'The Market for Automobiles in Cyprus'

Statistical Learning and Big Data

Matteo Orlandini [1] Riccardo Sommariva [1] Julian Tiedtke [2]

[1]PhD Economics, University of Côte d'Azur
[2]PhD Economics, Scuola Superiore Sant'Anna

May 23, 2022

Motivation

Motivation of our project is threefold:

1. Demand Analysis of the Automobile Market in Cyprus
2. Learn new statistical methods
3. Relate and discuss usefulness with regard to more traditional econometric techniques.

Data on the Automobile Market in Cyprus, 1989-2000[1]

- dataset provides information on units sold, total sales, retail and import prices and several product characteristics for different car models.
- 313 observations and 35 variables, where 26 are continuous and 8 nominal variables.

Table: Descriptive Statistics

	Percentile				
	0	25	50	75	100
Retail price	Mitsubishi Minica	Renault Clio	Opel Astra	Rover 416	BMW 730
	1782.15	6936.69	10829.35	16617.00	100240.38
Import price	Mitsubishi Minica	Honda Accord	Renault Megane	Mazda MPV	BMW 730
	1872.14	6903.37	10485.88	15289.11	75751.66
Sales	*i.a.* Fiat Panda	*i.a.* Mazda MX6	IVECO unknown	Toyota Landcruiser	Mazda 323
	10	33	102	383	7367
Engine size	Mitsubishi Minica	*i.a.* Rover 214	*i.a.* Honda Accord	Toyota unkown	Jeep Cherokee
	657	1396	1598	1986	3960

[1] Sofronis Clerides. *Gains from Trade in Used Goods: Evidence from the Global Market for Automobiles*. University of Cyprus Working Papers in Economics 6-2004. Dec. 2004. URL: https://ideas.repec.org/p/ucy/cypeua/6-2004.html.

Unsupervised Learning

Q: Are there different groups of cars in the market?
To answer this question we rely on clustering methods:

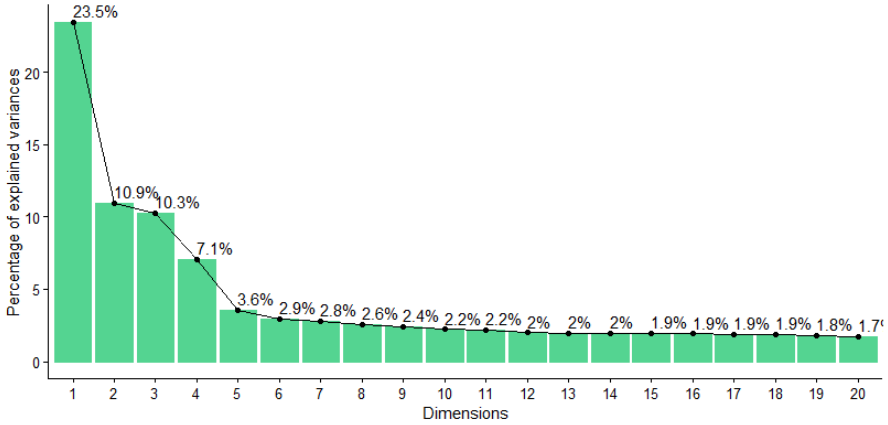
Problems:

- are all variables meaningful?
- mixed dataset

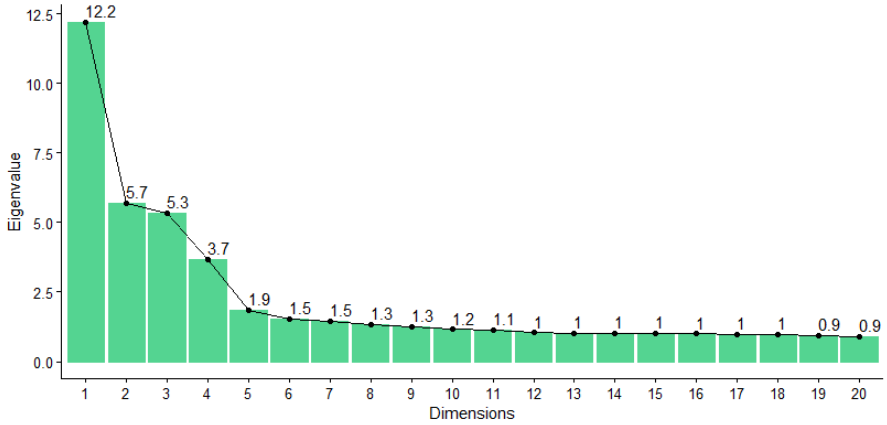
Approach:

1. extract meaningful information by relying on PCA
2. given mixed data, we rely on Factor analysis of mixed data (FAMD)
3. hierarchical clustering on results from factor analysis

FAMD

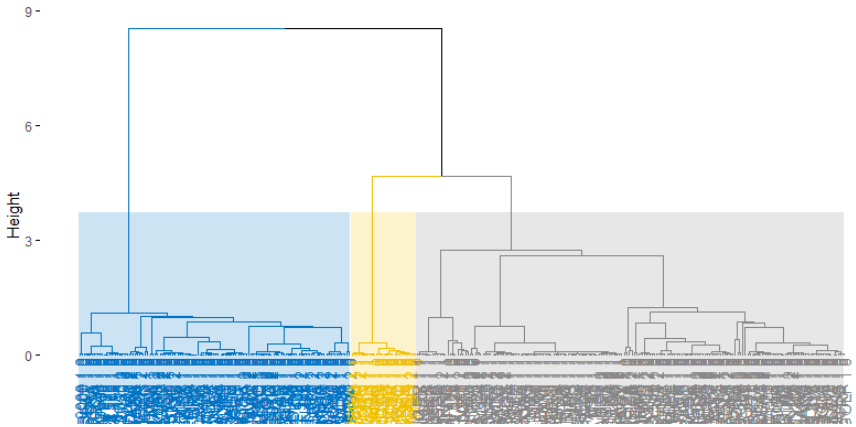


(a) Scree plot: explained variance

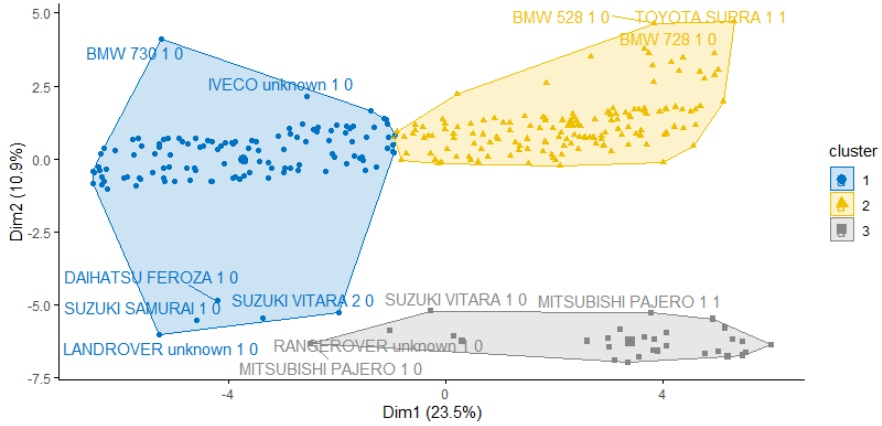


(b) Scree plot: eigenvalues

Hierarchical Clustering



(a) Dendogram



(b) Factor plot

Descriptive Statistics of Clusters

	Cluster 1	Cluster 2	Cluster 3
Retail price	15593.35	10476.37	14703.30
Unit sold	82.39	105.5	102.07
Sales	6566.05	11032.24	10101.45
aloga	14.98	17.55	25.97
Engine Power	23.29	67.9	79.22
Engine Capacities	1491.761	1751.477	2580.375
Age	0	2.06	1.31
Used (in %)	0	47	34
SUV (in%)	4	0	100

Supervised Learining: Lasso or Ridge?

$$maketotalsales_i = \beta_0 + \beta' x_i + \epsilon_i$$

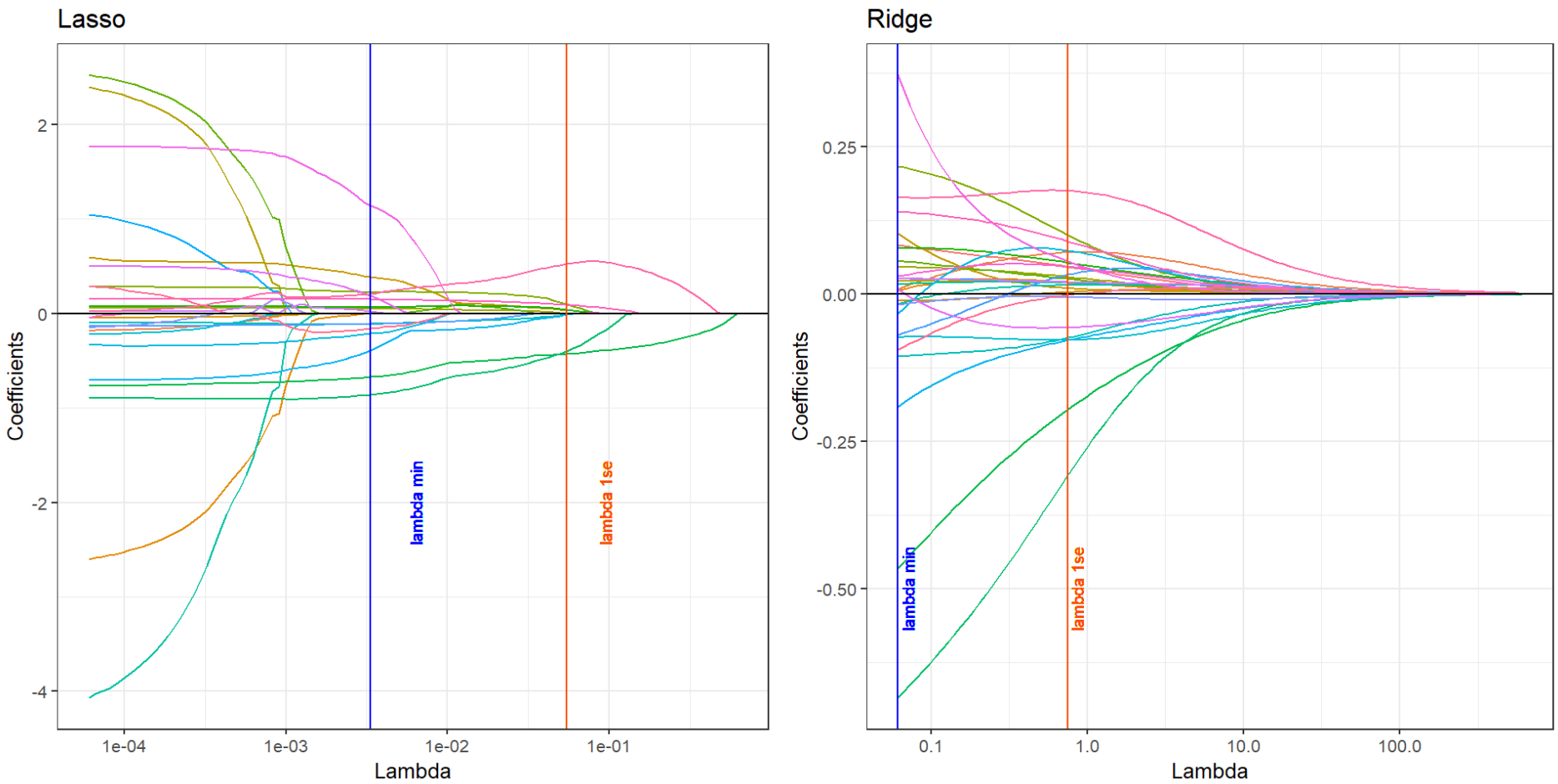
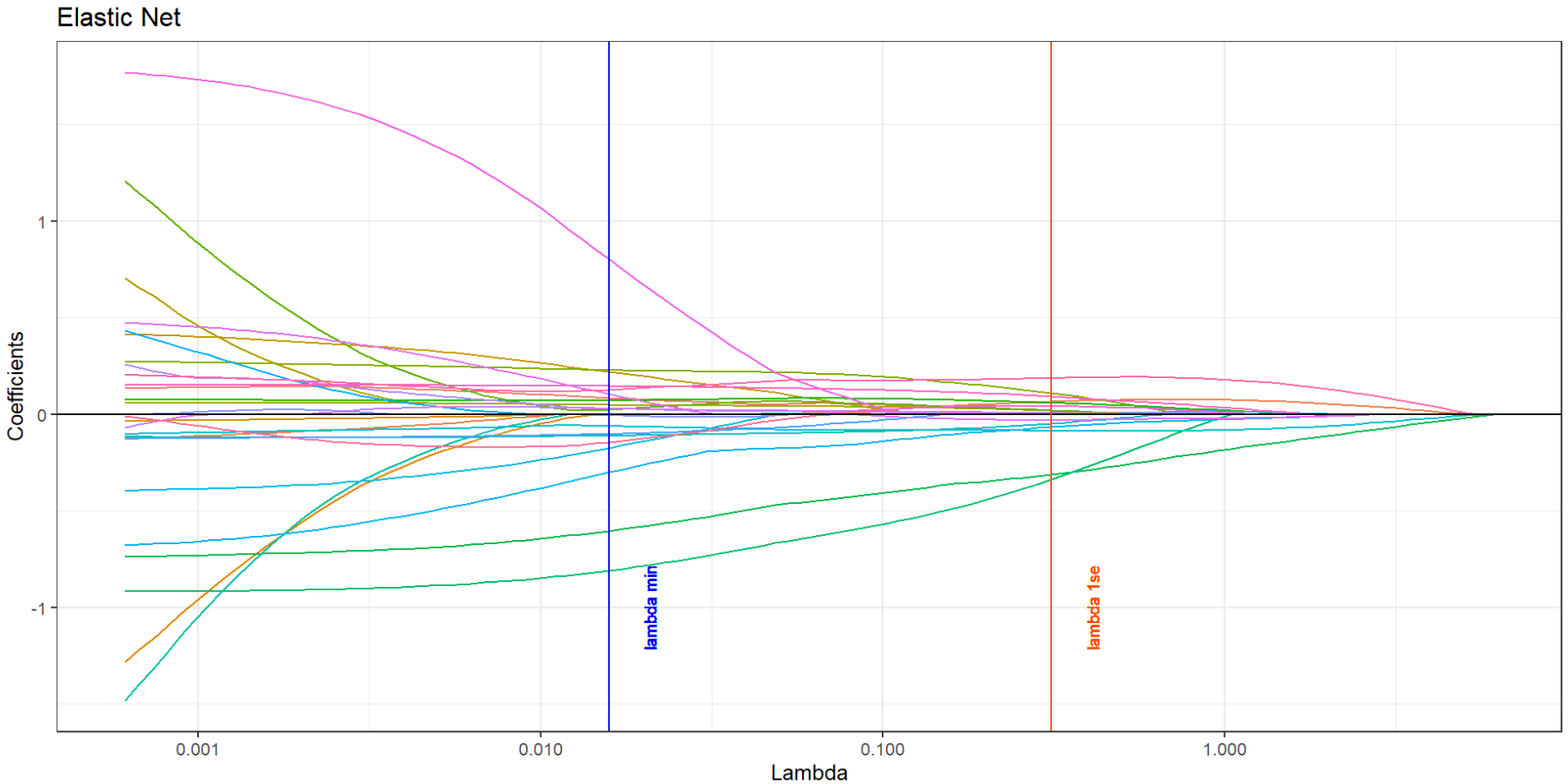


Figure: Lasso ($\alpha = 1$) vs Ridge ($\alpha = 0$)

... Elastic Net

α is the elastic net mixing parameter, with $0 \leq \alpha \leq 1$. The penalty is defined as

$$(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1$$



Mean-Squared Errors

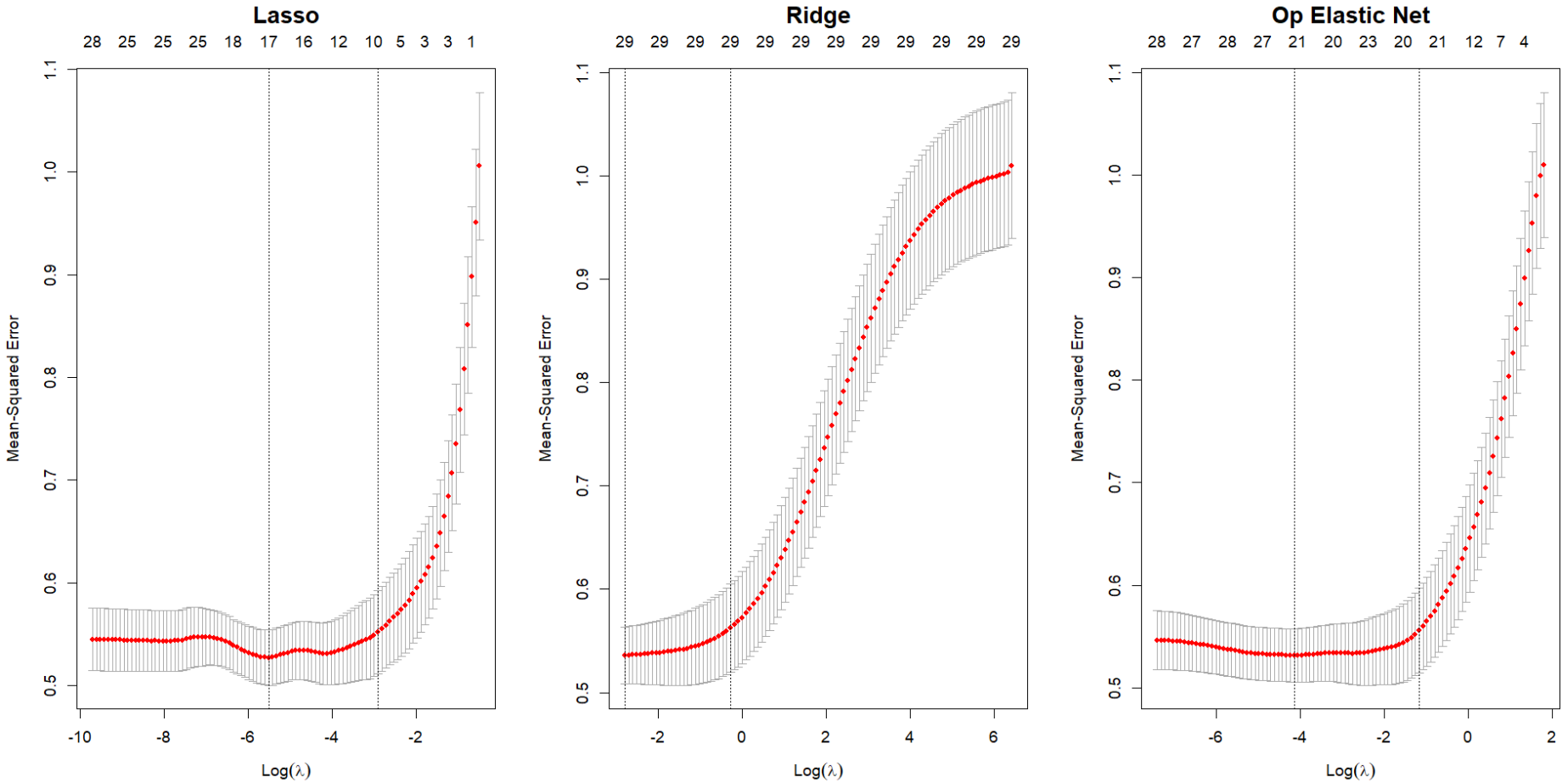


Figure: Lasso vs Ridge vs Optimal Elastic Net

Model Comparison

The elastic net with $\alpha = 0.1$ results to be the best specification for minimizing the RMSE.

Table: Models Performance Metrics

Statistic	Lasso	Ridge	Op. Elastic Net
RMSE	0.71	0.722	0.705
R Squared	0.509	0.493	0.512

Best subset selection

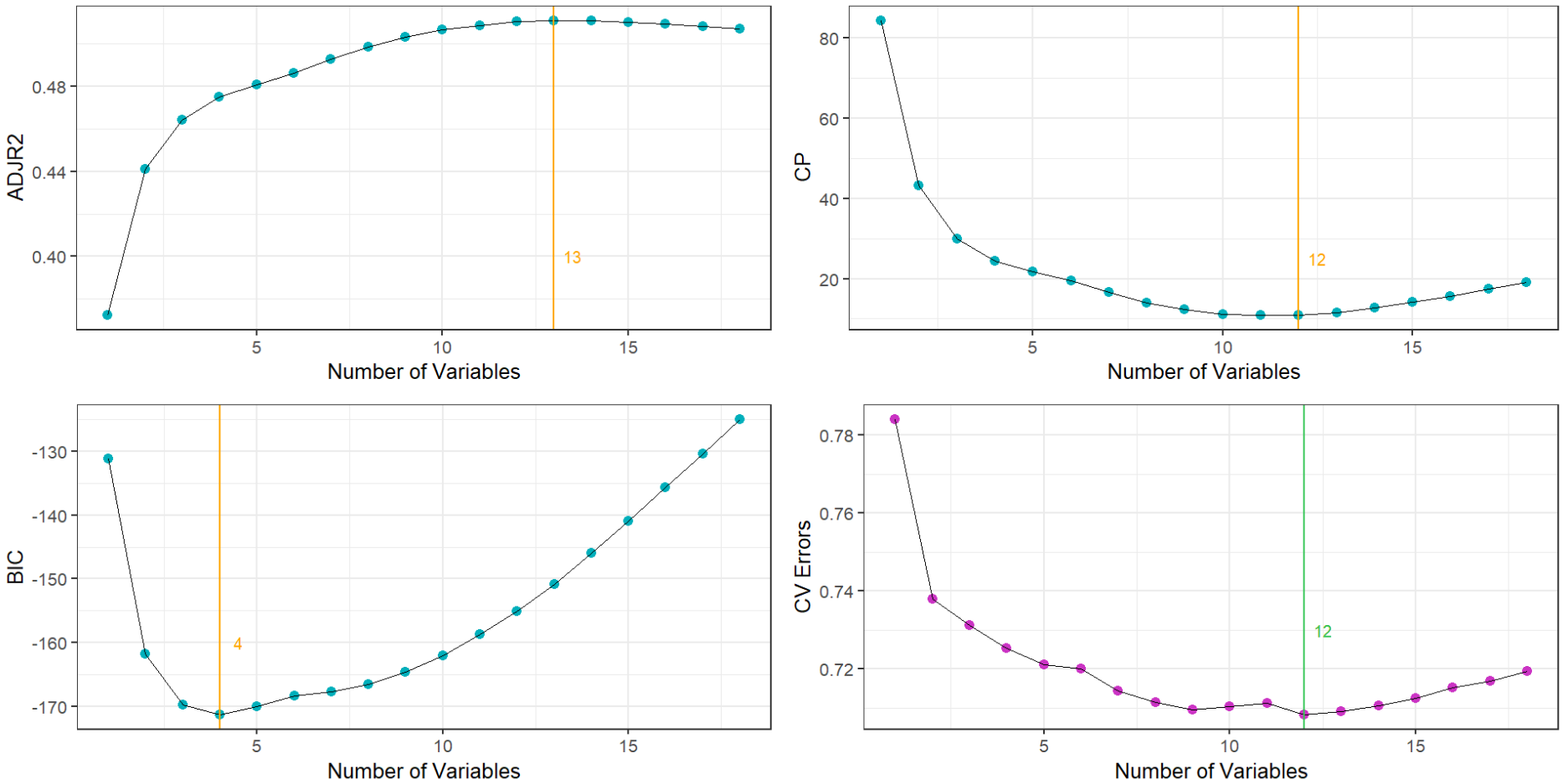


Figure: Selection Criteria

Best 12-Variable Model

	Total Sales:
Generation	−0.149
Used	0.255*
Suv	1.499***
Diesel	0.190*
Fourwd	−0.980***
Inxrate	−0.144***
EU	−0.773***
Constax	1.102***
Vat	−0.204
Advaltax	−0.643***
Retail Price	0.389
Import Price	−0.575**
Constant	−0.112
R^2	0.530
Adjusted R^2	0.511
F Statistic	27.334*** (df = 12; 291)

Stepwise Logistic

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.124e+02	7.127e+02	-0.158	0.8747	
generation1	1.997e+00	9.470e-01	2.109	0.0350	*
tottaxrate	1.076e+02	6.452e+02	0.167	0.8676	
cpi	1.270e+00	2.093e+00	0.606	0.5442	
gdp	-2.003e-08	1.408e-08	-1.422	0.1550	
unitsold	1.273e-02	6.194e-03	2.055	0.0399	*
refugee	3.090e+07	4.230e+07	0.731	0.4650	
vat	9.131e+01	1.510e+03	0.060	0.9518	
aloga	2.309e-01	1.939e-01	1.191	0.2338	
engpow	7.051e-03	2.554e-02	0.276	0.7825	
cylinder	-1.097e+00	1.198e+00	-0.916	0.3596	
diesel1	-4.373e-01	1.433e+00	-0.305	0.7602	
lnxrate	-6.121e+00	3.055e+01	-0.200	0.8412	
EU1	-6.339e+01	1.481e+03	-0.043	0.9659	
makesalestotal	-2.034e-05	4.984e-05	-0.408	0.6832	
impduty	3.090e+07	4.230e+07	0.731	0.4650	
constax	3.090e+07	4.230e+07	0.731	0.4650	
captariff	4.254e-03	2.388e-03	1.782	0.0748	.
fourwd1	-4.452e-01	2.057e+00	-0.216	0.8287	
advaltax	-3.090e+07	4.230e+07	-0.731	0.4650	
suv1	-1.087e+01	8.278e+00	-1.313	0.1893	

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.659e+01	2.759e+01	-0.964	0.3352	
lnxrate	-2.175e+00	9.608e-01	-2.264	0.0236	*
EU1	-3.670e+01	1.859e+03	-0.020	0.9842	
impduty	-2.163e+01	1.151e+01	-1.880	0.0602	.
captariff	1.757e-03	8.688e-04	2.023	0.0431	*
tottaxrate	3.716e+01	2.444e+01	1.520	0.1284	
advaltax	-3.916e+01	2.635e+01	-1.486	0.1373	

Complete Logistic (left) vs Stepwise Logistic (right)

Confusion Matrices

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  1
         1  6 15

               Accuracy : 0.8833
                 95% CI : (0.7743, 0.9518)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.004021

                  Kappa : 0.7287

 Mcnemar's Test P-Value : 0.130570

            Sensitivity : 0.8636
            Specificity : 0.9375
         Pos Pred Value : 0.9744
         Neg Pred Value : 0.7143
             Prevalence : 0.7333
         Detection Rate : 0.6333
   Detection Prevalence : 0.6500
      Balanced Accuracy : 0.9006

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  2
         1  6 14

               Accuracy : 0.8667
                 95% CI : (0.7541, 0.9406)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.0105

                  Kappa : 0.6842

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.8636
            Specificity : 0.8750
         Pos Pred Value : 0.9500
         Neg Pred Value : 0.7000
             Prevalence : 0.7333
         Detection Rate : 0.6333
   Detection Prevalence : 0.6667
      Balanced Accuracy : 0.8693

       'Positive' Class : 0
```

Confusion Matrix Complete (left) vs Confusion Matrix Stepwise (right)

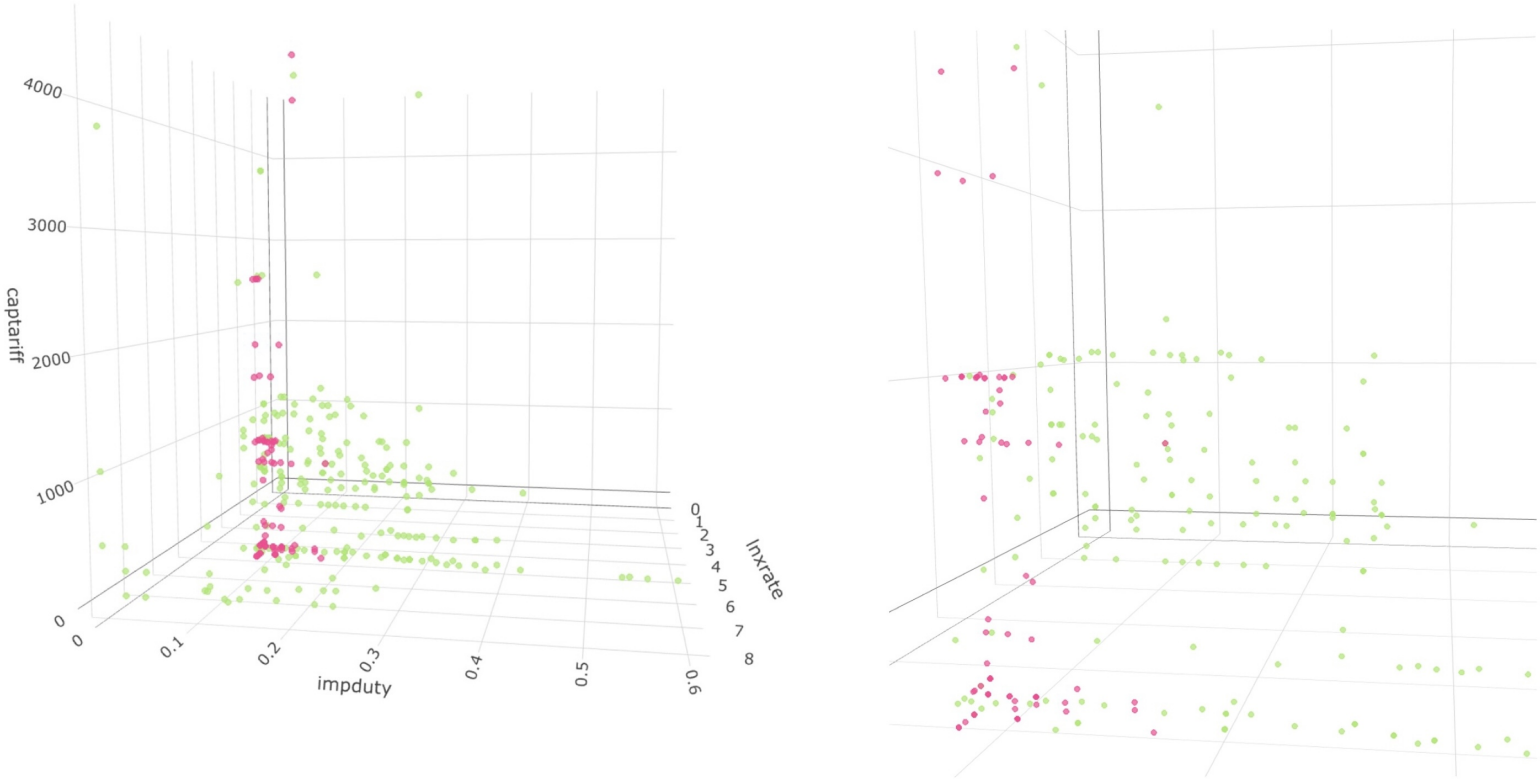
LDA and QDA

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 159  57
         1  18  10

               Accuracy : 0.6926
                 95% CI : (0.6306, 0.7499)
    No Information Rate : 0.7254
    P-Value [Acc > NIR] : 0.8877

                  Kappa : 0.0581

 Mcnemar's Test P-Value : 1.145e-05

            Sensitivity : 0.8983
            Specificity : 0.1493
         Pos Pred Value : 0.7361
         Neg Pred Value : 0.3571
             Prevalence : 0.7254
         Detection Rate : 0.6516
   Detection Prevalence : 0.8852
      Balanced Accuracy : 0.5238

       'Positive' Class : 0
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158   2
         1  19  65

               Accuracy : 0.9139
                 95% CI : (0.8714, 0.9459)
    No Information Rate : 0.7254
    P-Value [Acc > NIR] : 1.931e-13

                  Kappa : 0.7997

 Mcnemar's Test P-Value : 0.0004803

            Sensitivity : 0.8927
            Specificity : 0.9701
         Pos Pred Value : 0.9875
         Neg Pred Value : 0.7738
             Prevalence : 0.7254
         Detection Rate : 0.6475
   Detection Prevalence : 0.6557
      Balanced Accuracy : 0.9314

       'Positive' Class : 0
```

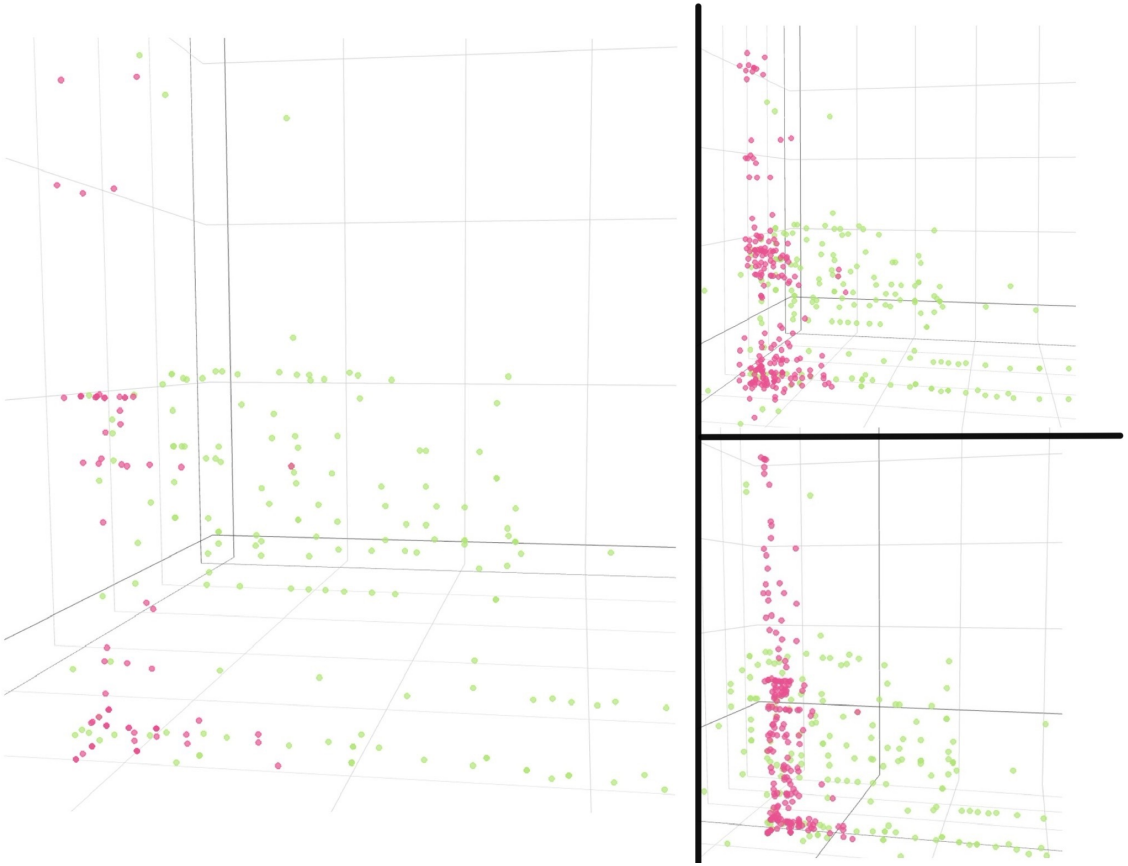
Confusion Matrix LDA (left) vs Confusion Matrix QDA (right)

Sample Rebalancing - 1



Data distribution (left) and zoom-in (right)

Sample Rebalancing - 2

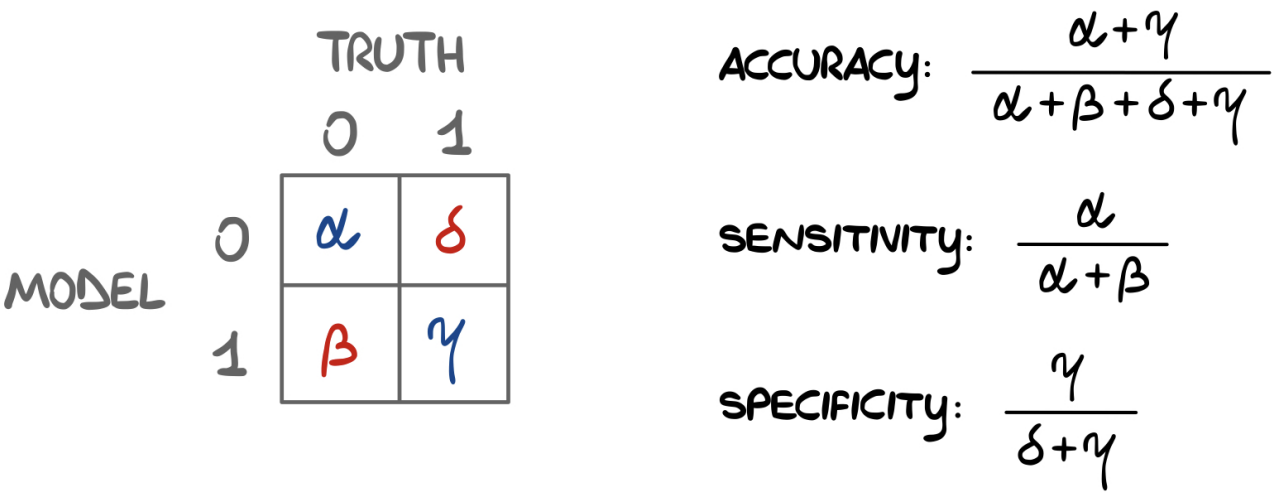


Zoom of the original plot (left), zoom of Gaussian rebalancing (up, right), zoom of convex rebalancing (down, right)

Comparison

	Standard	Oversampling	Undersampling	Gauss. Rebal.	Conv. Rebal.
Accuracy	0.6926	0.8418	0.8284	0.8371	0.8371
Sensitivity	0.8983	0.6893	0.6866	0.7014	0.6923
Specificity	0.1493	0.9944	0.9701	0.9729	0.9819

Comparison of the various techniques of sample rebalancing



Meaning of Accuracy, Sensitivity and Specificity.

References

- Hollander M., Wolfe D., Chicken E. - *Nonparametric Statistical Methods*, third edition, Wiley Series in Probability and Statistics.
- James G., Witten D., Hastie T., Tibshirani R. - *An Introduction to Statistical Learning*, second edition.
- Efron B., Hastie T. - *Computer Age Statistical Inference*, Cambridge University press.