CMPT 726 Assignment 2

Dhruv Patel, 301471961

1. Cross Validation

a) In order to get an unbiased estimate of how well our ML models and algorithms perform, it is typical to do a 60/20/20 split for our data: 60% train, 20% test and 20% validation. Several years ago this was widely considered best practice in machine learning. Do you still agree such ratios in the modern big data era? Why or why not? (Hint: consider different scenarios)

Answer: Clearly, in this modern big data era. We cannot fix any ratios to split into train,validation, and test dataset as we did several years ago. The decision to split the dataset should be done after considering various scenarios such as

- Availability of data
  - If we have more than enough data available to train the model. Then we can split the data into train, validation and test data.
- Representativeness of data
  - Suitable representation of the scenario, means data mostly covers all the possible cases. Else, we need to make sure. Training and test data have almost similar records in distribution.
- Computational cost on the data
  - Since there will always be a computational cost associated to train a model. Based on the model techniques and the size of the data. To optimise the cost, accordingly we can increase or decrease training data.

As a result, splitting of the records mainly depends on the data and the way to mode the problem. Some best estimations to split the data is

- If the model are more complex then we need more data for to train the model. (70%,15%,15%)
- If the model are all relatively simple and sample are large enough, then can split data (40%,30%,30%)
- If the model is complex and sample size is small, then we need to keep the training set pretty large.

b) Consider the hypothetical graph below of predictive error (y-axis) vs. model complexity (xaxis), and how test/training error varies as model complexity increases.

i) Which part of the plot means that the model is Overfitting on the data? (Choose A, B, or C)
Answer: After analysing the given hypothetical graph, we can say **part C** denotes that model is overfitted on data. Since, the gap between error on training set and test set is maximum. Because, model parameters were overly trained on training data so we have low training error and high error on unknown data.

ii) Which part of the plot means that the model is Underfitting on the data? (Choose A, B, or C)
Answer: Based on the hypothetical graph, **part A** denotes that model is under-fitted on data. Both training error and test error are high in that part. Since model capacity is less.

iii) Which part of the plot representing the ideal model complexity? (Choose A, B, or C)
Answer: Part B represents the ideal model complexity in the given hypothetical graph. Showing the ideal and best scenario where the gap between both the training and test error is minimum. And both do not not deviate that much.

c) For a decision tree model, whose train/test errors are in region A, which of the following is most likely to improve the model performance on real data? (Choose one)
i) Acquire more training data.
ii) Reduce the depth of the decision tree.
iii) Increase the depth of the decision tree.
Answer:
The problem with the decision tree model in region A, is that it's under fitted. So ideally option " i) Acquire more training data would not be helpful ". Even option " ii) Reduce the depth of the decision tree would make it even worse since reducing the depth will reduce the splits hence the model is not trained properly. Finally **iii) Increase the depth of the decision**" would help by categorising the data and splitting further, so will help our model to train better.

2. Regression

2.1 Getting started

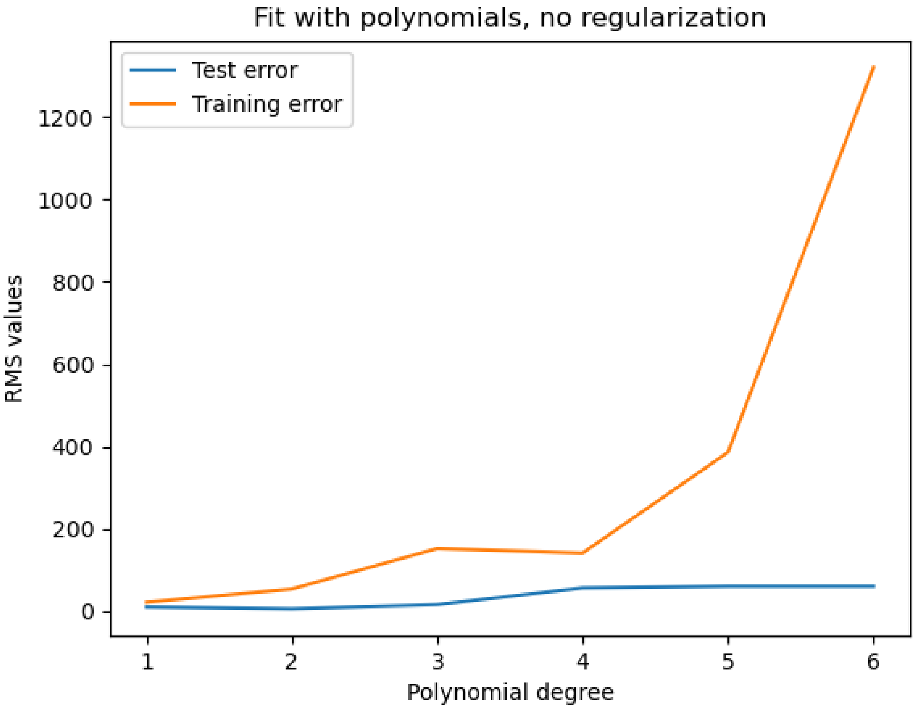
1. Which country had the highest child mortality rate in 1990? What was the rate
Answer:
Niger has the highest child mortality rate in 1990 = **313.7**

