Exploring Machine Learning

Cassandra Pengelly[1]

1. Introduction

Economists have long been interested in the discussion of what factors influence a person's income. In more recent times, machine learning techniques have become one tool

Machine learning race is a topic that has . . . One of the uses of machine learning is it can be used to classify data.

This paper is split into two main parts: part 1 3 applies machine learning techniques to the NIDS dataset 2 and part 2 4 makes use of sequel to manipulate the NIDS dataset. The machine learning section first compares the effectiveness of linear regression and regularised regression on predicting people's incomes. Then 5 classification algorithms - Linear Discriminant Analysis, Classification and Regression Trees, k-Nearest Neighbors, Support Vector Machines with a linear kernel and Random Forest - are evaluated on their accuracy in predicting a person's race.

2. Data

The data used for this assignment was sourced from Wave 5 of the National Income Dynamics Survey (NIDS) (*National income dynamics study 2017, wave 5 dataset* (2018)). The survey is a nationally representative household panel study, which started in 2008 with a group of over 28,000 individuals from 7,300 households. The same households are surveyed every 2 years for NIDS. The latest survey - wave 5 - was conducted in 2017. For wave 5, a total of 39,434 individuals were interviewed; 20,113 of which were part of the original study - wave 1 - and 2,016 were from a top-up sample. NIDS is funded by the Department of Planning, Monitoring and Evaluation and the survey is implemented by the Southern Africa Labour and Development Research Unit (SALDRU) at the University of Cape Town. The data set is comprehensive and covers topics relating to poverty, health, household composition, mortality, expenditure, income and employment.

Email address: 20346212@sun.ac.za (Cassandra Pengelly)

3. Machine Learning

3.1. Predicting Income

Econometrics often makes use of regression analysis to model economic phenomena, test economic hypotheses and to forecast economic activity Studenmund (2014: 2). A popular method in econometric regression modeling is that of Ordinary Least Squares; however, advances in machine learning have presented alternative/augmenting methods that may be (more) useful. One such augmenting method is K-fold cross validation, which evaluates the skill of machine learning models. As Rodriguez, Perez & Lozano (2009: 569) explain, in K-fold cross validation, a data set is randomly split into k number of groups, of similar sizes. The first group is considered a validation set and the method is fitted to the other $k - 1$ groups.

Below, 3.1 displays 3 different linear regressions of log of income using K-fold cross validation. For these regressions I built a function "linreg", which takes in a data frame, cleans and splits the data into a training (70% of the full data set) and a test set (30% of the full data set) and runs 3 different linear regressions, applying K-fold cross validation. The results of the regressions are then collected and stored in a list, which is returned by the function. I used k = 10, because empirically k=10 has been shown have test error rate estimates that have relatively low bias and variance (Kassambara (2018a)). I have also set seed in the function for reproducibility.

Based on the Mincerian wage equation, Regression 1 (see 3.1) regresses log of income on age, years of schooling and a dummy variable for if a person has a tertiary qualification or not. Regression 2 includes a variable for age-squared and the categorical variable race. Regression 3 includes a variable each for gender and marriage. The signs of the coefficients of the three regressions look fairly standard[1] and most of the coefficients are statistically significant at 1% and lower.

[1] This at least is a good indication that the data is fairly well cleaned and is usable for testing the machine learning techniques

Table 3.1: Log-Income Regression Output

	Reg 1	Reg 2	Reg 3
(Intercept)	6.097 ***	5.791 ***	5.526 ***
	(0.145)	(0.367)	(0.354)
age	0.019 ***	0.041 **	0.040 **
	(0.002)	(0.015)	(0.015)
school	0.127 ***	0.109 ***	0.104 ***
	(0.009)	(0.009)	(0.008)
tertiary	0.691 ***	0.624 ***	0.594 ***
	(0.048)	(0.046)	(0.044)
age2		-0.000	-0.000 *
		(0.000)	(0.000)
‘raceAsian/Indian‘		0.575 ***	0.554 ***
		(0.123)	(0.117)
raceColoured		0.023	0.008
		(0.048)	(0.046)
raceWhite		0.767 ***	0.777 ***
		(0.070)	(0.066)
married			0.317 ***
			(0.045)
male			0.477 ***
			(0.039)

