HW4

Classification on the Telco-churn dataset

Context
"Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs."

Content
Each row represents a customer, each column contains customer's attributes described on the column Metadata.
The data set includes information about:
Customers who left within the last month – the column is called Churn
Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
Demographic info about customers – gender, age range, and if they have partners and dependents

Feature Description

customerID
Customer ID

gender
Whether the customer is a male or a female

SeniorCitizen
Whether the customer is a senior citizen or not (1, 0)

Partner
Whether the customer has a partner or not (Yes, No)

Dependents
Whether the customer has dependents or not (Yes, No)

tenure
Number of months the customer has stayed with the company

PhoneService
Whether the customer has a phone service or not (Yes, No)

MultipleLines
Whether the customer has multiple lines or not (Yes, No, No phone service)

InternetService
Customer’s internet service provider (DSL, Fiber optic, No)

OnlineSecurity
Whether the customer has online security or not (Yes, No, No internet service)

OnlineBackup
Whether the customer has online backup or not (Yes, No, No internet service)

DeviceProtection
Whether the customer has device protection or not (Yes, No, No internet service)

TechSupport
Whether the customer has tech support or not (Yes, No, No internet service)

StreamingTV
Whether the customer has streaming TV or not (Yes, No, No internet service)

StreamingMovies
Whether the customer has streaming movies or not (Yes, No, No internet service)

Contract
The contract term of the customer (Month-to-month, One year, Two year)

PaperlessBilling
Whether the customer has paperless billing or not (Yes, No)

PaymentMethod
The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

MonthlyCharges
The amount charged to the customer monthly

TotalCharges
The total amount charged to the customer

Churn
Whether the customer churned or not (Yes or No)

Q1. Visualize the univariate distribution of each continuous feature, and the distribution of the target.

Q2. Split data into training (90%) and test set (10%) using random state = 0. Build a pipeline for dealing with categorical variables. Evaluate Logistic Regression, linear support vector machines and nearest neighbors using 5-fold cross-validation with their default parameters. How different are the results? How does scaling the continuous features with StandardScaler influence the results?

Q3. Tune the key parameters (C for Logistic Regression and linear support vector machines, and K for nearest neighbors) using GridSearchCV. Do the results improve?

Visualize the performance as function of the parameters for all three models.