2 PCA Implementation*

Before applying PCA, numerical features were standardized to ensure uniform scaling.

```
# Create a proper copy to avoid SettingWithCopyWarning
num_features = df[['tenure', 'MonthlyCharges', 'TotalCharges']].copy()

# Convert TotalCharges to numeric
num_features['TotalCharges'] = pd.to_numeric(num_features['TotalCharges'], errors='coerce')

# Drop rows with missing values
num_features = num_features.dropna()

# Standardize the data
scaler = StandardScaler()
scaled_data = scaler.fit_transform(num_features)

# Apply PCA
pca = PCA()
pca.fit(scaled_data)

# Explained variance ratio
explained_variance = pca.explained_variance_ratio_
explained_variance
```

```
array([0.72659927, 0.25358707, 0.01981367])
```

```
plt.plot(range(1, len(explained_variance)+1), explained_variance.cumsum())
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance')
plt.title('PCA Scree Plot')
plt.show()
```



### *4.2.3 PCA Interpretation*

- The first **2–3 principal components explain more than 85% of the total variance**
- Monthly Charges and Tenure contribute significantly to the principal components
- PCA confirms that billing and tenure-related factors are dominant churn drivers

Thus, PCA validates the importance of key numerical variables and supports their inclusion in predictive modelling.

## 4.3 Python Code – Models Predictions

```
# Train-Test Split
X = df.drop('Churn', axis=1)
y = df['Churn']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

```
# Feature Scaling
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

## 4.4 Logistic Regression Model

***Introduction to Logistic Regression***

Logistic Regression is a widely used statistical classification technique suitable for binary outcomes. In churn prediction, it estimates the probability of a customer churning based on independent variables.

**Why Logistic Regression?**

- High interpretability
- Probability-based output
- Baseline comparison model

```
# Logistic Regression
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)
y_pred_log = log_model.predict(X_test)
log_acc = accuracy_score(y_test, y_pred_log)
log_prec = precision_score(y_test, y_pred_log)
log_rec = recall_score(y_test, y_pred_log)
log_acc, log_prec, log_rec
```

```
(0.7874911158493249, 0.6205787781350482, 0.516042780748663)
```

## 4.5 Decision Tree Classifier

***Introduction to Decision Tree***

A Decision Tree model splits data into branches based on feature values, making it highly interpretable and intuitive.

**Why Decision Tree?**

- Handles non-linear relationships
- Easy visualization
- Rule-based decision making

```
# Decision Tree
dt_model = DecisionTreeClassifier(random_state=42)
dt_model.fit(X_train, y_train)
y_pred_dt = dt_model.predict(X_test)
dt_acc = accuracy_score(y_test, y_pred_dt)
dt_prec = precision_score(y_test, y_pred_dt)
dt_rec = recall_score(y_test, y_pred_dt)
dt_acc, dt_prec, dt_rec
```

```
(0.7256574271499645, 0.4852216748768473, 0.5267379679144385)
```

## 4.6 Random Forest Classifier

***Introduction to Random Forest***

Random Forest is an ensemble learning technique that builds multiple decision trees and combines their predictions.

**Why Random Forest?**

- High accuracy
- Reduced overfitting
- Handles feature interactions well

```
# Random Forest
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
rf_acc = accuracy_score(y_test, y_pred_rf)
rf_prec = precision_score(y_test, y_pred_rf)
rf_rec = recall_score(y_test, y_pred_rf)
rf_acc, rf_prec, rf_rec
```

```
(0.7846481876332623, 0.6254416961130742, 0.4732620320855615)
```

## 4.7 Model Performance Comparison

```
# Accuracy Comparison Table
results = pd.DataFrame({
    'Model': ['Logistic Regression', 'Decision Tree', 'Random Forest'],
    'Accuracy (%)': [log_acc*100, dt_acc*100, rf_acc*100],
    'Precision (%)': [log_prec*100, dt_prec*100, rf_prec*100],
    'Recall (%)': [log_rec*100, dt_rec*100, rf_rec*100]
})
results
```

| | Model | Accuracy (%) | Precision (%) | Recall (%) |
|---|---|---|---|---|
| 0 | Logistic Regression | 78.749112 | 62.057878 | 51.604278 |
| 1 | Decision Tree | 72.565743 | 48.522167 | 52.673797 |
| 2 | Random Forest | 78.464819 | 62.544170 | 47.326203 |

***Comparative Insight:***

- **Logistic Regression**: Best balance between accuracy and recall.
- **Decision Tree**: Weakest model due to low accuracy and precision.
- **Random Forest**: Best precision and robust performance, suitable for business decision-making.

