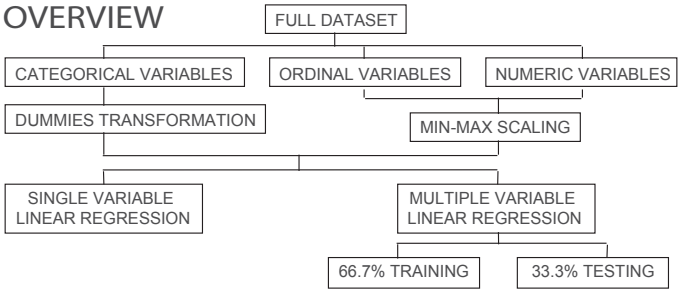
ASSIGNMENT 1

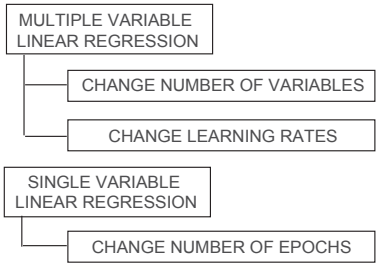
BUAN 6341.501 - Applied Machine Learning

Weiyang Sun

FULL DATASET LINEAR REGRESSION

OVERVIEW





Using Scikit-Learn and StatsModel linear regression packages, a baseline model consisting of 39 variables, with an adjusted R-square of 15.1% was established. This represents the proportion of variance in G3 explained by the 39 different variables.

MULTIPLE VARIABLE LINEAR REGRESSION

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.277
Model:	OLS	Adj. R-squared:	0.151
Method:	Least Squares	F-statistic:	2.201
Date:	Sun, 16 Sep 2018	Prob (F-statistic):	0.000185
Time:	13:15:33	Log-Likelihood:	-733.08
No. Observations:	264	AIC:	1546.
Df Residuals:	224	BIC:	1689.
Df Model:	39		
Covariance Type:	nonrobust		

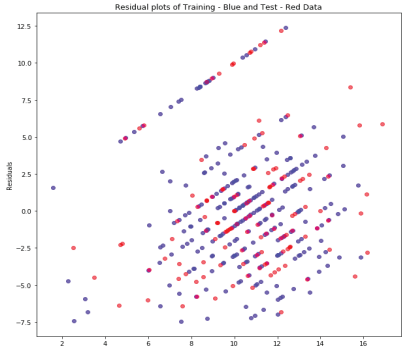
	coef	std err	t	P>|t|	[0.025	0.975]
const	14.8296	3.685	4.025	0.000	7.569	22.091
age	-3.4127	2.157	-1.582	0.115	-7.663	0.837
Medu	1.8506	1.630	1.136	0.257	-1.361	5.062
Fedu	0.3181	1.355	0.235	0.815	-2.352	2.988
traveltime	-0.3530	1.267	-0.279	0.781	-2.850	2.144
studytime	0.9694	1.041	0.931	0.353	-1.082	3.020
failures	-4.5599	1.294	-3.523	0.001	-7.111	-2.009
famrel	0.9902	1.240	0.799	0.425	-1.453	3.433
freetime	0.2365	1.221	0.194	0.847	-2.170	2.643
goout	-1.1693	1.190	-0.982	0.327	-3.515	1.176
Dalc	-1.7381	1.952	-0.891	0.374	-5.584	2.108
Walc	0.4853	1.322	0.367	0.714	-2.119	3.090
health	-0.6802	0.794	-0.857	0.392	-2.245	0.884
absences	3.2892	2.591	1.269	0.206	-1.817	8.395
school_GP	-0.1413	1.039	-0.136	0.892	-2.189	1.906
sex_F	-1.4196	0.625	-2.273	0.024	-2.651	-0.189
address_U	0.3341	0.705	0.474	0.636	-1.054	1.723
famsize_GT3	-0.9378	0.625	-1.501	0.135	-2.169	0.294
Pstatus_T	-1.0471	0.921	-1.137	0.257	-2.863	0.768
Mjob_at_home	1.4225	1.383	1.029	0.305	-1.303	4.148
Mjob_health	1.7344	1.278	1.357	0.176	-0.784	4.253
Mjob_other	0.4478	1.099	0.408	0.684	-1.717	2.613
Mjob_services	1.4933	1.007	1.484	0.139	-0.490	3.477
Fjob_at_home	-0.6236	1.693	-0.368	0.713	-3.960	2.713
Fjob_health	-0.1992	1.594	-0.125	0.901	-3.340	2.942
Fjob_other	-2.0558	1.167	-1.762	0.079	-4.355	0.244
Fjob_services	-1.5563	1.200	-1.297	0.196	-3.922	0.809
reason_course	-1.0145	1.115	-0.910	0.364	-3.212	1.183
reason_home	-0.0012	1.173	-0.001	0.999	-2.313	2.311
reason_reputation	-0.3307	1.146	-0.289	0.773	-2.590	1.928
guardian_father	-1.2776	1.339	-0.954	0.341	-3.917	1.362
guardian_mother	-1.4439	1.249	-1.156	0.249	-3.905	1.017
schoolsup_yes	-2.3361	0.805	-2.902	0.004	-3.922	-0.750
famsup_yes	-0.5902	0.608	-0.970	0.333	-1.789	0.608
paid_yes	-0.4284	0.619	-0.692	0.490	-1.648	0.791
activities_yes	-0.4495	0.561	-0.802	0.424	-1.554	0.655
nursery_yes	-0.6070	0.710	-0.855	0.393	-2.006	0.792
higher_yes	1.6274	1.443	1.128	0.261	-1.217	4.472
internet_yes	0.4813	0.797	0.604	0.547	-1.089	2.052
romantic_yes	-1.4539	0.629	-2.313	0.022	-2.692	-0.215

