Jared Roderick

In this problem we use the OnlineNews dataset available on Canvas. The dataset is about prediction the popularity of websites from 58 features. Discard the first two variables (url and timedelta) from the data. From the remaining variables use all but the last variable as predictors and the log of the last variable (log(shares)) as the response. Obtain all results using 5-fold-cross-validation, which is computed as follows:

1) Generate a random permutation of the data. Use this random permutation to split the data into 5 disjoint subsets of almost equal size (7928 or 7929 observations each).
2) For each fold i in (1, .., 5), train the model on all data except subset i and test it on subset i, obtaining test error ei.
3) Obtain the final test error as the average of the 5 test errors ei obtained above. Here, the errors ei could be the MSE or the R^2.

Report the results for the following models:

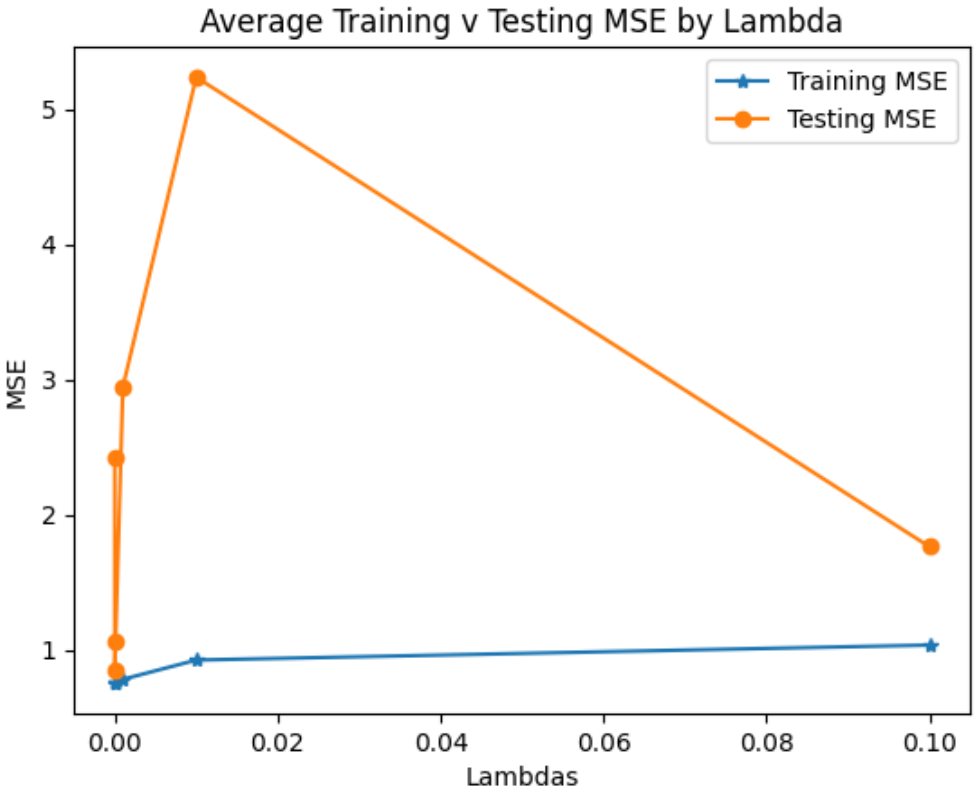
a) Null model. Report the average TRAIN and TEST MSE of the null model that always predicts training y_bar (average training y)

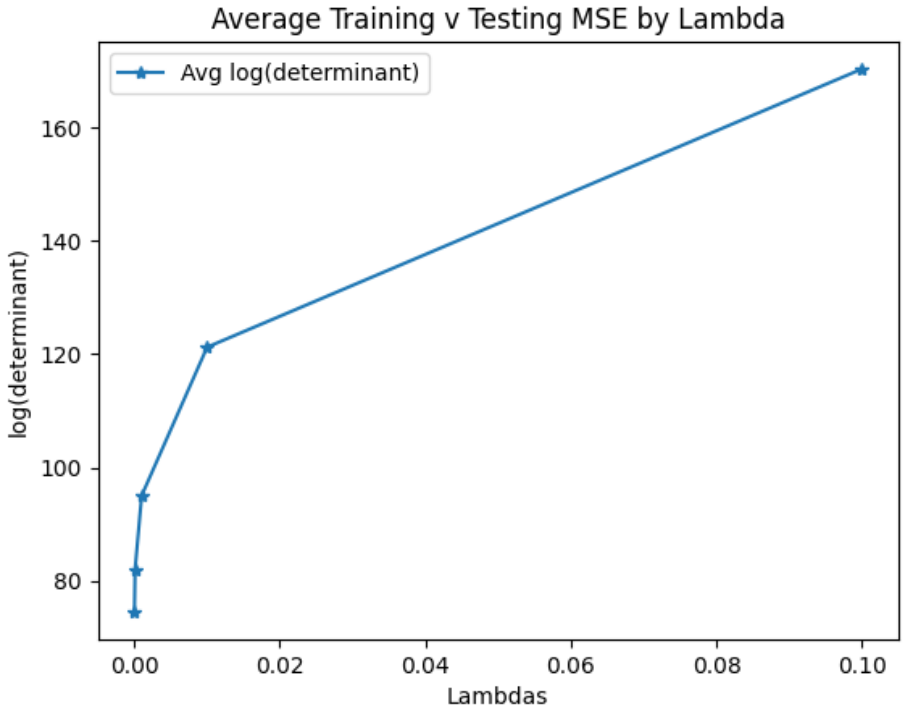
```
Average of the 5 training MSE values: 0.8657802926231689
Average of the 5 testing MSE values: 0.8658012627234397
```

