CS 6375
ASSIGNMENT -1
(Linear Regression using Gradient Descent)

Names of students in your group:

Siddhant Suresh Medar (ssm200002)
Adithya Sundararajan Iyer (asi200000)

Number of free late days used: _________0_________

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment:

1) Inductive Learning slides from E-Learning CS6375 Course Contents Page
2) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
3) https://towardsdatascience.com/stochastic-gradient-descent-clearly-explained-53d239905d31
4) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
5) https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
6) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
7) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html
8) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html
9) https://scikit-learn.org/stable/modules/generated/sklearn.metrics.explained_variance_score.html

Dataset used:

Computer Hardware Data Set

(https://archive.ics.uci.edu/ml/datasets/Computer+Hardware)

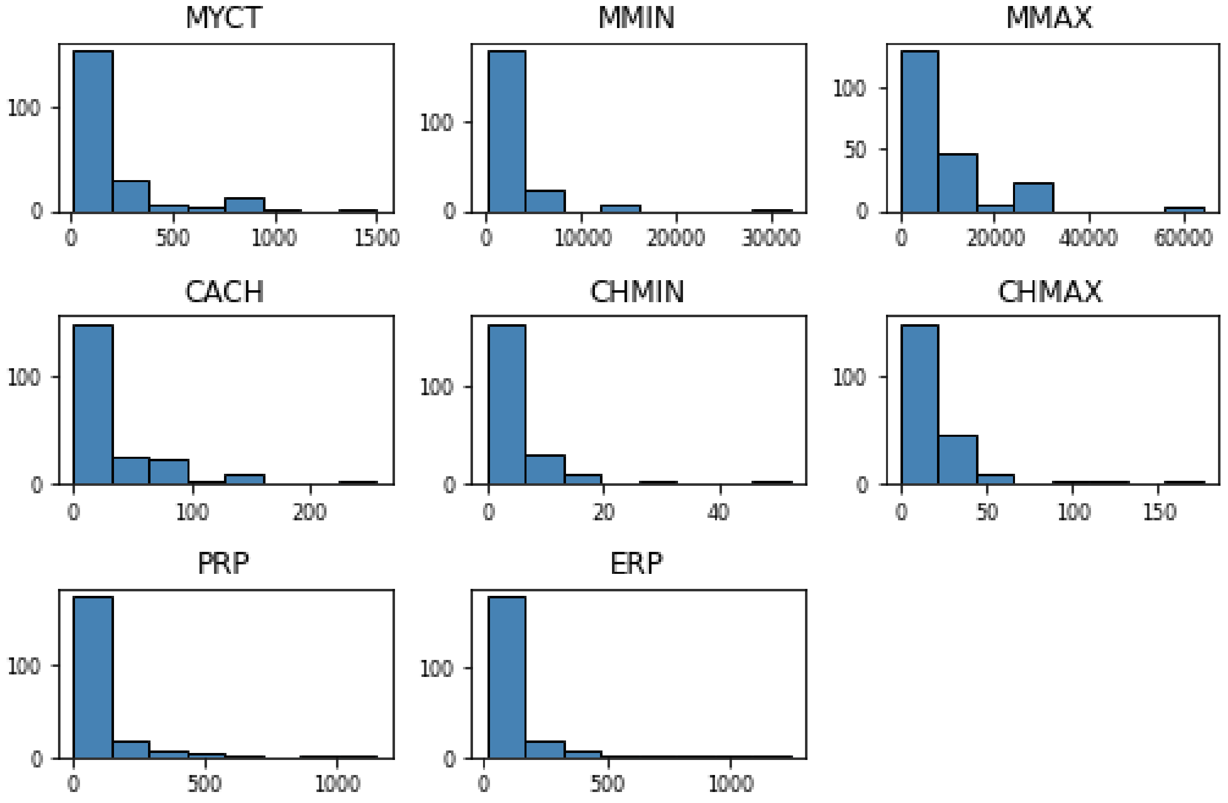
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

Data Preprocessing

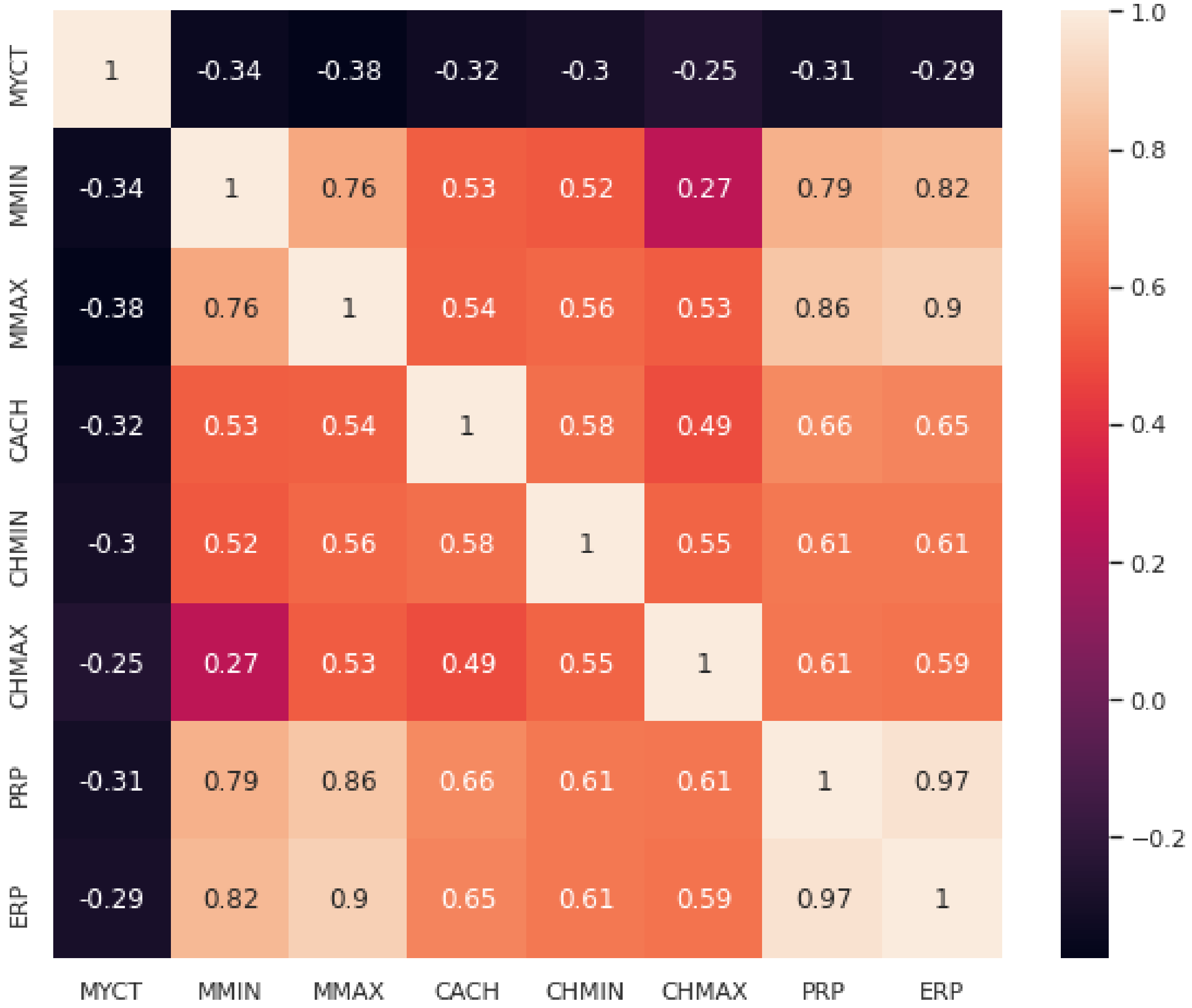
1. Check for null or NA values and remove, none found
2. Remove redundant rows, no duplicate values found
3. Categorical variables found but do not have correlation with outcome, hence only 8/10 attributes considered for training and prediction
4. Dataset normalized using Standard Scaler library