P-values are derived from hypotheses testing to determine if there is no relationship between X and Y (null) or if there is (alternate). If we reject the null, there is a 95% confidence interval that the coefficient does not include zero, therefore there is a relationship.

TEST RESULTS

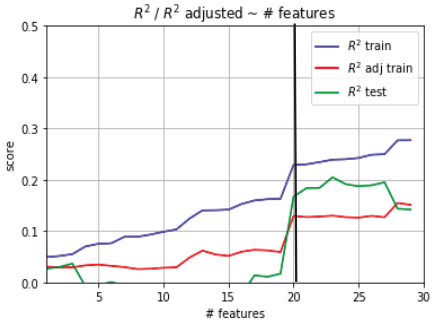
Mean Squared Error: 18.01
Mean Absolute Error: 3.24
Root Mean Squared Error: 4.24
R2 Statistics: 14.2%

RESIDUAL PLOTS



There are no distinct patterns displayed in the residual plots. This is indicative of a good model fit.

CHANGING # OF VARIABLES

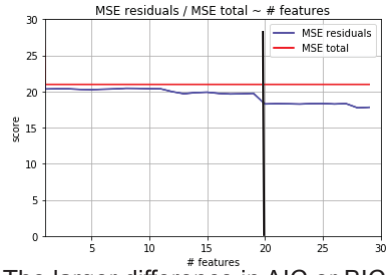


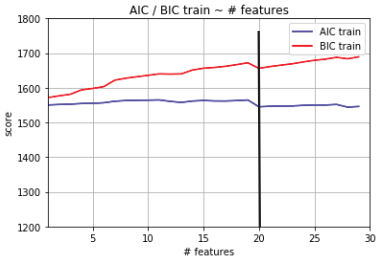
COEFFICIENTS INTERPRETATION

1) For every additional failure, there is an associated decrease in G3 by 4.56, controlling for other variables.
2) Female students as compared to Male, are associated with -1.42 difference in G3, controlling for other variables.
3) Students with school support as compared to those without, are associated with -2.34 difference in G3, controlling for other variables.
4) Students with romantic relationship as compared to those without, are associated with -1.45 difference in G3, controlling for other variables.

