Homework 1: Linear Regression and Neural Network Regression.

Anay Abhijit Joshi

CS 5173: **Deep Learning**
Level: **Undergraduate**

Step 1: Data

1) How many data samples are included in the dataset?

3047

```
[8] # Name: Anay Abhijit Joshi
    import pandas as pd

    # Let's specify the file path that conatins the data-set
    file_path = "cancer_reg.csv"
    # This dataset contains cancer mortality rates information. The goal of this project/challenge
    # is to predict the results of Cancer Mortality Rates. Therefore, the label is "TARGET_deathRate".
    # I will load the dataset into a pandas data-frame, and handle special-character encoding
    file = pd.read_csv(filepath_or_buffer=file_path, encoding="latin1")

[9] print(file.shape[0])

    3047
```

2) Which problem will this dataset try to address?

This dataset will predict the results of the Cancer Mortality Rates. For this problem, the “label” is “TARGET_deathRate”.

3) What is the minimum value and the maximum value in the dataset?

minimum value = 0.0

maximum value = 10170292

```
[10] print(file.min())
```

```
avgAnnCount                                      6.0
avgDeathsPerYear                                   3
TARGET_deathRate                                59.7
incidenceRate                                  201.3
medIncome                                      22640
popEst2015                                       827
povertyPercent                                   3.2
studyPerCap                                      0.0
binnedInc                         (34218.1, 37413.8]
MedianAge                                       22.3
MedianAgeMale                                   22.4
MedianAgeFemale                                 22.3
Geography                 Abbeville County, South Carolina
AvgHouseholdSize                              0.0221
PercentMarried                                  23.1
PctNoHS18_24                                     0.0
PctHS18_24                                       0.0
PctSomeCol18_24                                  7.1
PctBachDeg18_24                                  0.0
PctHS25_Over                                     7.5
PctBachDeg25_Over                                2.5
PctEmployed16_Over                              17.6
PctUnemployed16_Over                             0.4
PctPrivateCoverage                              22.3
PctPrivateCoverageAlone                         15.7
PctEmpPrivCoverage                              13.5
PctPublicCoverage                               11.2
PctPublicCoverageAlone                           2.6
PctWhite                                   10.199155
PctBlack                                         0.0
PctAsian                                         0.0
PctOtherRace                                     0.0
PctMarriedHouseholds                        22.99249
BirthRate                                        0.0
dtype: object
```

```
[12] float_or_int_only = file.select_dtypes([float, int])
     print(float_or_int_only.min().min())
```

```
0.0
```

```
[17] print(file.max())
```

```
avgAnnCount                          38150.0
avgDeathsPerYear                       14010
TARGET_deathRate                       362.8
incidenceRate                         1206.9
medIncome                             125635
popEst2015                          10170292
povertyPercent                          47.4
studyPerCap                      9762.308998
binnedInc                   [22640, 34218.1]
MedianAge                              624.0
MedianAgeMale                           64.7
MedianAgeFemale                         65.7
Geography                Zavala County, Texas
AvgHouseholdSize                        3.97
PercentMarried                          72.5
PctNoHS18_24                            64.1
PctHS18_24                              72.5
PctSomeCol18_24                         79.0
PctBachDeg18_24                         51.8
PctHS25_Over                            54.8
PctBachDeg25_Over                       42.2
PctEmployed16_Over                      80.1
PctUnemployed16_Over                    29.4
PctPrivateCoverage                      92.3
PctPrivateCoverageAlone                 78.9
PctEmpPrivCoverage                      70.7
PctPublicCoverage                       65.1
PctPublicCoverageAlone                  46.6
PctWhite                               100.0
PctBlack                           85.947799
PctAsian                           42.619425
PctOtherRace                       41.930251
PctMarriedHouseholds               78.075397
BirthRate                          21.326165
dtype: object
```

```
float_or_int_only = file.select_dtypes([float, int])
print(float_or_int_only.max().max())
```

```
10170292.0
```

4) How many features in each data samples?

33 Features

Excludes 1 Label

Total Columns: 34

```
[29] # Rows and Columns
     print(file.shape)
```

```
(3047, 34)
```

```
[30] # Total Columns
     print(file.shape[1])
```

```
34
```

```
[31] # Exclude 1 column of "LABEL" to get "FEATURES"
     print((file.shape[1]) - 1 )
```

```
33
```

5) Does the dataset have any missing information? E.g., missing features.