2. Which country had the highest child mortality rate in 2011? What was the rate?
Answer:
Sierra Leone has the highest child mortality rate in 2011 = **185.3**

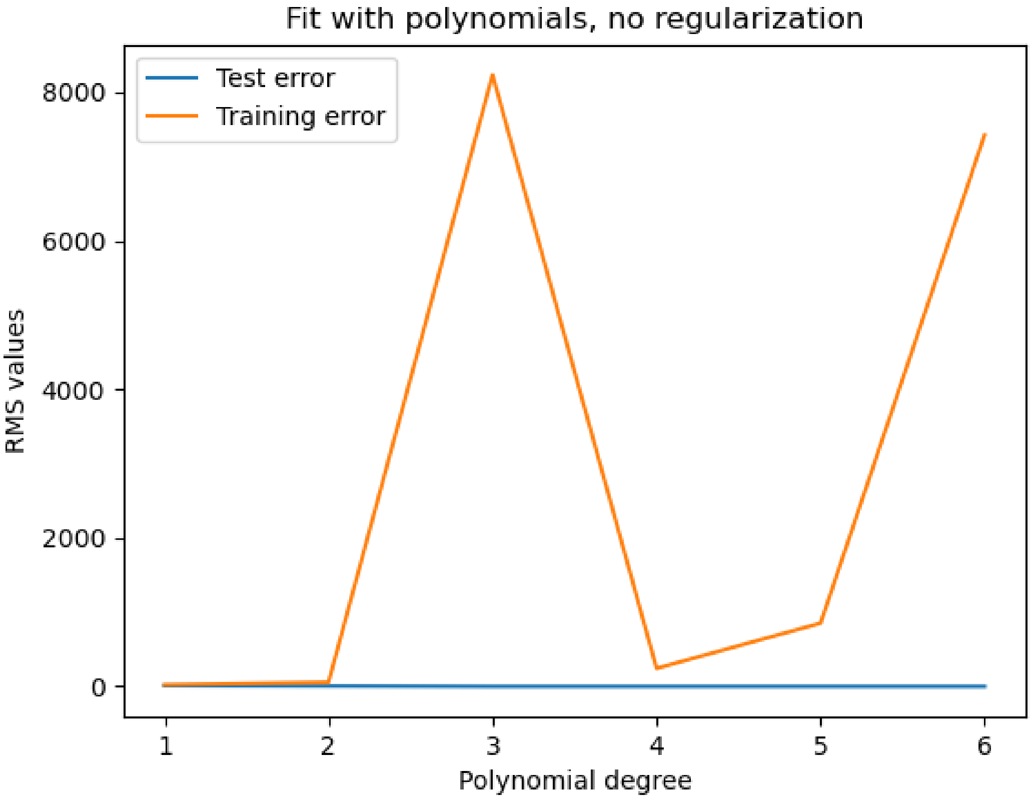
3. Some countries are missing some features (see original .xlsx/.csv spreadsheet). How is this handled in the function assignment2.load unicef data()?
Answer:
We replace those missing features (i.e. “_" or NaN values) with the mean of the column values which are not null.

