MACHINE LEARNING, STATISTICAL LEARNING AND ARTIFICIAL INTELLIGENCE PROJECT:

SUPERVISED LEARNING

UNIVERSITÀ DEGLI STUDI DI MILANO

STUDENT NAME: **SAINEY MANGA**
MAT: **943874**
PROGRAM: **MASTER'S IN DATA SCIENCE AND ECONOMICS**
COURSE CORDINATOR: **PROFESSOR SILVIA SALLINI**

TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION

MACHINE LEARNING, STATISTICAL LEARNING AND ARTIFICIAL INTELLIGENCE PROJECT:

SUPERVISED LEARNING

ABSTRACT

This report compares between linear models and non-linear models using techniques in regression analysis. The study focuses on the effect of a change in crime rate on the price of housing in a vicinity. The paper however, centers more on the performances of the various models and techniques applied. It is evident that some two variables might not be linearly related, i.e. they have a non-linear relationship and thus, performing a prediction on a linear model would not yield a robust estimate. This relationship could be verified through plotting our models as can be seen later in the paper.

We have basically set up two class of models i.e. Linear and non-linear models with more technical application in terms of a simple or multiple regression. Although the multiple regression model out-perform the simple regression model, but that is not our main goal as we would later see reasons why that is so.

In direct contrast of the simple linear model and its counterpart non-linear model, the non-linear model has a lower Root Mean Square Error (RMSE) of 8.713958 on the test set than a RMSE of 9.048147 the linear Model generated on the test set. A detailed analysis on other coefficients and terms of interest are captured on the findings section.

On the multiple regression part, the model where we included a polynomial term also reported a lower RMSE than the model without a polynomial term i.e. 5.886297 vs 5.920898 respectively. Generally, the report also indicates the relationship between crime rate and housing price. Both models report a negative relation between the two variables with housing price (MEDV) as our responds variable and crime rate as our target predictor. The simple linear model reports a coefficient of -0.41501 with a p-value of 3.07e-13 against its counterpart with a coefficient of -0.840935 and a p-value of 4.74e-12. We have also seen that including more predictors in the model has reported somewhat more interesting estimates. The multiple regression model without a non-linear term reports a coefficient of -0.034917 with a p-value of 0.433 compared to the multiple regression model with a non-linear term that reports a coefficient of 0.161963 with a p-value of 0.196773. From all the models above, we have notice that crime rate affects housing price negatively except in the final model where we include both the non-linear term and other predictors, but as we mention earlier, our main focus will be on the RMSE of the test set of our models which tells us how good a model performs.

TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION

MACHINE LEARNING, STATISTICAL LEARNING AND ARTIFICIAL INTELLIGENCE PROJECT:

SUPERVISED LEARNING

STATEMENT OF THE PROBLEM OR GOAL:

The relationship between crime rate and housing price; does a change in crime rate affect housing price? Basically, to predict the median price of a house given crime rate.

The goal of this report is to investigate the impact of a change in crime rate on housing price in Boston with emphasis on the comparison between linear models vs non-linear models, which model outperforms the other in regards with this case study.

The dataset was downloaded from this link http://lib.stat.cmu.edu/datasets/boston, it contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. The dataset contains 14 variables with crime rate and housing price as our target variables, it has 506 cases. We refer to MEDV as our housing price and CRIM as our crime rate. See the detail explanation of variables below:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq. Ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centers
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63) ^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

KEY FINDINGS:

I. The correlation matrix shows a negative correlation between crime rate and housing price.
II. Crime rate has a negative impact on housing price, where all models are consensus on their estimates.
III. The simple non-linear model outperforms the linear model in both the estimates and best linear fit.

TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION

MACHINE LEARNING, STATISTICAL LEARNING AND ARTIFICIAL INTELLIGENCE PROJECT:

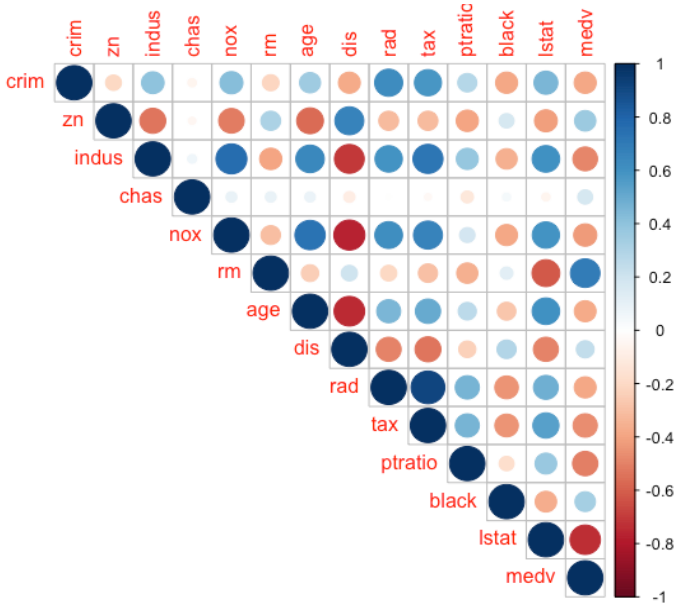
SUPERVISED LEARNING

IV. The multiple regression model with a polynomial term reports a lower RMSE than the one with a level term, but seems to contradict the trajectory of findings so far because it reports a positive relationship between CRIM and MEDV.

The sub-section below contains the main body of results corresponding to each of the key findings above.

- **MAIN RESULTS:**

I. The diagram below represents the correlation matrix which displays the correlation amongst all the variables in our dataset.



From the correlation matrix, we understand that there is a correlation of -0.3883046 between CRIM and MEDV which obviously tells us that a correlation is inherent. However, it also shows a correlation amongst CRIM with TAX and LSTAT, but it is convincing that it will not lead to a perfect collinearity. In fact, TAX and LSTAT affect our responds variable and would therefore do a great deal in our multiple regression model.

II. See the results of our simple linear model referred to as "lm.fit1":

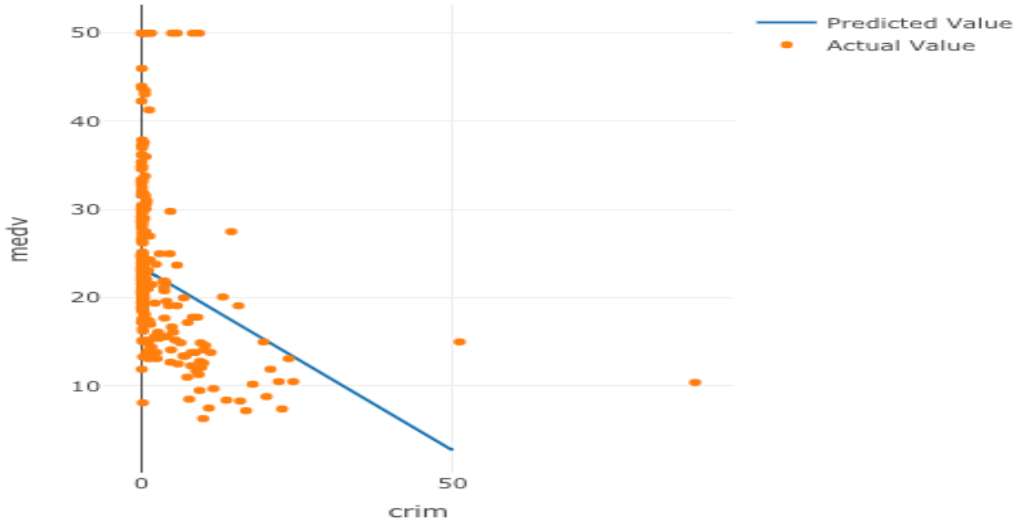
```
Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  23.45901    0.54298  43.204 <             2e-16 ***
crim         -0.41501    0.05387  -7.703              3.07e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.916 on 251 degrees of freedom
Multiple R-squared:  0.1912,  Adjusted R-squared:  0.188
F-statistic: 59.34 on 1 and 251 DF,  p-value: 3.067e-13
```