Feature Engineering

Histogram Plots for each column

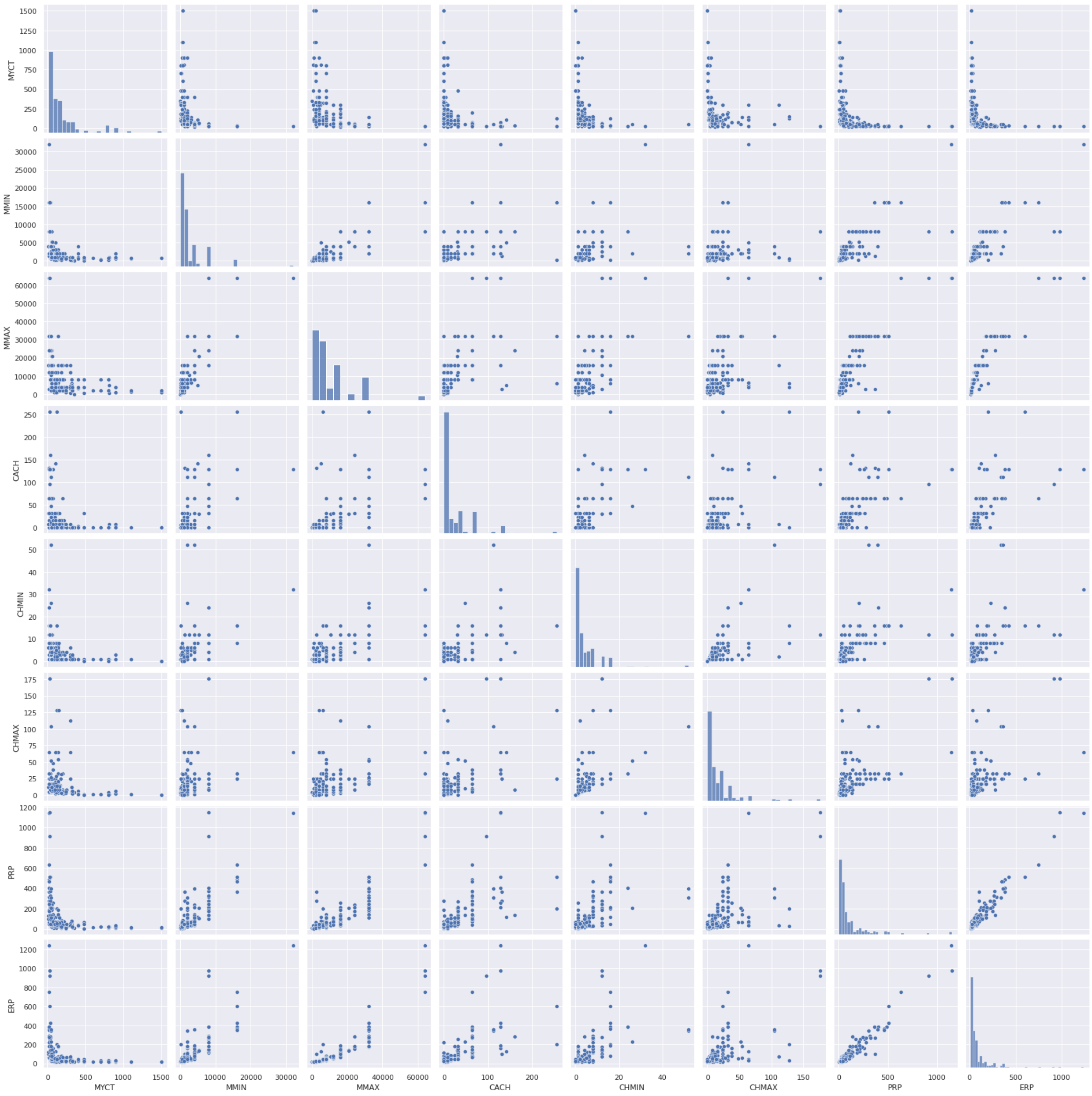


Heatmap Plot for Visual Correlation Matrix



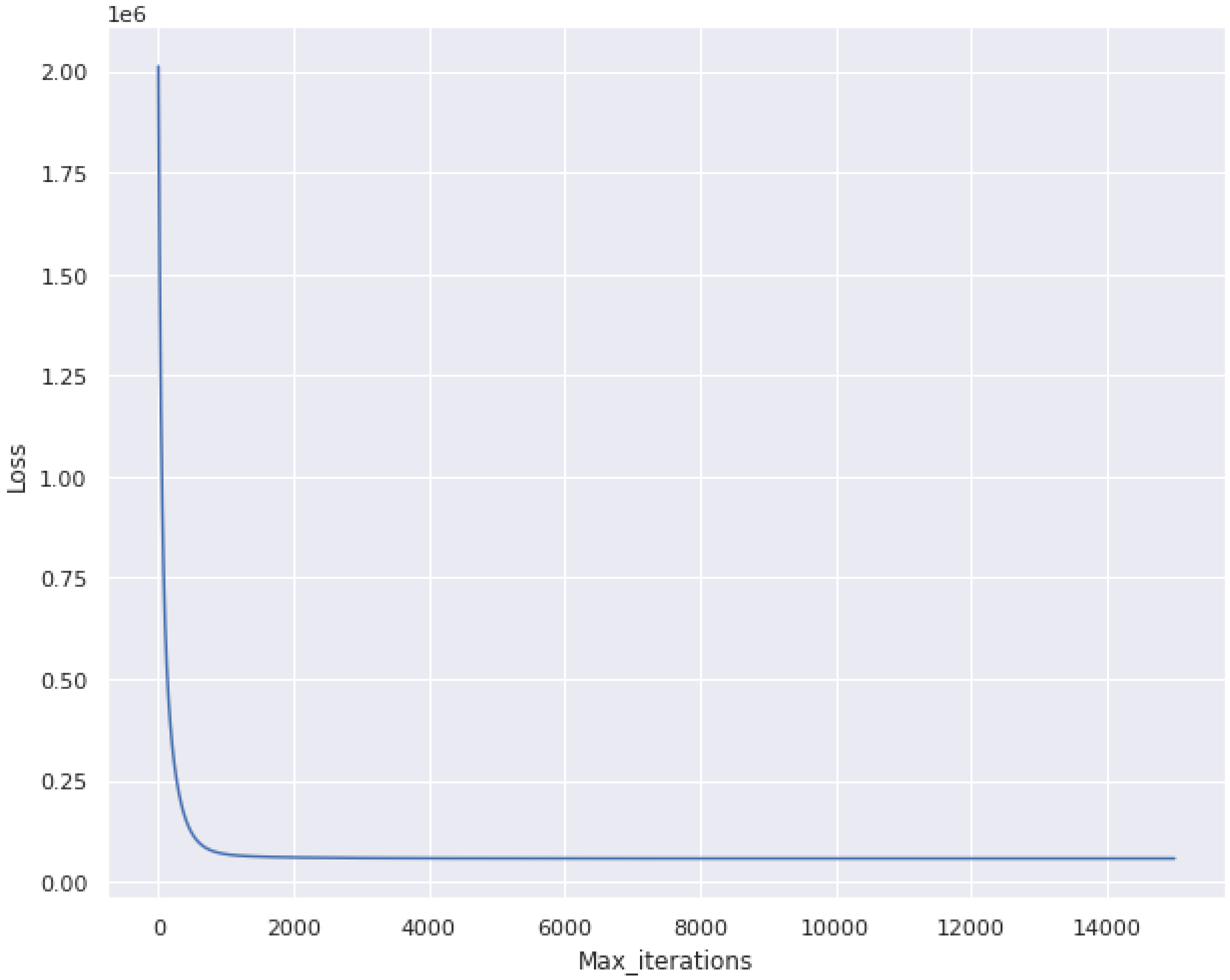
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP	ERP
MYCT	1	-0.34	-0.38	-0.32	-0.3	-0.25	-0.31	-0.29
MMIN	-0.34	1	0.76	0.53	0.52	0.27	0.79	0.82
MMAX	-0.38	0.76	1	0.54	0.56	0.53	0.86	0.9
CACH	-0.32	0.53	0.54	1	0.58	0.49	0.66	0.65
CHMIN	-0.3	0.52	0.56	0.58	1	0.55	0.61	0.61
CHMAX	-0.25	0.27	0.53	0.49	0.55	1	0.61	0.59
PRP	-0.31	0.79	0.86	0.66	0.61	0.61	1	0.97
ERP	-0.29	0.82	0.9	0.65	0.61	0.59	0.97	1

Pairplot to check Pairwise Relationships in Dataset



PART 1 – Implementing SGD regressor manually

Optimum Learning Rate(α=0.003) and epochs (15000 iterations)

