CIS 581 001 – COMPUTATIONAL LEARNING

MIDTERM PROJECT

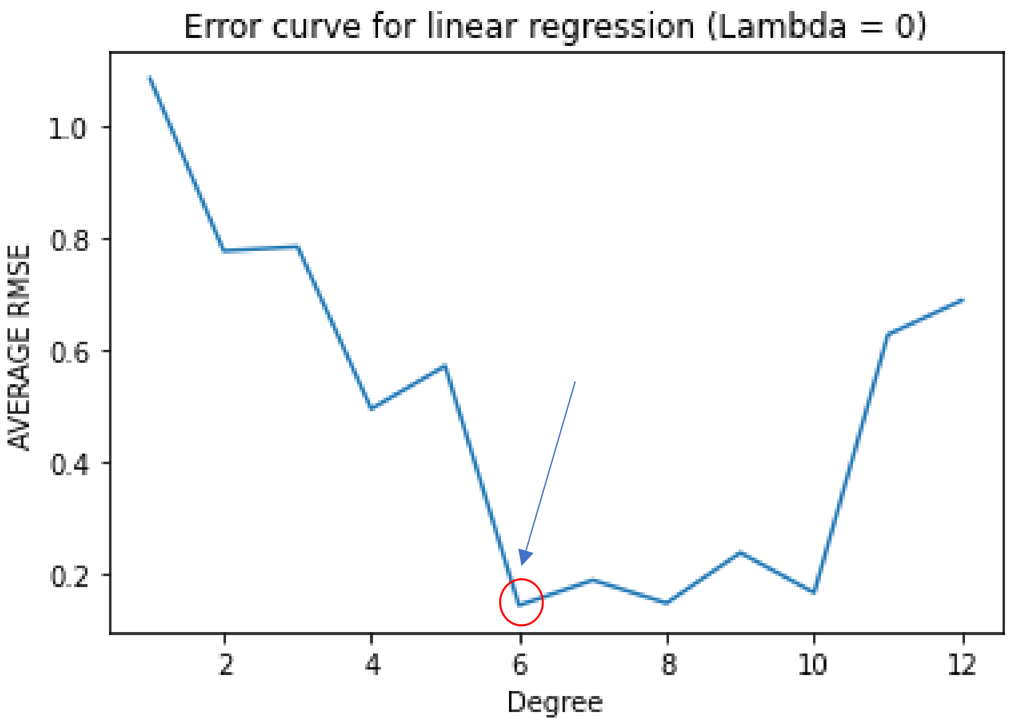
POLYNOMIAL CURVE FITTING REGRESSION FOR WORKING-AGE DATA

NAME	:	SURYA SUBRAMANI
UMID	:	69499602
DATE OF SUBMISSION	:	30-13-2023

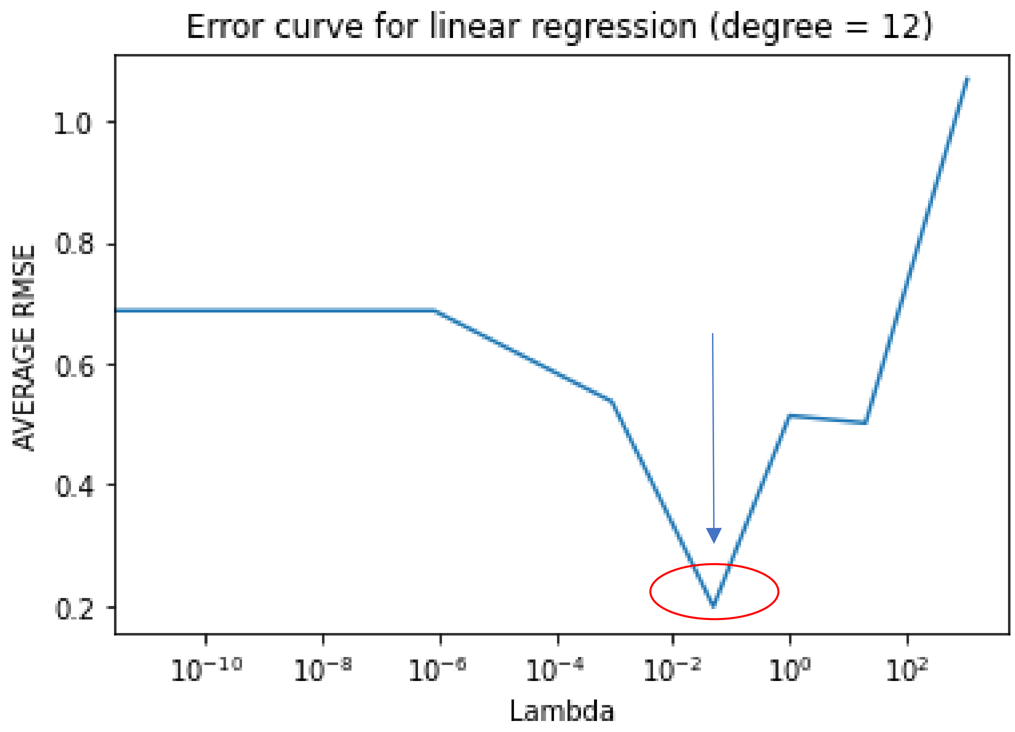
Polynomial Curve-Fitting Regression for Working-Age Data