2.2. Polynomial Regression

1. **Un-Normalised Plot**



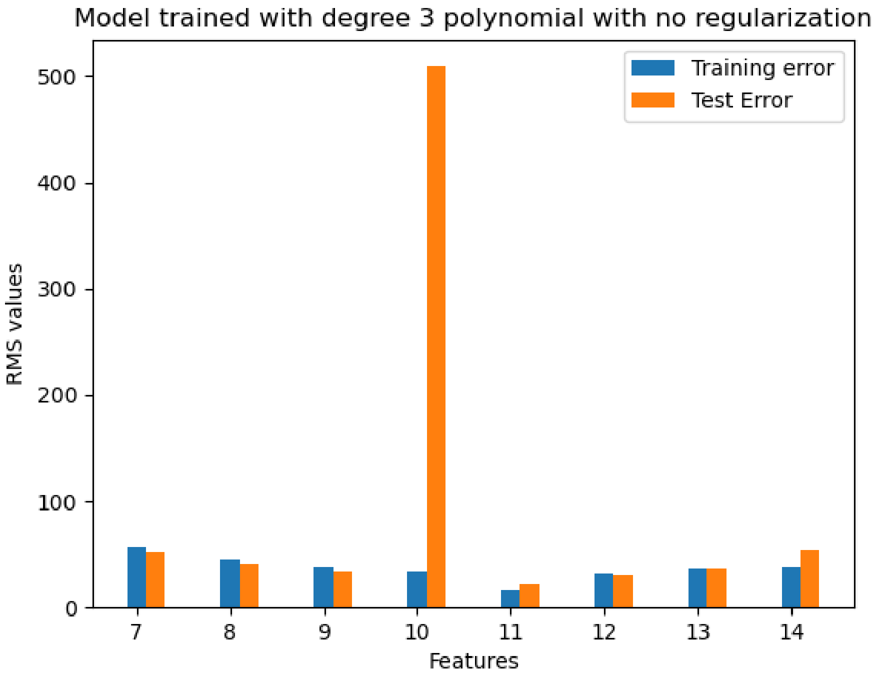
Normalized Plot



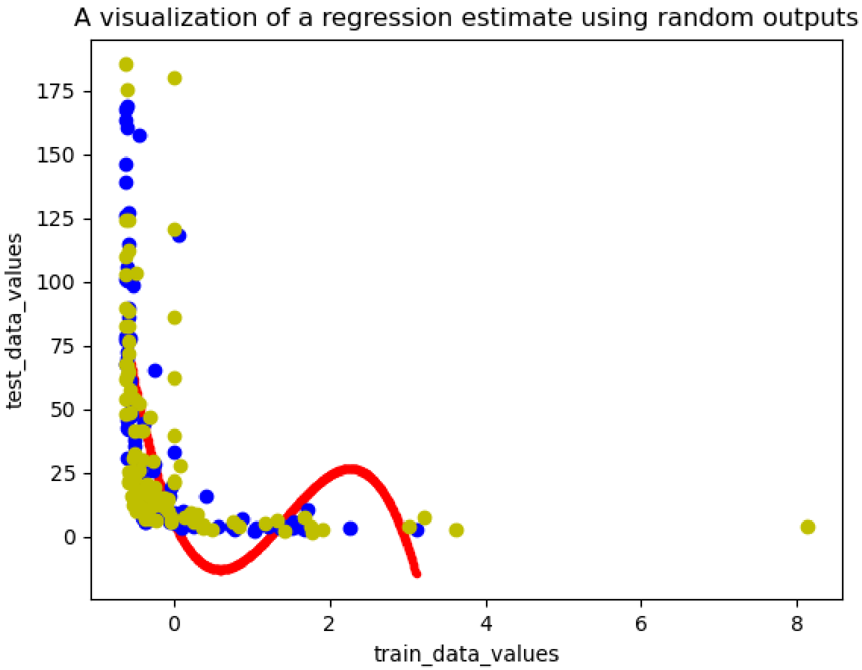
As, degree of the polynomial for our linear model increases the resultant RMS error should have been decreased. Since, as polynomial degree increases our model tries to fit to all the data points and as a result RMS should decrease.

However, here both training and testing error never decreases till polynomial degree 6. This plot suggests there is definitely something wrong with data. Normalizing the input data gives the plot below.