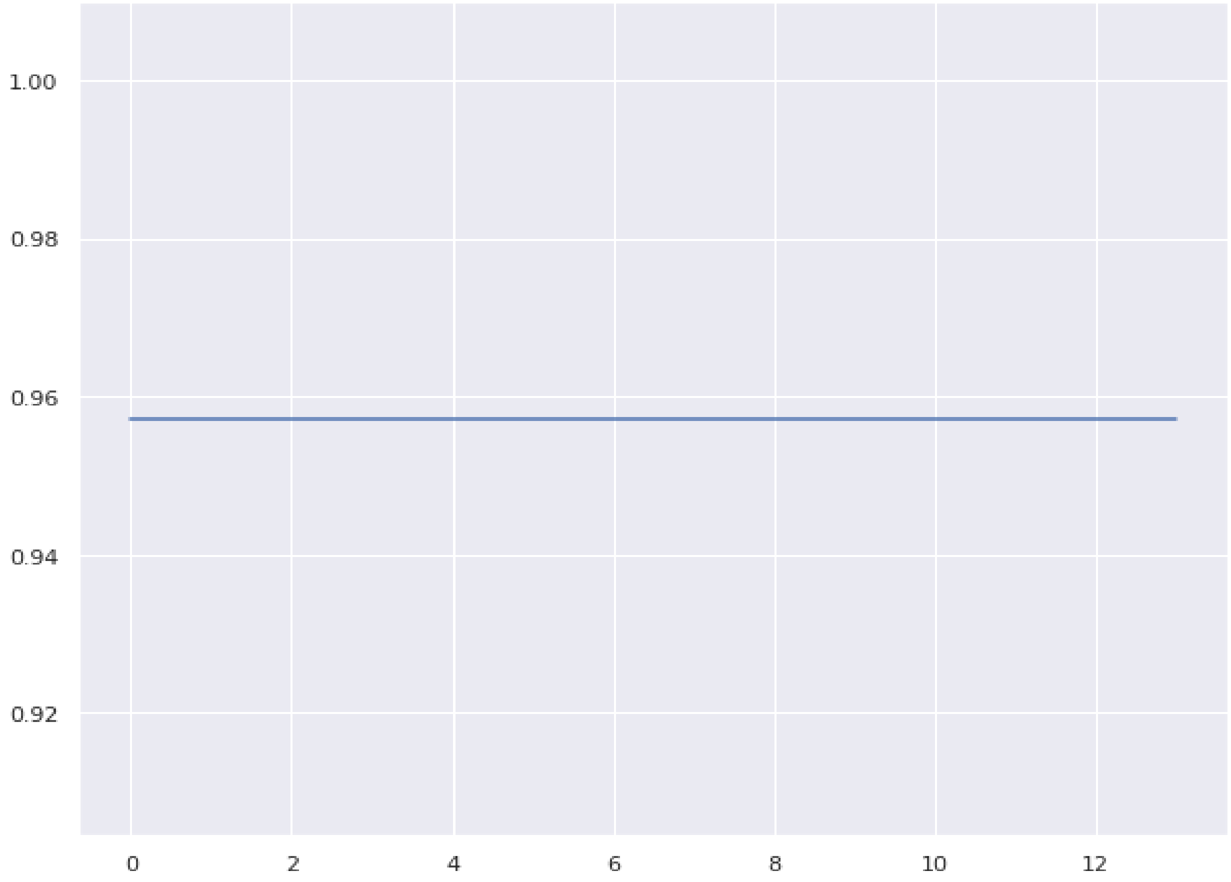


Error scores used:

- R^2 score – coefficient of determination
- MAE – Mean Absolute Error
- RMSE – Root Mean Squared Error
- EVS – Explained Variance Score

Error Log of Epoch Variations for fixed LR(α=0.003)

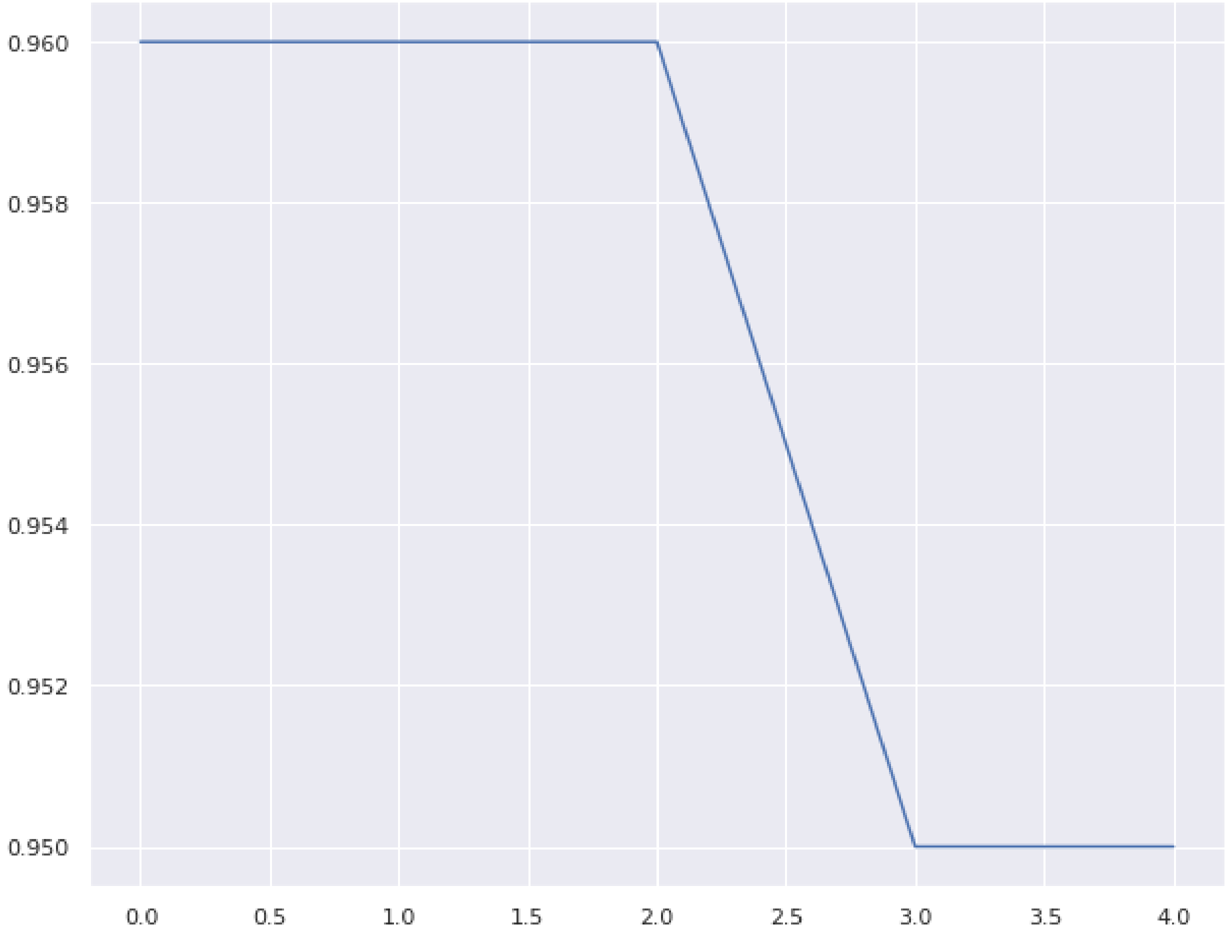
SlNo	Iterations	R^2 Score	MAE	RMSE	EVS
0	15000	0.957172610399	25.52664923942	54.8801227278	0.946392668976
1	17500	0.957172779916	25.52037231232	54.8593267532	0.946433051997
2	20000	0.95717280876	25.51773619243	54.8507670111	0.946449687797
3	22500	0.957172814091	25.51658697346	54.8470442278	0.946456927509
4	25000	0.957172815060	25.51609973310	54.8454677575	0.946459994310
5	27500	0.957172815248	25.51588214443	54.8447642338	0.946461363205
6	30000	0.957172815283	25.51578865526	54.8444620781	0.946461951205
7	32500	0.957172815289	25.51574771543	54.8443297868	0.946462208662
8	35000	0.957172815290	25.51573062114	54.8442745547	0.946462316155
9	37500	0.957172815291	25.51572322112	54.8442506466	0.946462362686
10	40000	0.957172815291	25.5157199901	54.8442402082	0.946462383002
11	42500	0.957172815291	25.51571862188	54.8442357880	0.946462391605
12	45000	0.957172815291	25.51571804087	54.844233911	0.946462395259
13	47500	0.957172815291	25.51571777851	54.8442330634	0.946462396908
14	50000	0.957172815291	25.51571767046	54.8442327143	0.946462397588



