Churn Prediction Project

Kaggle > Telecom Customer Churn

Churn Rate

classification → Yes/No

One-hot encoding

Scikit-Learn

$$\frac{1}{1+e^{-x}} = \text{Sigmoid}(x) = y$$

$$0 < y < 1$$

Risk Ratio

1) Data Preparation
2) Set validation Framework
3) EDA
4) Feature Importance
5) One-hot encoding
6) Logistic Regression
7) Model Interpretation
8) Using the model

Correlation Coefficient

$-1 \leq r \leq 1$ → Pearson correlation coef.

(*)

0 to -0.2	-0.2 to -0.5	-0.5 to -1
0 to 0.2	0.2 to 0.5	0.5 to 1
Low	Moderate	Strong / often

(*)

$r = 0$ zero correlation

$r < 0$ -ve "

$r > 0$ +ve "

Mutual Information (MI)

- used for feature selection
- high MI means strong relationship b/n 2 variable
- Zero MI means variables are independent
- It quantifies the statistical dependence b/n 2 variables by measuring how much knowing one variable reduces uncertainty over other (entropy)

One-Hot Encoding
↓
DictVectorizer()

#Sparse Matrix