

*Assignment report work on -*

**“PREDICTING CUSTOMER CHURN IN A  
TELECOMMUNICATION COMPANY”**

By-

**SANJANA UMRAO**

**Registration Number- 12107890**

## ABSTRACT

---

In this study, I aimed to develop an effective churn prediction model for a telecommunications company using machine learning algorithms. Churn prediction is a critical task for telecom companies to identify customers who are likely to churn, allowing them to take proactive measures to retain those customers. I utilized the given dataset containing various customer attributes such as demographics, service usage, and contract details to train and evaluate several machine learning models. The approach involved data preprocessing, exploratory data analysis (EDA), feature engineering and model training using logistic regression, gradient boosting and random forest classifiers. Then optimized these models through hyperparameter tuning to improve their predictive performance. The results demonstrate the effectiveness of the tuned models in accurately predicting customer churn providing valuable insights for telecom companies to mitigate churn and enhance customer retention strategies.

**KEYWORDS:** Churn Prediction, Telecommunications, Machine Learning, Logistic Regression, Gradient Boosting, Random Forest, Hyperparameter Tuning.

## DECLARATION STATEMENT

---

I hereby declare that the research work reported in the dissertation entitled "PREDICTING CUSTOMER CHURN IN A TELECOMMUNICATION INDUSTRY" in partial fulfilment of the recruitment process of the honorable organization SpeakX is an authentic work carried out by me. I have not submitted this work elsewhere for any degree or drive.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's and SpeakX's Policy on plagiarism, intellectual property rights and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.



*Signature of Candidate*

**Sanjana Umrao**

**Reg. No.- 12107890**

# TABLE OF CONTENTS

<b>CONTENTS</b>	<b>PAGE NO.</b>
Cover Page	i
Abstract	ii
Declaration Statement	iii
Table of Contents	iv
<b>CHAPTER 1: INTRODUCTION</b>	<b>1</b>
<b>CHAPTER 2: PRESENT WORK</b>	<b>9</b>
<b>CHAPTER 3: CONCLUSION AND FUTURE SCOPE</b>	<b>20</b>

# CHAPTER 1

## INTRODUCTION

---

Customer churn or the loss of clients or subscribers is a critical issue for businesses across various sectors especially for service-oriented industries like telecommunications. Understanding and predicting customer churn can provide valuable insights for retaining customers and enhancing overall business performance.

In today's competitive market, acquiring new customers is often more expensive than retaining existing ones. Therefore, identifying customers who are likely to churn and implementing strategies to retain them is essential for maintaining a stable customer base and ensuring steady revenue growth.

This project focuses on developing a predictive model for customer churn using machine learning techniques. By analyzing customer data, I aim to identify key factors influencing churn and build a model that can accurately predict whether a customer is likely to leave. The dataset used for this analysis includes various customer attributes such as demographic information, account details and service usage patterns.

My approach involves several steps: preprocessing the data to handle missing values and encode categorical variables, exploring the data to understand customer behavior and creating new features that may enhance the prediction model. I will then train multiple machine learning models including logistic regression, random forest and gradient boosting to find the best-performing model. Hyperparameter tuning will be applied to optimize the models further.

The ultimate goal of this project is to provide actionable insights and a reliable predictive model that businesses can use to proactively address customer churn, improve customer satisfaction and enhance long-term profitability.