1. The averages of the RMSE values obtained during the 6-fold CV for each case

   The below error charts shows the optimal **d = 6** and alpha = exp(-3) or 0.049



d	avg RMSE
1	1.083556162
2	0.775429312
3	0.783001685
4	0.493481844
5	0.570135013
6	0.142352625
7	0.187508475
8	0.146310074
9	0.236472911
10	0.1648808
11	0.625783576
12	0.686978331



l	avg RMSE
0	0.686978
1.39E-11	0.686978
2.06E-09	0.686978
8.32E-07	0.686958
0.000912	0.53732
0.049787	0.198978
1	0.513969
1096.633	1.069071
20.08554	0.501868

2. The optimal degree d∗ and regularization parameter λ∗ obtained via the 6-fold CV

d* = **6** with $\lambda = 0$
RMSE Average is **0.14235265**
$\lambda* = \mathbf{e^{-3}}$ with d = 12
RMSE Average is **0.198978**

3. The coefficient-weights of the d∗-degree polynomial and the λ∗-regularized 12-degree learned on all the training data

Weights for d* = **6** all training data

w1	w2	w3	w4	w5	w6
0.416333	3.93562	0.159123	-3.44054	0.00982416	0.627914

Weights for d = **12**, λ∗ = $\mathbf{e^{-3}}$, all training data

w1	w2	w3	w4	w5	w6
0.46638	2.84362	0.0968637	-1.19065	-0.00451745	-0.97337

w7	w8	w9	w10	w11	w12
0.0164566	0.274471	-0.000703978	0.106803	-0.0000783	-0.0306244

4. The training and test RMSE of that final, learned polynomials

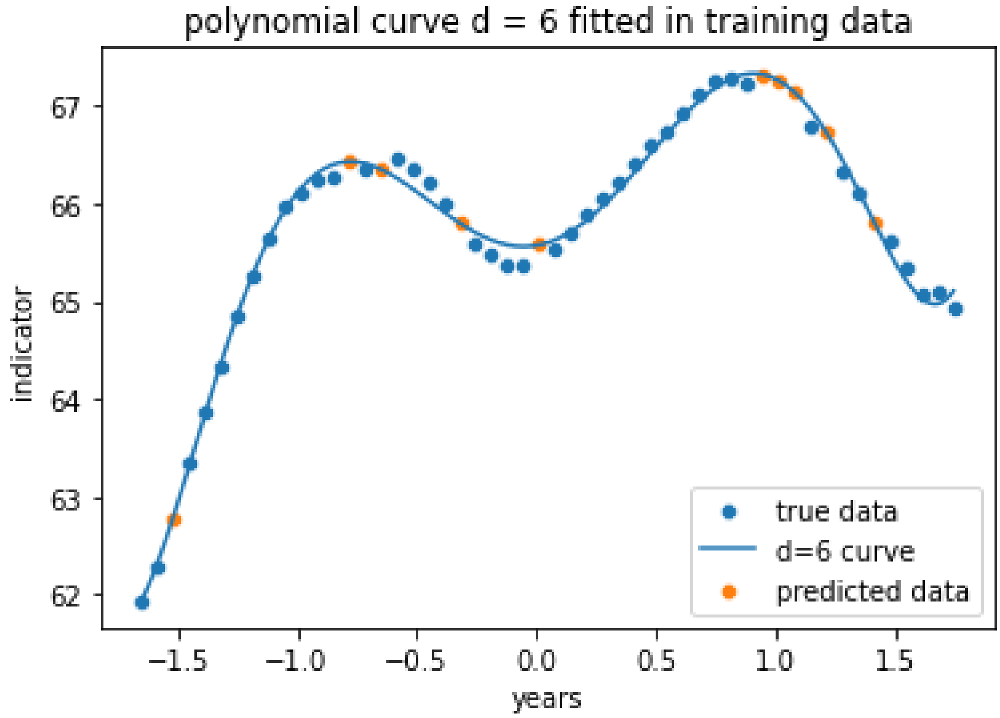
Training RMSE for d* = **6** is (Best Degree)
0.105401066732