FULL DATASET LINEAR REGRESSION

In the previous page, as we changed the number of variables, we discovered that 20 variables would provide us with the highest R-squared for both test and train, while balancing MSE, AIC and BIC.





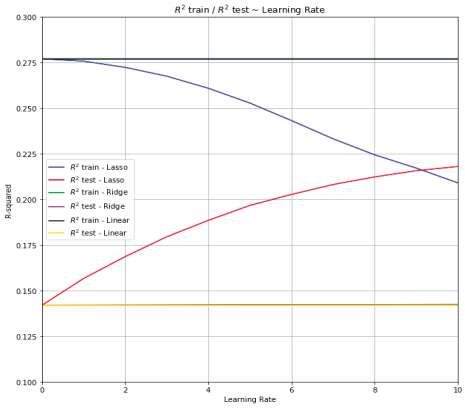
The AIC and BIC are two methods of assessing model fit penalized for the number of variables.

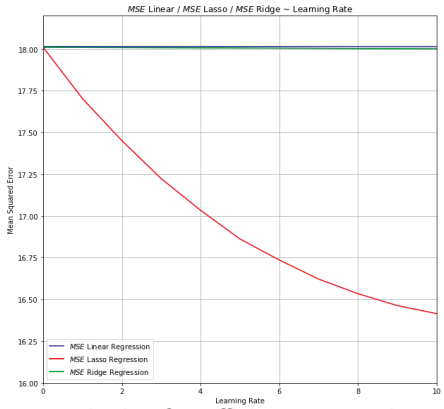
AIC tries to select the model that most adequately describes an unknown, high dimensional reality. BIC tries to find the true model among the set of candidates.

The larger difference in AIC or BIC indicates stronger evidence for one model over the other, therefore, the lower the score, the better the model is.

CHANGING LEARNING RATES

Having numerous variables creates a complicated model. This may result in increased variance and decreased bias (over-fitting). To overcome this, a simpler model has to be used or regularization is performed. In regularization, the same number of variables is kept, but the magnitude of the coefficients is reduced. Therefore, 2 new regression methods will be used to perform regularization: 1) Lasso regression and 2) Ridge regression.



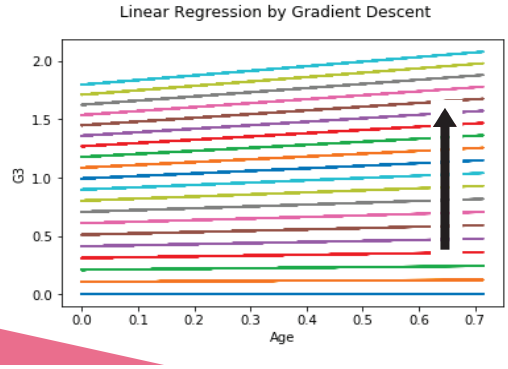


Note: Linear and ridge regression provides similar results. This is because Lasso which uses L1 regularization which reduces some coefficients to absolute zero - performing inherent feature selection. Whereas, Ridge uses L2 regularization which reduces the magnitude of the coefficients proportionately similar to Linear.

Since L2 regularization (Ridge) reduces the magnitude of coefficients proportionately, the weights of coefficients while being different from linear regression, maintains the same proportionality. This gives it the same effect as linear regression and retains the model complexity.

SINGLE VARIABLE LINEAR REGRESSION

In order to see the effects of gradient descent on linear regression rather than OLS, a function was written to replicate the mathematics taught in the lecture notes. This allowed the observation of gradient descent on linear regression over 1000 iterations.



