Compare Regression Methods.

Introduction

This report analyzes the application of Linear Regression, RANSAC Regression, Ridge, Lasso, ElasticNet, and Random Forest regression techniques, on the California housing dataset. The performance of these models was evaluated based on Mean Squared Error (MSE), R2 score, and training time. Hyperparameter tuning was conducted to assess its impact on model performance, and residual plots were used to analyze the distribution of errors.

Methodology

Dataset:

- For this assignment, we are using the California housing dataset, which we load using fetch_california_housing from sklearn.datasets.
- The California Housing Data has 20640 rows and 9 columns.
- There are no missing values in the dataset columns.
- All features were used for training and testing.
- The dataset was split into training and testing sets (80:20 ratio) and standardized using StandardScaler.

Evaluation Metrics:

The regression techniques were implemented in Python using the scikit-learn library and evaluated based on the following metrics.

- Mean Squared Error (MSE)Execution Time
- R2 score
- Training time

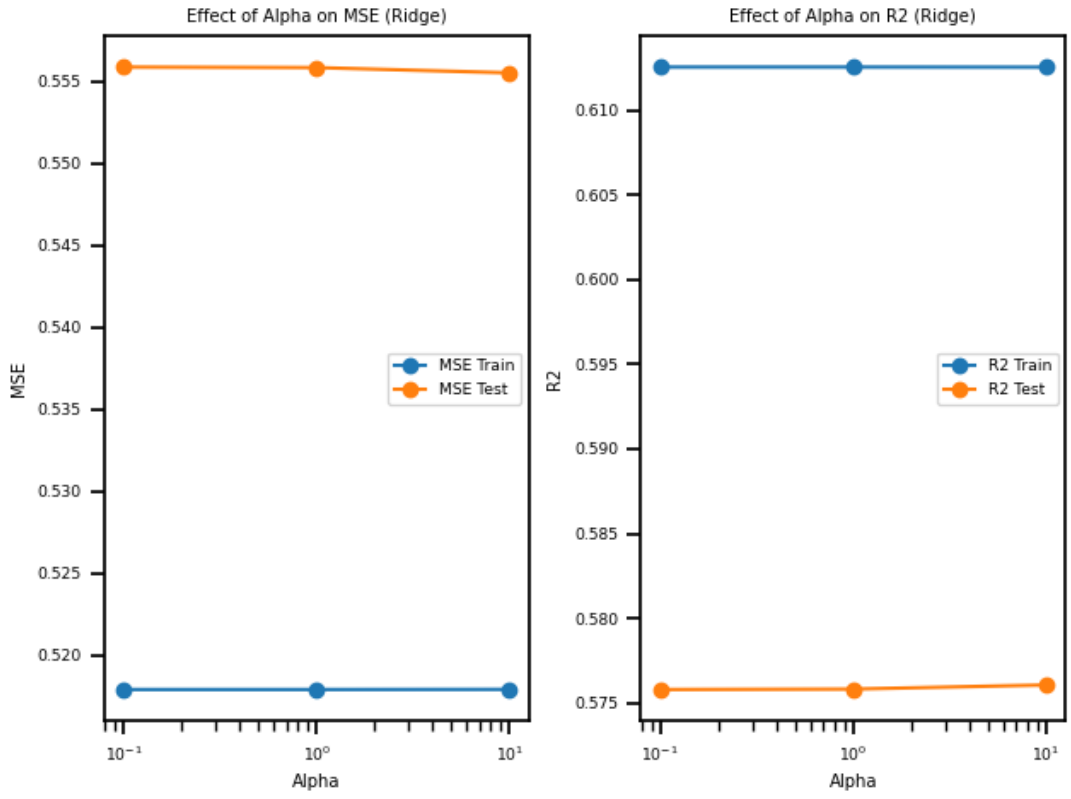
Results and Analysis:

