Question 1:

1.Ordinary least squares Linear Regression: fits a linear model with coefficients w = (w1, …, wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

The base linear model is consisted of 3 steps:

A) We create a 10-fold cross-validation scheme

B) Then we specify a range of hyperparameters to tune, in this case we have considered the number of features to select in the model with a range of 1 to 8 features to tune, and find the best hyper parameter.

C) And finally we perform grid search cross validation and pass on the estimator as the linear model

Here is the model's spec:

The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features.

RFE(estimator=LinearRegression(), n_features_to_select=6)

best hyper-parameters: {'n_features_to_select': 6}

2. For the second model we used Lasso which is a Linear Model trained with L1, tuned parameters = [{"alpha": alphas}] is considered, and cross-validation grid search is then performed to find the best parameters. Here are our model's specs:

Lasso(alpha=0.0001, max_iter=10000, random_state=0)

Best hyper-parameters: {'alpha': 0.0001}

3. And as the last and best model, we used a random forest regressor. A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Here are our model's specifications:

best model: RandomForestRegressor(max_depth=9, n_estimators=8)

best hyper-parameters: {'max_depth': 9, 'n_estimators': 8}

Question 2:

Regression Results

Dep. Variable:	Y	R-squared:	0.606
Model:	OLS	Adj. R-squared:	0.603
Method:	Least Squares	F-statistic:	172.4
Date:	Sat, 26 Feb 2022	Prob (F-statistic):	1.73e-175
Time:	14:36:24	Log-Likelihood:	-3387.4
No. Observations:	905	AIC:	6793.
Df Residuals:	896	BIC:	6836.
Df Model:	8		
Covariance Type:	non-robust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-18.6762	28.101	-0.665	0.506	-73.828	36.476
X1	0.1171	0.009	12.986	0.000	0.099	0.135
X2	0.0956	0.011	8.931	0.000	0.075	0.117
X3	0.0812	0.013	6.088	0.000	0.055	0.107
X4	-0.1409	0.043	-3.310	0.001	-0.224	-0.057
X5	0.3633	0.100	3.641	0.000	0.167	0.559
X6	0.0159	0.010	1.613	0.107	-0.003	0.035
X7	0.0167	0.011	1.472	0.141	-0.006	0.039
X8	0.1088	0.006	19.257	0.000	0.098	0.120

Omnibus:	2.765	Durbin-Watson:	1.964
Prob(Omnibus):	0.251	Jarque-Bera (JB):	2.791
Skew:	-0.134	Prob(JB):	0.248
Kurtosis:	2.957	Cond. No.	1.06e+05

As you can see above the coefficient of X4 is negative so if x4 instances increase the output will decrease.

Question 3:

The best hyper parameter for RandomForestRegressor are:

max_depth=9, and n_estimators=8.

The best hyper parameter for Lasso are:

alpha=0.0001, max_iter=10000, and random_state=0.

The best hyper parameter for Linear regression is:

n_features_to_select= 6.

Question 4:

The coefficient of determination (r squared) obtained from Linear Regression (base model) is 0.6049831074389158 while this value for the best model selected is calculated 0.9672032901965096, and we know the closer to 1 this value, the better estimations would be. Also, the final model RSS is calculated as 0 which conform the best model selection.

Question 6:

Program specification:

Python version: 3.10

Running in the command Line of windows 10 assuming in the directory: A2_Reg.py A2_Traindata.tsv A2_Testdata.tsv

Libraries used: NumPy, pandas, matplotlib, scikit-learn, statsmodels, sys, r

Acknowledgement:

https://scikit-learn.org/stable/modules/linear_model.html#polynomial-regression-extending-linear-models-with-basis-functions

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://www.youtube.com/watch?v=0B5eIE_1vpU

https://stackoverflow.com/questions/41524565/attributeerror-gridsearchcv-object-has-no-attribute-cv-results

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

https://www.statology.org/residual-sum-of-squares-in-python/

https://medium.datadriveninvestor.com/k-fold-cross-validation-6b8518070833

https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html