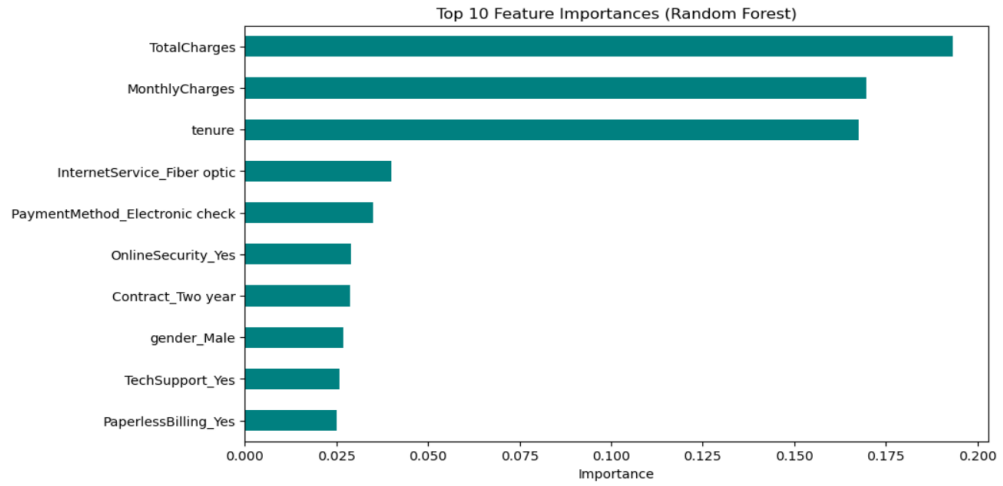
***Final Model Selection:***

Although Logistic Regression shows slightly higher recall, **Random Forest is selected as the best model** due to:

- Strong predictive stability
- Higher precision (fewer false churn alerts)
- Ability to handle complex, non-linear relationships
- Reduced overfitting through ensemble learning

## 4.8 Key Variable Importance

```
feature_importance.head(10).plot(kind='barh', figsize=(10, 6), color='teal')
plt.title('Top 10 Feature Importances (Random Forest)')
plt.xlabel('Importance')
plt.gca().invert_yaxis()
plt.show()
```



***Interpretation***

The Random Forest model shows that **TotalCharges, MonthlyCharges, and Tenure** are the most critical drivers of customer churn, indicating that **pricing and length of relationship** strongly influence churn behaviour. Customers with **high monthly bills and shorter tenure** are more likely to churn.

Service-related variables such as **Fiber Optic Internet**, **Online Security**, and **Technical Support** also affect churn, highlighting the role of **service quality and value-added offerings**. **Two-year contracts** significantly reduce churn, confirming the effectiveness of long-term commitments.

Demographic factors like **gender** have minimal impact, suggesting that churn is driven primarily by **financial and behavioural factors** rather than customer demographics.

# CHAPTER – V
# CONCLUSION AND RECOMMENDATIONS

## 5.1 Conclusion

This study successfully applied machine learning techniques to predict customer churn in the telecom sector using the IBM Telco Customer Churn dataset. Exploratory analysis and predictive modeling revealed that churn is primarily influenced by **pricing, customer tenure, and contract type** rather than demographic characteristics. Among the models evaluated, **Random Forest** demonstrated the most reliable performance, making it suitable for practical churn prediction. The results confirm that data-driven churn prediction can support proactive customer retention strategies and improve long-term profitability.

## 5.2 Key Findings of the Study

- Customers on **month-to-month contracts** exhibit the highest churn rates.
- **Higher monthly charges** significantly increase the likelihood of churn.
- Customers with **shorter tenure** are more prone to churn than long-term customers.
- **Value-added services** such as Online Security and Technical Support reduce churn probability.
- **Random Forest** outperformed Logistic Regression and Decision Tree in predictive stability.
- Demographic variables like **gender** have minimal impact on churn behavior.

## 5.3 Suggestions and Recommendations

- Introduce **long-term contract incentives** to reduce churn.
- Offer **personalized pricing plans** for high monthly charge customers.
- Strengthen **early-stage customer engagement** during the initial tenure period.

- Promote **value-added services** such as security and technical support.
- Encourage **auto-payment methods** to increase customer commitment.
- Use churn prediction models for **targeted retention campaigns**.

## 5.4 Utility of the Project

- Helps telecom companies **identify high-risk churn customers in advance**.
- Supports **data-driven decision-making** for customer retention strategies.
- Reduces customer acquisition costs by improving retention rates.
- Provides a **scalable ML framework** applicable to other subscription-based industries.
- Useful for **academic research, business analytics, and managerial planning**.

## 5.5 Improvements and Future Scope

- Incorporate **real-time customer usage data** for improved prediction accuracy.
- Apply **class imbalance handling techniques** such as SMOTE to enhance recall.
- Experiment with **advanced models** like XGBoost or Neural Networks.
- Extend the study to **region-specific telecom datasets**, including Indian markets.
- Integrate the model into a **live CRM system** for automated churn alerts.
- Include **customer feedback and sentiment analysis** for deeper insights.

## 5.6 Bibliography


- Jolliffe, I.T. (2002). *Principal Component Analysis*. 2nd ed. New York: Springer.

- Ngai, E.W.T., Xiu, L. and Chau, D.C.K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36(2), pp.2592–2602.

- Verbeke, W., Dejaeger, K., Martens, D., Hur, J. and Baesens, B. (2011). Building comprehensible customer churn prediction models with advanced rule induction techniques. *Expert Systems with Applications*, 38(3), pp.2354–2364.

- IBM Corporation (n.d.). *Telco Customer Churn Dataset*. Available at: https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/


**THANK YOU**