Error Log of various Learning Rates for fixed Iterations (=15000)

SlNo	LR(α)	R^2 Score	MAE	RMSE	EVS
0	0.003	0.957172621674	25.52642699684	54.8790749199	0.946394711746
1	0.0015	0.957140565604	25.62907692646	55.2307461803	0.945717426802
2	0.00075	0.95670209178	25.7593056678	55.9008202628	0.944457948229
3	0.000375	0.954967734119	25.89714565067	56.6620442898	0.943071207459
4	0.0001875	0.948059108631	27.51813488562	60.7232832299	0.93674371651

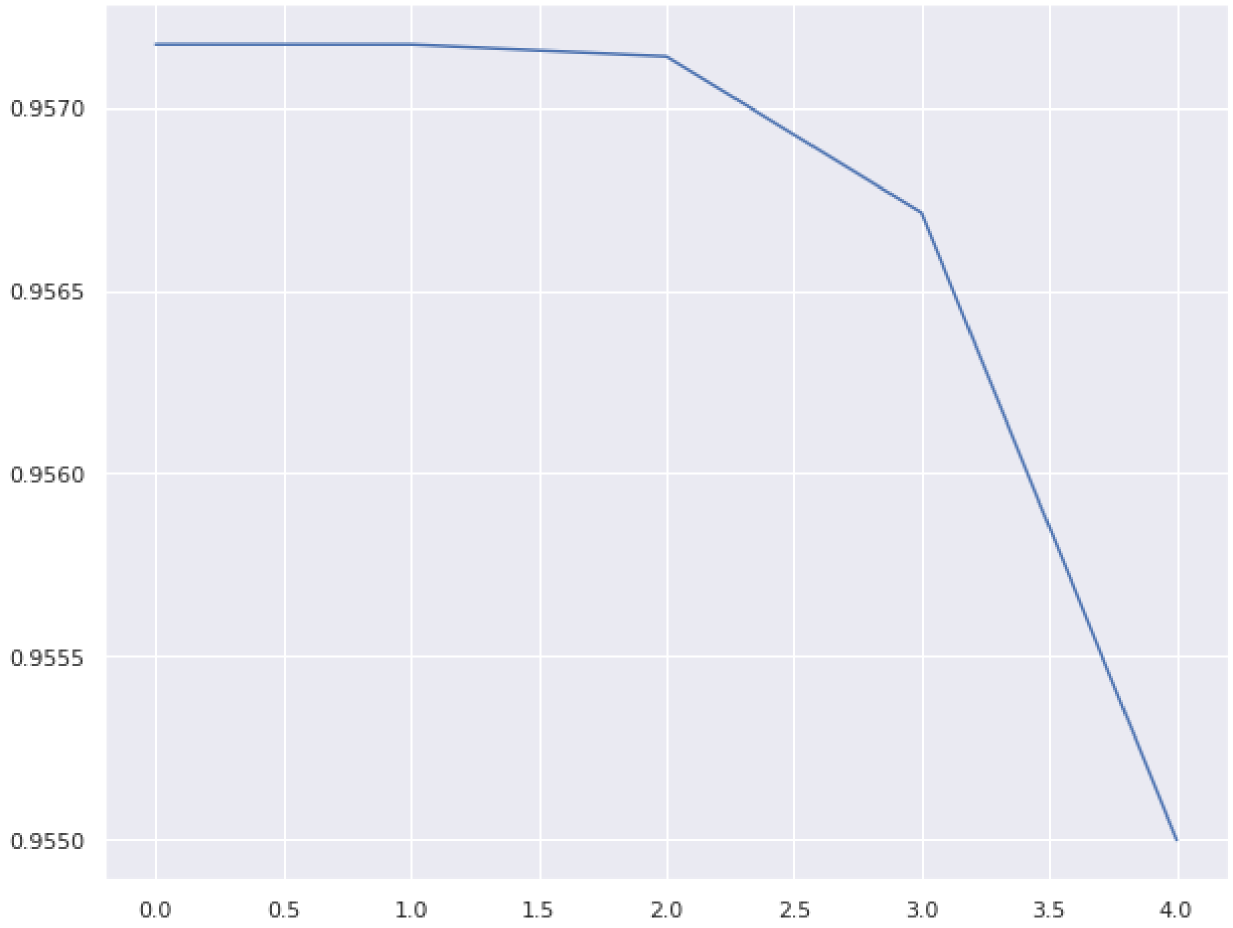
Graph of R^2 score at different trials (0,1,2,3,4) of learning rate



Error Log of various Learning Rates for fixed Iterations (=30000)

SlNo	LR(α)	R^2 Score	MAE	RMSE	EVS
0	0.003	0.957172815283	25.51578849325	54.8444615538	0.946461952224
1	0.0015	0.957172616848	25.52657266002	54.8795424057	0.946393798317
2	0.00075	0.957139743491	25.62997047277	55.2341948385	0.945710968008
3	0.000375	0.956712628948	25.74735647152	55.8630077198	0.944534172478
4	0.0001875	0.95499911213	25.87462331295	56.6314304672	0.943127614705

Graph of R^2 score at different trials (0,1,2,3,4) of learning rate



Manual SGD Regressor performance result: R^2 score = 95.72%

PART 2 – Implementing SGD regressor using Scikit-learn Library

Model provided with learning rate (eta0) values ranging from 0.001-0.1 and number of epochs (max_iter) values ranging from 500-50000. GridSearchCV loops through these values to find the best estimators as hyperparameters.

```
Result: SGDRegressor(eta0=0.02,
 learning_rate='constant', max_iter=30000)
```

