Churn Machine Learning Using Python Report

IDS 472 Business Data Mining – SP 2025

1) Introduction

For telecom firms, “Churn” prediction is an essential component of customer relationships management. Significant revenues losses are caused by customer attrition, particularly if valuable clients are lost. The goal of this project is to create prediction models that use account information, service usage, and customer demographics to identify probable churners. Supporting the telecom company in putting proactive retention measures into practice is the goal.

2) Data Description

a) Dataset Summary:

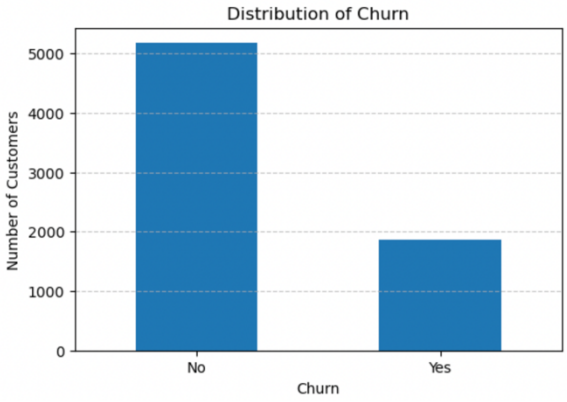
The dataset contains 7043 record and 21 features including:

- **Service:** phone, tech support, streaming services and internet
- **Demographics:** gender, partner, dependents, senior citizen
- **Charges:** monthly and total
- **Account Information:** tenure, contract type, paperless billing, payment method
- **Target Variables:** Churn (Yes/No)

b) Preprocessing Steps:

1. Load and Clean Data:
    - Dataset Telco_Customer_Churn.csv is loaded into Jupyter Notebooks
    - The TotalCharges column is converted to numeric values, with invalid entries replaced by “NaN”.
    - Missing values in TotalCharges are handle by filling with the median value of the whole column.
2. Perform statistical summaries and visualizations to understand patterns
    - A bar chart base on the distribution of churned and non-churned number of customers with dataset is created.

- Statistical summaries (mean, standard deviation, min, max) of 3 columns which critical for churn analysis for 2 category churn and not churn is generated (tenure, MonthlyCharges, TotalCharges)



Statistical summaries:

		mean	std	min	max
Churn					
No	MonthlyCharges	61.265124	31.092648	18.25	118.75
	TotalCharges	2552.882494	2327.590132	18.80	8672.45
	tenure	37.569965	24.113777	0.00	72.00
Yes	MonthlyCharges	74.441332	24.666053	18.85	118.35
	TotalCharges	1531.796094	1890.822994	18.85	8684.80
	tenure	17.979133	19.531123	1.00	72.00

3. Transform categorical variables (one-hot encoding, label encoding)
    - Converting “Yes”, “No” values of label-encoded column Churn into numeric values
    - Transforming categorical columns into one-hot encoded variables
4. Split data into training (80%) and testing (20%) sets
    - The dataset is split into 80% training and 20% testing subsets.
    - The Churn column is set as the target variable so exclude from features that used for modeling

3) Feature Engineering

a. Create new features if needed (e.g., tenure groups, monthly spend per service)
    i. Tenure Groups:
        1. Customers were binned into three groups (0-12 months); (13-24 months); 25+ months for easier analysis
    ii. Monthly Spend Per Service:
        1. Observe services that generate fees include: 'PhoneService_Yes', 'InternetService_Fiber optic', 'StreamingTV_Yes', 'StreamingMovies_Yes'
        2. Calculate sum count paid service of each customer
        3. If they don’t have any paid service, set the value to $0

b. Use correlation analysis and feature importance to select key predictors
    i. Measures absolute correlation relationship between churn and categorical feature
    ii. Ranks them from strongest to weakest
    iii. Displays the top 10 key predictors

```
Top Correlated Features:
tenure                                  0.345593
InternetService_Fiber optic             0.312656
PaymentMethod_Electronic check          0.309214
Contract_Two year                       0.302209
StreamingMovies_No internet service     0.228929
TechSupport_No internet service         0.228929
DeviceProtection_No internet service    0.228929
OnlineBackup_No internet service        0.228929
OnlineSecurity_No internet service      0.228929
InternetService_No                      0.228929
dtype: float64
```

4) Methodology

a. Three of the methods
    - Logistic Regression (baseline model)
    - Random Forest (handles complex patterns)
    - k-Nearest Neighbors (k-NN) (distance-based prediction)

2. Performance Evaluation for each model
    - Accuracy
    - Precision, Recall, F1 Score
    - AUC-ROC
    - Confusion Matrix

3. Select Best Model By AUC

5) Results

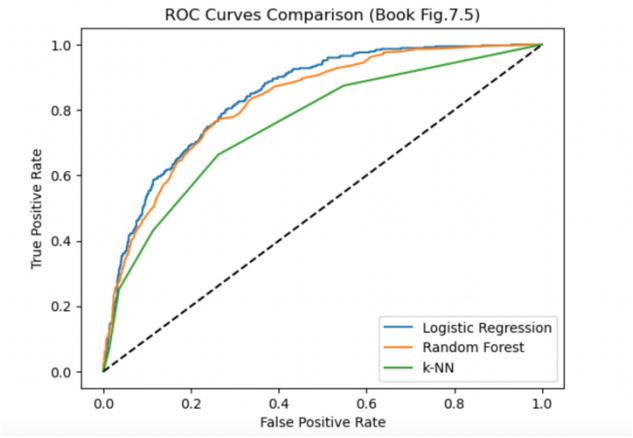
a. Model Performance Summary
    i. The table below compares the performance metrics of Logistic Regression, Random Forest, and k-NN models.

b. Best Model:
    i. From table metric compare between three models and ROC Curves, logistic regression outperforms the other two models in term of predicting

churn. To be more specific, it returns the highest accuracy about 0.804 and AUC about 0.842

ii. The ROC curve also illustrates the same that Logistic Regression showing the best performance that its line colored in blue above other lines.

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	0.804	0.655	0.553	0.600	0.843
Random Forest	0.786	0.620	0.497	0.552	0.826
k-NN	0.766	0.579	0.430	0.494	0.755



6) Insights

Top Predictors of Churn:

a. Tenure
   i. Impact: New customers more likely to churn
b. Fiber Internet Service
   i. Impact: higher churn compared to DSL customers
c. Electronic Check Payments
   i. Impact: higher churn rate about 58% compared to 15% for autopay
d. Month-to-Month Contracts
   i. Impact: 28% churn vs 3% for 2-year contracts
e. No Movie Streaming
   i. Impact: 22% higher churn among fiber customers

7) Business Improvement Ideas & Limitation:

Improvement:

- Encourage long-term agreements to lower month-to-month customer attrition
- Increase retention, combine value-added service (such tech help & online security) with savings
- Use onboarding efforts to target new users during the crucial churn window of their first year
- Keep an eye on high spenders with short stay; they might leave since they feel their money isn't worth it

Limitation:

- Temporal constraints: Churn patterns may change but the dataset is static.
- No information about complaints or customer satisfaction is provided.

- For more in-depth analysis, future research might use call center records or time-series activity.