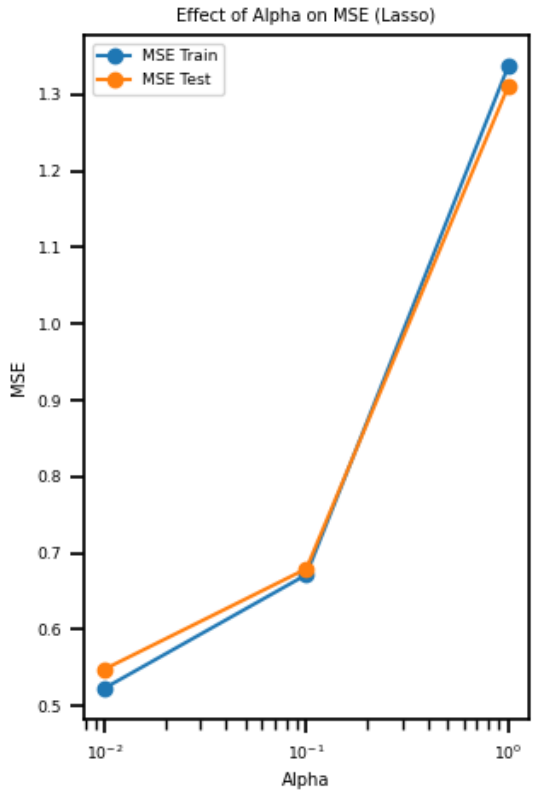
MSE Train	MSE Test	R2 Train	R2 Test	Training Time	Regressor	Alpha	L1 Ratio	e-sti mator	max_ depth
0.51793	0.55589	0.61255	0.57579	0.023213	Linear	NaN	NaN	NaN	NaN
10.64980	0.86251	-6.96677	0.34180	0.143190	RANSAC	NaN	NaN	NaN	NaN
0.51793	0.55589	0.61255	0.57579	0.009424	Ridge	0.1	NaN	NaN	NaN

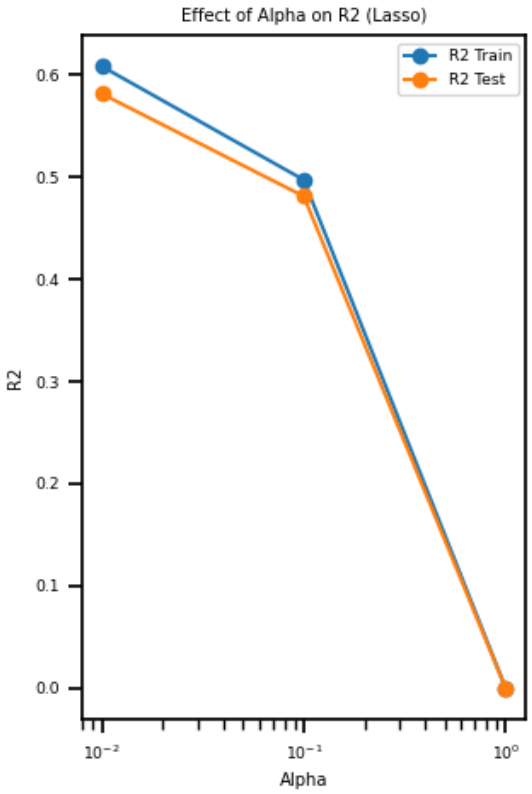
0.51793	0.55586	0.61255	0.57582	0.004347	Ridge	1.0	NaN	NaN	NaN
0.51794	0.55554	0.61254	0.57606	0.002079	Ridge	10.0	NaN	NaN	NaN
0.52338	0.54826	0.60847	0.58162	0.144909	Lasso	0.01	NaN	NaN	NaN
0.67184	0.67963	0.49742	0.48136	0.006814	Lasso	0.1	NaN	NaN	NaN
1.33678	1.31070	0.00000	-0.00022	0.004245	Lasso	1.0	NaN	NaN	NaN
0.52030	0.55138	0.61078	0.57923	0.114131	ElasticNet	0.01	0.2	NaN	NaN
0.521155	0.54995	0.61014	0.58032	0.127043	ElasticNet	0.01	0.5	NaN	NaN
0.522354	0.54881	0.60924	0.58119	0.175815	ElasticNet	0.01	0.8	NaN	NaN
0.591101	0.60128	0.55782	0.54115	0.071676	ElasticNet	0.1	0.2	NaN	NaN
0.627268	0.63586	0.53076	0.51477	0.055846	ElasticNet	0.1	0.5	NaN	NaN
0.663801	0.67190	0.50343	0.48726	0.013563	ElasticNet	0.1	0.8	NaN	NaN
0.916309	0.90884	0.31454	0.30644	0.008647	ElasticNet	1.0	0.2	NaN	NaN
1.058553	1.04423	0.20813	0.20313	0.006860	ElasticNet	1.0	0.5	NaN	NaN
1.33678	1.31070	0.00000	-0.00022	0.005990	ElasticNet	1.0	0.8	NaN	NaN
0.03770	0.25742	0.97180	0.80356	10.783016	Random Forest	NaN	NaN	50.0	NaN
0.17286	0.29905	0.87069	0.77179	6.025193	Random Forest	NaN	NaN	50.0	10.0
0.03944	0.25930	0.97049	0.80112	10.849714	Random Forest	NaN	NaN	50.0	20.0
0.03544	0.25451	0.97349	0.80578	24.480964	Random Forest	NaN	NaN	100.0	NaN
0.17003	0.29417	0.87281	0.77552	10.563100	Random Forest	NaN	NaN	100.0	10.0
0.03724	0.25386	0.97214	0.806276	19.795605	Random Forest	NaN	NaN	100.0	20.0

