UNIVERSITEIT VAN TILBURG

Introduction to Data Science: Group Assignment

Group number: 10
Ernani Hazbolatow 2023708/304318
Julia van Bon 2014511/872721
Andrey Peshev 2023683/138297
Sarah Via 2035640/135598

April 2020

Answers Short Questions

What were the cross-validation errors (CVEs) for the optimal ridge-penalized and LASSO- penalized regression models that you selected in Section 2.1?

The optimal ridge-penalized yields to a cross-validation error of 1.143765 and the cross-validation error for the Lasso-penalized regression model is 1.139137.

What were the test-set mean squared errors (MSEs) for the unpenalized, ridge-penalized, and LASSO-penalized regression models that you selected in Section 2.1?

The mean square error (MSE) for the unpenalized regression model is 1.570859. The ridge-penalized regression model yielded to an MSE of 1.321605. The MSE for the Lasso-penalized regression model is 1.301016.

Which modeling approach performed best in the regression task? Provide a statistical justification for your answer.

The Lasso-penalized regression model is the best fit model to predict "Decision making" of young people. The model gives out the lowest MSE (1.301016) which means that the predicted values of the dependent variable have a low average variance to the true observations than with the unpenalized and ridge-penalized model.

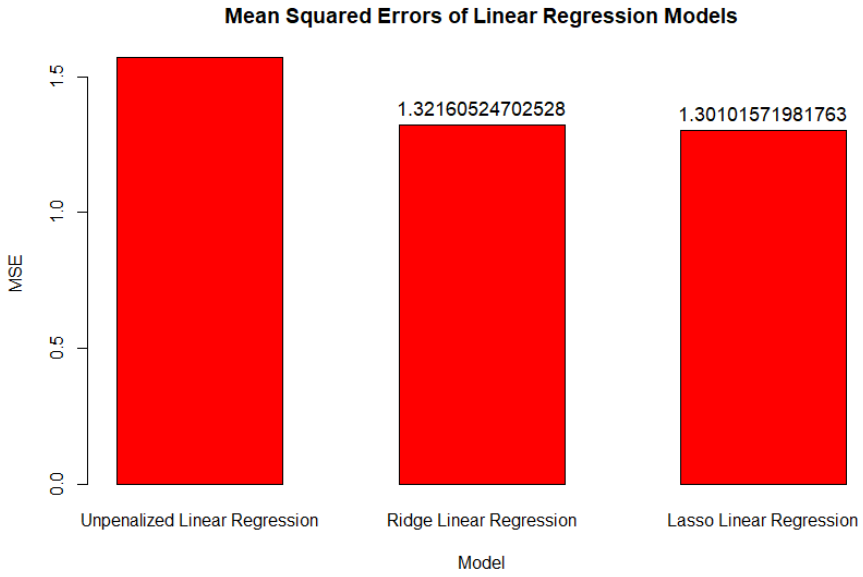

Figure 1: MSE per model

What were the cross-validation errors (CVEs) for the optimal ridge-penalized and LASSO- penalized multinomial logistic regression models that you selected in Section 2.2?

The optimal CVE for the ridge model is 2.091565 and for the LASSO-penalized is 2.085149.

What were the test-set cross-entropy errors (CEEs) for the unpenalized, ridge-penalized, and LASSO-penalized multinomial logistic regression models that you selected in Section 2.2?

The CEE for the unpenalized multinomial logistic regression is 1.340459, 1.008105 for the ridge-penalized, and 1.013170 for the LASSO-penalized model.

Which modeling approach performed best in the classification task? Provide a statistical justification for your answer.

The ridge model had the lowest CEE because it made predictions with higher confidence compared to the other two. This means that it had a higher probability of putting a person into a group (for example "i am often running late") if he really belongs to that group. The CEE takes into consideration the accuracy of the predictions but also with how much confidence the model puts people into groups.

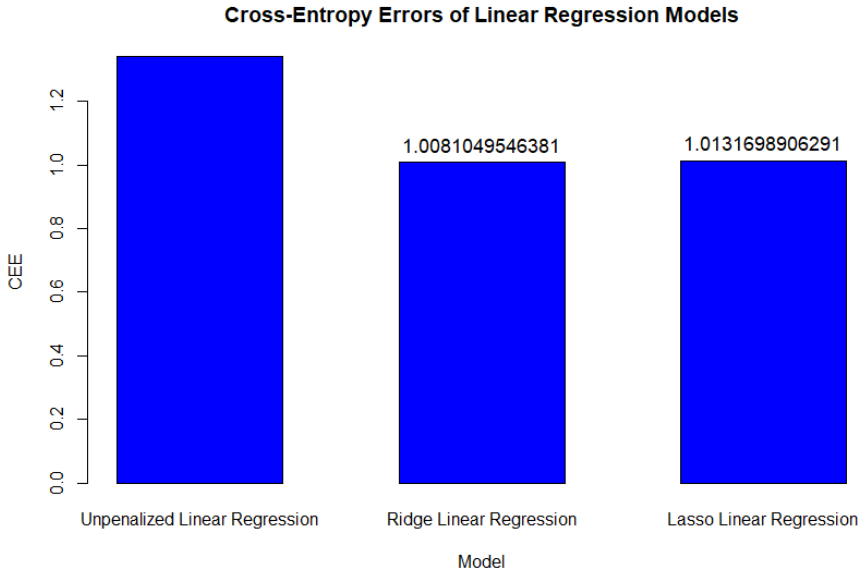

Figure 2: CEE per model

What were the predicted class memberships generated by your chosen model in Section 2.3?

