**Report on Telecom Customer Churn Prediction**

**1. Project Overview**

The primary objective of this project is to develop a predictive model that can identify customers at risk of churning for a telecommunications company. By analysing customer data, we aim to enable the company to take proactive measures to retain these customers.

**2. Data Collection and Preprocessing**

The dataset used for this project was sourced from Kaggle and contains information on 7,043 customers, including 21 feature columns and one target column (`Churn`). The preprocessing steps included:

1. Handling missing values: Imputed missing `TotalCharges` values with the median.

2. Encoding categorical variables: Converted categorical features to numerical using one-hot encoding.

3. Normalizing numerical features: Applied standard scaling to numerical features for better model performance.

**3. Exploratory Data Analysis (EDA)**

**3.1. Initial Summary Statistics**

- \*\*Numerical Features\*\*:

- `tenure`: Number of months a customer has stayed with the company.

- `MonthlyCharges`: Monthly charges for the customer.

- `TotalCharges`: Total charges incurred by the customer.

- \*\*Categorical Features\*\*:

- `gender`, `Partner`, `Dependents`, `PhoneService`, `MultipleLines`, `InternetService`, `OnlineSecurity`, `OnlineBackup`, `DeviceProtection`, `TechSupport`, `StreamingTV`, `StreamingMovies`, `Contract`, `PaperlessBilling`, `PaymentMethod`.

**3.2. Univariate Analysis**

- \*\*Categorical Unordered\*\*: Features like `gender`, `Partner`, `Dependents` showed balanced distributions.

- \*\*Categorical Ordered\*\*: `Contract` type revealed most customers are on a month-to-month contract.

- \*\*Numerical Variables\*\*:

- `tenure`: Right-skewed distribution.

- `MonthlyCharges` and `TotalCharges`: Varied distributions indicating diverse billing amounts among customers.

**3.3. Bivariate Analysis**

- \*\*Numerical vs. Numerical\*\*:

- Strong correlation between `TotalCharges` and `MonthlyCharges`.

- \*\*Categorical vs. Categorical\*\*:

- Higher churn observed in customers with month-to-month contracts and those without tech support.

- \*\*Numerical vs. Categorical\*\*:

- Customers with higher `MonthlyCharges` tended to churn more.

**4. Feature Engineering**

Based on the EDA findings, key features were selected using various methods:

1. \*\*Correlation Matrix\*\*: Identified strongly correlated features with `Churn`.

2. \*\*Univariate Feature Selection (SelectKBest)\*\*: Selected top features based on statistical tests.

3. \*\*Recursive Feature Elimination (RFE)\*\*: Further refined feature set using a RandomForest classifier.

4. \*\*Feature Importance from RandomForest\*\*: Extracted important features directly from the model.

Selected Features:

- `Contract\_Month-to-month`

- `Contract\_Two year`

- `tenure`

- `MonthlyCharges`

- `TotalCharges`

- `PaymentMethod\_Electronic check`

- `InternetService\_Fiber optic`

- `OnlineSecurity\_Yes`

- `TechSupport\_Yes`

**6. Model Building and Evaluation**

A RandomForest classifier was chosen due to its robustness and ability to handle both numerical and categorical features. The dataset was split into training and testing sets with an 80-20 ratio.

\*\*Model Performance\*\*:

- \*\*Accuracy\*\*: 79%

- \*\*Precision\*\*: 82%

- \*\*Recall\*\*: 74%

- \*\*F1-Score\*\*: 78%

These metrics indicate a well-balanced model with good precision and recall, ensuring a reliable identification of customers at risk of churning.

**7. Challenges Faced**

1. \*\*Imbalanced Data\*\*: The target variable (`Churn`) was imbalanced. Techniques like stratified sampling and evaluation metrics suitable for imbalanced data (precision, recall) were employed.

2. \*\*Feature Selection\*\*: Balancing the trade-off between feature relevance and redundancy required multiple iterations of feature selection methods.

3. \*\*Model Overfitting\*\*: Ensured model generalizability by tuning hyperparameters and using cross-validation techniques.

**8. Conclusion**

The project successfully developed a churn prediction model with significant accuracy and reliability. Key insights from EDA helped in feature engineering and improving model performance. The model can assist the telecommunications company in implementing targeted retention strategies, potentially reducing churn rates and improving customer satisfaction.