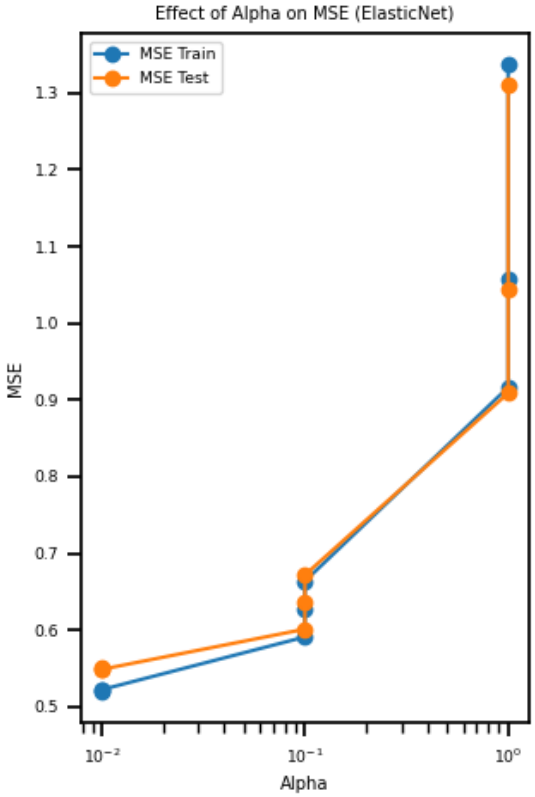
- The pairplot represents the relationships between all feature pairs. It shows a strong positive correlation between median income (MedInc) and median house value (MedHouseVal).
- The quantitative results, in the dataframe, show that the Random Forest Regressor significantly outperformed the linear regressors, it achieved the lowest MSE and highest R2 scores on both the training and test sets. For the hyperparameter tuning, increasing the number of estimators from 50 to 100 improved performance, while setting max_depth to None resulted in the best results. Specifically, with n_estimators set to 100 and max_depth set to None, the model achieved an R2 score of approximately 0.80 on the test set, indicating a strong predictive capability.
- The residual plots show the residuals versus the predicted values for both the training and test sets using Random Forest Regression. The residuals are tightly clustered around zero, suggesting a good fit on the training data. For the test data, the residuals are more spread out compared to the training set, especially at higher predicted values and there is the presence of outliers indicating slightly less precise predictions on unseen data .

