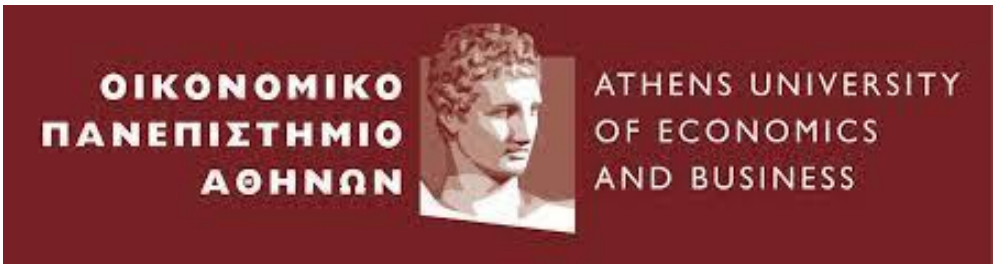

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY

MSc BUSINESS ANALYTICS

"Assignment 1 in course: Statistics for Business Analytics II"

Full Name: STAMATIOS SIDERIS

Register Number: f2822113

ATHENS, 2022

Table of Contents

Introduction

The dataset refers to telemarketing phone calls to sell long-term deposits. It consists of 39883 observations and 20 variables. 9 variables are numeric and 11 are character. The purpose is to find which variables contribute to a successful contact (the client subscribes to the product).

Data Cleaning

The dataset has zero blank values and 12520 unknown values. Variable pdays is excluded as most of the observations were not contacted before offering no significance to our model. Character variables are changed to type factor and separated to their respected levels and integer variables are changed to numeric.

Model Selection

Subscriptions (the response variable of our model) is a binary variable taking values of 0-1, counting the number of successes in a sequence of 39883 observations. This highly indicates that our data are following the **binomial distribution** and the appropriate generalized linear model to follow is the **logistic regression** and the link function **logit** as its output will be a binary value (0 or 1) rather than a numerical value.

Lasso is a regression analysis method that performs variable selection by minimizing its equation with respect to a penalty constraint. The penalty is proportional to a value λ. We perform **cross-validation** to find a reasonable value for λ that minimizes the mean classification error with respect to the constraint. We choose a grid of λ values and compute the cross-validation error rate for each value of λ. We then select the tuning parameter value for which the cross-validation error is smallest and choose the λ that is 1 standard deviation from the minimum value of λ to be more parsimonious. In **Figure 1**, are presented the misclassification errors as log(λ) increases. The intermittent lines show the minimum log(λ) (on the left) and the log(λ) 1se from minimum (on the right). For a λ 1se from minimum equal to 0.00059 variables age, cons.conf.idx and euribor3m are excluded from our model as they are penalized the most.

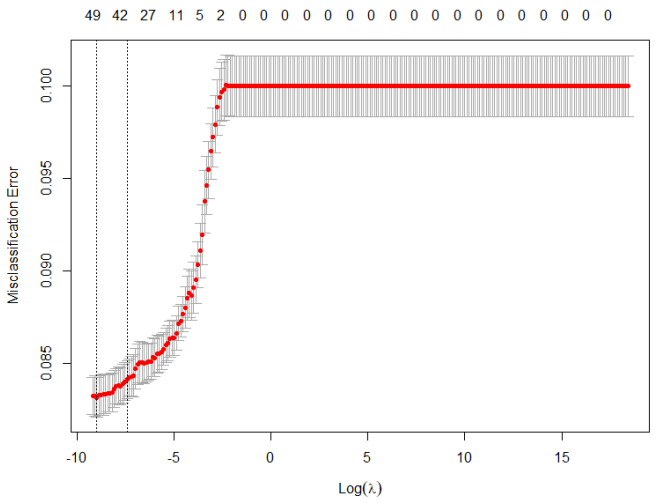

Figure 1 Misclassification Error as log(λ) increases

Stepwise regression with bidirectional elimination is used to check if the deletion or addition of a variable, according to a criterion, gives the most statistically significant improvement of the fit and repeats the

process until none improves the model to a statistically significant extent. We use AIC and BIC criterion to check which variables are eliminated and the result for both procedures is the same as variables loan, housing and education are excluded from our model.

Wald test is used to check the hypothesis Ho that a variable's coefficient is equal to zero, which means that it is statistically insignificant for our model. We perform wald test for all the remaining factor variables and the Ho is not rejected only for variable marital, which we exclude from our model.

Finally, we check remaining numeric variables' collinearity by plotting their correlations, in **Figure 2**. It is observed that a high correlation (close to 1) exists between variables emp.var.rate, cons.price.idx and nr.employed and so we exclude variables emp.var.rate and cons.price.idx from our model as they are equally explained by variable nr.employed and so we avoid the problem of multicollinearity.

