19MIS1018_SWE4012_LAB-2_ML-MISSING VALUES ENCODING

August 9, 2022

NAME : B DEVI PRASAD

REG NO: 19MIS1018

SLOT: L13+14

FACULTY: Prof. BHARADWAJA KUMAR

1 Handling Missing Data in ML Modelling

Deleting missing data In my opinion, if the missing value percentage is above a certain threshold (say, 60%), it does not make much sense to try and impute them because it would likely influence our predictions due to the biased estimations. Deletion of the rows or columns with unknown values would be better suited. For illustrative purposes, suppose the data set looks like this (missing instances are denoted with the NaN notation):

[1]:

```
!pip install category_encoders
import category_encoders as ce
import pandas as pd
```

```
Requirement already satisfied: category_encoders in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (2.5.0)
Requirement already satisfied: pandas>=1.0.5 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
category_encoders) (1.3.4)
Requirement already satisfied: patsy>=0.5.1 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
category_encoders) (0.5.2)
Requirement already satisfied: statsmodels>=0.9.0 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
category_encoders) (0.13.2)
Requirement already satisfied: scipy>=1.0.0 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
category_encoders) (1.8.0)
Requirement already satisfied: scikit-learn>=0.20.0 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
category_encoders) (1.0.2)
Requirement already satisfied: numpy>=1.14.0 in
```

```
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
category_encoders) (1.21.4)
Requirement already satisfied: pytz>=2017.3 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
pandas>=1.0.5->category_encoders) (2021.3)
Requirement already satisfied: python-dateutil>=2.7.3 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
pandas>=1.0.5->category_encoders) (2.8.2)
Requirement already satisfied: six in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
patsy>=0.5.1->category_encoders) (1.16.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
scikit-learn>=0.20.0->category_encoders) (3.1.0)
Requirement already satisfied: joblib>=0.11 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
scikit-learn>=0.20.0->category_encoders) (1.1.0)
Requirement already satisfied: packaging>=21.3 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
statsmodels>=0.9.0->category_encoders) (21.3)
Requirement already satisfied: pyparsing!=3.0.5,>=2.0.2 in
c:\users\admin\appdata\local\programs\python\python39\lib\site-packages (from
packaging>=21.3->statsmodels>=0.9.0->category_encoders) (3.0.6)
```

[2]:

```
dataset=pd.read_csv('titanic_dataset.csv')
dataset
```

[2]:

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …         …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name  gender   Age  SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                               Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                 …     …   …    …
```

```
886                              Montvila, Rev. Juozas    male  27.0      0
887                       Graham, Miss. Margaret Edith  female  19.0      0
888             Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare Cabin Embarked
0        0         A/5 21171   7.2500   NaN        S
1        0          PC 17599  71.2833   C85        C
2        0  STON/O2. 3101282   7.9250   NaN        S
3        0            113803  53.1000  C123        S
4        0            373450   8.0500   NaN        S
..     …               …      …   …      …
886      0            211536  13.0000   NaN        S
887      0            112053  30.0000   B42        S
888      2        W./C. 6607  23.4500   NaN        S
889      0            111369  30.0000  C148        C
890      0            370376   7.7500   NaN        Q

[891 rows x 12 columns]
```

[3]:

```
encoder= ce.BinaryEncoder(cols=['Cabin'],return_df=True)
dataset=encoder.fit_transform(dataset)
dataset=dataset.drop(['Cabin_1'],axis=1)
dataset
```

[3]:

```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …       …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                   Name  gender   Age  SibSp  \
0                              Braund, Mr. Owen Harris    male  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0      1
2                               Heikkinen, Miss. Laina  female  26.0      0
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                             Allen, Mr. William Henry    male  35.0      0
..                                                 …     …   …     …
886                              Montvila, Rev. Juozas    male  27.0      0
```

```
887                       Graham, Miss. Margaret Edith  female  19.0      0
888           Johnston, Miss. Catherine Helen "Carrie"  female   NaN      1
889                              Behr, Mr. Karl Howell    male  26.0      0
890                                Dooley, Mr. Patrick    male  32.0      0