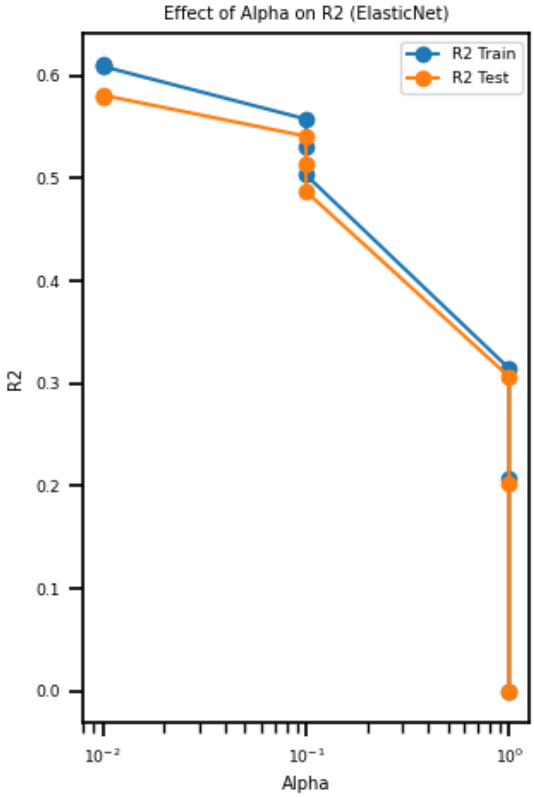
Plots

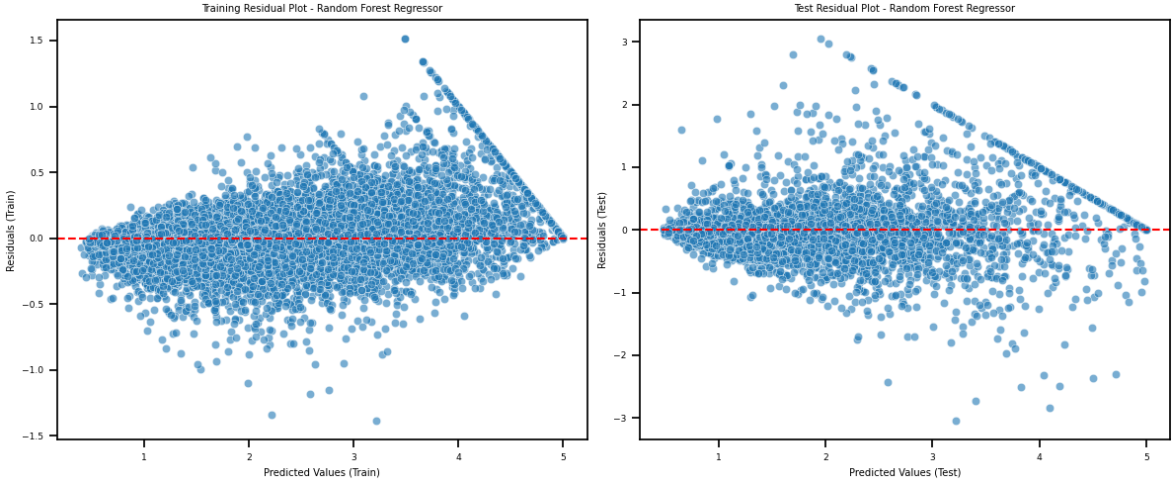


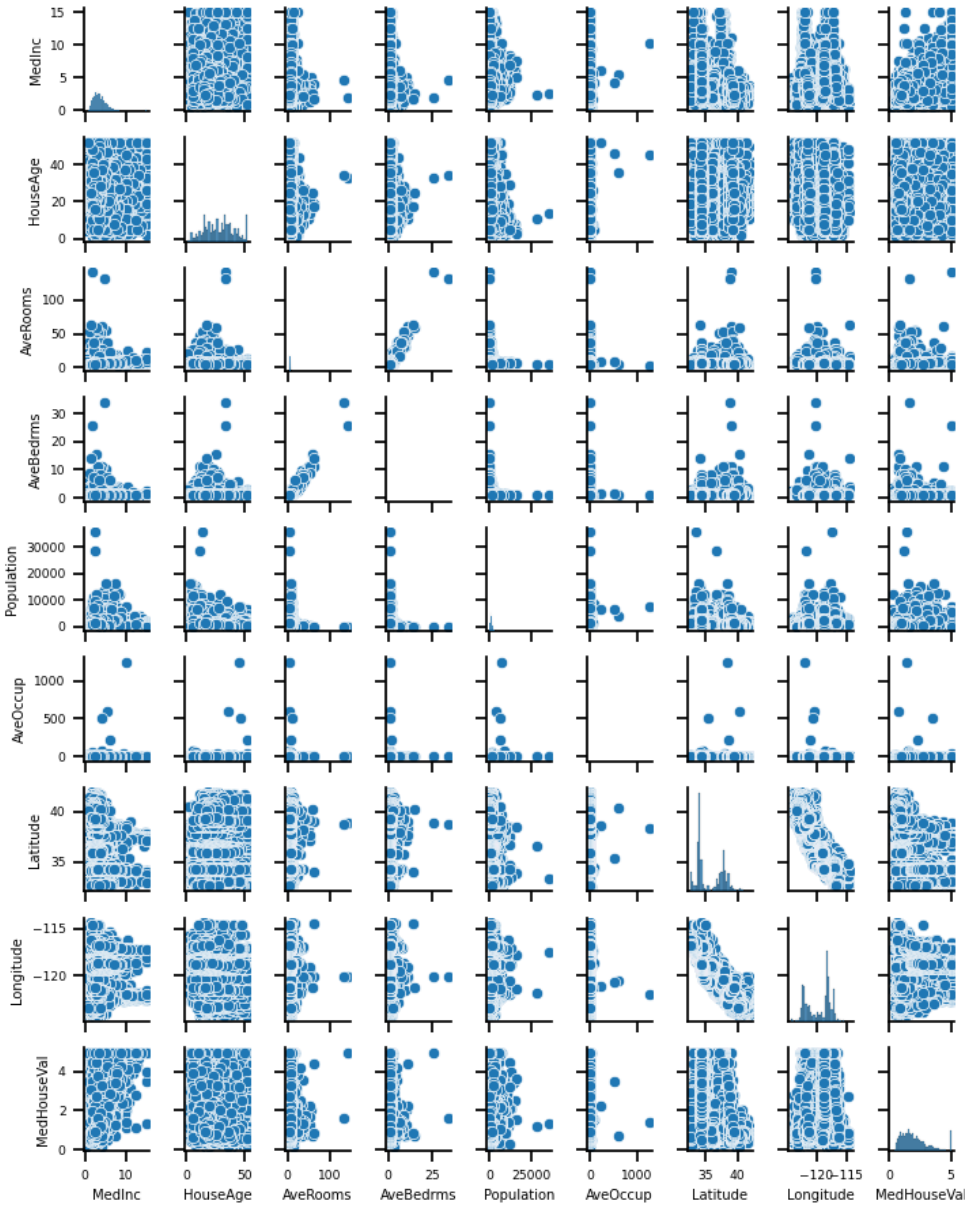

Effect of Alpha on MSE (Lasso)
MSE Train
MSE Test
MSE
Alpha


Effect of Alpha on R2 (Lasso)
R2 Train
R2 Test
R2
Alpha


Effect of Alpha on MSE (ElasticNet)
MSE Train
MSE Test
MSE
Alpha


Effect of Alpha on R2 (ElasticNet)
R2 Train
R2 Test
R2
Alpha


Training Residual Plot - Random Forest Regressor
Residuals (Train)
Predicted Values (Train)
Test Residual Plot - Random Forest Regressor
Residuals (Test)
Predicted Values (Test)


MedInc
HouseAge
AveRooms
AveBedrms
Population
AveOccup
Latitude
Longitude
MedHouseVal