b) OLS regression computed analytically by solving the following normal equations: $[(1/N)\, X^T X + \lambda I_p)]\beta = (1/N)\, X^T Y$ where $\lambda \in \{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and N is the number of observations in X. Report in a table the average TRAINING and TEST R^2 and MSE as well as THEIR STANDARD DEVIATIONS obtained from the 5 folds. On the same graph, plot the average TRAINING and TEST MSE vs λ as two separate curves. Also plot the average value of the logarithm of the determinant of $[(1/N)\, X^T X + \lambda I_p)]$ vs λ.

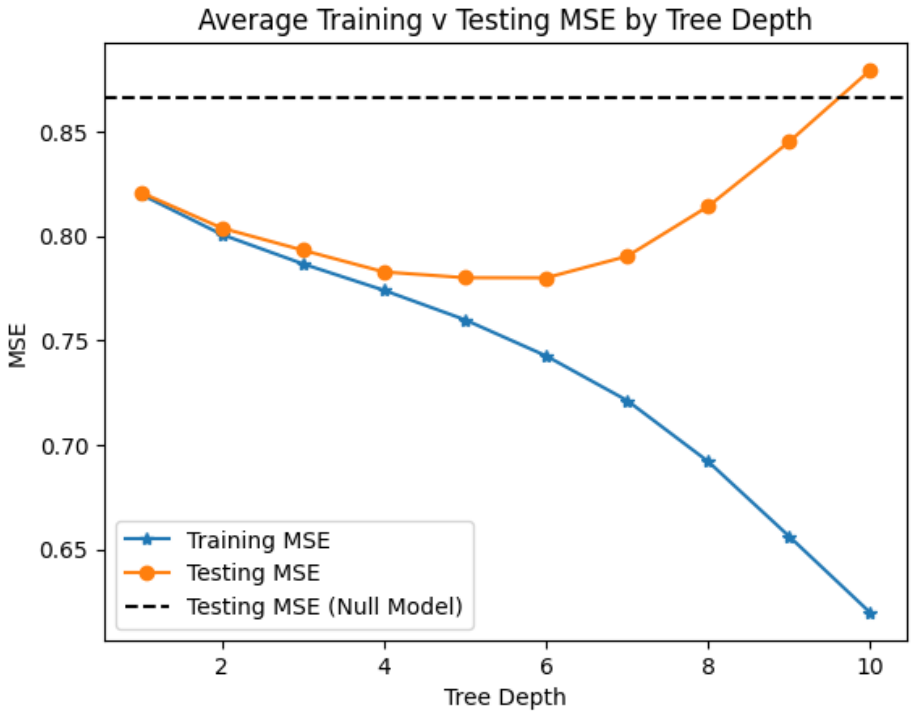
	lambda	Avg Training MSE	Std Dev Training MSE	Avg Testing MSE	Std Dev Testing MSE
0	0.00000	0.755628	0.005801	2.421421	3.330085
1	0.00001	0.755608	0.003438	0.850005	0.186003
2	0.00010	0.756115	0.004440	1.062526	0.608706
3	0.00100	0.780463	0.001491	2.942941	4.314659
4	0.01000	0.923086	0.005984	5.235734	8.609780
5	0.10000	1.033956	0.003530	1.761516	1.421265

Avg Training R-Squared	Std Dev Training R-Squared	Avg Testing R-Squared	Std Dev Testing R-Squared	Avg Log(Determinant)
0.127233	0.001802	-1.774155	3.795762	NaN
0.127256	0.001741	0.019495	0.206611	74.457382
0.126660	0.001635	-0.221987	0.685090	81.995261
0.098539	0.002611	-2.438725	5.069438	94.995345
-0.066244	0.004907	-5.026411	9.897742	121.251906
-0.194275	0.006909	-1.046303	1.671633	170.435783