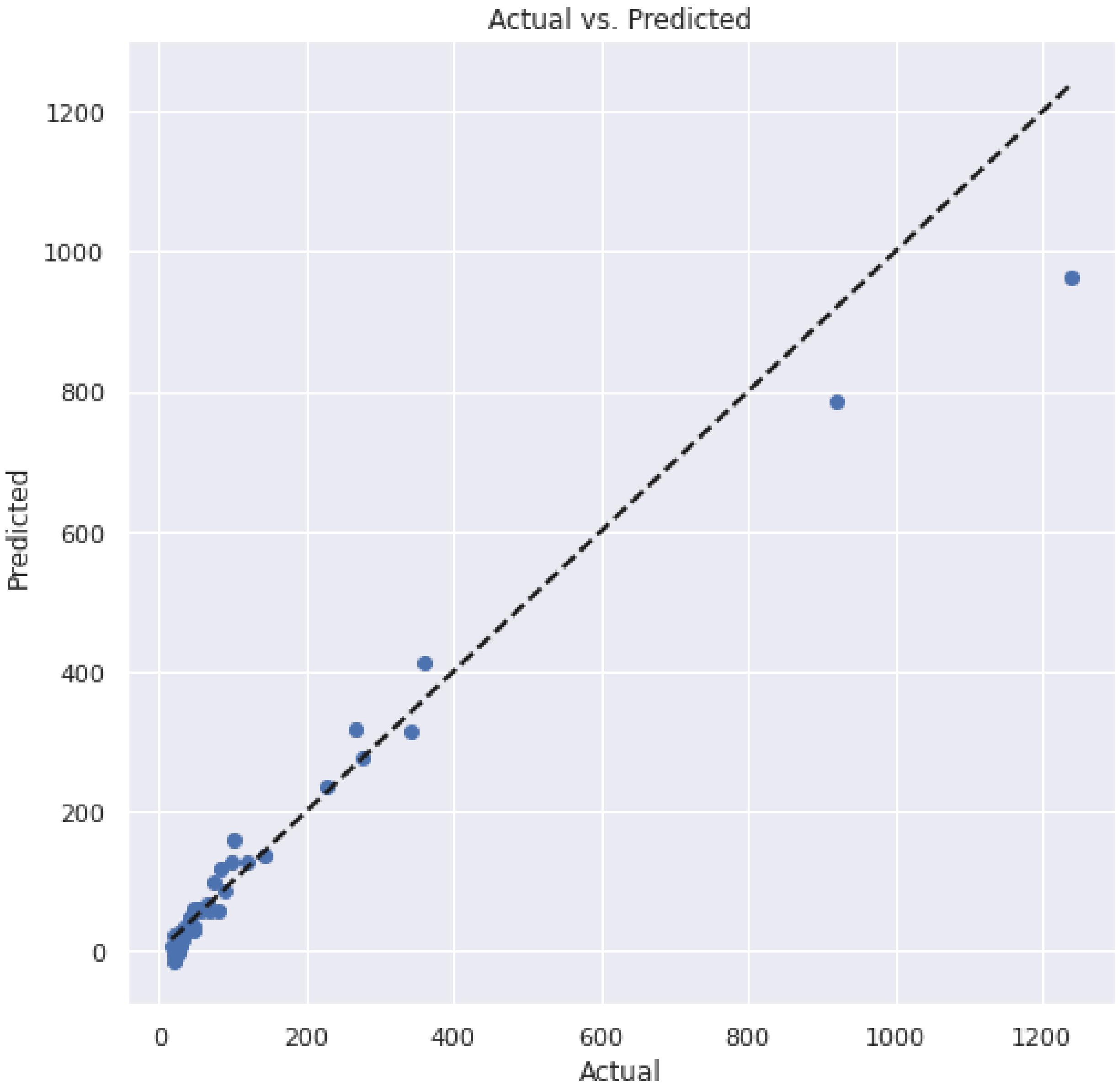
Calculating the metrics for model trained using ML Library:

R2 Score: 0.9426297591728554
Mean absolute error: 30.693156077522872
Root Mean squared error: 66.38940030040736
Explained Variance Score: 0.9225820701019357

Conclusion: We obtained a r2 score of 95.16% using scikit library which is less than that of our custom SGD regressor's score using GridSearchCV

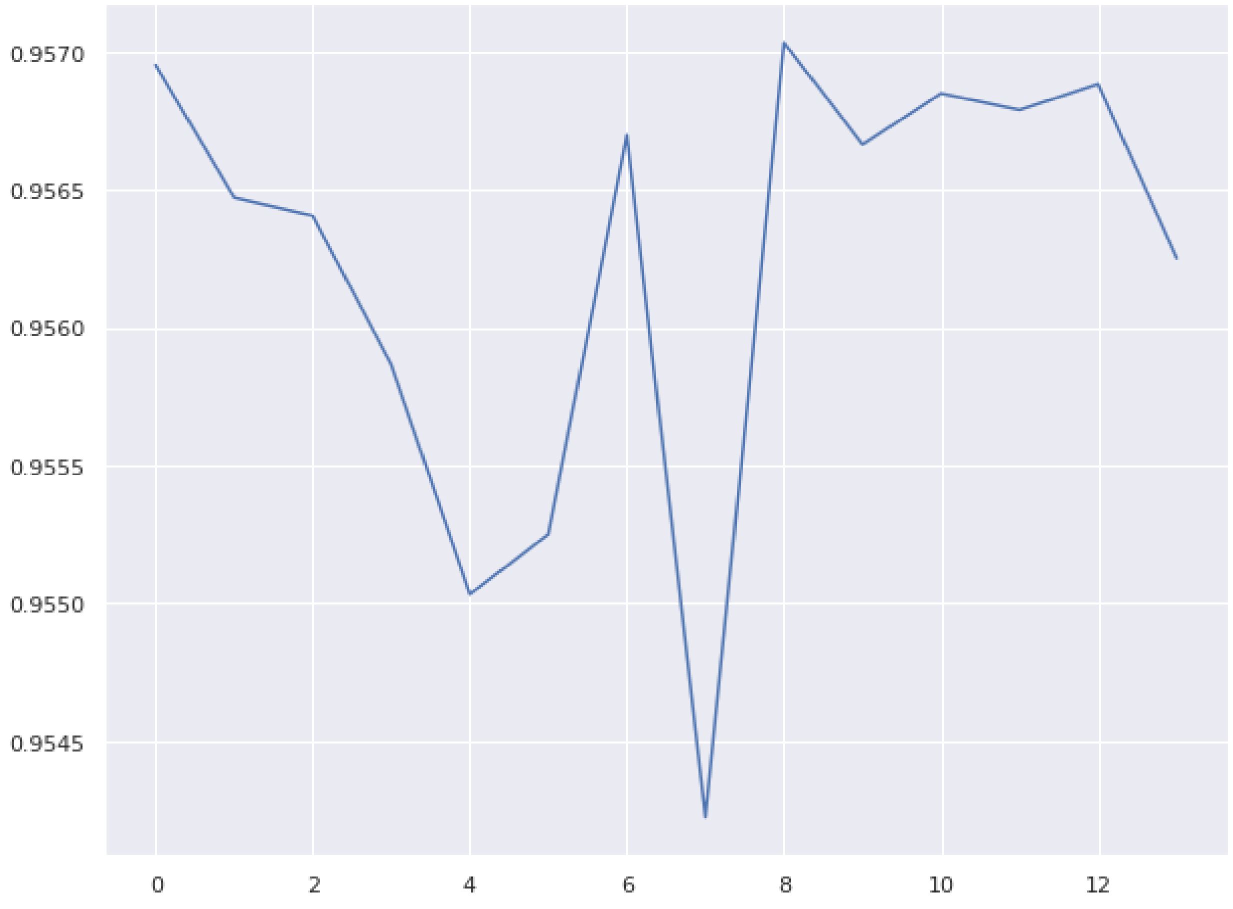
Comparing model performance with the same hyperparameters used as in manual part

