Machine Learning Project Report: Customer Churn Prediction

1. Problem Definition

Objective: Predict whether a customer will churn (leave the telecom company).

Type of Problem: Binary Classification

Business Goal: Identify customers likely to churn so that retention strategies can be applied to reduce loss.

2. Data Collection

Dataset Used: Telco Customer Churn Dataset

Source: Public dataset hosted on GitHub

Data Characteristics:

- Customer demographics
- Account details
- Subscription and service usage
- Churn label: Yes/No

3. Data Exploration (EDA)

Key Activities:

- Checking for nulls and data types
- Summary statistics (mean, median, etc.)
- Target class distribution (churn vs no churn)
- Visualizations: count plots, heatmaps, correlation matrix

Findings:

- Class imbalance exists (more non-churners)
- Some columns had missing or invalid values (e.g., TotalCharges)

4. Data Preprocessing

Steps Performed:

- Dropped irrelevant columns (e.g., customerID)
- Converted 'TotalCharges' to numeric and removed NaNs
- Encoded categorical features using Label Encoding

- Target column 'Churn' mapped to binary (Yes: 1, No: 0)
- Features scaled using StandardScaler
- Data split: 80% training, 20% testing (stratified)

5. Feature Engineering

Techniques Used:

- Feature selection using SelectKBest (top 10 features)
- Dimensionality reduction for model efficiency

6. Model Selection

Model Chosen: Random Forest Classifier

Why Random Forest?

- Handles both numerical and categorical data well
- Robust to overfitting due to ensemble of trees
- Can calculate feature importance

Baseline Comparison: Considered Decision Trees; Random Forest outperformed in cross-validation.

7. Model Training

Hyperparameter Tuning:

- Used GridSearchCV with cross-validation (cv=5)
- Parameters tuned:
  - n_estimators: [100, 200]
  - max_depth: [5, 10, None]

Best Estimator Selected: Output of the grid search with the highest F1 score.

8. Model Evaluation

Metrics Used:

Machine Learning Project Report: Customer Churn Prediction

- Confusion Matrix
- Classification Report (Precision, Recall, F1-Score)
- ROC-AUC Score

Results:

- High F1 score for both classes
- ROC-AUC score indicated strong separability
- No major signs of overfitting

Final Outcome

- Model saved as `best_rf_model.pkl`
- Ready for deployment or integration in business tools for churn prediction.

Tools & Libraries

- Python
- Pandas, NumPy
- Scikit-learn
- Seaborn, Matplotlib
- Jupyter Notebook (for presentation)

Prepared By

[Your Name]

[Date]