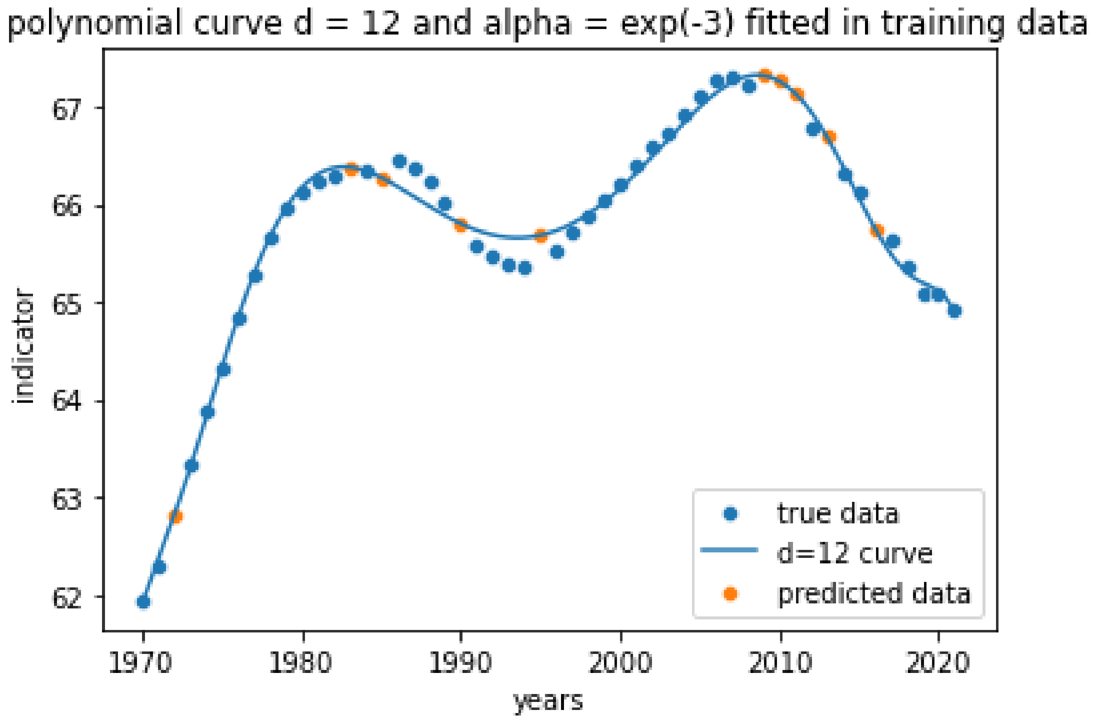
Testing RMSE for d* = **6** is (Best Degree)
0.11432570919500

Training RMSE for d = **12**, λ∗ = $\mathbf{e^{-3}}$ is
0.12755963811596235

Testing RMSE for d = **12**, λ∗ = $\mathbf{e^{-3}}$ is
0.12863909044553795

5. The 2 plots containing all the training data along with the resulting polynomial curves for d∗ and λ∗, for the range of years 1968-2023 as input





6. Brief discussion of your findings and observations.

From the data and the graph, we can able to observe that for every 20 years the age indicator goes to the peak and starts descending. This repeats for every 30 years with an increase in indicator. With the training data year ranges from 1970 to 2021.

Criteria	Year	True Values
count	42	42
mean	1994.80952	65.70398
std	15.209493	1.181847
min	1970	61.93886
25%	1981.25	65.35731
50%	1995	65.98761
75%	2005.75	66.36326
max	2021	67.29843

(the above shown Data is before scaling)

To reduce the standard deviation, which can affect the model training, The input matrix is scaled to minimum values using (X-mean)/std.

Criteria	Year	True Values
count	42.00	42.00
mean	0.00	65.70
std	1.00	1.18
min	-1.63	61.94
25%	-0.89	65.36
50%	0.01	65.99
75%	0.72	66.36
max	1.72	67.30

(the above shown Data is after scaling)

Repeating the same process inside cross validation for each training test pair

To train model and calculate weights I have used the below equation

$W = (X^T.X + lambda*I).X^T.y$

Which is the derivation of

$$L(\mathbf{w}) = \frac{1}{2}\sum_{l=1}^{m}\left(y^{(l)} - \sum_{i=0}^{d} w_i \left(x^{(l)}\right)^i\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

The randomized data is benefitted for training since it is a time series data. If it is sorted based on training the cross validation will leads to information leak. Prophet and XGB can be a better models for this data to train and forecast.