LR: 0.003 Iterations= 15000
R2 Score: 0.9565421203737108
Mean absolute error: 25.47748191911172
Root Mean squared error: 51.82665486241274
Explained Variance Score: 0.9521341027181411



Error Log of Epoch Variations for fixed LR(α=0.003)

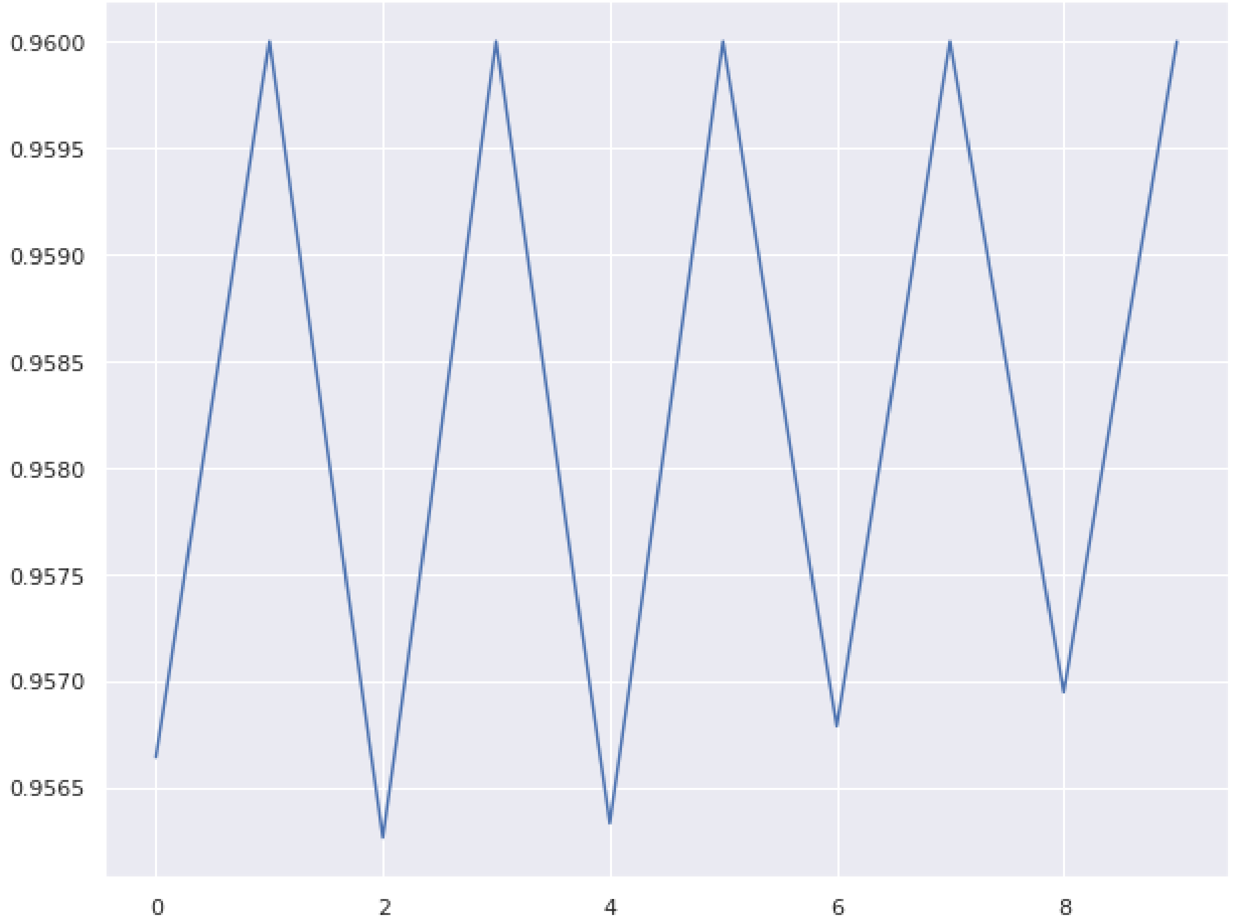
SlNo	Iterations	R^2 Score	MAE	RMSE	EVS
0	15000	0.956542120373	25. 4774819191	54. 8266548624	0. 95213410271
1	17500	0.956952512809	25.59392807145	54.1855701909	0.947641219317
2	20000	0.956471519270	25.99351321665	56.83475104137	0.942684199743
3	22500	0.956405783762	26.18684809482	58.5179681920	0.939236814162
4	25000	0.955865633395	25.68461130319	52.60741421817	0.950769628453
5	27500	0.955032806788	25.62997170564	50.22067472102	0.954925957154
6	30000	0.955250553082	25.79416301917	50.76103845609	0.954189393423
7	32500	0.956699036340	25.63614204248	53.82061946605	0.948447569676
8	35000	0.954225043749	26.05364056943	49.93717463993	0.955671697044
9	37500	0.957033155617	25.83464524771	55.14027004551	0.946027256631
10	40000	0.956664723522	25.90995952362	55.9377796991	0.944500662315
11	42500	0.956849110160	25.86020412549	55.64029768059	0.944935469159
12	45000	0.956789997765	25.65515485384	53.71324042685	0.948735128826
13	47500	0.956882909972	25.72057496177	54.524942634	0.947204113484
14	50000	0.956252071268	26.1479768005	58.40769090113	0.939389031301