Yes, the dataset has the missing information.

PctSomeCol18_24	**2285**
PctEmployed16_Over	**152**
PctPrivateCoverageAlone	**609**

```
[42] print(file.isnull().sum())
```

```
avgAnnCount                    0
avgDeathsPerYear               0
TARGET_deathRate               0
incidenceRate                  0
medIncome                      0
popEst2015                     0
povertyPercent                 0
studyPerCap                    0
binnedInc                      0
MedianAge                      0
MedianAgeMale                  0
MedianAgeFemale                0
Geography                      0
AvgHouseholdSize               0
PercentMarried                 0
PctNoHS18_24                   0
PctHS18_24                     0
PctSomeCol18_24             2285
PctBachDeg18_24                0
PctHS25_Over                   0
PctBachDeg25_Over              0
PctEmployed16_Over           152
PctUnemployed16_Over           0
PctPrivateCoverage             0
PctPrivateCoverageAlone      609
PctEmpPrivCoverage             0
PctPublicCoverage              0
PctPublicCoverageAlone         0
PctWhite                       0
PctBlack                       0
PctAsian                       0
PctOtherRace                   0
PctMarriedHouseholds           0
BirthRate                      0
dtype: int64
```

```
[43] print((file.isnull().sum())[file.isnull().sum() > 0])
```

```
PctSomeCol18_24             2285
PctEmployed16_Over           152
PctPrivateCoverageAlone      609
dtype: int64
```

6) What is the label of this dataset?

TARGET_deathRate

7) How many percent of data will you use for **training, validation** and **testing**?

80% - 10% - 10%

8) What kind of data pre-processing will you use for your training dataset?

In the given dataset, for data pre-processing, first, I filled the missing values in the columns with the column's data's "mean". This was done to make sure that no data is being lost. Moreover, the categorical features, such as "Geography" and "binnedInc" were label encoded using "LabelEncoder()" to convert these feature values into numerical data and then, use it. In addition, I also used "numpy's mathematical log function" for the features and the target variable ("TARGET_deathRate"), for transformation of the skewed data to reduce skewness and make the data distribution more symmetric. Further, I also used "StandardScaler()" for standardizing the features by removing the mean and scaling to unit variance. Finally, I used this data by splitting it into 80%-20%-20% for training, validation, and testing, respectively, for a balanced model evaluation.

Step 2: Model

Model	Test R-squared	MSE
Linear regression	0.770306897015671	0.005662627419371372
DNN-16 *(LR=0.001)*	0.8526817893566186	0.004149031329956375
DNN-30-8 *(LR=0.001)*	0.8152757268053347	0.005202525835337959
DNN-30-16-8 *(LR=0.001)*	0.8182953390336587	0.005117482270900689
DNN-30-16-8-4 *(LR=0.001)*	0.7761951143111193	0.006303181924793659

Step 5: Model Selection

Epochs = 100

Batch Size = 64

R-Squared value

Model	LR: 0.1 (R^2)	LR: 0.01 (R^2)	LR: 0.001 (R^2)	LR: 0.0001 (R^2)
Linear regression	-	-	-	-
DNN-16	0.7609437147162864	0.8462586399544709	0.8526817893566186	0.8567260630482327
DNN-30-8	0.6768140614561169	0.8189500334814785	0.8152757268053347	0.8150193659723532
DNN-30-16-8	0.6775849782102178	0.8090008832305213	0.8182953390336587	0.8157842936758392
DNN-30-16-8-4	0.7301531324384998	0.7875299336077706	0.7761951143111193	0.7762262288035106

MSE (Mean Squared Error)

Model	LR: 0.1 (MSE)	LR: 0.01 (MSE)	LR: 0.001 (MSE)	LR: 0.0001 (MSE)
Linear regression	-	-	-	-
DNN-16	0.006732718330837954	0.004329931219997128	0.0041490313299556375	0.004035129469621198
DNN-30-8	0.009102123753496269	0.005099043628700995	0.005202525835337959	0.005209745914397893
DNN-30-16-8	0.009080411856836763	0.005379248879072817	0.005117482270900689	0.005188202691784384
DNN-30-16-8-4	0.0075998961901140565	0.005983951055943129	0.006303181924793659	0.006302305624415157