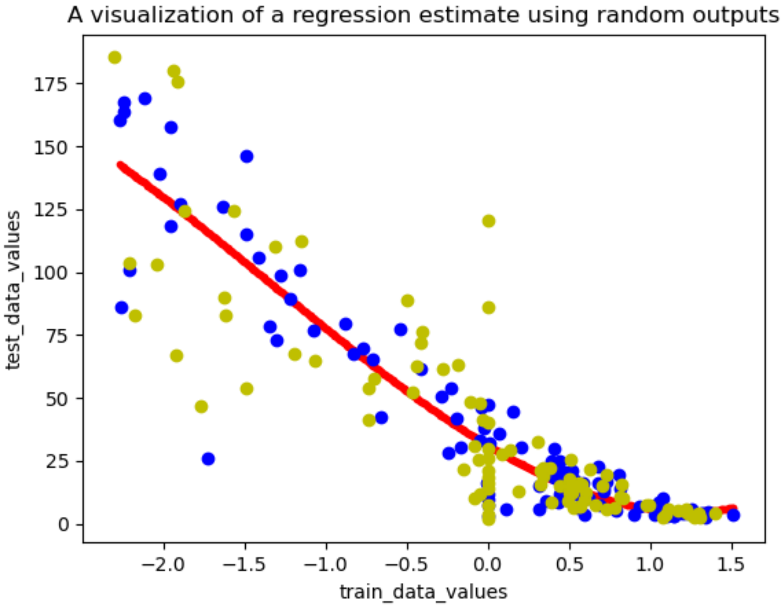
b. Bar chart for training error and test error (in RMS error) for each of the 8 features



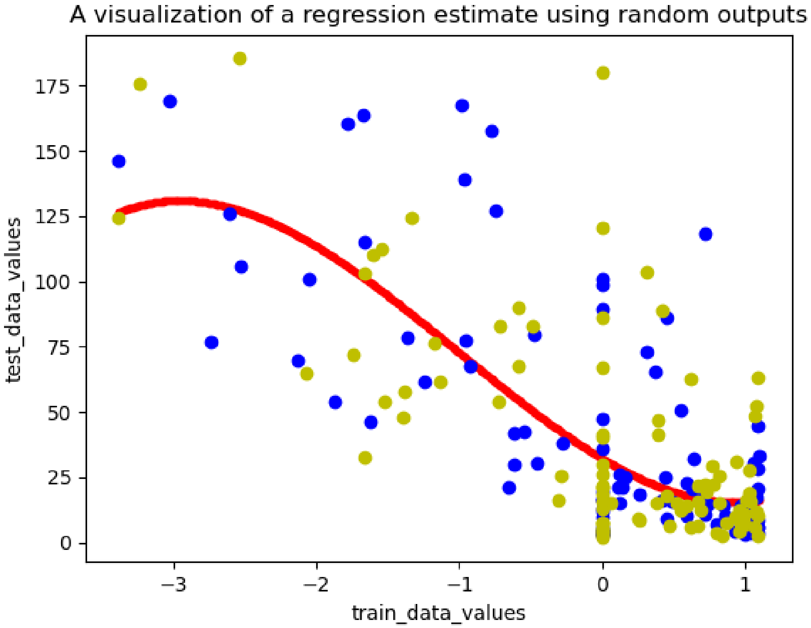
3 Degree polynomial fit to feature: GNI (The outlier data point in test set, shoots up the test error)



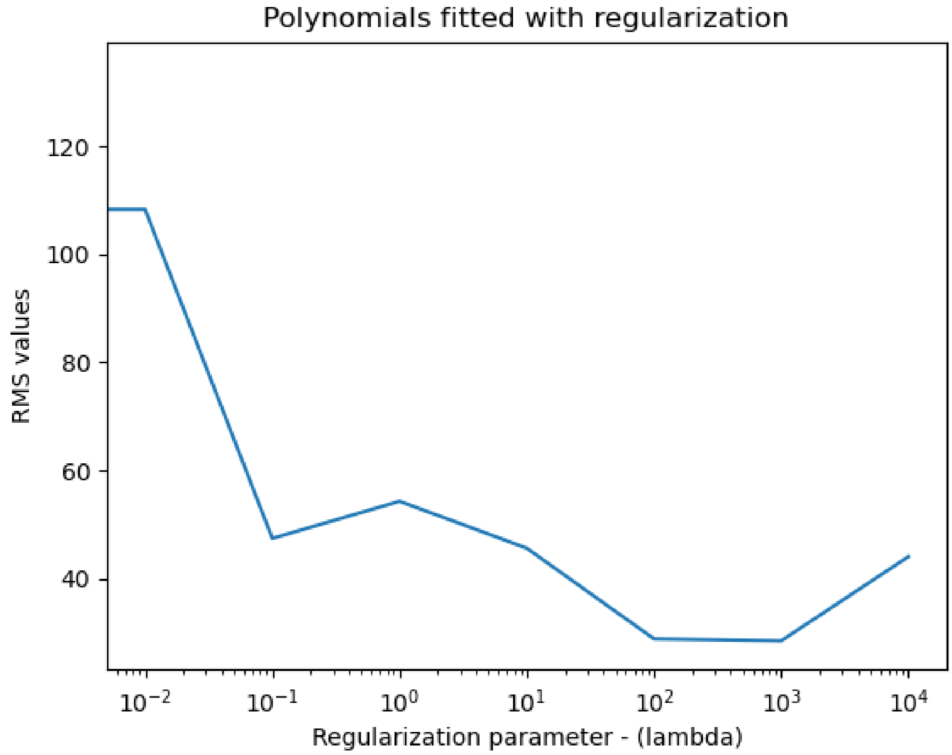
3 Degree polynomial fit to feature: Life expectancy



