Machine Learning 441 Assignment 1

David Nicolay (26296918) 6 August 2025

Task 1: Analytics Base Table

1. 581012
2. 61
3. Continuous features
4. Target feature data quality report:

Task 2: Data Quality Issues

Feature	Data Quality Issue	Justification
A1	Missing Values	Many records have nulls, impacting model accuracy.
A2	High Cardinality	Too many unique values may cause overfitting.

Task 3: Addressing The Data Quality Issues

References

Table 1: Data Quality Report for Continuous Features

Feature	Count	Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	
0	581012	0	1978	2054845.65	3104928.15	3271134.43	3311628.60	3496222.05	426
2	581012	0	576099	0.00	145.49	389.92	318.12	652.53	
5	581012	0	581012	-173.07	6.99	46.42	29.91	68.97	
7	581012	0	5811	0.00	1106.00	8158.11	1997.00	3328.00	51016
10	581012	0	255	0.00	119.00	142.53	143.00	168.00	
13	581012	0	2	0.00	0.00	0.05	0.00	0.00	
14	581012	0	2	0.00	0.00	0.44	0.00	1.00	
16	581012	0	1	0.00	0.00	0.00	0.00	0.00	
18	581012	0	2	0.00	0.00	0.50	0.00	1.00	
21	581012	0	2	0.00	0.00	0.01	0.00	0.00	
22	581012	0	2	0.00	0.00	0.01	0.00	0.00	
23	581012	0	2	0.00	0.00	0.02	0.00	0.00	
24	581012	0	2	0.00	0.00	0.00	0.00	0.00	
25	581012	0	2	0.00	0.00	0.01	0.00	0.00	
26	581012	0	2	0.00	0.00	0.00	0.00	0.00	
27	581012	0	2	0.00	0.00	0.00	0.00	0.00	
28	581012	0	2	0.00	0.00	0.00	0.00	0.00	
29	581012	0	2	0.00	0.00	0.06	0.00	0.00	
30	581012	0	2	0.00	0.00	0.02	0.00	0.00	
31	581012	0	2	0.00	0.00	0.05	0.00	0.00	
32	581012	0	2	0.00	0.00	0.03	0.00	0.00	
33	581012	0	2	0.00	0.00	0.00	0.00	0.00	
34	581012	0	2	0.00	0.00	0.00	0.00	0.00	
35	581012	0	2	0.00	0.00	0.00	0.00	0.00	
36	581012	0	2	0.00	0.00	0.01	0.00	0.00	
37	581012	0	2	0.00	0.00	0.00	0.00	0.00	
38	581012	0	2	0.00	0.00	0.01	0.00	0.00	
39	581012	0	2	0.00	0.00	0.02	0.00	0.00	
40	581012	0	2	0.00	0.00	0.00	0.00	0.00	
41	581012	0	2	0.00	0.00	0.06	0.00	0.00	
42	581012	0	2	0.00	0.00	0.10	0.00	0.00	
43	581012	0	2	0.00	0.00	0.04	0.00	0.00	
44	581012	0	2	0.00	0.00	0.00	0.00	0.00	
45	581012	0	2	0.00	0.00	0.00	0.00	0.00	
46	581012	0	2	0.00	0.00	0.00	0.00	0.00	
47	581012	0	2	0.00	0.00	0.00	0.00	0.00	
48	581012	0	2	0.00	0.00	0.20	0.00	0.00	
49	581012	0	2	0.00	0.00	0.05	0.00	0.00	
50	581012	0	2	0.00	0.00	0.04	0.00	0.00	
51	581012	0	2	0.00	0.00	0.09	0.00	0.00	
52	581012	0	2	0.00	0.00	0.08	0.00	0.00	
53	581012	0	2	0.00	0.00	0.00	0.00	0.00	
54	581012	0	2	0.00	0.00	0.00	0.00	0.00	
55	581012	0	2	0.00	0.00	0.00	0.00	0.00	
56	581012	0	2	0.00	0.00	0.00	0.00	0.00	
57	581012	0	2	0.00	0.00	0.03	0.00	0.00	
58	581012	0	2	0.00	0.00	0.02	0.00	0.00	
59	581012	0	2	0.00	0.00	0.02	0.00	0.00	
60	581012	0	581012	1.00	145253.75	290506.50	290506.50	435759.25	58
61	581012	0	1	1.00	1.00	1.00	1.00	1.00	

Table 2: Data Quality Report for target feature

Feature	Count	Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
T	580952	60	7	1.00	1.00	2.05	2.00	2.00	7.00	1.40