10 RANDOM FEATURES	10 CHOSEN FEATURES
Health	Travel Time
Mother's Job	StudyTime
Family Support	Failures
Reason	Health
Travel Time	Absences
Address	School Support
Father's Job	Family Support
Freetime	Paid
Nursery	Higher
Daily Alcoholic Consumption	School

PARTIAL DATASET LINEAR REGRESSION

10 RANDOM FEATURES

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.093
Model:	OLS	Adj. R-squared:	0.027
Method:	Least Squares	F-statistic:	1.399
Date:	Sun, 16 Sep 2018	Prob (F-statistic):	0.132
Time:	23:03:44	Log-Likelihood:	-762.99
No. Observations:	264	AIC:	1564.
Df Residuals:	245	BIC:	1632.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	15.3602	1.876	8.186	0.000	11.664	19.056
health	-0.9621	0.824	-1.167	0.244	-2.586	0.662
Dalc	-0.8113	1.449	-0.560	0.576	-3.666	2.043
traveltime	-1.0910	1.267	-0.861	0.390	-3.587	1.405
freetime	0.0547	1.180	0.046	0.963	-2.270	2.379
Mjob_at_home	-1.9164	1.106	-1.733	0.084	-4.094	0.261
Mjob_health	0.3562	1.317	0.271	0.787	-2.237	2.950
Mjob_other	-1.5497	0.935	-1.657	0.099	-3.392	0.293
Mjob_services	-0.0565	0.941	-0.060	0.952	-1.911	1.798
famsup_yes	-0.7444	0.590	-1.261	0.209	-1.907	0.418
reason_course	-1.4008	0.742	-1.888	0.060	-2.862	0.061
reason_other	-1.0132	1.192	-0.850	0.396	-3.362	1.335
reason_reputation	-0.4098	0.782	-0.524	0.601	-1.950	1.130
address_U	0.3976	0.698	0.569	0.570	-0.978	1.773
Fjob_at_home	-1.3049	1.716	-0.761	0.448	-4.684	2.074
Fjob_health	-1.1143	1.641	-0.679	0.498	-4.346	2.117
Fjob_other	-2.0116	1.155	-1.742	0.083	-4.287	0.263
Fjob_services	-1.8712	1.199	-1.561	0.120	-4.232	0.490
nursery_yes	-0.7765	0.717	-1.083	0.280	-2.189	0.636

10 random variables were selected to create a linear regression model, an adjusted R-squared of 2.7% was obtained as compared to the baseline of 15.1%. This decrease in performance was because the independent variables selected have little ability to explain the variance of G3.

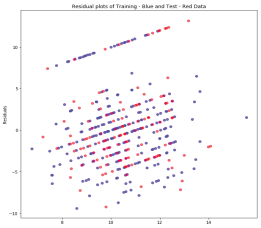
TEST RESULTS

Mean Squared Error: 21.74 ▼
Mean Absolute Error: 3.54 ▼
Root Mean Squared Error: 4.66 ▼
R2 Statistics: -3.6% ▼

Note: It is possible to obtain a negative R-square if the fit is worse than fitting a horizontal line.

RESIDUAL PLOTS

There are no distinct patterns displayed in both the test and training residual plots.

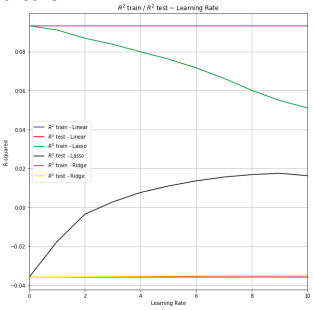


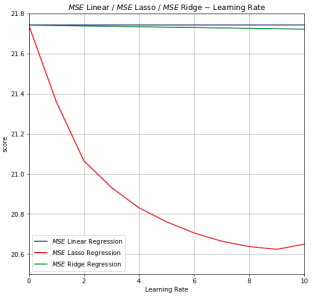
COEFFICIENT INTERPRETATION

1) Students whose mother's job is others as compared to a teacher, is associated with a -1.55 difference in G3, controlling for other variables.

CHANGING LEARNING RATE

The best performing model is Lasso regression. This is because of L1 regularization performs an inherent feature selection. The main problem with lasso regression is when correlated variables exists. It will retain only one variable and sets other correlated variables to zero. That will possibly lead to some loss of information resulting in lower accuracy.