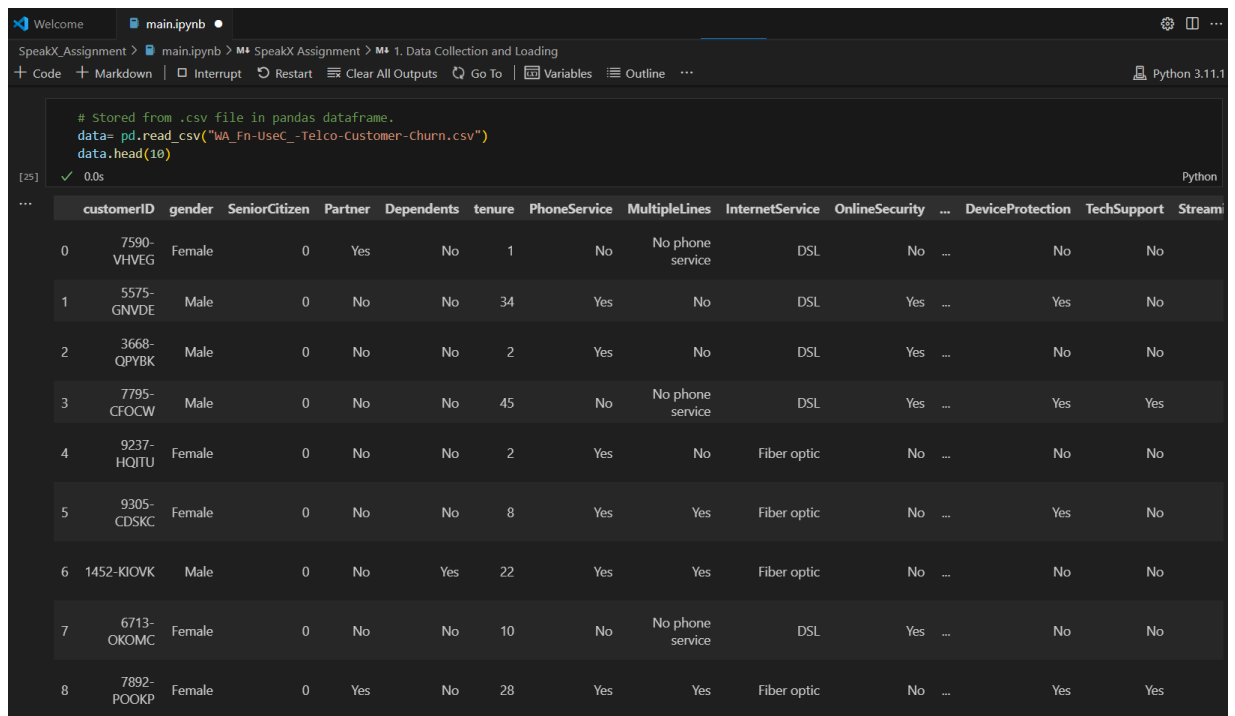
## CHAPTER 2

### PRESENT WORK

In this section, I conducted a comprehensive analysis and implementation of machine learning models for predicting customer churn in the telecom industry. The primary objective was to evaluate various machine learning algorithms to accurately predict churn and identify key factors influencing customer retention. The models examined included traditional algorithms like Logistic Regression, Gradient Boosting, Random Forest and Random Forest post hyperparameter tuning. At last, all the four models were evaluated on the different classification metrics as the problem statement involves the binary classification task.

### 3.1. DATA COLLECTION AND LOADING

The data was sourced from a Kaggle dataset whose link is given below. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>. It includes various customer-related attributes. The initial steps involved loading the dataset and then preparing it for analysis.



```
# Stored from .csv file in pandas dataframe.
data= pd.read_csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")
data.head(10)
```

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	...	DeviceProtection	TechSupport	Stream
0	7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL	No	...	No	No	
1	5575-GNVDE	Male	0	No	No	34	Yes	No	DSL	Yes	...	Yes	No	
2	3668-QPYBK	Male	0	No	No	2	Yes	No	DSL	Yes	...	No	No	
3	7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL	Yes	...	Yes	Yes	
4	9237-HQITU	Female	0	No	No	2	Yes	No	Fiber optic	No	...	No	No	
5	9305-CDSKC	Female	0	No	No	8	Yes	Yes	Fiber optic	No	...	Yes	No	
6	1452-KIOVK	Male	0	No	Yes	22	Yes	Yes	Fiber optic	No	...	No	No	
7	6713-OKOMC	Female	0	No	No	10	No	No phone service	DSL	Yes	...	No	No	
8	7892-POOKP	Female	0	Yes	No	28	Yes	Yes	Fiber optic	No	...	Yes	Yes	

Figure 1. Dataset is Loaded to the Notebook.

```

# Information regarding the dataset.
data.info()

[26] ✓ 0.0s

... <class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   customerID            7043 non-null   object
1   gender                 7043 non-null   object
2   SeniorCitizen          7043 non-null   int64
3   Partner                7043 non-null   object
4   Dependents             7043 non-null   object
5   tenure                 7043 non-null   int64
6   PhoneService           7043 non-null   object
7   MultipleLines          7043 non-null   object
8   InternetService        7043 non-null   object
9   OnlineSecurity         7043 non-null   object
10  OnlineBackup           7043 non-null   object
11  DeviceProtection       7043 non-null   object
12  TechSupport            7043 non-null   object
13  StreamingTV            7043 non-null   object
14  StreamingMovies        7043 non-null   object
15  Contract               7043 non-null   object
16  PaperlessBilling        7043 non-null   object
17  PaymentMethod          7043 non-null   object
18  MonthlyCharges         7043 non-null   float64
19  TotalCharges           7043 non-null   object
20  Churn                  7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB

```

Figure 2. Initial details of the dataset.