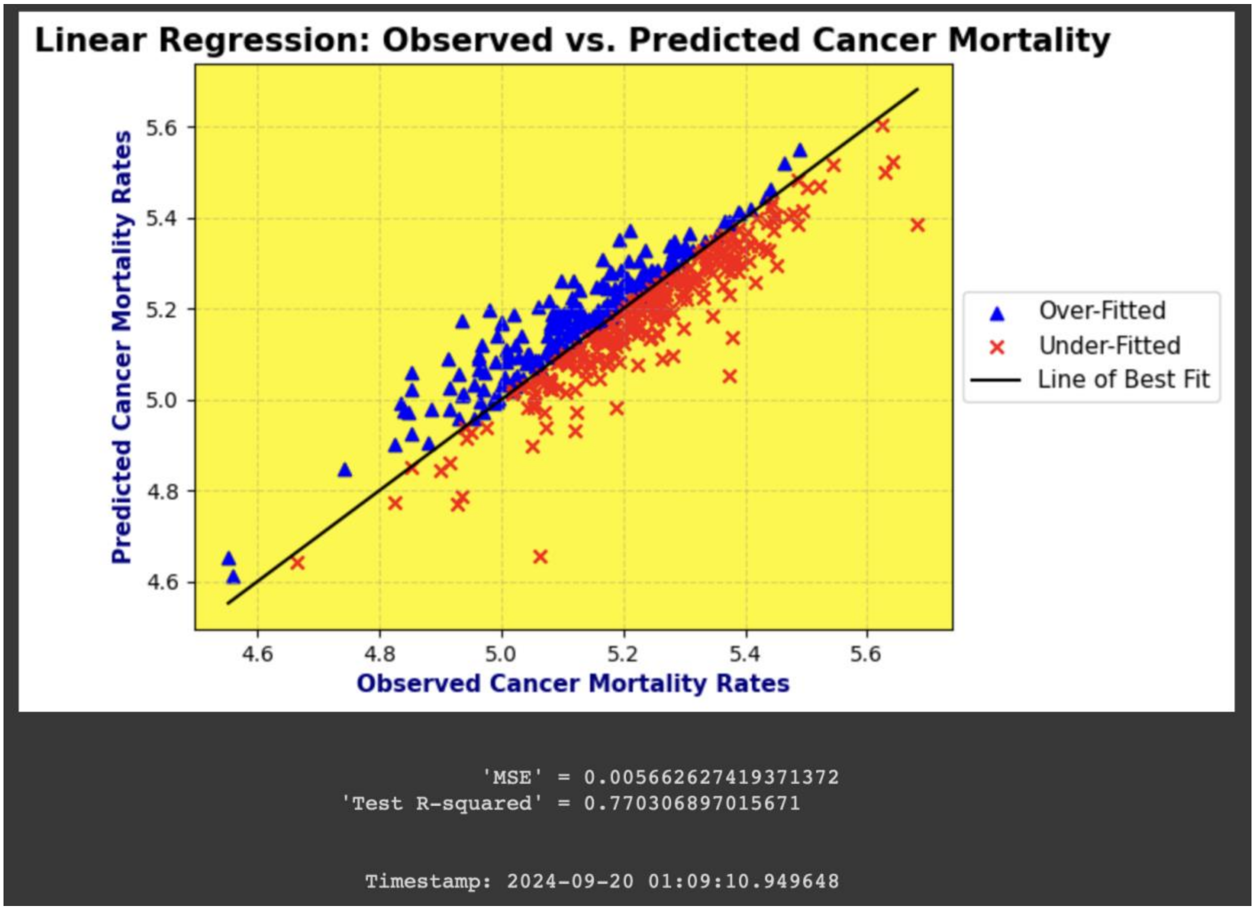
Step 6: Model Performance

As shown in the table in STEP 2, for **Linear Regression** model, the **MSE (Mean Squared Error)** value was **0.770306897015671** and **Test R-Squared value** was **0.005662627419371372**

