Phase 2: Prep

Joseph Nelson Farrell & Michael Massone
DS 5220 Supervised Machine Learning
Northeastern University
Professor Steven Morin, PhD

Overview

This slide deck will communicate the following:

- Machine learning attributes configuration
  - Our findings related to attributes that will be included in machine learning and attributes that will be excluded from machine learning
- Numerical column data transformations
  - The specific transformations that will be applied to ML numerical columns.
- Categorical column data transformations
  - The specific transformations that will be applied to ML categorical columns.
- Transformation outcome
  - The outcome of the numerical and categorical transformations on the train set.

Data

Dataset : Customer Segmentation

Origins : https://www.kaggle.com/datasets/vetrirah/customer?select=Train.csv

Size : (8068, 11)

- Instances: 8068
- Attributes: 11

Description :

- The dataset contains information about customers of an automobile company segmented into 4 classes (target).

Attributes

Attribute Name	AttributeType	Percent Missing Values	ML Attribute Designation
index	Numerical - Discrete	0.0%	non_ML
ID	Numerical - Discrete	0.0%	non_ML
Gender	Categorical - Nominal	0.0%	ML
Ever_Married	Categorical - Nominal	1.7%	ML
Age	Numerical - Discrete	0.0%	ML
Graduated	Categorical - Nominal	0.9%	ML
Profession	Categorical - Nominal	1.6%	ML
Work_Experience	Numerical - Discrete	10.0%	ML
Spending_Score	Categorical - Ordinal	0.0%	ML
Family_Size	Numerical - Discrete	4.1%	ML
Var_1	Categorical - Nominal	0.9%	ML

ML Attribute Selection

ML Attributes

Numerical Attributes:

- Age
- Work_Experience
- Family_Size

Categorical Attributes:

- Gender
- Ever_Married
- Graduated
- Profession
- Spending_Score
- Var_1

Total Attributes:

- 9

Non-ML Attributes

Non-ML Attributes List:

- ID

Missingness Drop List:

- None
- No attributes missing > 20 % of the observations

ML Attributes Drop List:

- None
- No attributes were identified during EDA to exclude.

Total Attributes:

- 1

Attribute Transformation

Categorical

Data Imputation:

- Fill missing values with the most frequent

Data Encoding:

- Target encoding
- Transforms single categorical column into n columns, where n equals the number of target classes.
- Values represent the probability of the original value being a member of a particular class.

Numerical

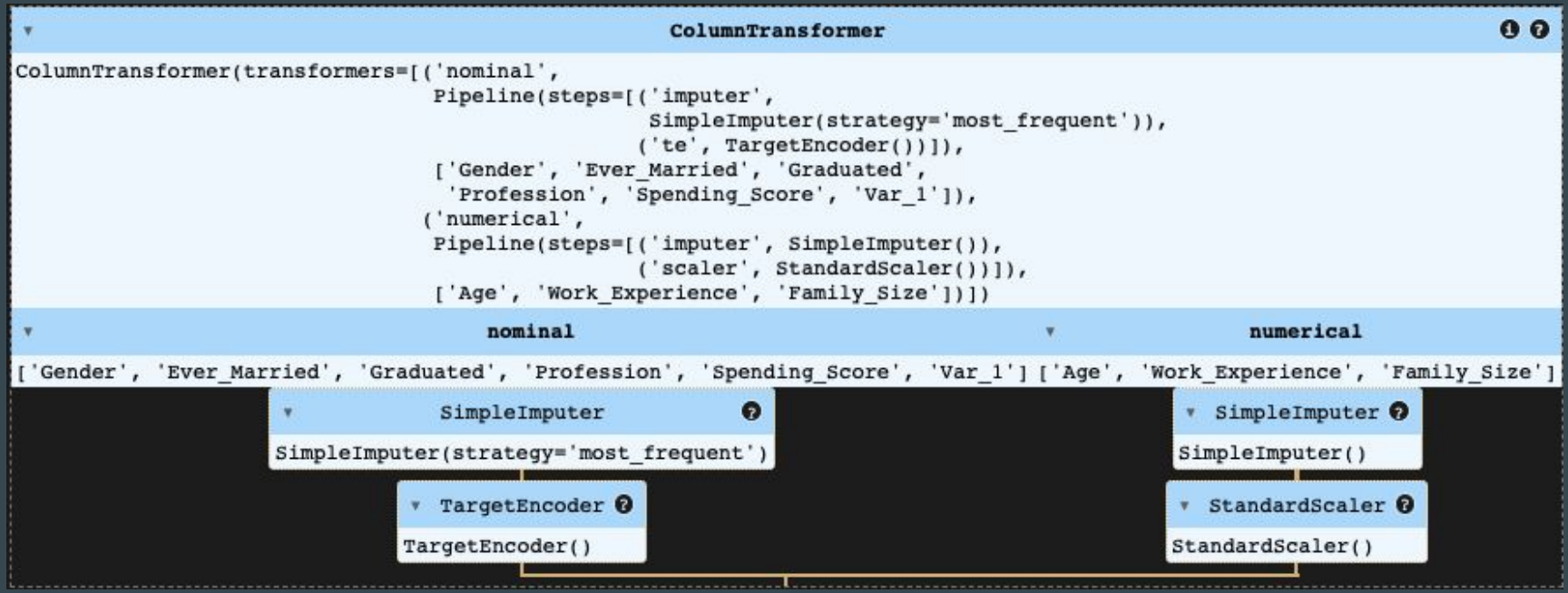
Data Imputation:

- Fill missing values with the mean.