### 3.2. DATA PREPROCESSING

Data preprocessing was essential to ensure the dataset was ready for model training. Key steps included:

- Converting Numerical Columns: The TotalCharges column, which was incorrectly formatted as an object, was converted to a numeric type.
- Dropping Unwanted Columns: The customerID column was dropped as it did not contribute to the prediction.
- Encoding Categorical Columns: Categorical variables were converted to numerical values using techniques such as one-hot encoding. This increased the number of features from 21 to 43.
- Handling Missing Values: Missing values were addressed, primarily in the TotalCharges column, by filling them with the mean of the column.

```

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

# Handled missing values by replacing with the mean of the column.
data['TotalCharges'].fillna(data['TotalCharges'].mean(), inplace=True)

# Dropped the 'customerID' column.
data.drop('customerID', axis=1, inplace=True)

# Encoded binary categorical variables.
binary_columns = ['gender', 'Partner', 'Dependents', 'PhoneService', 'PaperlessBilling', 'Churn']
for col in binary_columns:
    data[col] = data[col].apply(lambda x: 1 if x == 'Yes' or x == 'Male' else 0)

# Encoded non-binary categorical variables using one-hot encoding.
dataset = pd.get_dummies(data)

# Displayed the first few rows of the preprocessed dataset.
print(dataset.head())

```

Figure 3. Code snippet for data preprocessing.

### 3.3. EXPLORATORY DATA ANALYSIS (EDA)

EDA provided initial insights into the data, highlighting the characteristics and patterns within the dataset:

- *Imbalance in Churn Data:* It was observed that the dataset is imbalanced, with approximately 75% of customers not churning.

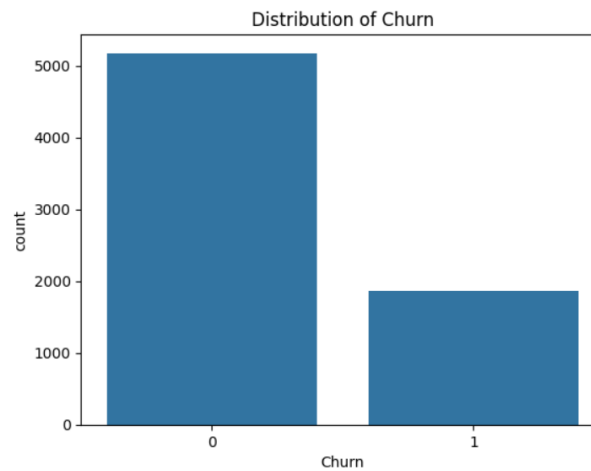


Figure 4. Distribution of Churn.

- *Univariate Analysis:*
  - *Tenure Distribution:* The distribution of tenure was multimodal, indicating varying lengths of customer relationships.
  - *Monthly Charges:* The monthly charges approximated a normal distribution after the initial bars, suggesting a central average charge for most customers.
  - *Total Charges:* The distribution was positively skewed, indicating a smaller group of customers with significantly higher charges.

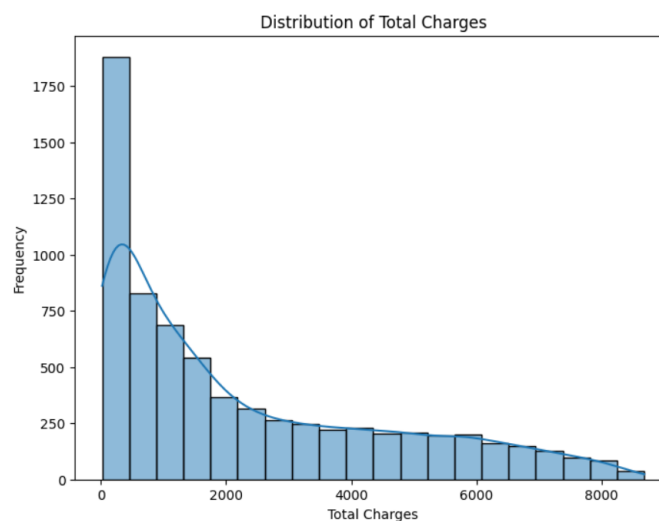


Figure 5. Positively skewed distribution for Total Charges.



