# CUSTOMER CHURN PREDICTION:

A MACHINE LEARNING APPROACH FOR SYRIATEL.

# BUSINESS UNDERSTANDING.

Telecommunications companies like SyriaTel face a major challenge: customer churn — the phenomenon where customers stop using their services after a certain period. Every customer that leaves represents a loss in potential revenue and increases the cost of customer acquisition.

The goal of this project is to build a machine learning classifier that can predict whether a customer is likely to churn (i.e., stop doing business with SyriaTel "soon"). If successful, such a system could be used to proactively identify at-risk customers and take steps to retain them — through promotions, targeted outreach, or improved customer service

# DATA UNDERSTANDING.

Data Source:

- The dataset consists of customer records from SyriaTel, a telecommunications company.
- Each entry represents an individual customer and includes features that capture various aspects of their interaction with the company.

- The primary target variable is a field indicating whether or not a customer has churned.
- This data is structured to capture patterns that may correlate with customer churn, making it highly useful for our goal of predicting which customers are likely to leave the company.

# DATA PREPARATION.

The following steps were taken to prepare the data for modelling:

- Missing Values: Checked and confirmed that no missing values were present in the dataset.
- Duplicates: Verified that there were no duplicate rows based on a unique identifier
- Data Types & Unique Values: Reviewed the data types and unique values for all columns to understand the structure and detect potential issues.
- Outliers: Identified and handled outliers using boxplots to ensure numeric features were within reasonable ranges.
- Categorical Data Encoding: Applied one-hot encoding to transform categorical variables into numeric format, increasing the feature set from 21 to 69 columns.
- Unnecessary Columns: Removed the 'phone number' column, which was a unique identifier not useful for modeling.
- Data Type Conversion: Converted all boolean columns from True/False to 0/1 integers to ensure consistency with numeric features.
- Final Dataset Shape: The cleaned dataset now contains 3333 rows and 69 columns, ready for modeling.

# MODELING.

Modeling methods used;

- Logistic regression; Used to predict whether a customer will churn or not. It's simple, fast and provides interpretable probabilities for classification.
- Linear regression; Helps explore relationships between features and a numerical target like charges.
- Multiple linear regression; This is used for exploratory analysis with multiple features affecting a continuous outcome.
- Polynomial regression; This is used to model non-linear patterns in relationships.
- Ridge & Lasso; Used to improve model performance by reducing overfitting and managing multicollinearity.
- Random Forest; It handles categorical and numerical features, captures complex patterns and gives high accuracy with built in importance.

# MODEL COMPARISON SUMMARY

| Model | $R^2$ Score | MAE | MSE | RMSE | Notes |
|---|---|---|---|---|---|
| Linear Regression | 1.0000 | 0.0026 | 0.0000 | 0.0029 | Perfect fit (likely due to a direct formula relationship) |
| Multiple Linear Regression | 1.0000 | 0.0026 | 0.0000 | 0.0029 | Also perfect — confirms strong deterministic relationship |
| Polynomial Regression | 1.0000 | 0.0026 | 0.0000 | 0.0029 | No improvement over linear models (expected) |
| Ridge Regression | 1.0000 | 0.0037 | 0.0000 | 0.0045 | Excellent accuracy with slight regularization |
| Lasso Regression | 0.9999 | 0.0807 | 0.0101 | 0.1005 | Slight drop in accuracy; useful for feature selection |

# Final Model Selected

**Random Forest Classifier** was selected as the final model. It showed the best combination of predictive performance and practical usefulness in identifying customer churn risk.

**Random Forest** had a strong ROC AUC score of **0.94**, indicating excellent class separability.

Its accuracy was also high at **94%**.

It achieved **very high precision (0.97)** for classifying churners (class 1), meaning most flagged churners were truly at risk.

While recall (0.59 for churners) wasn't perfect, it still outperformed simpler models and offered a good balance.

This makes Random Forest the most **business-relevant** choice: it detects churners reliably while minimizing unnecessary alerts.

The final Random Forest model was trained using training data, validated using a validation split, and finally tested on a **completely separate holdout test set**, preventing data leakage.

**Final Test Results:**
- **Accuracy**: 94%
- **ROC AUC Score**: 0.94
- **Confusion Matrix**:

# Business Implications

The Random Forest model can help the business:

- Flag customers likely to churn in advance
- Target those customers with retention campaigns
- Reduce customer loss and improve revenue retention by providing meaningful utility in solving the problem of customer churn.

- **Identify high-risk churn customers early**, allowing for proactive retention efforts.
- **Understand key drivers of churn**, such as day/evening usage and customer service interactions.
- **Segment users** into behavior-based clusters using K-Means, enabling targeted marketing or service personalization.

# BUSINESS RECOMMENDATIONS.

Use model insights to build a churn alert system

Target high risk customers with retention offers

Monitor churn reduction impact overtime

Monitor model performance overtime. Regularly track the model's accuracy and recalibrate as customer behavior changes.

# THANK YOU

I APPRECIATE YOUR TIME AND ATTENTION.