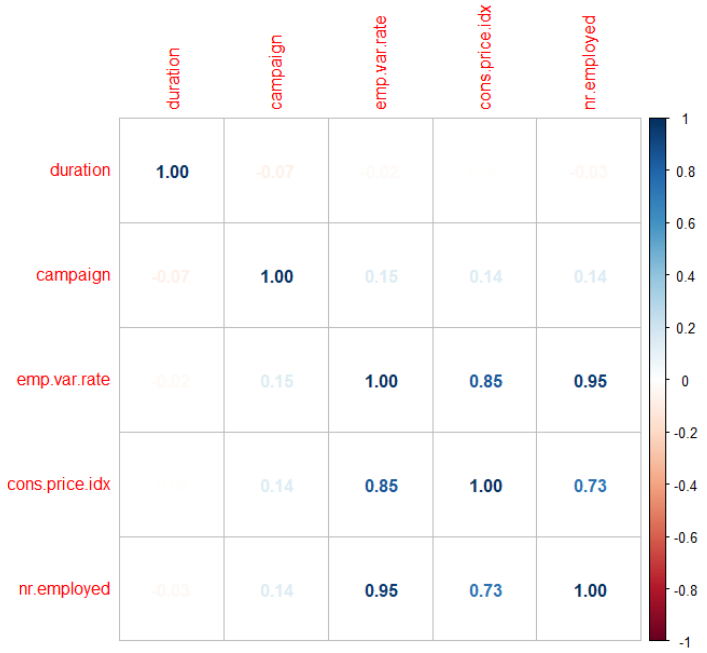

Figure 2 Correlation plot of numeric variables

Interpretation

Our final model is:

logit(SUBSCRIBED) = 76.67 – 0.35*jobbluecolar – 0.19*jobentrepreneur – 0.11*jobhousemaid – 0.05*jobmanagement + 0.25*jobretired – 0.12*jobselfemployed - 0.25*jobservices + 0.37*jobstudent – 0.06*jobtechnician – 0.05*jobunemployed – 0.04*jobunknown – 0.32*defaultunknown – 7.49*defaultyes – 0.26*contacttelephone + 0.51*monthaug + 0.10*monthdec + 0.42*monthjul + 0.58*monthjun + 1.21*monthmar – 0.71*monthmay – 0.02*monthnov + 0.31*monthoct – 0.26*monthsep – 0.11*day_of_weekmon + 0.03*day_of_weekthu + 0.07*day_of_weektue + 0.14*day_of_weekwed + 0.44*poutcomenonexistent + 1.76*poutcomesuccess + 0.005*duration – 0.04*campaign – 0.02*nr.employed

The model has high null deviance of 25925 meaning that the null model does not explain our data well.

The model has high residual deviance of 15681 meaning that the proposed model does not explain our data well, but it explains them better than the null model.

The intercept is equal to 76.67 meaning that for a customer working as an admin, no default credit, been contacted via mobile phone, in month April, on day Friday, with previous marketing campaign being a failure and zero call duration, number of contacts and number of employees the log odds of subscribing is 76.67.

Having all other variables values fixed:

Having a job as a blue-colar multiplies the actual odds of subscription by $\mathbf{e}^{-0.35} \simeq$ 0.70.

Having a job as an entepreneur multiplies the actual odds of subscription by $\mathbf{e}^{-0.19} \simeq$ 0.83.

Having a job as a manager multiplies the actual odds of subscription by $\mathbf{e}^{-0.05} \simeq$ 0.95.

Being retired multiplies the actual odds of subscription by $\mathbf{e}^{0.25} \simeq$ 1.28.

Being self-employed multiplies the actual odds of subscription by $\mathbf{e}^{-0.12} \simeq$ 0.89.

Having a job offering services multiplies the actual odds of subscription by $\mathbf{e}^{-0.25} \simeq$ 0.78.

Being a student multiplies the actual odds of subscription by $\mathbf{e}^{0.37} \simeq$ 1.45.

Having a job as a technician multiplies the actual odds of subscription by $\mathbf{e}^{-0.06} \simeq$ 0.94.

Being unemployed multiplies the actual odds of subscription by $\mathbf{e}^{-0.05} \simeq$ 0.95.

Not knowing the potential subscriber's job multiplies the actual odds of subscription by $\mathbf{e}^{-0.04} \simeq$ 0.96.

Not knowing if the potential subscriber has a credit in default multiplies the actual odds of subscription by $\mathbf{e}^{-0.32} \simeq$ 0.73.

Having a credit in default multiplies the actual odds of subscription by $\mathbf{e}^{-7.49} \simeq$ 0.0005.

Having been contacted via telephone multiplies the actual odds of subscription by $\mathbf{e}^{-0.26} \simeq$ 0.77.

Having been last contacted in month August multiplies the actual odds of subscription by $\mathbf{e}^{0.51} \simeq$ 1.67.