     Parch            Ticket     Fare  Cabin_0  Cabin_2  Cabin_3  Cabin_4  \
0        0         A/5 21171   7.2500        0        0        0        0
1        0          PC 17599  71.2833        0        0        0        0
2        0  STON/O2. 3101282   7.9250        0        0        0        0
3        0            113803  53.1000        0        0        0        0
4        0            373450   8.0500        0        0        0        0
..     …                …      …      …      …      …      …
886      0            211536  13.0000        0        0        0        0
887      0            112053  30.0000        1        0        1        0
888      2        W./C. 6607  23.4500        0        0        0        0
889      0            111369  30.0000        1        0        1        0
890      0            370376   7.7500        0        0        0        0

     Cabin_5  Cabin_6  Cabin_7 Embarked
0          0        0        1        S
1          0        1        0        C
2          0        0        1        S
3          0        1        1        S
4          0        0        1        S
..       …      …      …      …
886        0        0        1        S
887        0        1        1        S
888        0        0        1        S
889        1        0        0        C
890        0        0        1        Q

[891 rows x 18 columns]
```

```
[4]: encoder= ce.OrdinalEncoder(cols=['gender'],return_df=True,
                                mapping=[{'col':'gender',
     'mapping':{'male':1,'female':2}}])
     dataset= encoder.fit_transform(dataset)
     dataset
```

```
[4]:      PassengerId  Survived  Pclass  \
     0              1         0       3
     1              2         1       1
     2              3         1       3
     3              4         1       1
     4              5         0       3
     ..           …       …     …
     886          887         0       2
```

```
887          888          1       1
888          889          0       3
889          890          1       1
890          891          0       3

                                                  Name  gender   Age  SibSp  \
0                               Braund, Mr. Owen Harris       1  22.0      1
1    Cumings, Mrs. John Bradley (Florence Briggs Th…       2  38.0      1
2                                Heikkinen, Miss. Laina       2  26.0      0
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)       2  35.0      1
4                              Allen, Mr. William Henry       1  35.0      0
..                                                 …     …   …    …
886                               Montvila, Rev. Juozas       1  27.0      0
887                        Graham, Miss. Margaret Edith       2  19.0      0
888            Johnston, Miss. Catherine Helen "Carrie"       2   NaN      1
889                               Behr, Mr. Karl Howell       1  26.0      0
890                                 Dooley, Mr. Patrick       1  32.0      0

     Parch            Ticket     Fare  Cabin_0  Cabin_2  Cabin_3  Cabin_4  \
0        0         A/5 21171   7.2500        0        0        0        0
1        0          PC 17599  71.2833        0        0        0        0
2        0  STON/O2. 3101282   7.9250        0        0        0        0
3        0            113803  53.1000        0        0        0        0
4        0            373450   8.0500        0        0        0        0
..     …               …      …      …      …      …      …
886      0            211536  13.0000        0        0        0        0
887      0            112053  30.0000        1        0        1        0
888      2        W./C. 6607  23.4500        0        0        0        0
889      0            111369  30.0000        1        0        1        0
890      0            370376   7.7500        0        0        0        0

     Cabin_5  Cabin_6  Cabin_7 Embarked
0          0        0        1        S
1          0        1        0        C
2          0        0        1        S
3          0        1        1        S
4          0        0        1        S
..       …      …      …      …
886        0        0        1        S
887        0        1        1        S
888        0        0        1        S
889        1        0        0        C
890        0        0        1        Q

[891 rows x 18 columns]
```

[5]:
```
dataset['Age']=dataset['Age'].fillna(dataset['Age'].mean())
dataset
```

[5]:
```
     PassengerId  Survived  Pclass  \
0              1         0       3
1              2         1       1
2              3         1       3
3              4         1       1
4              5         0       3
..           …         …     …
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                  Name  gender        Age  \
0                              Braund, Mr. Owen Harris       1  22.000000
1    Cumings, Mrs. John Bradley (Florence Briggs Th…       2  38.000000
2                               Heikkinen, Miss. Laina       2  26.000000
3         Futrelle, Mrs. Jacques Heath (Lily May Peel)       2  35.000000
4                             Allen, Mr. William Henry       1  35.000000
..                                                 …     …        …
886                              Montvila, Rev. Juozas       1  27.000000
887                       Graham, Miss. Margaret Edith       2  19.000000
888           Johnston, Miss. Catherine Helen "Carrie"       2  29.699118
889                              Behr, Mr. Karl Howell       1  26.000000
890                                Dooley, Mr. Patrick       1  32.000000

