Mushroom Toxicity Analysis

Written by: Ted Haley

January, 2018

Introduction

Mushrooms are a delicious edible fungus that are used in a variety of foods, however some varieties of fungus can be toxic to humans. The mushroom dataset that I will analyze in this report includes various features of toxic and non-toxic mushrooms. The goal of this project is to determine which features are relevant or irrelevant when determining the toxicity of a mushroom with a particular set of features.

Analysis

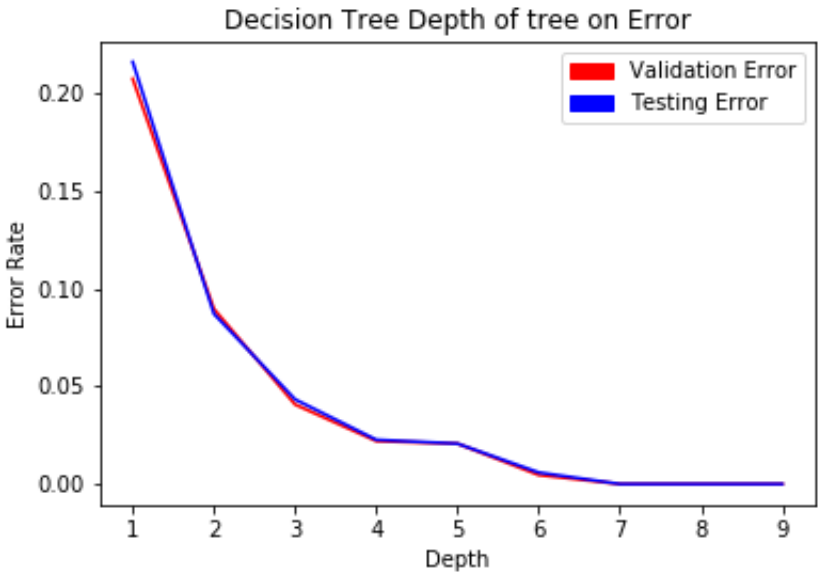
The mushroom dataset includes 22 features and a single binary classifier (toxic or non-toxic). Predictors include cap shape, cap surface, odor, stalk shape, colour, etc. In this analysis, I will try to determine which features are irrelevant in classifying the toxicity of the mushroom.

The data that I imported consists of categorical labels. I encoded these labels onto a numeric scale so I can interpret them analytically using various feature and model selection tools.

Before I initiate a feature selection, I will first look at the existing parameters to see how they perform when classifying the toxic mushrooms. Below is a decision tree that is fit over various depths.

I decided to use a decision tree classifier to perform my initial exploratory data analysis because decision trees tend to overfit on data with a large number of features. That being said, the decision tree should see the largest improvement with a reduction in the number of features.

Figure 1: Decision Tree:



	depth	train error	validation error
0	1.0	0.207485226527	0.216149679961
1	2.0	0.0896257386737	0.0871491875923
2	3.0	0.0407091267236	0.0433284096504
3	4.0	0.0219960604071	0.0226489414082
4	5.0	0.0206828627708	0.0206794682422
5	6.0	0.00459619172685	0.00590841949778
6	7.0	0.0	0.0
7	8.0	0.0	0.0
8	9.0	0.0	0.0

From the results above, we can see that both the training error and validation error are 0 with a depth of 7. This is useful to know, as a combination of 7 individual features resulted in an error of 0. We could expect the number of useful features to be in around 7.

The following analysis are the results of a Recursive Feature Elimination (RFE). The RFE is using a ridge estimator that performs L2 regularization on the training data to determine the optimal number of features to use. The plot below shows the relationship between the number of features and the training and validation error of the model.

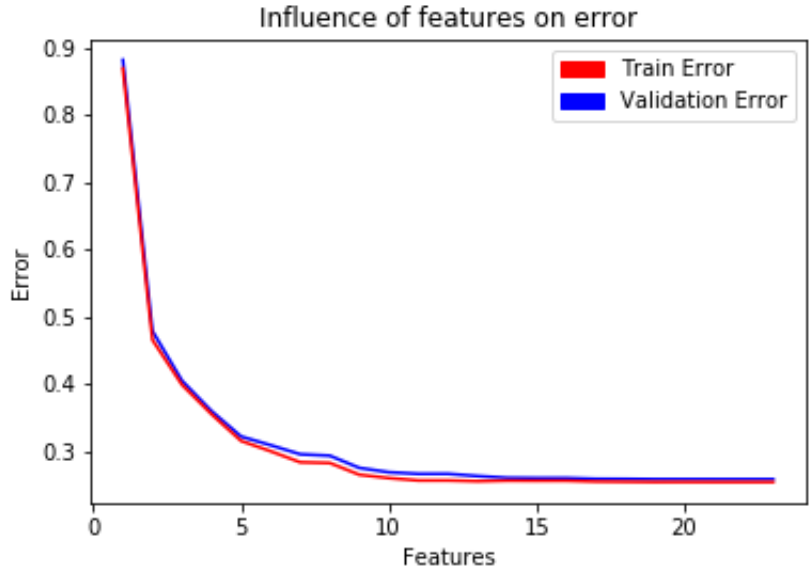


Figure 2: RFE:

	features	train error	validation error
0	1.0	0.881805642249	0.869236811191
1	2.0	0.478494561118	0.465251723791
2	3.0	0.404468082084	0.39845283888
3	4.0	0.359335835286	0.354663468112
4	5.0	0.321498379846	0.314970750738
5	6.0	0.308644758658	0.299618420856
6	7.0	0.295133029948	0.283170441735
7	8.0	0.293006196929	0.282147437654
8	9.0	0.274829474256	0.264653939612
9	10.0	0.268333762844	0.259623693334
10	11.0	0.266025130889	0.256114550796
11	12.0	0.266023796027	0.256099830682
12	13.0	0.262696630603	0.255003419829
13	14.0	0.26005011983	0.255887297916
14	15.0	0.259706851221	0.25570928417
15	16.0	0.25970814586	0.255733978807
16	17.0	0.258482606441	0.254527971506
17	18.0	0.258367443286	0.254177909078
18	19.0	0.257996804855	0.25386891806
19	20.0	0.257953636853	0.253901244393
20	21.0	0.257943770665	0.253829518712
21	22.0	0.257943770665	0.253829518712
22	23.0	0.257943770665	0.253829518712

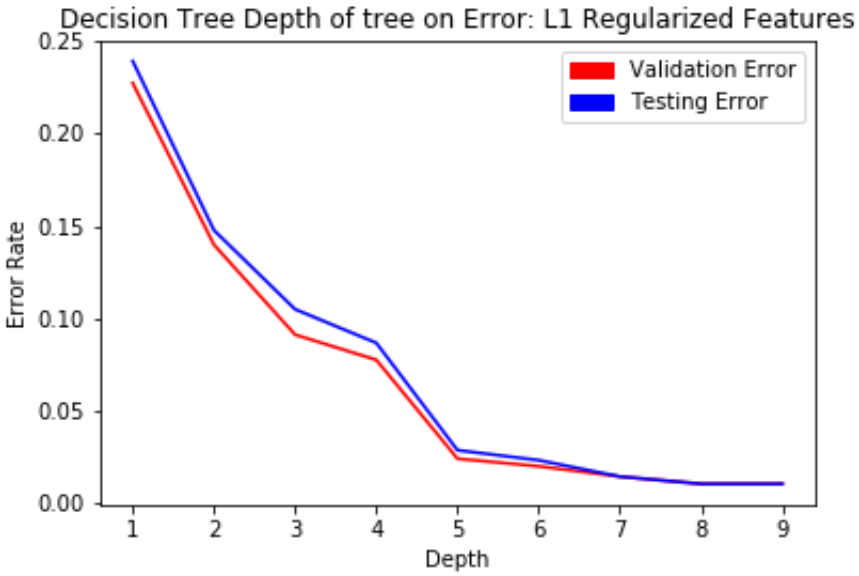
We can see from the results above that there is no significant decrease in validation or training error when we include more than approximately 9 or 10 features. This seems reasonable as the max depth for the decision tree was 7. To confirm our suspisions, we will perform a regularized logistic regression for both the L1 regualarization and L2 regularization below.

```
L1 Features Selected: 8
L1 Score: 0.954702117184
L1 Test Error Rate: 0.0452978828163

L2 Features Selected: 7
L2 Score: 0.944854751354
L2 Test Error Rate: 0.055145248646
```

As I had suspected, the L1 and L2 logistic regression resulted in 8 and 7 features selected, respectively. For the L1, L2, and RFE reductions I have found above, I will now apply the selected features to the training data and refit the decision tree classifier from figure 1.

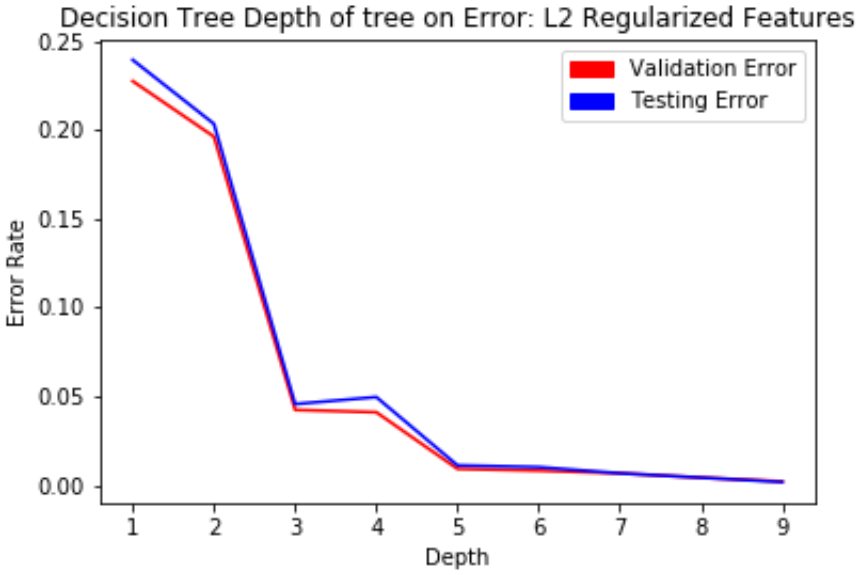
Figure 3: Decision Tree with L1 Features:



	depth	train error	validation error
0	1.0	0.227347340775	0.23929098966
1	2.0	0.13985554826	0.147710487445
2	3.0	0.0911030860144	0.104874446086
3	4.0	0.0774786605384	0.0866568193008
4	5.0	0.0239658568615	0.028557360906
5	6.0	0.0198621142482	0.0231413096997
6	7.0	0.0142810242942	0.014278680453
7	8.0	0.0103414313854	0.0103397341211
8	9.0	0.0103414313854	0.0103397341211

We can see above that the L1 Regularized features actually led to an increase in the error rate with respect to the unregularized dataset (marginal increase). This is an interesting result as I would have expected the validation error to be lower.

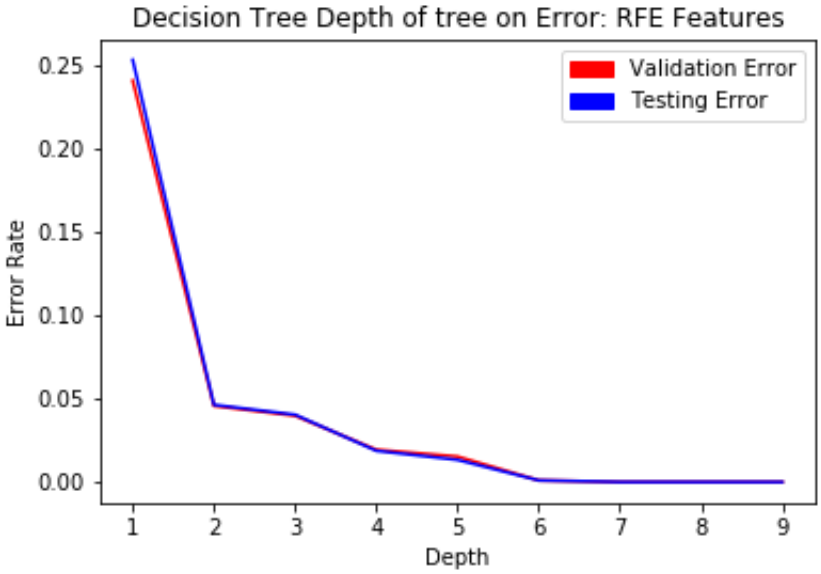
Figure 4: Decision Tree with L2 Features:



	depth	train error	validation error
0	1.0	0.227347340775	0.23929098966
1	2.0	0.196158896914	0.203348104382
2	3.0	0.0425147734734	0.0457902511078
3	4.0	0.0412015758372	0.0497291974397
4	5.0	0.00935653315824	0.0113244707041
5	6.0	0.00837163493106	0.0103397341211
6	7.0	0.00689428759028	0.00689315608075
7	8.0	0.00443204202232	0.00443131462334
8	9.0	0.00229809586343	0.00196947316593

We can see above that the L2 reduced features performed well with the decision tree of depth 3. This is an improvement over the decision tree using all 23 features shown in figure 1. This reduced model shows that the 7 features selected in conjunction with a decision tree of depth 3 has a 95% accuracy for both the training and validation data.

Figure 5: Decision Tree with RFE features:



	depth	train error	validation error
0	1.0	0.240643466842	0.253077301822
1	2.0	0.0456336178595	0.0462826193993
2	3.0	0.0397242284964	0.0403741999015
3	4.0	0.0193696651346	0.0187099950763
4	5.0	0.0152659225213	0.01329394387
5	6.0	0.000984898227183	0.000984736582964
6	7.0	0.0	0.0
7	8.0	0.0	0.0
8	9.0	0.0	0.0

The decision tree using the RFE features shows the most improvement over the full feature set, achieving a 100% accuracy score for both the training and validation data with a tree of depth 7. As previously mentioned, the optimal tree depth using all features was 7, so this proves that the combination of these 10 features selected from the RFE feature selection in conjunction with a depth of 7, results in the best model and feature combination.

Conclusion

As the analysis above has shown, there are a variety of features from the mushroom dataset that are irrelvant when determining if a mushroom is toxic. The model and features that result in the best overall combination are a decision tree of depth 7, with 10 features selected from RFE using the ridge estimator, resulting in a classification accuracy fo 100%.