Error log of various learning rates for fixed iterations (=15000)

SlNo	LR(α)	R^2 Score	MAE	RMSE	EVS
0	0.003	0.956640825671	25.69740932056	54.1351827560	0.947891526963
1	0.0015	0.956263920777	25.77364799237	56.0271333517	0.944261961023
2	0.00075	0.956331156605	25.7735298892	55.6436296669	0.94497324373
3	0.000375	0.956785203719	25.74990960733	55.6736549367	0.944898911889
4	0.0001875	0.956945792422	25.73419414543	55.6275830875	0.944973891875

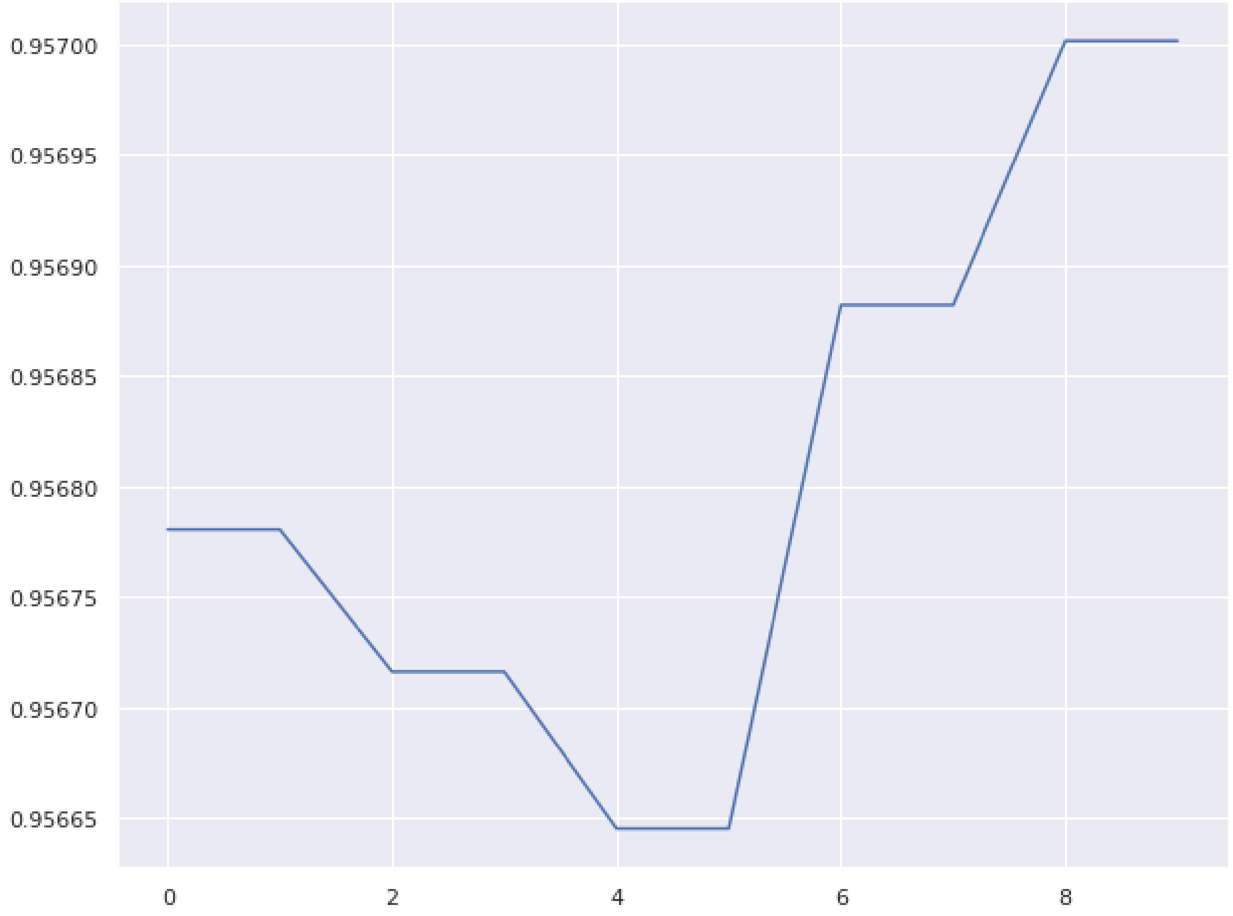
Graph of R^2 score at different trials (0,1,2,3,4) of learning rate



Error log of various learning rates for fixed iterations (=150000)

SlNo	LR(α)	R^2 Score	MAE	RMSE	EVS
0	0.003	0.956780368464	25.81688007988	54.9652418284	0.946376045071
1	0.0015	0.956716079937	25.73949250215	55.0059028139	0.94624445046
2	0.00075	0.956645224135	25.79057871112	56.1910902298	0.943877893974
3	0.000375	0.956881890948	25.7466705892	55.6396609266	0.944959022578
4	0.0001875	0.957001530409	25.71783795648	55.5331705772	0.9451487248

Graph of R^2 score at different trials (0,1,2,3,4) of learning rate



Library Linear Regressor performance result: R^2 score = 95.70%

Are you satisfied that the package has found the best solution? How can you check? Explain.

Yes, we are satisfied that we have found an optimal solution as is seen by the obtained R^2 score of the manual approach which is very close to that obtained with the Scikit-learn Library implementation of the Linear Regressor.
We can check this by comparing the various metrics used apart from R^2 score (co-efficient of determination) as well such as MAE (mean absolute error), RMSE (root mean squared error), and EVS (explained variance score).