General steps for data preprocessing:

1. transform and separate target values ('y' -> [0, 1])

2. drop unnecessary features

3. group / bin features:

- group job

- group education

- bin age

4. Encode values:

- age:

(1) leave as it is,

(2) delete rows with outliers (GV),

(3) bin and WoE encode

- job:

(1) OrdinalEncoder/OneHotEncoder,

(2) WoE encode with 'unknown' as category,

(3) impute 'unknown' + WoE encode,

(4) group freq<2000 (GV)

- marital:

(1) OrdinalEncoder/OneHotEncoder,

(2) WoE encode with 'unknown' as category,

(3) impute 'unknown' + WoE encode

- education:

(1) OrdinalEncoder/OneHotEncoder,

(2) WoE encode with 'unknown' as category,

(3) impute 'unknown' + WoE encode,

(4) group freq<4000 (GV)

- default:

(1) drop,

(2) WoE encode with 'unknown' as category,

(3) impute 'unknown' + WoE encode,

(4) unknown->no [+ 1/0 encode]

- housing:

(1) yes->1, no->0, unknown->1(yes),

(2) WoE encode with 'unknown' as category,

(3) impute + WoE encode,

(4) drop

- loan:

(1) yes->1, no->0, unknown->0(no),

(2) WoE encode with 'unknown' as category,

(3) impute + WoE encode,

(4) drop

- contact:

(1) cellular->1, telephone->0,

(2) WoE encode

- month:

(1) month number encode,

(2) WoE encode,

(3) (???)num\_of\_positives encode,

(4) drop,

(5) (???)cyclical encode

- day\_of\_week:

(1) dow number encode,

(2) WoE encode,

(3) (???)num\_of\_positives encode,

(4) drop,

(5) (???)cyclical encode

- duration: ???

- campaign:

(1) do nothing,

(2) cutoff with Q3+1.5\*IQR only,

(3) cutoff and WoE encode

- pdays:

(1) do nothing/replace 999->-1,

(2) cutoff with Q3+1.5\*IQR only,

(3) cutoff and WoE encode,

(4) bin and WoE encode,

(5) drop

- previous:

(1) do nothing,

(2) group [3..7]->3 only,

(3) group [3..7] and WoE encode,

(4) group [1..7]->1 only,

(5) group[1..7] and WoE encode,

(6) drop >=4 (GV)

- poutcome:

(1) encode with nonexistent->0, success=1, failure=-1,

(2) WoE encode,

(3) drop

- cons.conf.idx:

(1) leave as it is,

(2) delete rows with outliers (GV)

- all other features remain as they are (for logistic regression model possibly drop 2 out of emp.var.rate, euribor3m and nr.employed as they have correlation >0.9 to each other which leads to multicollinearity problem)

5. Impute:

- job: (1) cross-impute education->job, (2) most\_frequent, (3) KNN, (4) IterativeImputer

- education: (1) cross-impute job->education, (2) most\_frequent, (3) KNN, (4) IterativeImputer

- marital: (1) cross-impute age[\_binned]->marital, (2) most\_frequent, (3) KNN, (4) IterativeImputer

- default: (1) most\_frequent, (2) KNN, (3) IterativeImputer

- housing: (1) most\_frequent, (2) KNN, (3) IterativeImputer

- loan: (1) most\_frequent, (2) KNN, (3) IterativeImputer

6. MinMaxScaler / StandardScaler if necessary before using in model

7. [!!!] Oversample positive class

Alternative encodings: (1) mean target value for category, (2) num\_of\_positives

Preprocessing scenarios

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Scenario 1 (DS) | Scenario 2 (DS) | Scenario 3 (GV) | Scenario 4 |
| age | bin and WoE encode | leave as it is | drop rows with outliers |  |
| job | WoE encode with 'unknown' as category | impute 'unknown' + WoE encode | group if frequency <2000 |  |
| marital | WoE encode with 'unknown' as category | impute 'unknown' + WoE encode | drop rows with ‘unknown’ |  |
| education | WoE encode with 'unknown' as category | impute 'unknown' + WoE encode | group if frequency <4000 |  |
| default | WoE encode with 'unknown' as category | unknown->no + WoE encode | drop |  |
| housing | WoE encode with 'unknown' as category | yes->1, no->0, unknown->1(yes) | yes->1, no->0, unknown->1(yes) |  |
| loan | WoE encode with 'unknown' as category | yes->1, no->0, unknown->0(no) | yes->1, no->0, unknown->0(no) |  |
| contact | WoE encode | cellular->1, telephone->0 | leave as it is |  |
| month | WoE encode | WoE encode | leave as it is |  |
| day\_of\_week | WoE encode | WoE encode | leave as it is |  |
| duration | drop | drop | ffill? |  |
| campaign | cutoff to 6 and WoE encode | cutoff to 6 only | leave as it is |  |
| pdays | bin and WoE encode | 999->-1 + cutoff to 13 only | drop |  |
| previous | cutoff to 3 and WoE encode | cutoff to 3 only | drop rows where >=4 |  |
| poutcome | WoE encode | nonexistent->0, success=1, failure=-1 | drop |  |
| emp.var.rate | leave as it is | leave as it is | leave as it is |  |
| cons.price.idx | leave as it is | leave as it is | leave as it is |  |
| cons.conf.idx | leave as it is | leave as it is | drop rows with outliers |  |
| euribor3m | leave as it is | leave as it is | leave as it is |  |
| nr.employed | leave as it is | leave as it is | leave as it is |  |