Caravan Insurance Challenge

Hard voting and soft voting bagging techniques

Alessandro Della Siega, Davide Capone, Gabriele Pintus, Giulio Modesti

February 16, 2024

University of Trieste

The aim of this is project is to apply a Bagging technique called Soft Voting on a real problem, discussing the results and compare it to others baseline models.

Table of contents

Dataset Overview

Dataset Overview

- The dataset contains informations on **customers** of an **insurance company**.
- In particular, it includes product usage data and socio-demographic data derived by zip area codes.
- The goal is to **predict potential buyers of caravan insurance policy** and give explanation why.

Dataset Overview

Variables beginning with:

- *M*: refer only to demographic statistics of the postal code;
- *P* and *A*: refer to product ownership and insurance statistics in the postal code.

MFALLEEN	% of singles
ABRAND	Number of fire policies

Table 1: Example of two variables.

The target variable is ***CARAVAN***: number of mobile home policies, but the interpretation is binary (0: no polices, 1: otherwise).

Train-test split

The original dataset has shape (9822×87), but then it is **splitted** into train set and test set using the *ORIGIN* variable.

- Training set has 5822 rows.
- Test set has 4000 rows ($\approx 40\%$ of the original dataset).

In the following, we conduct the EDA part only on the training set, the test set is only used for testing the models.

EDA, preprocessing

Variables type

- The majority of the variables are of type **ordinal categorical**;
- the only variables with a truly **numeric** interpretation are *MAANTHUI* and *MGEMOMV*;
- Only two variables were purely **categorical**;

Spotting multicolinearity

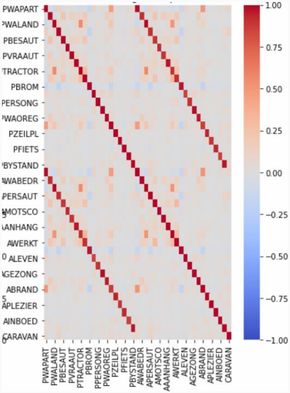


Figure 1: Zooming on the upper-right correlation plot.

Examining the correlation plot reveals significant **multicollinearity** among certain variables: we will eliminate variables with high correlation coefficients with each other.

One-hot encoding, features scaling

- **One-hot encoding** is applied to the variable *MOSTYPE* because it is purely categorical nominal (*L0* customer subtype category, 41 levels in total).
- We conducted **Min-Max Normalization** on each variable, considering that each variable might belong to a different domain. The target variable *CARAVAN* was not normalized, as it is a binary variable.

Target variable distribution

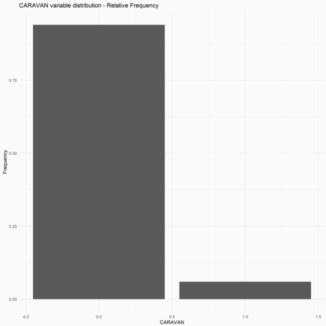
The distribution plot of the target variable (*CARAVAN*) reveals a **highly unbalanced dataset**: $\approx 6\%$ of observations belonging to the positive class.

Table 2: Proportions of positive and negative classes in the training set.

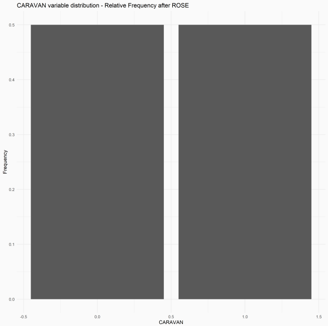
Class	Proportion
negative	94%
positive	6%

In order to **mitigate the class imbalance**, we utilized the *ROSE* (Random Over-Sampling Examples) technique to generate synthetic samples of the minority class through **oversampling**.

ROSE oversampling



(a) Original target variable distribution.



(b) Target variable distirbution after ROSE.

The target variable in the training set is now **perfectly balanced**.

Logistic Regression

Complete model

The train set includes 10948 observations and 94 variables (one of them is the target variable).