10 CHOSEN FEATURES

OLS Regression Results

Dep. Variable:	G3	R-squared:	0.140
Model:	OLS	Adj. R-squared:	0.106
Method:	Least Squares	F-statistic:	4.105
Date:	Mon, 17 Sep 2018	Prob (F-statistic):	2.94e-05
Time:	00:43:32	Log-Likelihood:	-756.06
No. Observations:	264	AIC:	1534.
Df Residuals:	253	BIC:	1573.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	8.9153	1.707	5.221	0.000	5.553	12.278
traveltime	-1.1593	1.184	-0.979	0.329	-3.492	1.173
studytime	0.0309	0.962	0.032	0.974	-1.864	1.925
failures	-5.0995	1.173	-4.349	0.000	-7.409	-2.790
health	-0.6666	0.775	-0.860	0.391	-2.193	0.860
absences	1.9663	2.334	0.843	0.400	-2.630	6.562
schoolsup_yes	-1.6681	0.758	-2.201	0.029	-3.161	-0.176
famsup_yes	-0.3806	0.589	-0.646	0.519	-1.541	0.780
paid_yes	-0.3386	0.581	-0.583	0.560	-1.482	0.805
higher_yes	2.5197	1.374	1.834	0.068	-0.186	5.225
school_GP	0.8832	0.886	0.996	0.320	-0.863	2.629

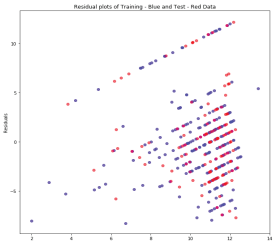
Selecting 10 variables through an intuitive thought process, an adjusted R-squared value of 10.6% was obtained as compared to the baseline of 15.1%. While the performance of the model was decreased, it's became simpler. Test results obtained outperforms the base model test results.

TEST RESULTS

Mean Squared Error: 17.39 ▲
Mean Absolute Error: 3.11 ▲
Root Mean Squared Error: 4.17 ▲
R2 Statistics: 17.2% ▲

There are no distinct patterns displayed in both the test and training residual plots.

RESIDUAL PLOTS

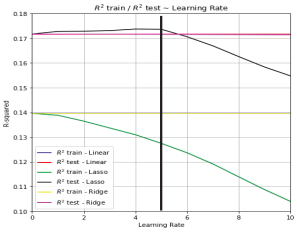


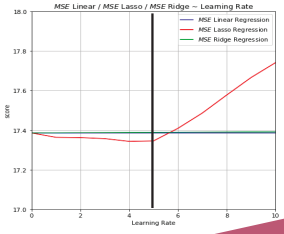
COEFFICIENT INTERPRETATION

1) Every additional failure, is associated with a decrease in G3 by 5.1 points, controlling for all other variables.
2) Students who have school support as compared to those without, are associated with a -1.67 difference, controlling for all other variables.

CHANGING LEARNING RATE

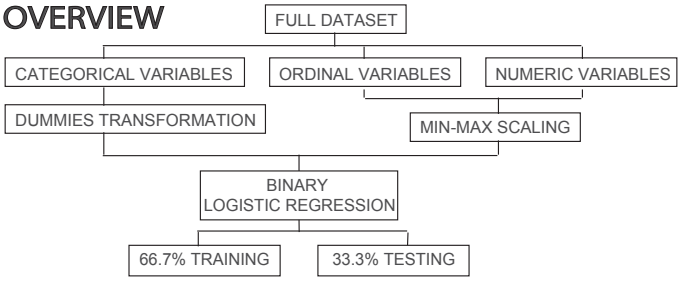
The best performing model is Lasso regression.The desired learning rate should be set at about 5 to obtain the highest R-squared with the lowest corresponding MSE.

