**📊 Predicting Customer Churn for a Telecom Company Using Logistic Regression**

**By [Shubham Jha]**

**1️⃣ Introduction**

Customer churn — when a customer stops using a company’s product or service — is a major challenge in the telecom industry. Retaining existing customers is often **cheaper than acquiring new ones**, so predicting churn can save companies significant revenue.

In this project, I built a **Logistic Regression** model to predict churn for a telecom company, aiming for a **simple, interpretable baseline** before exploring more complex models.

**2️⃣ Problem Statement**

The goal is to answer a key business question:

**"Can we predict which customers are at risk of churning?"**

If we can identify high-risk customers in advance, the company can take proactive steps, such as offering discounts or improving service quality, to retain them.

**3️⃣ Dataset Overview**

The dataset contains telecom customer details like:

* **Demographics:** Gender, Senior Citizen, Partner, Dependents
* **Service Information:** Internet service, Contract type, Payment method
* **Account Information:** Monthly charges, Total charges, Tenure
* **Target Variable:** Churn (Yes/No)

**4️⃣ Data Preprocessing**

Before training the model, I cleaned and prepared the data:

1. **Handled Missing Values** – Dropped/filled missing rows.
2. **Encoded Categorical Features** – Converted string categories into numerical form using one-hot encoding.
3. **Converted Data Types** – Changed TotalCharges from object to numeric (removed invalid entries).
4. **Feature Scaling** – Standardized numerical columns to ensure fair model weighting.

**5️⃣ Model Building: Logistic Regression**

Logistic Regression is ideal as a **baseline model** because:

* It’s interpretable (coefficients show feature importance)
* Works well for binary classification
* Quick to train

python

CopyEdit

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import train\_test\_split

from sklearn.metrics import accuracy\_score, confusion\_matrix, classification\_report

# Train-test split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Model training

log\_model = LogisticRegression(max\_iter=1000)

log\_model.fit(X\_train, y\_train)

# Predictions

y\_pred = log\_model.predict(X\_test)

**6️⃣ Model Evaluation**

I used **accuracy**, **confusion matrix**, and **classification report** to assess performance:

* **Accuracy:** ~80%
* **Precision & Recall:** Balanced performance
* **Confusion Matrix:** Shows the breakdown of true/false positives and negatives

Example:

| **Metric** | **Value** |
| --- | --- |
| Accuracy | 0.80 |
| Precision | 0.76 |
| Recall | 0.72 |
| F1-Score | 0.74 |

**7️⃣ Key Insights from Coefficients**

Logistic Regression coefficients indicate which features push churn probability **up** or **down**.

Some **strong churn indicators**:

* **Month-to-month contracts** → Higher churn
* **Electronic check payment method** → Higher churn
* **Low tenure customers** → Higher churn

Some **loyalty factors**:

* **Two-year contracts** → Lower churn
* **Higher tenure** → Lower churn

**8️⃣ Business Recommendations**

Based on these findings, the telecom company should:

✅ **Encourage longer-term contracts** — Offer discounts for yearly or multi-year plans  
✅ **Target month-to-month customers** — Special offers, loyalty perks  
✅ **Focus on early-stage customers** — Improve onboarding and engagement during first few months  
✅ **Reduce reliance on electronic checks** — Promote auto-pay or credit card payments

**9️⃣ Conclusion**

This Logistic Regression model provided:

* **80% accuracy**
* **Clear insights** into churn drivers
* **Actionable strategies** for reducing churn

While Logistic Regression is a great starting point, the next step could be testing **Random Forest** or **XGBoost** for potentially better accuracy while balancing interpretability.