Use a logistic model using all the variables:

```
glm(formula = CARAVAN ~ ., family = binomial(link = "logit"),
    data = train_rose)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.758e+00  8.404e-01  -6.851 7.32e-12 ***
MOSTYPE.1    7.198e+00  5.382e+00   1.337 0.181094
MOSTYPE.2    4.951e+00  3.423e+00   1.446 0.148074
MOSTYPE.3    6.259e+00  3.020e+00   2.072 0.038238 *
####### TRUNCATED OUTPUT #######
PFIETS       6.029e+00  1.190e+00   5.065 4.08e-07 ***
PINBOED     -1.680e+00  1.103e+00  -1.523 0.127797
PBYSTAND     1.726e+00  4.259e-01   4.053 5.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 15177  on 10947  degrees of freedom
Residual deviance: 11897  on 10854  degrees of freedom
AIC: 12085
Number of Fisher Scoring iterations: 15
```

Variance Inflation Factor

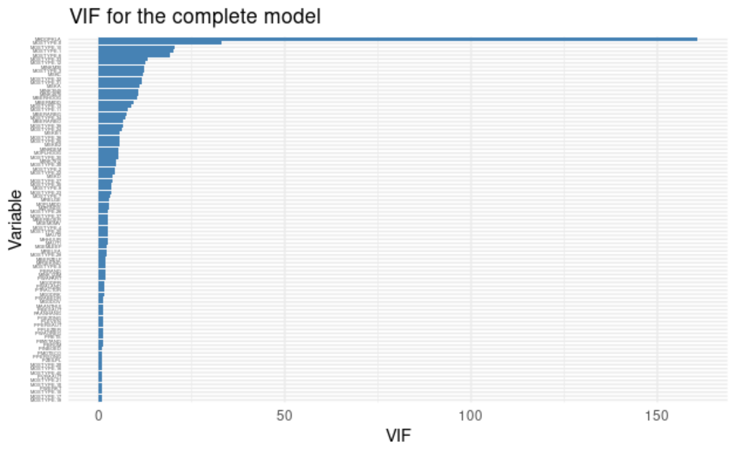


Figure 2: Variance Inflation Factor Barplot

Reduced model

The reduced train set includes only 29 variables (one of them is the target variable). The features where selected using VIF information, magnitude of the coefficients, respective p-value and real meaning.

```
glm(formula = CARAVAN ~ ., family = "binomial", data = train_rose[,
    vars])
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.970447   0.244575 -20.323  < 2e-16 ***
MOSTYPE.5     0.987707   2.306797   0.428 0.668526
MOSTYPE.11    1.393457   1.176395   1.185 0.236209
MOSTYPE.36    5.598584   0.972320   5.758 8.51e-09 ***
####### TRUNCATED OUTPUT #######
PPLEZIER      6.177039   0.679209   9.094  < 2e-16 ***
PFIETS        7.393364   1.115308   6.629 3.38e-11 ***
PBYSTAND      2.194998   0.405270   5.416 6.09e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 15177  on 10947  degrees of freedom
Residual deviance: 12608  on 10920  degrees of freedom
AIC: 12664
Number of Fisher Scoring iterations: 5
```

Stepwise regression

Using only the selected variables, we built a model using a stepwise regression. Both directions were used and BIC criterion was taken into consideration.

```
glm(formula = CARAVAN ~ (variables...) , family = "binomial", data = train_rose[,
    vars])
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.64807    0.16589 -28.018  < 2e-16 ***
MOSTYPE.36   5.39975    0.96818   5.577 2.44e-08 ***
MOSTYPE.38   5.53579    0.83983   6.592 4.35e-11 ***
MOSTYPE.39   2.97858    0.83748   3.557 0.000376 ***
##### TRUNCATED OUTPUT #####
PWAOREG      2.67156    0.42184   6.333 2.40e-10 ***
PBRAND       0.89802    0.12476   7.198 6.11e-13 ***
PPLEZIER     6.25528    0.68387   9.147  < 2e-16 ***
PFIETS       7.38016    1.10747   6.664 2.67e-11 ***
PBYSTAND     2.25831    0.40124   5.628 1.82e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ”. 0.1 ”  1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 15177  on 10947  degrees of freedom
Residual deviance: 12634  on 10929  degrees of freedom
AIC: 12672
Number of Fisher Scoring iterations: 5
```

Interpretation

The most important features in order to determine the target variable are:

-PPLEZIER Contribution boat policies, if PPLEZIER increases by 1 (one category), the log odds of the target varaible to be 1 increases by 6.25/9 = 0.69

-PFIETS Contribution bicycle policies, $\rightarrow$ 7.38/9 = 0.82

-PPERSAUT Contribution car policies, $\rightarrow$ 2.07/9 = 0.23.

-MOSTYPE.36 Couples with teens, $\rightarrow$ 5.39.

-MOSTYPE.38 Traditional families, $\rightarrow$ 5.5

Bayes Information Criteria

Let's compare the threee models using the BIC.

```
               df      BIC
complete_model 94 12771.04
reduced_model  28 12868.38
stepwise_model 19 12810.30
```

Test Set Confusion Matrix

Let's test the stepwise model on the test set.