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Typically, economists are interested in evaluating the performance of a model, which can be done by assessing how well the model predicts the outcome variable. A useful statistical metric for measuring the performance of a regression model is the Root Mean Squared Error (RMSE) (Kassambara (2018b)).The RMSE measures the average error performed by the model in predicting the outcome for an observation. The mathematical formula for the RMSE is given by

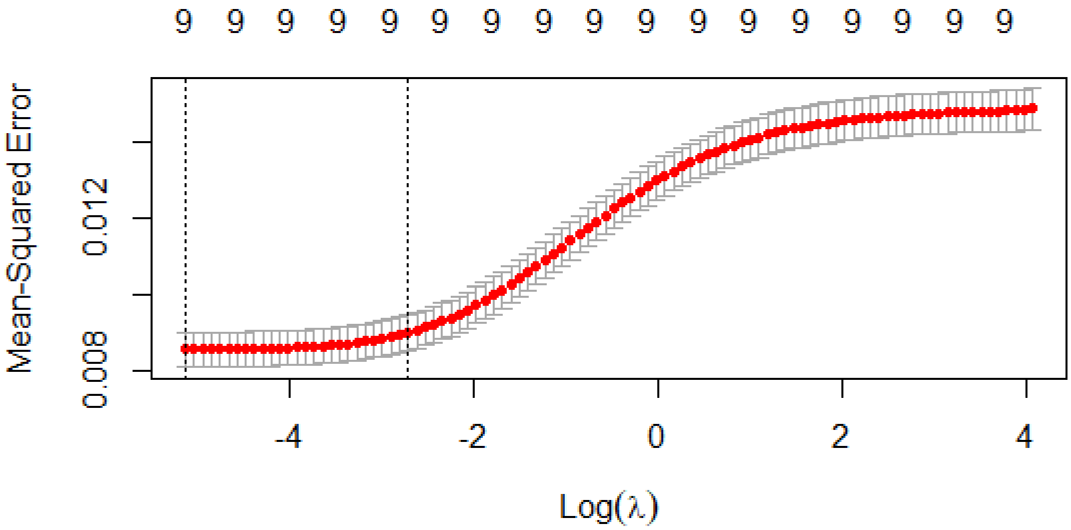
$$RMSE = \sqrt{(observeds - predicteds)^2/N)}$$

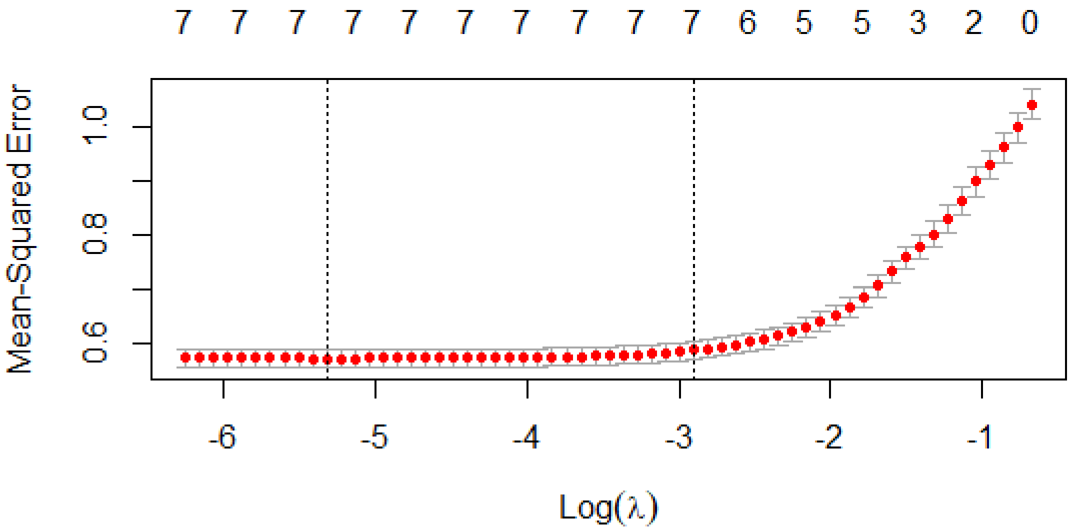
This implies lower the RMSE, the better the model performs.