The predicted class membership for N=10 on punctuality generated by the ridge multinomial logistic regression model model were:

```
      1
[1,] "i am often running late"
[2,] "i am often early"
[3,] "i am often early"
[4,] "i am always on time"
[5,] "i am often running late"
[6,] "i am always on time"
[7,] "i am always on time"
[8,] "i am always on time"
[9,] "i am often early"
[10,] "i am always on time"
```

Figure 3: Classification of N=10

What were the predicted probabilities of class membership generated by your chosen model in Section 2.3?

The predicted probabilities of class membership for N=10 on punctuality generated by the ridge multinomial logistic regression model were:

	i am always on time	i am often early	i am often running late
Steve	0.3154235	0.2916942	0.3928823
Suzy	0.3325202	0.4027942	0.2646856
Sara	0.3706833	0.3853839	0.2439327
Bob	0.3773891	0.3056449	0.3169660
Billy	0.2999377	0.3576139	0.3424484
Brenda	0.4023525	0.3221203	0.2755272
Adam	0.4211265	0.3217016	0.2571719
Alice	0.3772171	0.3345823	0.2882006
Anne	0.2856418	0.4133692	0.3009890
Dave	0.3887009	0.2913438	0.3199553

Figure 4: Class probabilities for N=10

Compare the predicted class assignments you computed in Section 2.2 to the true class memberships represented in the file: "candidates.rds". Would you say that your classifier did a "good" job of grouping these 10 young people? Explain why or why not?

The classifier managed only to predict 5 out of 10 participants in their true class. Whether the classifier did a good or bad job depends on what other methods of classification exist and how they perform. If the other methods perform worse, then this classifier has obviously done a good job. If the other methods perform better, then this classifier has done a poor job. From a selection standpoint, the classifier did a poor job with 5 correct predictions.

	pred	true	Freq
1	i am always on time	i am always on time	3
2	i am often early	i am always on time	3
3	i am often running late	i am always on time	0
4	i am always on time	i am often early	0
5	i am often early	i am often early	0
6	i am often running late	i am often early	0
7	i am always on time	i am often running late	2
8	i am often early	i am often running late	0
9	i am often running late	i am often running late	2

Figure 5: Table contrasting predicted vs. tre for N=10

Examine the predicted probabilities of class membership that you computed in Section 2.2. Do these predicted probabilities shed any light on the good/poor performance of your classifier? Why do you think so?

From the predicted probabilities, we can see that the classifier has trouble classifying a person into one class or the other with confidence. Sometimes, this

difference is minuscule. Take index 583 for example, the probabilities for class "I am always on time" and "I am often running late" is roughly $P = 0.015$. Other times, this is quite big. However, the classifier almost never manages to reach high skewed probabilities for one class over the other. This might explain the bad performance.

	i am always on time	i am often early	i am often running late
726	0.2983616	0.2922942	0.40934420
497	0.3550366	0.4249436	0.22001986
417	0.3954812	0.2457263	0.35879251
807	0.4657337	0.3865367	0.14772957
172	0.2594226	0.4588242	0.28175321
693	0.3342853	0.4173646	0.24835016
55	0.4051453	0.2643537	0.33050096
118	0.2182064	0.6388635	0.14293010
99	0.4600577	0.2480774	0.29186489
634	0.4783825	0.2606518	0.26096569
60	0.4317150	0.3342057	0.23407928
583	0.3680957	0.2785044	0.35339992
808	0.2865716	0.5097890	0.20363947
797	0.3911199	0.4596351	0.14924498
776	0.3232628	0.3317651	0.34497208
589	0.4056564	0.3429843	0.25135932
784	0.4268304	0.3346779	0.23849173
101	0.3435345	0.4209588	0.23550667
346	0.3822579	0.4106039	0.20713814
618	0.4237818	0.3082466	0.26797159
129	0.3683070	0.3145431	0.31714995
642	0.3055614	0.3124718	0.38196681
691	0.4064689	0.3345636	0.25896750
231	0.4047984	0.2330164	0.36218521
574	0.4150679	0.1298078	0.45512425
680	0.3771213	0.2353980	0.38748068
359	0.3329830	0.4604250	0.20659208

Figure 6: Inspection of Classifier