superMarket_regression

April 25, 2018

```
In [195]: import pandas as pd
          import numpy as np
          from sklearn.linear_model import LinearRegression,Ridge,Lasso
          from sklearn.model_selection import train_test_split
          from sklearn.metrics import mean_squared_error
```

```
In [196]: df = pd.read_csv('train.csv')
          df.columns
```

```
Out[196]: Index([u'Item_Identifier', u'Item_Weight', u'Item_Fat_Content',
                 u'Item_Visibility', u'Item_Type', u'Item_MRP', u'Outlet_Identifier',
                 u'Outlet_Establishment_Year', u'Outlet_Size', u'Outlet_Location_Type',
                 u'Outlet_Type', u'Item_Outlet_Sales'],
                dtype='object')
```

```
In [197]: df.head()
```

```
Out[197]:   Item_Identifier  Item_Weight Item_Fat_Content  Item_Visibility  \
          0           FDA15         9.30          Low Fat         0.016047
          1           DRC01         5.92          Regular         0.019278
          2           FDN15        17.50          Low Fat         0.016760
          3           FDX07        19.20          Regular         0.000000
          4           NCD19         8.93          Low Fat         0.000000

                          Item_Type  Item_MRP Outlet_Identifier  \
          0                   Dairy  249.8092            OUT049
          1             Soft Drinks   48.2692            OUT018
          2                    Meat  141.6180            OUT049
          3   Fruits and Vegetables  182.0950            OUT010
          4               Household   53.8614            OUT013

             Outlet_Establishment_Year Outlet_Size Outlet_Location_Type  \
          0                       1999      Medium               Tier 1
          1                       2009      Medium               Tier 3
          2                       1999      Medium               Tier 1
          3                       1998         NaN               Tier 3
          4                       1987        High               Tier 3
          Outlet_Type  Item_Outlet_Sales
0   Supermarket Type1          3735.1380
1   Supermarket Type2           443.4228
2   Supermarket Type1          2097.2700
3       Grocery Store           732.3800
4   Supermarket Type1           994.7052
```

```
In [198]: # preprocessing fillna
          df['Item_Identifier'] = df['Item_Identifier'].fillna(df['Item_Identifier'].max())
          df['Item_Weight'] = df['Item_Weight'].fillna(df['Item_Weight'].mean())
          df['Item_Fat_Content']=df['Item_Fat_Content'].fillna(df['Item_Fat_Content'].max())
          df['Item_Visibility']=df['Item_Visibility'].fillna(df['Item_Visibility'].mean())
          df['Item_Type']=df['Item_Type'].fillna(df['Item_Type'].max())
          df['Item_MRP']=df['Item_MRP'].fillna(df['Item_MRP'].mean())
          df['Outlet_Identifier']=df['Outlet_Identifier'].fillna(df['Outlet_Identifier'].max())
          df['Outlet_Establishment_Year']=df['Outlet_Establishment_Year'].fillna(df['Outlet_Esta
          df['Outlet_Size']=df['Outlet_Size'].fillna(df['Outlet_Size'].max())
          df['Outlet_Location_Type']=df['Outlet_Location_Type'].fillna(df['Outlet_Location_Type'
          df['Outlet_Type']=df['Outlet_Type'].fillna(df['Outlet_Type'].max())
          df['Item_Outlet_Sales']=df['Item_Outlet_Sales'].fillna(df['Item_Outlet_Sales'].mean())
```

```
In [199]: # replace and format
          df['Item_Fat_Content'] = df['Item_Fat_Content'].replace('low fat', 'Low Fat')
          df['Item_Fat_Content'] = df['Item_Fat_Content'].replace('LF', 'Low Fat')
          df['Item_Fat_Content'] = df['Item_Fat_Content'].replace('reg', 'Regular')
```

```
In [200]: strData = [df['Item_Identifier'],df['Item_Fat_Content'],df['Item_Type'], df['Outlet_Id
          f = pd.DataFrame(strData)
          f = f.T
          f.head()
```

```
Out[200]:   Item_Identifier Item_Fat_Content              Item_Type Outlet_Identifier  \
          0           FDA15          Low Fat                  Dairy            OUT049
          1           DRC01          Regular            Soft Drinks            OUT018
          2           FDN15          Low Fat                   Meat            OUT049
          3           FDX07          Regular  Fruits and Vegetables            OUT010
          4           NCD19          Low Fat              Household            OUT013

            Outlet_Size Outlet_Location_Type        Outlet_Type
          0      Medium               Tier 1  Supermarket Type1
          1      Medium               Tier 3  Supermarket Type2
          2      Medium               Tier 1  Supermarket Type1
          3       Small               Tier 3      Grocery Store
          4        High               Tier 3  Supermarket Type1
```

```
In [201]: mylist = list(df.select_dtypes(include=['object']).columns)
          df = pd.get_dummies(df, prefix= mylist)
          df.head()
          # df['Item_Identifier'] = pd.get_dummies(f['Item_Identifier'])
# df['Item_Type'] = pd.get_dummies(f['Item_Type'])
# df['Item_Fat_Content']= pd.get_dummies(f['Item_Fat_Content'])
# df['Outlet_Identifier']= pd.get_dummies(f['Outlet_Identifier'])
# df['Outlet_Size']= pd.get_dummies(f['Outlet_Size'])
# df['Outlet_Location_Type']= pd.get_dummies(f['Outlet_Location_Type'])
# df['Outlet_Type']= pd.get_dummies(f['Outlet_Type'])
```