Table 3.2 reports the RMSES for both the training data and the test data. We can see that the RMSEs decreased from regression 1 to regression 3 for both the training and test data. This indicates that regression 3 is a better model than both regressions 1 and 2 (and that regression 2 has better predictive power than regression 1). The RMSEs for the regression based on the training data are lower than for the test data. However, the RMSEs are close enough between the two data sets for all 3 models that the out-of-sample performance is fair; it does not seem that any of the models have been overfitted to the training data.

Table 3.2: Regression RMSEs and Observations

Regression	RMSE Train	RMSE Test
Reg 1	0.86	0.91
Reg 2	0.83	0.88
Reg 3	0.79	0.82



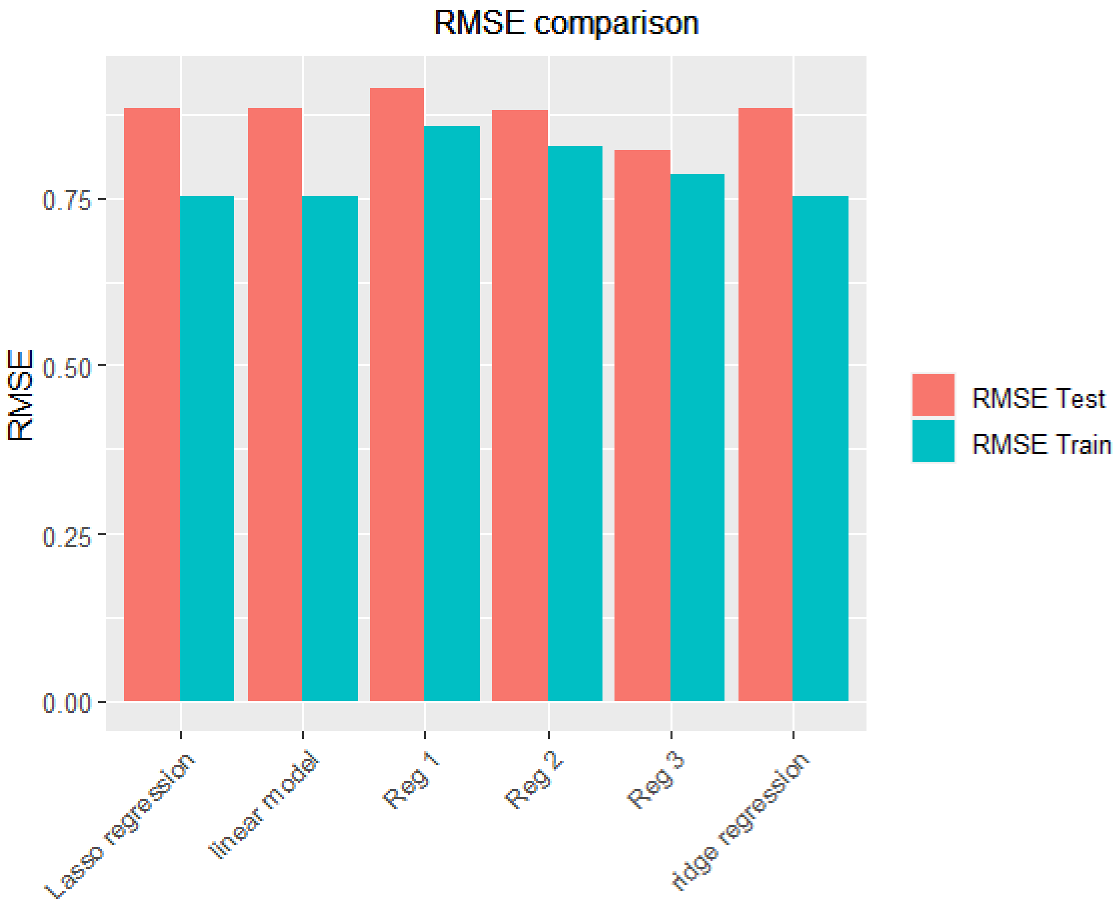


tibble [3 x 1] (S3: tbl_df/tbl/data.frame) $ Reg: num [1:3] 0.915 0.881 0.822 [1] "RMSE on training set: 0.752131288720079" [1] "RMSE on test set: 0.882932895421073" [1] "RMSE on training set: 0.752842149603987" [1] "RMSE on test set: 0.88481632436016" [1] "RMSE on training set: 0.752681521998011" [1] "RMSE on test set: 0.885305140817852" [1] "RMSE on training set: 0.760957159147364" [1] "RMSE on test set: 0.734724499416784" [1] "RMSE on training set: 0.761176361760862" [1] "RMSE on test set: 0.735047813129022" [1] "RMSE on training set: 0.76097313627485" [1] "RMSE on test set: 0.734938202183037"

Table 3.3: Model RMSEs

Dataset	Lasso regression	linear model	Reg 1	Reg 2	Reg 3	ridge regression
Test	0.89	0.88	0.91	0.88	0.82	0.88
Train	0.75	0.75	0.86	0.83	0.79	0.75

If we wanted to compare all 6 models visually we could look at the bar graph below .


RMSE comparison
RMSE
0.00
0.25
0.50
0.75
Lasso regression
linear model
Reg 1
Reg 2
Reg 3
ridge regression
RMSE Test
RMSE Train