Having been las contacted in month December multiplies the actual odds of subscription by $\mathbf{e}^{0.10} \simeq$ 1.11.

Having been last contacted in month July multiplies the actual odds of subscription by $\mathbf{e}^{0.42} \simeq$ 1.52.

Having been last contacted in month June multiplies the actual odds of subscription by $\mathbf{e}^{0.58} \simeq$ 1.79.

Having been last contacted in month March multiplies the actual odds of subscription by $\mathbf{e}^{1.21} \simeq$ 3.35.

Having been last contacted in month May multiplies the actual odds of subscription by $\mathbf{e}^{-0.71} \simeq$ 0.49.

Having been las cotacted in month November multiplies the actual odds of subscription by $\mathbf{e}^{-0.02} \simeq$ 0.98.

Having been last contacted in month October multiplies the actual odds of subscription by $\mathbf{e}^{0.31} \simeq$ 1.36.

Having been las cotacted in month September multiplies the actual odds of subscription by $\mathbf{e}^{-0.26} \simeq$ 0.77.

Having been last contacted on day Monday multiplies the actual odds of subscription by $\mathbf{e}^{-0.11} \simeq$ 0.90.

Having been last contacted on day Thursday multiplies the actual odds of subscription by $\mathbf{e}^{0.03} \simeq$ 1.03.

Having been last contacted on day Tuesday multiplies the actual odds of subscription by $\mathbf{e}^{0.07} \simeq$ 1.07.

Having been last contacted on day Wednesday multiplies the actual odds of subscription by $\mathbf{e}^{0.14} \simeq$ 1.15.

Not knowing the outcome of the previous marketing campaign multiplies the actual odds of subscription by $\mathbf{e}^{0.44} \simeq$ 1.55.

Being the result of previous marketing campaign a success multiplies the actual odds of subscription by $\mathbf{e}^{1.76} \simeq$ 5.81.

An increase by 1 unit to the call duration multiplies the actual odds of subscription by $\mathbf{e}^{0.005} \simeq$ 1.005.

An increase by 1 unit to number of contacts performed during this campaign and for this client multiplies the actual odds of subscription by $\mathbf{e}^{0.04} \simeq$ 1.04.

An increase by 1 unit to number of employees working multiplies the actual odds of subscription by $\mathbf{e}^{-0.02} \simeq$ 0.98.

Goodness of Fit

For Deviance = -2[LM - LS] where LM is the maximum log likelihood of the model, and LS is the maximum log likelihood of an "ideal" model that fits as well as possible, the greater the deviance, the worse the model fits compared to the "best case". If our final model is good, then, its Deviance should follow a X^2 distribution with n − (p + 1) degrees of freedom, where n is the sample size, and p is the number of parameters. As the pchisq test for lower.tails=FALSE is equal to 1 we cannot reject the null hypothesis that our final model is correct.

We compare our model to the constant-only model to see if the chosen variables significantly improve our model. By performing the pchisq for lower.tail=FALSE on the difference of the deviance of the 2 models we observe that the result is 0 meaning that our model is superior to the null model.

Residuals

We are plotting the Pearson residuals (**Figures 3**) and Deviance residuals (**Figures 4**) of our model against the independent variables. By performing an outliers test we observe 2 observations that perform as outliers and so we limit the shown values of y-axis so that the rest of observations and patterns are clearly observed on the plot. From the plots it is observed that residuals of variable duration are facing a problem of non-linearity which might could be fixed adding its quadratic term in our model.