Data Scaling:

- Standard scaling
- De-mean
- Whiten

Finalized Column Transformer:



Data Sample - Pre Transformations

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1
917	465905	Female	No	32	Yes	Artist	9.0	Low	1.0	Cat_6
3398	462903	Male	Yes	72	Yes	Entertainment	NaN	Average	2.0	Cat_6
2045	467901	Female	No	33	Yes	Entertainment	1.0	Low	4.0	Cat_6
8060	463613	Female	Yes	48	Yes	Artist	0.0	Average	6.0	Cat_6
4604	459859	Female	Yes	28	No	Doctor	9.0	Low	1.0	Cat_7
...	...	...	...	...	...	...	...	...	...	...
3822	463101	Female	No	27	No	Homemaker	8.0	Low	1.0	Cat_6
5864	467844	Male	No	37	Yes	Healthcare	0.0	Low	2.0	Cat_6
3589	460706	Female	No	27	No	Engineer	6.0	Low	6.0	Cat_4
1489	464339	Male	No	26	No	Artist	0.0	Low	2.0	Cat_6
2661	459407	Female	No	37	Yes	Doctor	0.0	Low	3.0	Cat_6

Shape: (6454, 10)

Data Sample - Post Transformations

	0	1	2	3	4	5	6	7	8	9	...	17	18	19	20	21	22	23	24	25	26
0	0.246725	0.237006	0.250948	0.265321	0.243218	0.147315	0.117767	0.491637	0.243003	0.267905	...	0.184781	0.133056	0.407085	0.232460	0.237425	0.283613	0.246501	-0.695320	1.942754	-1.227022
1	0.242490	0.224609	0.238557	0.294344	0.245240	0.285555	0.328461	0.140718	0.243003	0.267905	...	0.295321	0.452007	0.069321	0.232460	0.237425	0.283613	0.246501	1.703982	0.000000	-0.560068
2	0.243851	0.238758	0.249785	0.267605	0.251069	0.146765	0.117233	0.484873	0.246483	0.267266	...	0.183260	0.132564	0.403355	0.231114	0.234355	0.285701	0.248828	-0.635337	-0.513120	0.773838
3	0.245887	0.238714	0.253897	0.261501	0.243990	0.286529	0.325829	0.143627	0.246287	0.266077	...	0.290604	0.453880	0.077353	0.233433	0.236045	0.284289	0.246230	0.264401	-0.820105	2.107745
4	0.245887	0.238714	0.253897	0.261501	0.243990	0.286529	0.325829	0.143627	0.241845	0.169885	...	0.183341	0.137358	0.402381	0.255554	0.255434	0.195799	0.293140	-0.935250	1.942754	-1.227022
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6449	0.246725	0.237006	0.250948	0.265321	0.243218	0.147315	0.117767	0.491637	0.246861	0.166344	...	0.184781	0.133056	0.407085	0.232460	0.237425	0.283613	0.246501	-0.995233	1.635769	-1.227022
6450	0.245189	0.223096	0.239132	0.292583	0.251069	0.146765	0.117233	0.484873	0.246483	0.267266	...	0.183260	0.132564	0.403355	0.231114	0.234355	0.285701	0.248828	-0.395407	-0.820105	-0.560068
6451	0.246725	0.237006	0.250948	0.265321	0.243218	0.147315	0.117767	0.491637	0.246861	0.166344	...	0.184781	0.133056	0.407085	0.307015	0.216637	0.096543	0.379651	-0.995233	1.021801	2.107745
6452	0.243553	0.222782	0.236034	0.297630	0.245563	0.146791	0.123775	0.483813	0.241845	0.169885	...	0.183341	0.137358	0.402381	0.233433	0.236045	0.284289	0.246230	-1.055215	-0.820105	-0.560068
6453	0.245887	0.238714	0.253897	0.261501	0.245563	0.146791	0.123775	0.483813	0.246287	0.266077	...	0.183341	0.137358	0.402381	0.233433	0.236045	0.284289	0.246230	-0.395407	-0.820105	0.106885

Outcome Shape: (6454, 27)

- The 6 categorical attributes have been expanded to 24 attributes.
- Number of categorical attributes * Number of target classes = 24
- 24 + Numerical Attributes = 27

Acknowledgements

1. Steven Morin PhD., DS 5220 Class Materials
2. Sklearn.org

Thank you!