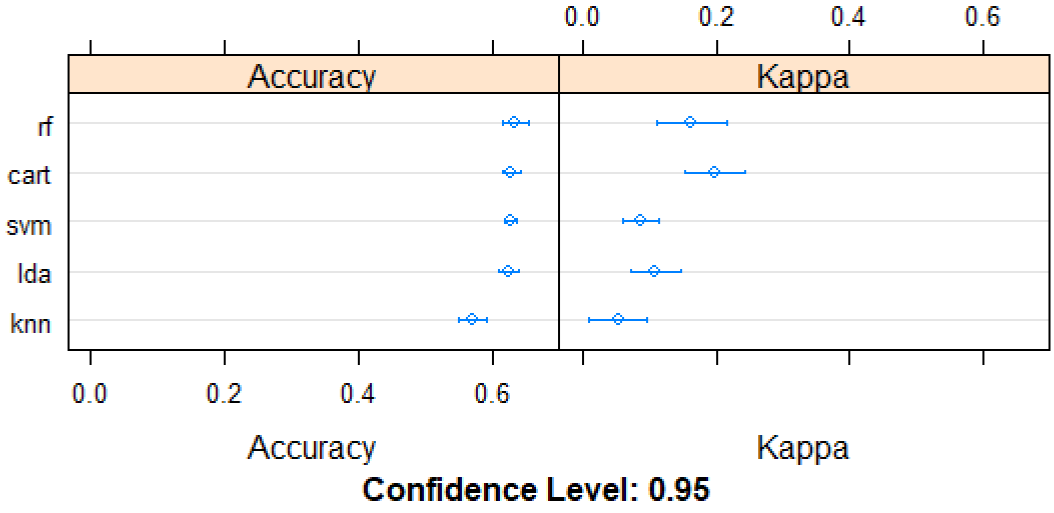


Figure 3.1: Machine Learning applied to unbalanced data

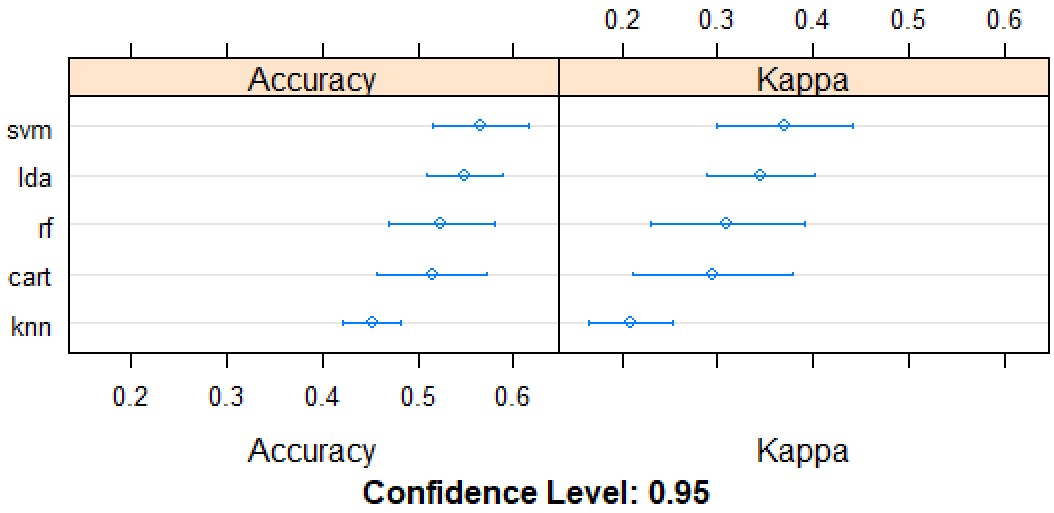


