Group Name: Individual

Name: Maria Contractor

Email, Country, College, Specialization: mariacontractor5@gmail.com , USA , University of Central Florida, Data Science

Batch Code: LISUM22

Submission Date: August 2, 2023

Submitted to: Data Glacier

Group Project: Week 9

Instructions:

1. Data cleansing and transformation done on the data.

2. Try at least 2 techniques to clean the data (for NA values : mean/median/mode/Model based approach to handle NA value/WOE and like this try different techniques to identify and handle outliers as well)

Steps

1. Data Cleansing and Transformation Done on the Data & Techniques

Performed an analysis on the data based on numerical and categorical data. From this, I was able to point out all of the outliers in the numerical data through box and whisker plots. Furthermore, I was able to determine how many unknown variables there were in each of the categorical data distributions.

From this analysis, I cleansed and transformed the data using two techniques. My first technique was imputing the categorical data, where I focused on age, job, and education. Each of these factors allowed me to decrease the amount of unknown variables by predicting where some of them may be placed based on the existing data.

Before:

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown
admin.	77	151	499	3329	1	363	5753	249
blue-collar	2318	1426	3623	878	8	453	94	454
entrepreneur	137	71	210	234	2	135	610	57
housemaid	474	77	94	174	1	59	139	42
management	100	85	166	298	0	89	2063	123
retired	597	75	145	276	3	241	285	98
self-employed	93	25	220	118	3	168	765	29
services	132	226	388	2682	0	218	173	150
student	26	13	99	357	0	43	170	167
technician	58	87	384	873	0	3320	1809	212
unemployed	112	34	186	259	0	142	262	19
unknown	52	22	31	37	0	12	45	131

After:

education job	basic.4y	basic.6y	basic.9y	high.school	illiterate	professional.course	university.degree	unknown
admin.	77	151	499	3329	1	363	5753	249
blue-collar	2366	1448	3654	878	8	453	94	454
entrepreneur	137	71	210	234	2	135	610	57
housemaid	516	77	94	174	1	59	139	0
management	100	85	166	298	0	89	2186	0
retired	601	75	145	276	3	243	286	112
self-employed	93	25	220	118	3	168	765	29
services	132	226	388	2832	0	218	173	0
student	26	13	99	357	0	43	170	167
technician	58	87	384	873	0	3320	1809	212
unemployed	112	34	186	259	0	142	262	19
unknown	0	0	0	37	0	10	44	117

After this, I cleansed the data by switching all of the unknown variables into NaN variables to make it easier to read/adjust if necessary.