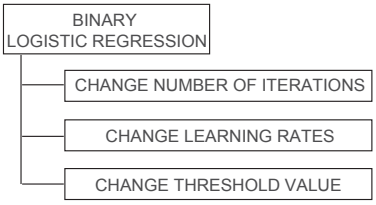




FULL DATASET LOGISTIC REGRESSION

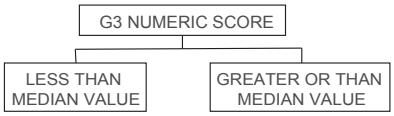
OVERVIEW





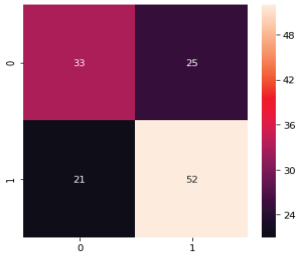
Using a self-created logistic regression function, a baseline model consisting of 39 variables was established, with an accuracy score of 65%. This represents the proportion of correct predictions relative to the total number of data points.

BINARY CLASSIFICATION

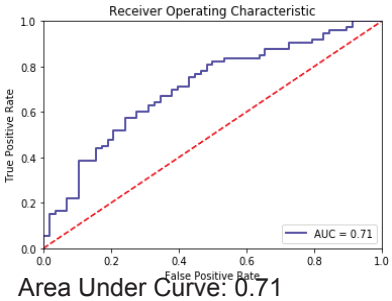


The median value for G3 is at 11 points. The intention for adopting this classification approach is to determine students who are underperforming (below 50%) relative to the entire student sample.

TEST RESULTS



Accuracy: 65%
Misclassification Rate: 35%
True Positive Rate: 71.2%
False Positive Rate: 43.1%
Specificity: 56.9%
Precision: 67.5%



Area Under Curve: 0.71

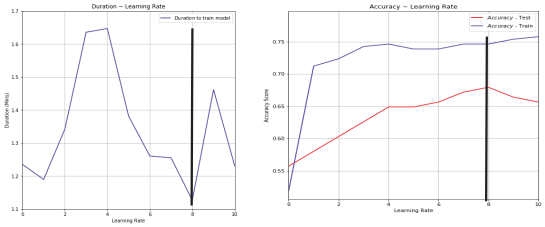
This means that the model has a 71% chance of being able to distinguish between the positive and negative classes

UNDERSTANDING AUC-ROC

AUC - ROC curve is a performance measurement for classification problems at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It shows much a model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s.

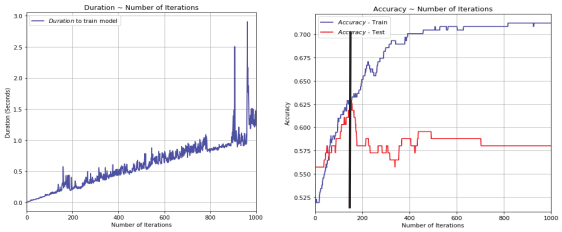
Note: True Positive Rates and Specificity are inversely proportional to each other. Decreasing threshold settings for classification, will result in more positive values increasing True Positive Rates but decreases the specificity (when it is actually no, how often does it predict no).

CHANGING LEARNING RATES



There are times where a larger learning rates leads to an increased in training duration. This is likely due to skipping through the minima due to large learning rates resulting in multiple attempts to return back. This is reflected with a decrease in accuracy.

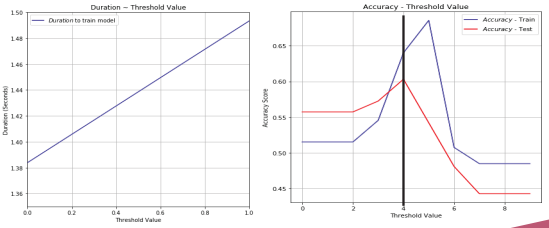
CHANGING # OF ITERATIONS



Fluctuations are likely due to noise expressed by random changes in values provided by the logistic regression. By increasing the number of iterations, accuracy increases initially but declines later. This is likely due to over-fitting.

CHANGING THRESHOLD VALUES

Changing the threshold values have positive effect on the accuracy of the model. However, after a certain point, there is a trade-off between wrongly classifying a negative as a positive or vice-versa. This is the trade-off between specificity and sensitivity.