Linear Regression Model

Test R-Squared = 0.770306897015671

Mean Squared Error (MSE) = 0.005662627419371372



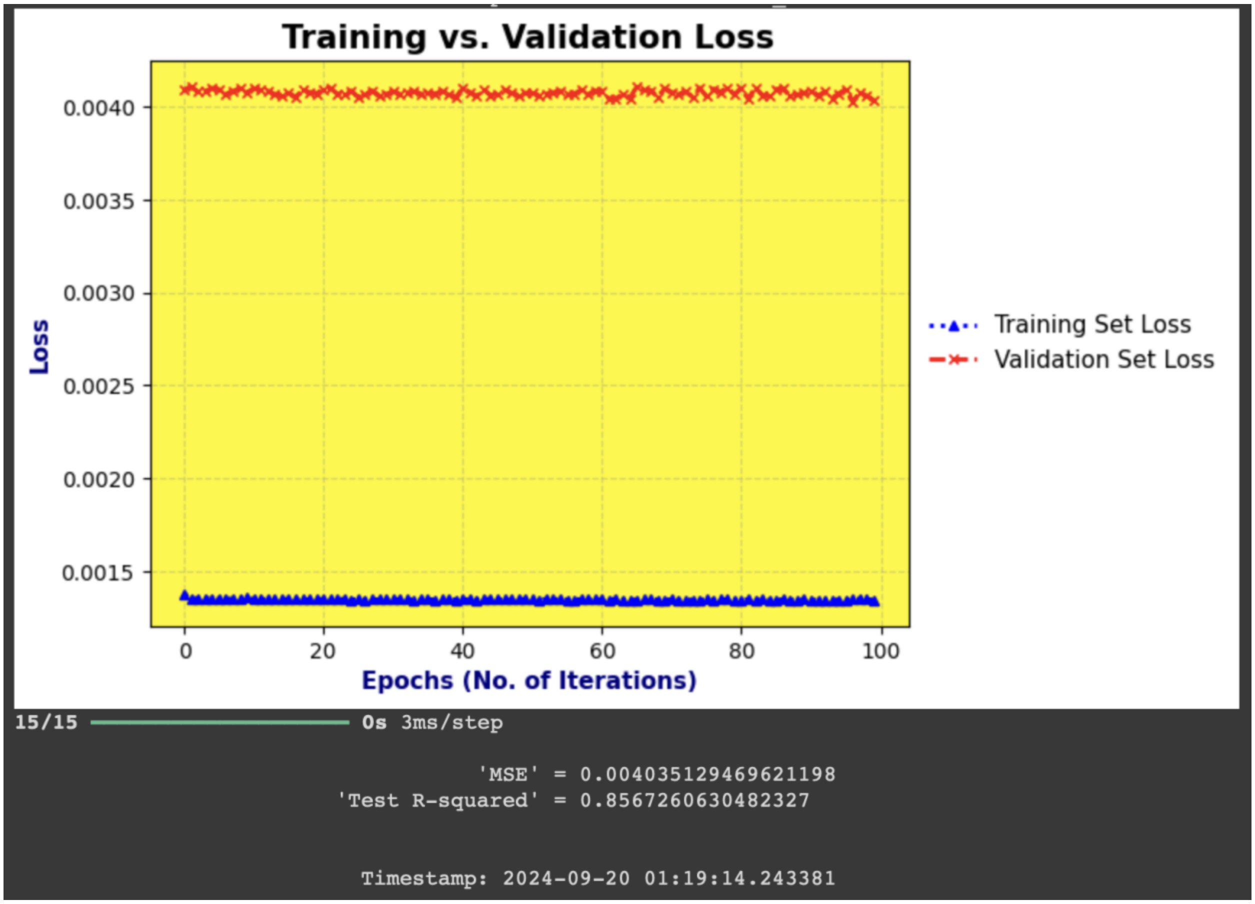
The best performing model, with respect to all the obtained values of R-Squared value and Mean Squared Error (MSE) value is as follows -

DNN-16

Learning Rate (LR) = 0.0001

Epochs = 100
Batch-Size = 64

Mean Squared Error (MSE) = 0. 004035129469621198
R-Squared = 0.8567260630482327



I set the number of training iterations over the dataset, or "EPOCHS" to "100"; and the "BATCH-SIZE" to "64" for evaluating the performance of <u>all the models with different learning rates</u>. Therefore, these were consistent across all the models…