Figure 3.2: Machine Learning applied to balanced (undersampled) data

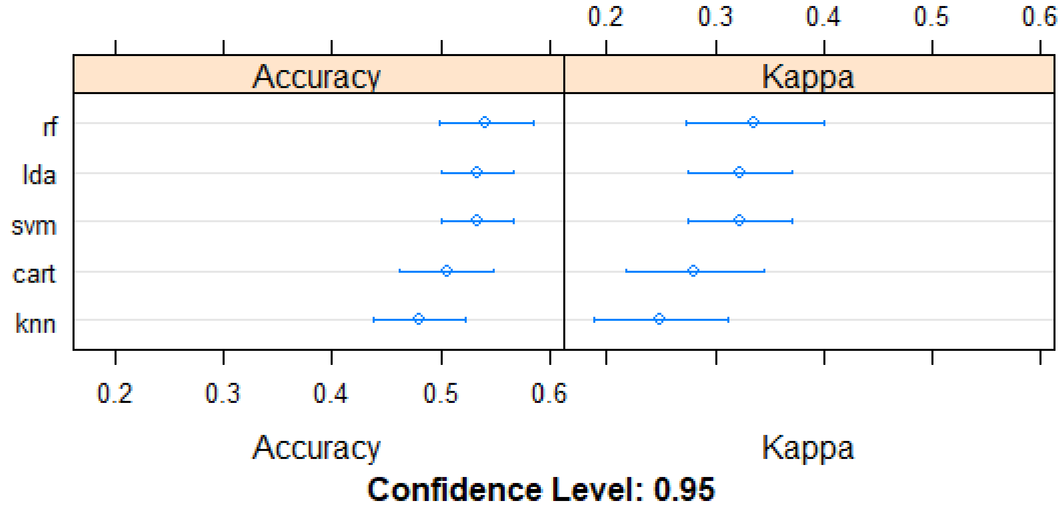
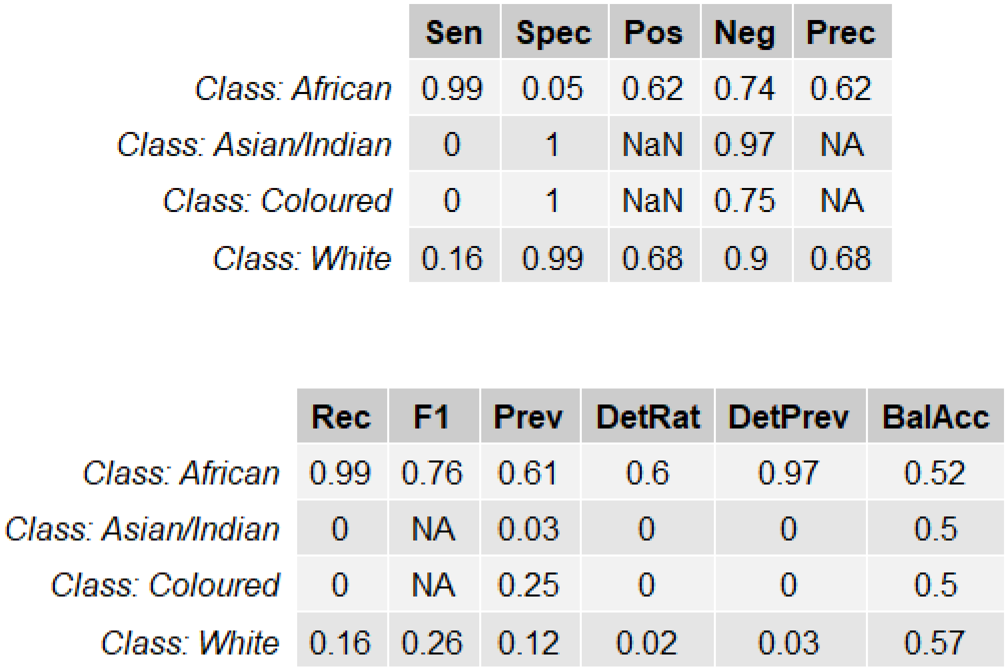


Figure 3.3: Machine Learning applied to balanced data

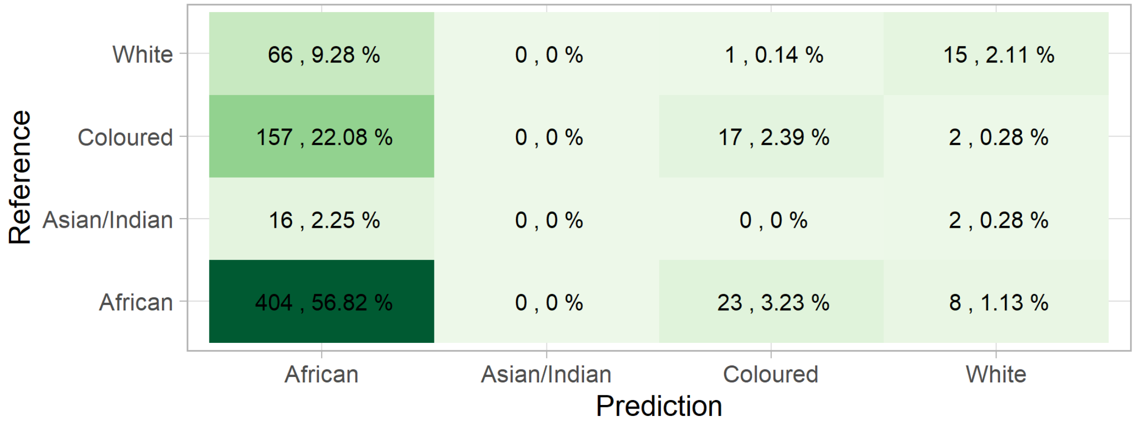
The graph below displays the confusion matrix and the tables report the relevant statistics.



	Sen	Spec	Pos	Neg	Prec
Class: African	0.99	0.05	0.62	0.74	0.62
Class: Asian/Indian	0	1	NaN	0.97	NA
Class: Coloured	0	1	NaN	0.75	NA
Class: White	0.16	0.99	0.68	0.9	0.68

	Rec	F1	Prev	DetRat	DetPrev	BalAcc
Class: African	0.99	0.76	0.61	0.6	0.97	0.52
Class: Asian/Indian	0	NA	0.03	0	0	0.5
Class: Coloured	0	NA	0.25	0	0	0.5
Class: White	0.16	0.26	0.12	0.02	0.03	0.57

Figure 3.4: SVM Statistics



	Statistics
Accuracy	0.61
Kappa	0.1

Figure 3.5: Confusion Matrix

4. Sequel

5. Conclusion

Katzke (2017)

References

10 Kassambara, A. 2018a. Cross-validation essentials in r.

Kassambara, A. 2018b. Linear regression essentials in r. [Online], Available: http://www.sthda.com/english/articles/40-regression-analysis/165-linear-regression-essentials-in-r/.

Katzke, N.F. 2017. *Texevier: Package to create elsevier templates for rmarkdown.* Stellenbosch, South Africa: Bureau for Economic Research.

National income dynamics study 2017, wave 5 dataset. 2018. Cape Town, South Africa: Department of Planning, Monitoring,; Evaluation [funding agency] & DataFirst [distributor]. [Online], Available: https://doi.org/10.25828/fw3h-v708.

Rodriguez, J.D., Perez, A. & Lozano, J.A. 2009. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence.* 32(3):569–575.

Studenmund, A.H. 2014. *Using econometrics a practical guide.* Pearson.

Appendix

Appendix A

Some appendix information here

Appendix B