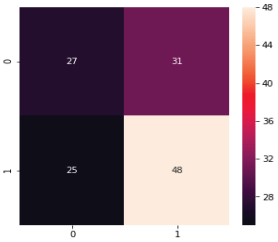
PARTIAL DATASET LOGISTIC REGRESSION

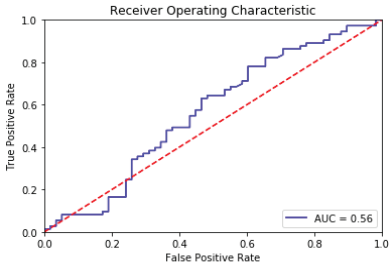
10 RANDOM FEATURES	10 CHOSEN FEATURES
Health	Travel Time
Mother's Job	StudyTime
Family Support	Failures
Reason	Health
Travel Time	Absences
Address	School Support
Father's Job	Family Support
Freetime	Paid
Nursery	Higher
Daily Alcoholic Consumption	School

10 RANDOM FEATURES

Using 10 randomly selected variables, the model achieved a 57% accuracy. This is a decrease in performance as compared to the baseline model.

TEST RESULTS



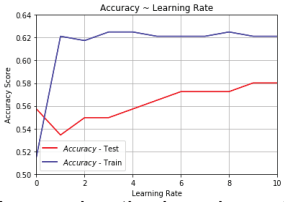


Accuracy: 57%
Misclassification Rate: 43%
True Positive Rate: 65.8%
False Positive Rate: 53.4%
Specificity: 46.6%
Precision: 60.8%

The model has a 56% chance of being able to distinguish between the positive and negative classes. At lower threshold values, the model has higher specificity and lower sensitivity. This is why the blue line was below the red line initially. The model is predicting negative class as a positive class.

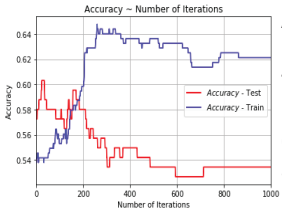
CHANGING LEARNING RATES

While training accuracy consistently increased with increasing learning rates, test accuracy decreased initially.



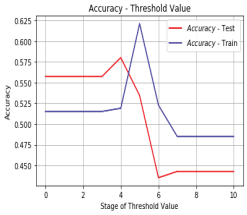
Increasing the learning rate allows the model to traverse saddle point plateaus rapidly thereby increasing performance.

CHANGING # OF ITERATIONS



As discussed previously, that increasing iterations can cause test accuracies to decline because of over-fitting. As compared to the training accuracies which increased.

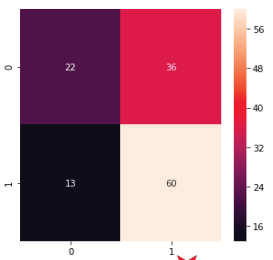
CHANGING THRESHOLD VALUES

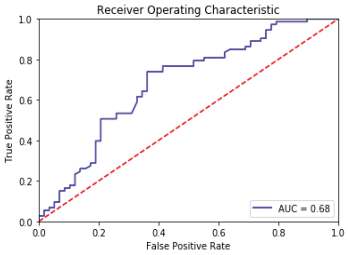


The ideal threshold value is at 0.4 which has the highest test accuracy. It is tempting to assume that the classification threshold should always be 0.5, but thresholds are problem-dependent, and must be tuned individually.

10 CHOSEN FEATURES

Using 10 selected variables, the model achieved a 63% accuracy. This is a decrease in performance as compared to the baseline model but the model is simpler.

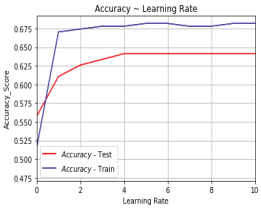




Accuracy: 63%
Misclassification Rate: 37%
True Positive Rate: 82.2%
False Positive Rate: 62.1%
Specificity: 37.9%
Precision: 62.5%

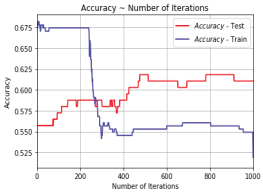
The model has a 68% chance of being able to distinguish between the positive and negative classes. It also consistently predicts the correct classes.