- Bivariate Analysis:
  - *Churn vs. Tenure:* A boxplot analysis showed that customers with shorter tenures are more likely to churn. The boxplot for tenure identifies certain small outliers which are greater than the upper fence. Also the median tenure for 'Churned' customers is very less depicting they are much likely to churn.
  - *Churn vs. Monthly Charges:* A clear distinction in the monthly charges between churned and non-churned customers was observed. The median monthly charges as well as minimum value is less for 'Not Churned' customers depicting they are less likely to churn.
  - *Churn vs. Total Charges:* Similar to monthly charges, total charges also depicted a difference between the two groups. The total charges column is showing outliers mostly due to the large scale values.
- Multivariate Analysis:
  - Relationship between all the columns of the dataset was analysed using heatmap which is nothing but the visualised version of correlation matrix.
- Other Useful Insights:
  - Gender-wise proportion of customers is almost same for the company.
  - Female customers churned more than men but on very slight margin.
  - Company's customer domain include only around 15% as senior citizens and in which nearly 50% churned.
  - Around 75% of non-senior citizens stick with the company.
  - The number of such customers are more who are living as partners.
  - Among customers who were living with their partners, 70% were happy with the services of company which is more in comparison to the customers not living with partners.
  - There are only around 30% of customers who have dependants and out of those nearly 82% stick to the services provide by the telecom company.
  - Around 90% of the customers have opted for the phone service and out of those 25% were at the risk of churning.
  - At the same time around 90% percent of customers are retaining with the company if they haven't opted for phone service.
  - The churn count is more in customers who have access to paperless billing than compared to those who have not. The main reason could be that paperless billing slightly imposes that the customer is more advanced and might be using internet more where he/she could be available to other rival telecom company services.

- Churning rate is very less in customers whose payment method is set to automatic either by bank or by credit card.
- Churning rate is observed high in customers with month-to-month subscription that to 1 year or 2 year subscription.

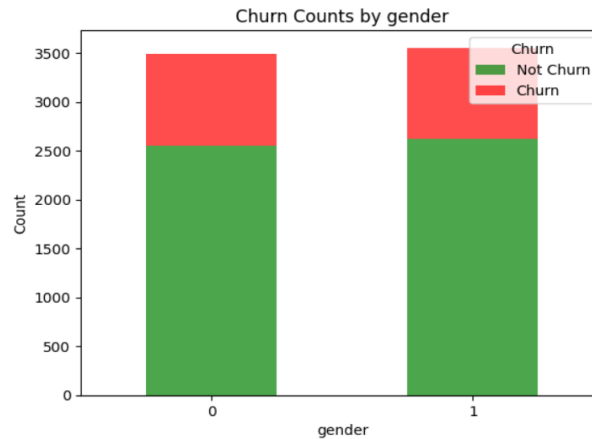


Figure 6. Churn count by gender.

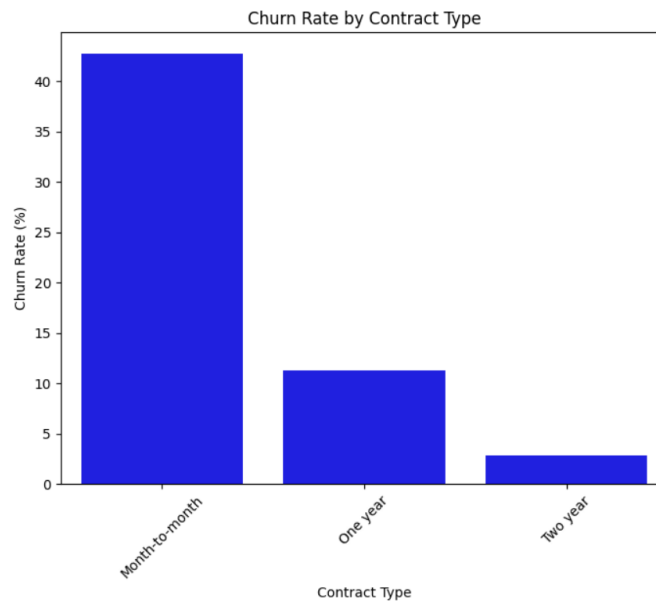


Figure 7. Churn rate by contract type.

### 3.4. FEATURE ENGINEERING

To enhance the predictive power of the model, two new features were engineered:

- Number of Services: This feature represents the total number of services subscribed to by each customer. It is hypothesized that customers with more services may have higher loyalty.
- Monthly Charges per Service: This feature calculates the average monthly charge per service for each customer, which may indicate higher value customers who are less likely to churn.

```
customers_with_2_or_more_services = dataset[dataset['TotalServices'] >= 2]

# Count the number of customers who churned after availing 2 or more services
num_churned_after_2_or_more_services = customers_with_2_or_more_services[customers_with_2_or_more_services['Churn'] == 1].shape[0]

print("Number of customers who churned after availing 2 or more services:", num_churned_after_2_or_more_services)

[40] ✓ 0.0s
... Number of customers who churned after availing 2 or more services: 1731
```

Figure 8. Code snippet of Analysis using Feature Engineering.