Out[201]:

```
   Item_Weight  Item_Visibility  Item_MRP  Outlet_Establishment_Year  \
0         9.30         0.016047  249.8092                       1999
1         5.92         0.019278   48.2692                       2009
2        17.50         0.016760  141.6180                       1999
3        19.20         0.000000  182.0950                       1998
4         8.93         0.000000   53.8614                       1987

   Item_Outlet_Sales  Item_Identifier_DRA12  Item_Identifier_DRA24  \
0          3735.1380                      0                      0
1           443.4228                      0                      0
2          2097.2700                      0                      0
3           732.3800                      0                      0
4           994.7052                      0                      0

   Item_Identifier_DRA59  Item_Identifier_DRB01  Item_Identifier_DRB13  \
0                      0                      0                      0
1                      0                      0                      0
2                      0                      0                      0
3                      0                      0                      0
4                      0                      0                      0

               ...                Outlet_Size_High  Outlet_Size_Medium  \
0              ...                               0                   1
1              ...                               0                   1
2              ...                               0                   1
3              ...                               0                   0
4              ...                               1                   0

   Outlet_Size_Small  Outlet_Location_Type_Tier 1  \
0                  0                            1
1                  0                            0
2                  0                            1
3                  1                            0
4                  0                            0

   Outlet_Location_Type_Tier 2  Outlet_Location_Type_Tier 3  \
0                            0                            0
1                            0                            1
2                            0                            0
3                            0                            1
4                            0                            1
```

```
   Outlet_Type_Grocery Store  Outlet_Type_Supermarket Type1  \
0                          0                              1
1                          0                              0
2                          0                              1
3                          1                              0
4                          0                              1

   Outlet_Type_Supermarket Type2  Outlet_Type_Supermarket Type3
0                              0                              0
1                              1                              0
2                              0                              0
3                              0                              0
4                              0                              0

[5 rows x 1602 columns]
```

```
In [202]: df['Outlet_Establishment_Year'] = 2018 - df['Outlet_Establishment_Year']
          df['Outlet_Establishment_Year'].head()

Out[202]: 0    19
          1     9
          2    19
          3    20
          4    31
          Name: Outlet_Establishment_Year, dtype: int64
```

```
In [203]: x = df
          x = x.drop('Item_Outlet_Sales',axis=1)
          y = df['Item_Outlet_Sales']
```

```
In [204]: train_x,test_x,train_y,test_y = train_test_split(x,y)
```

```
In [205]: train_x.head()

Out[205]:       Item_Weight  Item_Visibility  Item_MRP  Outlet_Establishment_Year  \
          8438     9.300000         0.088932  143.3786                         31
          6534    15.600000         0.035561  112.1518                         19
          1194    12.857645         0.032750  112.1518                         33
          2105    11.800000         0.014075  176.8344                         31
          3017    12.857645         0.253948  223.8404                         33

                Item_Identifier_DRA12  Item_Identifier_DRA24  Item_Identifier_DRA59  \
          8438                      0                      0                      0
          6534                      0                      0                      0
          1194                      0                      0                      0
          2105                      0                      0                      0
          3017                      0                      0                      0
```

```
      Item_Identifier_DRB01  Item_Identifier_DRB13  Item_Identifier_DRB24  \
8438                      0                      0                      0
6534                      0                      0                      0
1194                      0                      0                      0
2105                      0                      0                      0
3017                      0                      0                      0

                  ...                Outlet_Size_High  Outlet_Size_Medium  \
8438              ...                               1                   0
6534              ...                               0                   1
1194              ...                               0                   1
2105              ...                               1                   0
3017              ...                               0                   0

      Outlet_Size_Small  Outlet_Location_Type_Tier 1  \
8438                  0                            0
6534                  0                            1
1194                  0                            0
2105                  0                            0
3017                  1                            1

      Outlet_Location_Type_Tier 2  Outlet_Location_Type_Tier 3  \
8438                            0                            1
6534                            0                            0
1194                            0                            1
2105                            0                            1
3017                            0                            0

      Outlet_Type_Grocery Store  Outlet_Type_Supermarket Type1  \
8438                          0                              1
6534                          0                              1
1194                          0                              0
2105                          0                              1
3017                          1                              0

      Outlet_Type_Supermarket Type2  Outlet_Type_Supermarket Type3
8438                              0                              0
6534                              0                              0
1194                              0                              1
2105                              0                              0
3017                              0                              0

[5 rows x 1601 columns]
```

In [206]:

```
model =Ridge()
# model.fit(train_x['Item_MRP'].values.reshape(-1,1),train_y)
model.fit(train_x,train_y)
```

Out[206]:

```
Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
             normalize=False, random_state=None, solver='auto', tol=0.001)

In [207]: predicted = model.predict(test_x)

In [208]: coef1 = df['Item_MRP'].corr(df['Item_Weight'])
          coef2 = df['Item_MRP'].corr(test_y)
          coef1, coef2

Out[208]: (0.02475610129707686, 0.5594689675161207)

In [216]: max(model.coef_)

Out[216]: 2063.6704122734827

In [217]: model.intercept_

Out[217]: -574.3408700685627

In [210]: model.score(test_x,test_y)

Out[210]: 0.46767579277790206

In [211]: # model score manual // R squared
          sstot= sum((test_y - np.mean(test_y))**2)
          ssres = sum((test_y - predicted)**2)
          rs2 = 1-(ssres/sstot)
          rs2

Out[211]: 0.4676757927779014

In [212]: mean_squared_error(predicted,test_y)

Out[212]: 1464805.011700845

In [213]: # mean squared error manual
          mse = np.mean((predicted - test_y)**2)
          mse

Out[213]: 1464805.011700845
```