CHANGING LEARNING RATES



The ideal learning rate should be set at 4. Any further increase would not yield significant changes in test or train accuracies.

CHANGING # OF ITERATIONS

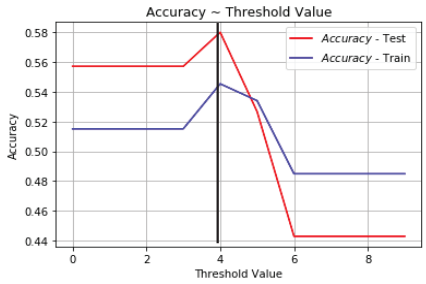


Increasing the number of iterations appears to have an adverse effect on the training set instead of the test set. The ideal iteration should be set at ~200.

This could be an effect of learning rate. A good minimal of the cost function may have been found in the earlier iterations but then jumped out of it due to stochastic gradient noise.

PARTIAL DATASET LOGISTIC REGRESSION

CHANGING THRESHOLD



The ideal threshold value is 0.4. This would allow us to obtain the highest train and test accuracy as compared to the rest.

Using derived metrics from the confusion matrix, the chosen features model performs worse in 5 out of 6 as compared to the baseline model. However, by plotting out the train and test errors, it can be seen that the chosen model provides a better generalization of the problem as compared to the baseline model. Therefore, test accuracies out-performs training accuracies, making this a good model.

SUMMARY

LINEAR REGRESSION

By experimenting, it was discovered that 20 features with a learning rate of 5 (Lasso) would provide the highest R-squared values while balancing for a relatively low Mean Squared Error and AIC & BIC value for both training and test datasets. In order to change the learning rates, lasso and ridge regression had to be deployed. This allowed for the usage of L1 and L2 regularization. There are disadvantages and advantages to the different types of regularization. However, ridge and linear regression are similar in their proportionality. Learning rates seem to have a minimum effect on linear regressions because Scikit Learn uses OLS.

BASELINE
Mean Squared Error: 18.01
Mean Absolute Error: 3.24
Root Mean Squared Error: 4.24
R2 Statistics: 14.2%

10 RANDOM
Mean Squared Error: 21.74
Mean Absolute Error: 3.54
Root Mean Squared Error: 4.66
R2 Statistics: -3.6%

10 CHOSEN
Mean Squared Error: 17.39
Mean Absolute Error: 3.11
Root Mean Squared Error: 4.17
R2 Statistics: 17.2%

LOGISTIC REGRESSION

By experimenting, it was discovered that a learning rate of 8, with an approximate 180 - 200 iterations and a threshold value setting of 0.4 would obtain the highest test accuracy for the logistic regressions. Care must be given to the various trade-offs as a result of adopting these settings. This setting holds true regardless of whether the variables are randomly chosen, specifically chosen or all just fed in. While in terms of the metrics shown below, the baseline model seems to be the best, however, by comparing the test and train results, it is advisable to adopt the chosen model. Based on the test and train error results, the chosen model provides a good genralization of the problem, allowing for the test results to outperform the training results.

BASELINE
Accuracy: 65%
Misclassification Rate: 35%
True Positive Rate: 71.2%
False Positive Rate: 43.1%
Specificity: 56.9%
Precision: 67.5%

10 RANDOM
Accuracy: 57%
Misclassification Rate: 43%
True Positive Rate: 65.8%
False Positive Rate: 53.4%
Specificity: 46.6%
Precision: 60.8%

10 CHOSEN
Accuracy: 63%
Misclassification Rate: 37%
True Positive Rate: 82.2%
False Positive Rate: 62.1%
Specificity: 37.9%
Precision: 62.5%