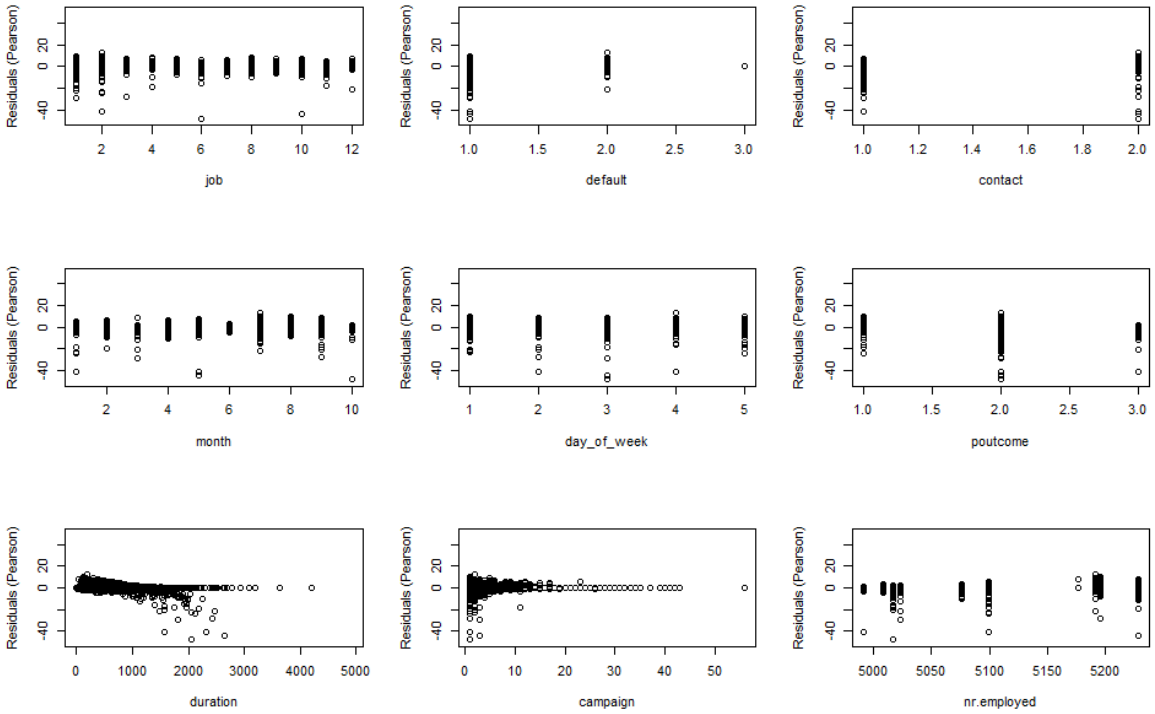

Figure 3 Pearson Residuals against independent variables

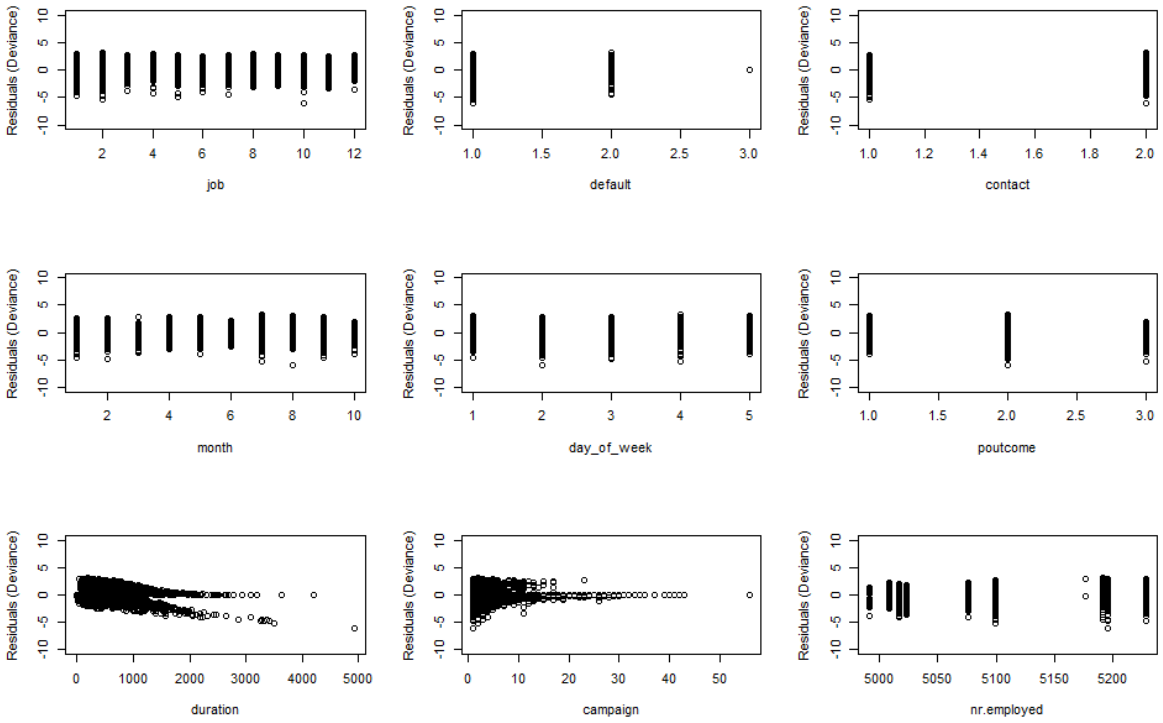

Figure 4 Deviance Residuals against independent variables

Conclusion

The chosen distribution followed by our data is the binomial and we choose the logistic regression followed by the logit link function to link our reference variable to its predictors. The model shows that being retired or a student increases the probability of subscription. The best months to be contacted are December, June, July, March, October and the best days Thursday, Tuesday, Wednesday. If a client had a subscription in the past, it is highly possible to re-subscribe. The model consists of high residual deviance but performs better than the constant model. Residuals assumption of non-linearity needs to be fixed by adding polynomial terms to the model.