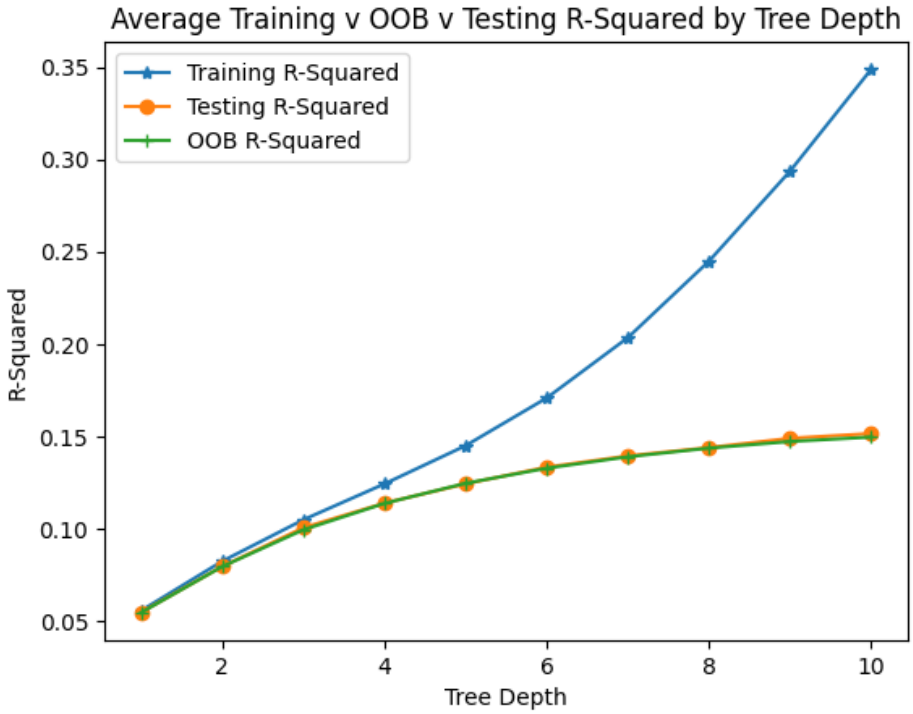
c) Regression tree of maximum depth 1, 2, …, up to 10, for a total of 10 regression trees. On the same plot, plot the average TRAINING and TEST R^2 vs the tree depth as two separate curves. On another plot, plot the average TRAINING and TEST MSE vs the tree depth, and show the null model from a) as a horizontal line.



d) Random forest regression with 300 trees and mazimum depths 1, 2,..., up to 10. Report the average TRAINING, OOB, and TEST R^2 and MSE and THEIR STANDARD DEVIATION in each case. On the same plot, plot the average TRAINING, OOB, and TEST R^2 vs the maximum depth as three separate curves. How does the average OOB R^2 compare to the test R^2.

	Avg Training MSE	Avg Training MSE St Dev	Avg OOB MSE	Avg OOB MSE St Dev	Avg Testing MSE	Avg Testing MSE St Dev
0	0.817822	0.002823	0.818708	0.002779	0.818530	0.009873
1	0.794612	0.003089	0.797126	0.003093	0.797001	0.012575
2	0.775141	0.001819	0.779849	0.001804	0.778813	0.007097
3	0.758295	0.002457	0.767457	0.002541	0.767276	0.009227
4	0.740368	0.005235	0.758028	0.005550	0.758266	0.022359
5	0.717974	0.005086	0.750915	0.005265	0.750208	0.019570
6	0.689676	0.000832	0.745579	0.001136	0.745099	0.005890
7	0.653766	0.002042	0.741500	0.002452	0.741120	0.007536
8	0.611475	0.001521	0.738403	0.000657	0.736782	0.003985
9	0.563697	0.003790	0.736336	0.003076	0.734548	0.011006

Avg Training R-Squared	Avg Training R-Squared St Dev	Avg OOB R-Squared	Avg OOB R-Squared St Dev	Avg Testing R-Squared	Avg Testing R-Squared St Dev
0.055393	0.000994	0.054369	0.000938	0.054516	0.003089
0.082198	0.000680	0.079295	0.000595	0.079359	0.003034
0.104688	0.001770	0.099250	0.001696	0.100373	0.005762
0.124147	0.002014	0.113565	0.002257	0.113736	0.005082
0.144854	0.000920	0.124459	0.000935	0.124167	0.004625
0.170709	0.000956	0.132661	0.000911	0.133219	0.004124
0.203401	0.001740	0.138832	0.001563	0.139315	0.005361
0.244876	0.001250	0.143542	0.001050	0.143825	0.004813
0.293724	0.001178	0.147117	0.000678	0.148848	0.002651
0.348914	0.002759	0.149507	0.000602	0.151423	0.004430



The average OOB and test R^2 are nearly identical at each depth, as seen by the orange and green lines on the plot above.