     SibSp  Parch            Ticket     Fare  Cabin_0  Cabin_2  Cabin_3  \
0        1      0         A/5 21171   7.2500        0        0        0
1        1      0          PC 17599  71.2833        0        0        0
2        0      0  STON/O2. 3101282   7.9250        0        0        0
3        1      0            113803  53.1000        0        0        0
4        0      0            373450   8.0500        0        0        0
..     …    …               …      …      …      …
886      0      0            211536  13.0000        0        0        0
887      0      0            112053  30.0000        1        0        1
888      1      2        W./C. 6607  23.4500        0        0        0
889      0      0            111369  30.0000        1        0        1
890      0      0            370376   7.7500        0        0        0

     Cabin_4  Cabin_5  Cabin_6  Cabin_7 Embarked
0          0        0        0        1        S
1          0        0        1        0        C
2          0        0        0        1        S
3          0        0        1        1        S
```

```
4           0        0        0        1        S
..        …      …       …       …      …
886         0        0        0        1        S
887         0        0        1        1        S
888         0        0        0        1        S
889         0        1        0        0        C
890         0        0        0        1        Q

[891 rows x 18 columns]
```

```
[7]: encoder= ce.OrdinalEncoder(cols=['Embarked'],return_df=True,
                                mapping=[{'col':'Embarked',
     'mapping':{'C':1,'S':2,'Q':3}}])
     dataset= encoder.fit_transform(dataset)
     dataset
```

```
[7]:      PassengerId  Survived  Pclass  \
     0              1         0       3
     1              2         1       1
     2              3         1       3
     3              4         1       1
     4              5         0       3
     ..            …        …      …
     886          887         0       2
     887          888         1       1
     888          889         0       3
     889          890         1       1
     890          891         0       3

                                                        Name  gender        Age  \
     0                               Braund, Mr. Owen Harris       1  22.000000
     1    Cumings, Mrs. John Bradley (Florence Briggs Th…       2  38.000000
     2                                Heikkinen, Miss. Laina       2  26.000000
     3          Futrelle, Mrs. Jacques Heath (Lily May Peel)       2  35.000000
     4                              Allen, Mr. William Henry       1  35.000000
     ..                                                  …     …         …
     886                               Montvila, Rev. Juozas       1  27.000000
     887                        Graham, Miss. Margaret Edith       2  19.000000
     888            Johnston, Miss. Catherine Helen "Carrie"       2  29.699118
     889                               Behr, Mr. Karl Howell       1  26.000000
     890                                 Dooley, Mr. Patrick       1  32.000000

          SibSp  Parch             Ticket     Fare  Cabin_0  Cabin_2  Cabin_3  \
     0        1      0          A/5 21171   7.2500        0        0        0
     1        1      0           PC 17599  71.2833        0        0        0
     2        0      0   STON/O2. 3101282   7.9250        0        0        0
     3        1      0             113803  53.1000        0        0        0
```

```
4        0      0            373450   8.0500        0        0        0
..     …    …              …      …      …      …      …
886      0      0            211536  13.0000        0        0        0
887      0      0            112053  30.0000        1        0        1
888      1      2        W./C. 6607  23.4500        0        0        0
889      0      0            111369  30.0000        1        0        1
890      0      0            370376   7.7500        0        0        0

     Cabin_4  Cabin_5  Cabin_6  Cabin_7  Embarked
0          0        0        0        1       2.0
1          0        0        1        0       1.0
2          0        0        0        1       2.0
3          0        0        1        1       2.0
4          0        0        0        1       2.0
..       …      …      …      …       …
886        0        0        0        1       2.0
887        0        0        1        1       2.0
888        0        0        0        1       2.0
889        0        1        0        0       1.0
890        0        0        0        1       3.0

[891 rows x 18 columns]
```

[ ]: