ACTIVITY #2

Feature Selection

TOPICS

2.1 Data Exploration

2.2 Remove variables with High Variable Correlation

2.3 Variable Chi-square with High p-value

LIBRARIES

1. import numpy as np
2. import pandas as pd
3. import matplotlib.pyplot as plt
4. import seaborn as sns
5. from sklearn import preprocessing
6. from sklearn.feature_selection import chi2

2.1

Data Exploration and Transform

2.1 Data exploration

1. Read .csv file
   - read_csv("https://raw.githubusercontent.com/srivatsan88/YouTubeLI/master/dataset/churn_data_st.csv",sep=",")
2. View Data Array Shape
   - # Variables
   - # Samples
3. Remove
   - 'customerID'
4. View Variable info
   - Info()
     - Data Type / # non null
5. Fill NA
   - fillna()

2.2

Remove variables with High Variable Correlation

2.2 Remove variables with High Variable Correlation

1. Create data frame of continuous data columns
   - columns = ['tenure','ServiceCount', 'MonthlyCharges','TotalCharges']
2. Calculate correlation between variables
   - Corr()
3. Plot Headmap
   - Sns.heatmap()
4. Reduce Corr() to Lower Matrix
   - lower = pd.DataFrame(np.tril(dataCorr, -1),columns = dataCorr.columns)
5. Drop columns if correlation value > 0.6
   - to_drop = [column for column in lower if any(lower[column] > 0.6)]
   - df.drop(to_drop, inplace=True, axis=1)
6. Show statistics
   - Describe()

2.2 Results

	ServiceCount	TotalCharges
0	2	29.85
1	4	1889.50
2	4	108.15
3	4	1840.75
4	2	151.65
...	...	...
7038	8	1990.50
7039	7	7362.90
7040	2	346.45
7041	3	306.60
7042	7	6844.50

[7043 rows x 2 columns]

	ServiceCount	TotalCharges
count	7043.000000	7016.000000
mean	5.446259	2282.589168
std	1.964916	2265.506114
min	1.000000	18.800000
25%	4.000000	401.925000
50%	6.000000	1397.100000
75%	7.000000	3792.325000
max	9.000000	8684.800000

2.3

Remove Variable with High p-value from Chi-square

2.3 Remove Variable with High p-value from Chi-square

1. Create data frame of discrete data columns
   - columns = ['Churn']
2. Data Transform (Category to number)
   - LabelEncoder()
3. Separate Variables and Output
   - Output = [‘Churn’]
   - Variables = ['gender','Contract','PaperlessBilling’]
4. Calculate Chi-Square, p_value
   - Chi_table = chi2(Variables,Output)
   - Print(Chi_table)
5. Select insignificant variables with p_value > 0.05 (5%)
   - p_value = Chi_table[1]
   - to_drop = [column for column in lower if any(p_value[column] > 0.05)] #5% significant
6. Create final data table
   - Continuous data, category data

2.3 Results

	ServiceCount	TotalCharges	gender	PaperlessBilling	Contract	Churn
0	2	29.85	0	1	0	0
1	4	1889.50	1	0	0	0
2	4	108.15	0	1	1	1
3	4	1840.75	1	0	0	0
4	2	151.65	0	1	1	1
...	...	...	...	...	...	...
7038	8	1990.50	1	1	0	0
7039	7	7362.90	1	1	0	0
7040	2	346.45	0	1	0	0
7041	3	306.60	0	1	1	1
7042	7	6844.50	2	1	0	0