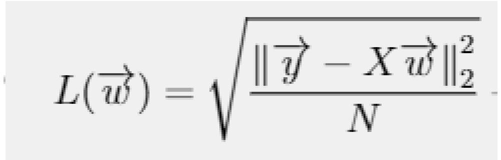
3 Degree polynomial fit to feature: Literacy



2.3 Regularized Polynomial Regression



After seeing the resultant plot of **polynomial_regression_reg.py** script, we can observe that the RMS value is minimal at lambda range (100,1000). So it will be optimal to choose any value between that range.
(Note): For the validation/testing phase use the loss function without the regularizer. And below formula, has been used.



$$L(\overrightarrow{w}) = \sqrt{\frac{\| \overrightarrow{y} - X\overrightarrow{w} \|_2^2}{N}}$$

3. Probabilistic Modeling and Bayes' Rule

a) Assume the probability of being infected with Malaria disease is 0.01. The probability of test positive given that a person is infected with Malaria is 0.95 and the probability of test positive given the person is not infected with Malaria is 0.05.
Answer:
Given,
P(malaria) = 0.01 and so P(no malaria) = 1 - P(malaria) = 1 - 0.01 = 0.99
P(positive | malaria) = 0.95
P(positive | no malaria) = 0.05

(a) Calculate the probability of test positive.

P(positive) = P(malaria) ∗ P(positive | malaria) + P(no malaria) ∗ P(positive | no malaria)
= (0.01 ∗ 0.95) + (0.99 ∗ 0.05)
= 0.0095 + 0.0495
= 0.059

(b) Use Bayes' Rule to calculate the probability of being infected with Malaria given that the test is positive.

P(malaria | positive) = P(malaria, positive) / P(positive)
= P(malaria) ∗ P (positive | malaria) / P(positive)
= (0.01 ∗ 0.95) / 0.059
= 0.16101694915

b) Suppose P(rain today) = 0.30, P(rain tomorrow) = 0.60, P(rain today and tomorrow) = 0.25. Given that it rains today, what is the probability it will rain tomorrow?
Answer:
Given,
P(rain today) = 0.30
P(rain tomorrow) = 0.60
P(rain today ∩ tomorrow) = 0.25

So, P(rain tomorrow | rain today) = P(rain today ∩ rain tomorrow) / P (rain today)
= (0.25 / 0.30)
= 0.83333333333

c) A biased die has the following probabilities of landing on each face:

Face	1	2	3	4	5	6
P(face)	0.2	0.2	0.1	0.1	0.1	0.3

i) I win if the die shows odd. What is the probability that I win? Compare this to a fair die (i.e., a die with equal probabilities for each face).

Answer:
P(win) = P(1) + P(3) + P(5) (i.e adding probabilities for all odd numbers)
= 0.2 + 0.1 + 0.1
= 0.4

P(win) of a fair die would be equal, so probability will be for each.

P(win) for a fair die = P(1) + P(3) + P(5)
= 3 ∗ (1/6)
= 0.5

ii) What is the entropy of this die? Compare this to a fair die.

Answer:
Entropy = - $\Sigma\ p_i \log_e (p_i)$, where i is from 1 to 6
= - $[(2* (0.2) \log_e (0.2)) + (3*(0.1) \log_e (0.1)) + (0.3) \log_e (0.3)]$
= - (- 0.64377516497 - 0.69077552789 - 1.20397280433)
= 2.53852349719

Entropy for a fair die is = - $\Sigma\ p_i \log_e (p_i)$, where i is from 1 to 6
= - $(6*(1/6) \log_e ())$
= 1.79175946923