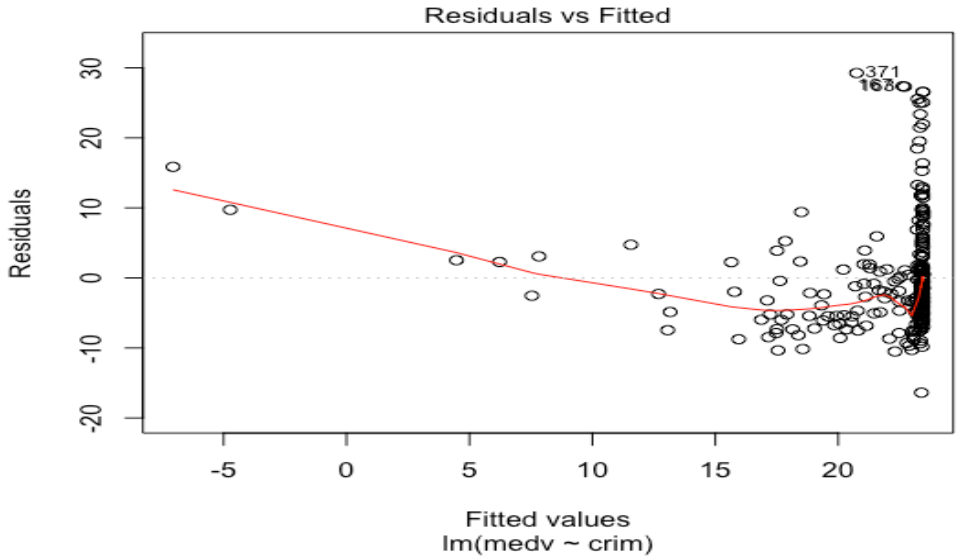
MACHINE LEARNING, STATISTICAL LEARNING AND ARTIFICIAL INTELLIGENCE PROJECT:

SUPERVISED LEARNING

A unit increase in crime rate will decrease the median price of a house by around 42% with a reported p-value of 3.067e-13. We reserve the conclusions as to whether these estimates are significant or not until later. The RMSE of the test set of the model(lm.fit1) is 9.048147 which seems high and our conclusions will be centered on RMSE to know which model best fit the dataset. The plot of the model is shown below:



Below is also the diagnosis plot which says MEDV and CRIM does not have a linear relationship:



We have also notice that the crime rate spreads between 0 and 20 with a mean of 3.61352 and a max of 88.97620, so the distribution is skewed to the left.

TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION

SUPERVISED LEARNING

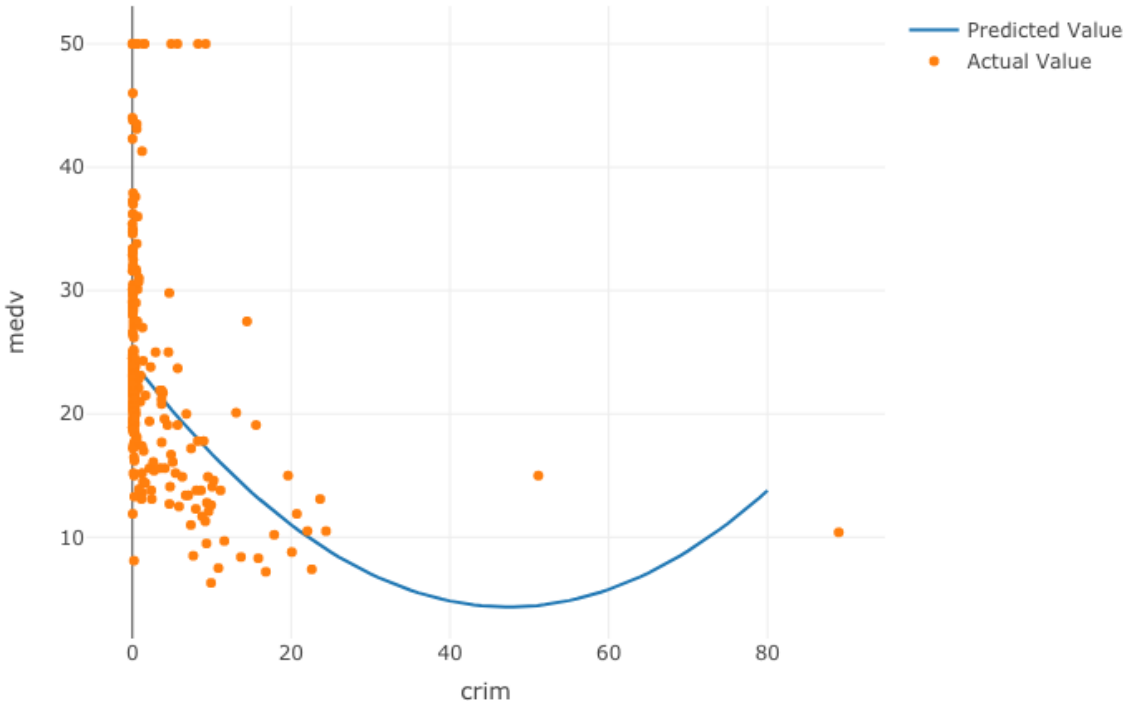
III. The non-linear model in direct comparison with the linear model shows the results below:

```
Coefficients:
             Estimate Std. Error t value    Pr(>|t|)
(Intercept) 24.274018   0.562338  43.166  < 2e-16 ***
crim        -0.840935   0.115753  -7.265   4.74e-12 ***
I(crim^2)    0.008874   0.002152   4.123  5.09e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.675 on 250 degrees of freedom
Multiple R-squared:  0.2427,  Adjusted R-squared:  0.2367
F-statistic: 40.06 on 2 and 250 DF,  p-value: 8.074e-16
```

A unit increase in crime rate will reduce housing by 84% with a p-value of 4.74e-12, but our main focus here is the RMSE of the test set of the model i.e. 8.713958, which is lower than that of the linear model. Therefore, one can conclude that the latter model more suitably fits the dataset or predicts the dataset better.

See the plot of the model below:



TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION

SUPERVISED LEARNING

IV. Finally, we examine the results of the last two models where we tried to compare two multiple regression models. The variables RM, TAX and LSTAT were included in both models as predictors, but the last model is somewhat different with a non-linear term of crime rate. The idea is to investigate whether adding predictors to our previous models would yield different results and perhaps better estimates.
The former reports the following estimates:

```
Coefficients:
             Estimate   Std. Error t value    Pr(>|t|)
(Intercept) -1.624511   4.124261  -0.394    0.694
crim        -0.034917   0.044475  -0.785    0.433
rm           5.292967   0.568646   9.308 < 2e-16 ***
tax         -0.007238   0.002415  -2.997    0.003 **
lstat       -0.500867   0.065618  -7.633 4.94e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.982 on 248 degrees of freedom
Multiple R-squared:  0.6836,  Adjusted R-squared:  0.6784
F-statistic: 133.9 on 4 and 248 DF,  p-value: < 2.2e-16
```

A unit increase in crime rate would reduce housing price by about 35% with a p-value of 0.433. The model reports a RMSE of 5.920898 on the test set which is lower than all the previous models and we prefer a model that reports a lower RMSE on the test set. However, we cannot compare this model with the simple regression model that has a polynomial term of crime rate because the three variables we included play a key role in producing the results above.