### 3.5. MODEL BUILDING AND EVALUATION

Four machine learning models were chosen for churn prediction:

1. Logistic Regression:
  - *Explanation:* Logistic Regression is a widely used statistical model that is simple, interpretable, and effective for binary classification problems. It models the probability of the default class (churn) using a logistic function.
  - *Performance:* Without hyperparameter tuning, the logistic regression model achieved an accuracy of 81.69%. The precision was 67.69%, recall was 58.98%, and the F1 score was 63.04%.
2. Gradient Boosting:
  - *Explanation:* Gradient Boosting is an ensemble technique that builds models sequentially, each new model correcting the errors made by the previous ones. It is known for its high performance with imbalanced datasets.
  - *Performance:* This model achieved an accuracy of 80.70%. The precision was 66.45%, recall was 54.69%, and the F1 score was 60.00%.
3. Random Forest:
  - *Explanation:* Random Forest is another ensemble technique that builds multiple decision trees and merges them together to get a more accurate and stable prediction. It is robust and effective for large datasets.
  - *Performance:* Initially, the Random Forest model achieved an accuracy of 78.99%. The precision was 64.00%, recall was 47.18%, and the F1 score was 54.32%.
4. Random Forest with Hyperparameter Tuning:
  - *Explanation:* After hyperparameter tuning using GridSearchCV, which exhaustively searches over a specified parameter grid to find the optimal model parameters, the Random Forest model's performance improved significantly.
  - *Performance:* The best model achieved an accuracy of 80.98%. The precision increased to 68.04%, recall to 53.08%, and the F1 score to 59.64%.

### 3.6. FINAL RESULT

At last I reached out at these 2 important result on the working of the models -

- Logistic regression emerges as the best performer in terms of accuracy and F1 score, making it the most reliable model for churn prediction in this analysis. However, there is a trade-off between precision and recall, indicating the potential need for further optimization or combining models to improve recall rates.

- The gradient boosting and tuned random forest models are also good alternatives.

TABLE 1. METRICS COMPARISON

S No	Comparative Analysis of Metrics				
	ML Models	Accuracy	Precision	Recall	F1-Score
1.	Logistic Regression	0.817	0.677	0.591	0.631
2.	Gradient Boosting	0.808	0.677	0.528	0.593
3.	Random Forest	0.791	0.64	0.472	0.543
4.	Random Forest Post Tuning	0.809	0.68	0.53	0.596

### 3.7. CHALLENGES FACED

Several challenges were encountered during the project:

- Imbalanced Dataset: The imbalance in the dataset required me to careful handling to ensure that the models were not biased towards the majority class which is nothing but Not\_Churned class.
- Feature Engineering: Also there was difficulty in identifying and creating meaningful features that could improve model performance as it requires domain knowledge and experimentation.
- Model Tuning: Hyperparameter tuning was computationally intensive and time-consuming but necessary to enhance model performance of the random forest classifier.

As logistic regression was best performing model it is also being run in .py python script file.

## **CHAPTER 3**

### **CONCLUSION AND FUTURE SCOPE**

---

#### **5.1 CONCLUSION**

In this study, I successfully developed machine learning models to predict customer churn, focusing on logistic regression, gradient boosting and random forest algorithms. My approach included thorough data preprocessing, exploratory data analysis and feature engineering resulting in insightful findings about customer behavior. The logistic regression model provided a strong baseline, while gradient boosting and a hyperparameter-tuned random forest model offered improved performance highlighting the importance of model tuning. Moving forward, the exploration of advanced features, deep learning models and real-time prediction systems, along with a focus on model interpretability and robust deployment can significantly enhance the predictive accuracy and practical application of churn prediction models.

#### **5.2 FUTURE SCOPE**

While my study successfully identified key predictors of customer churn and demonstrated the effectiveness of various machine learning models, there are several avenues for future research and improvements. Advanced feature engineering, including the exploration of additional features like customer interaction history and sentiment analysis of customer could provide deeper insights and improve model performance. Additionally, implementing advanced deep learning architectures such as Recurrent Neural Networks (RNNs) might capture complex patterns and temporal dependencies in the data potentially enhancing prediction accuracy.