On the other hand, the multiple regression model with a non-linear term reports the following results:

```
Coefficients:
             Estimate Std. Error   t value        Pr(>|t|)
(Intercept)  0.226734   4.253917   0.053 0.957536
crim         0.161963   0.125136   1.294 0.196773
I(crim^2)   -0.003187   0.001894  -1.682 0.093768 .
rm           5.186548   0.570079   9.098  < 2e-16 ***
tax         -0.010251   0.002999  -3.418 0.000739 ***
lstat       -0.533135   0.068133  -7.825 1.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.963 on 247 degrees of freedom
Multiple R-squared:  0.6871,  Adjusted R-squared:  0.6808
F-statistic: 108.5 on 5 and 247 DF,  p-value: < 2.2e-16
```

Interestingly, this model reports a positive correlation between crime rate and housing price and a corresponding p-value of 0.196773. The RMSE of its test set is 5.886297 which is almost the same as the contrast model. Obviously, it is curiosity that led us into this model as the simple model with a polynomial crime rate is in itself a multiple regression model, though without the added three variables.

CONCLUSION:

In conclusion, we have notice that an increase in crime rate would reduce housing price. Crime rate has a negative impact on housing price.

Considering the simple regression models, the non-linear model performs better than the linear model as it produces a lower RMSE.

Finally, the multiple regression models have similar results in terms of their RMSEs. They report different signs of crime rate coefficient which could be due to the fact we have many parameters in the last model as adding too many parameters in a model could lead to undesirable results.

MACHINE LEARNING, STATISTICAL LEARNING AND ARTIFICIAL INTELLIGENCE PROJECT:

SUPERVISED LEARNING

APPENDICES:

```
>boston<- read.csv('http://lib.stat.cmu.edu/datasets/boston', sep = ';')
>head(boston)
>attach(boston)
>summary(crim)
>summary(medv)
## CORRELATION MATRIX
>library(corrplot)
>corr_matrix<-cor(Boston)
>corrplot(corr_matrix, type="upper")
>any(is.na(Boston))

## SPLITTING THE DATA INTO TRAINING AND TEST SET

>smp_size<-floor(0.50*nrow(Boston))
>set.seed(1)
>train_ind<-sample(seq_len(nrow(Boston)), size=smp_size)
>train<-Boston[train_ind, ]
>test<-Boston[-train_ind, ]
>lm.fit1=lm(medv~crim,data=train)  ## SIMPLE LINEAR MODEL
>summary(lm.fit1)
### PREDICT MODEL
>require(Metrics)
>pred<- predict(lm.fit1, test)
>rmse(pred,test[,14 ])
### PLOT MODEL
>require(ggplot2)
>require(plotly)
>dat <- data.frame(crim = (1:50),
              medv = predict(lm.fit1, data.frame(crim = (1:50))))
plot_ly() %>%
  add_trace(x=~crim, y=~medv, type="scatter", mode="lines", data = dat, name = "Predicted
Value") %>%
  add_trace(x=~crim, y=~medv, type="scatter", data = test, name = "Actual Value")


### MULTIPLE REGRESSION MODEL
>lm.fit2=lm(medv~crim+rm + tax + lstat,data=train)
```


TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION

SUPERVISED LEARNING

```
>summary(lm.fit2)
>pred1<-predict(lm.fit2, test)
>rmse(pred1, test[,14])

#####  NON-LINEAR MODELS
>lm.fit3=lm(medv~crim+I(crim^2),data=train).  ### SIMPLE NON-LINEAR
>summary(lm.fit3)
>dat <- data.frame(crim = (1:80),
              medv = predict(lm.fit3, data.frame(crim = (1:80))))
plot_ly() %>%
  add_trace(x=~crim, y=~medv, type="scatter", mode="lines", data = dat, name = "Predicted
Value") %>%
  add_trace(x=~crim, y=~medv, type="scatter", data = test, name = "Actual Value")

>pred2<-predict(lm.fit3, test)
>rmse(pred2, test[,14])
#### MULTIPLE REGRESSION WITH A NON-LINEAR MODEL
>lm.fit4=lm(medv~crim+I(crim^2)+rm + tax + lstat,data=train)
>summary(lm.fit4)
>pred3<-predict(lm.fit4, test)
>rmse(pred3, test[, 14])
```


TOPIC: LINEAR MODELS VS NON-LINEAR MODELS IN REGRESSION