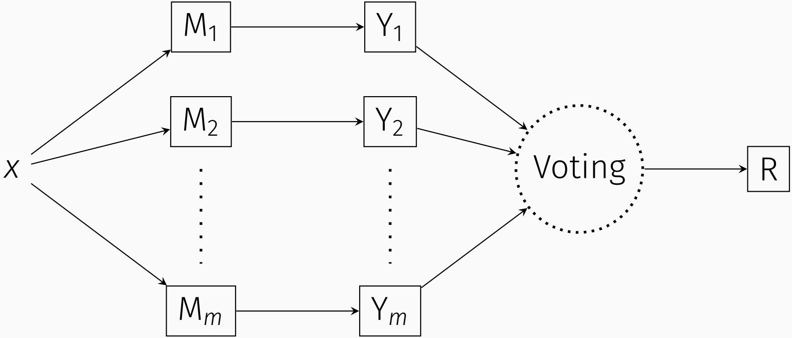
```
           Reference
Prediction    0    1
         0 2576 1186
         1   84  154

               Accuracy : 0.6825
                 95% CI : (0.6678, 0.6969)
    No Information Rate : 0.665
    P-Value [Acc > NIR] : 0.009728
                  Kappa : 0.1047
 Mcnemars Test P-Value : < 2.2e-16
            Sensitivity : 0.1149
            Specificity : 0.9684
         Pos Pred Value : 0.6471
         Neg Pred Value : 0.6847
             Prevalence : 0.3350
         Detection Rate : 0.0385
   Detection Prevalence : 0.0595
      Balanced Accuracy : 0.5417
```

Bagging

Bagging

An ensemble technique which has been proved to provide stability

1. Generate m bootstrap samples from $\mathcal{D}_{train}$
2. Fit m models, one for sample
3. Select a voting mechanism to combine all the predictions



Briefly - Hard voting

Every of the m voter express his opinion on the data point to belong to class 1.

- m_1 voters think the data point belongs to class 1
- m_0 voters think the data point belongs to class 0

Majority wins

Briefly - Soft voting

Every of the m voter express his **level of confidence** on the data point to belong to class 1.

- m voters think the point belongs to class 1 with a lof $\geq l_1$
- $m-1$ voters think the point belongs to class 1 with a lof $\geq l_2 \geq l_1$

  ...
- 1 voter thinks the point belongs to class 1 with a lof $\geq l_m$

The overall level of confidence is the weighted average:

$$l = \frac{m}{m} l_1 + \frac{m-1}{m}(l_2 - l_1) + \cdots + \frac{1}{m}(l_m - l_{m-1})$$

Eventually the point is classified in 1 if the level of confidence is greater than 50%, in 0 otherwise.

proof

Define $F_Y(a,b) = \int_a^b f_Y(t)\mathrm{dt}$
Suppose wlog that $\eta_1 \leq \eta_2 \leq \cdots \leq \eta_3$.

$$
\begin{aligned}
S_m &= F_Y(0,\eta_1) + F_Y(0,\eta_2) + \cdots + F_Y(0,\eta_m) = \\
&= F_Y(0,\eta_1) \\
&+ F_Y(0,\eta_1) + F(\eta_1,\eta_2) \\
&+ F_Y(0,\eta_1) + F(\eta_1,\eta_2) + F(\eta_2,\eta_3) \\
&\vdots \\
&+ F_Y(0,\eta_1) + F(\eta_1,\eta_2) + \cdots + F(\eta_{m-1},\eta_m) = \\
&= mF_Y(0,\eta_1) + (m-1)F_Y(\eta_1,\eta_2) + \cdots + F(\eta_{m-1},\eta_m)
\end{aligned}
$$

Our case

Every model is a logistic regression
We assume all the covariates to be indepent one another

The parameters vector $\hat{\boldsymbol{\beta}}$ is estimated via maximum likelihood, therefore:

$$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}_\beta, \Sigma_\beta | \mathsf{x}),$$

Our case

The linear predictor $\hat{\eta}$ is a linear combination of the vector of parameters, therefore:

$$\hat{\eta} \sim \mathcal{N}(\mu_\eta, \sigma_\eta^2 | \mathsf{x})$$

where:

- $\mu_\eta = \mathsf{x}^T \boldsymbol{\mu}_\beta$
- $\sigma_\eta^2 = \mathsf{x}^T \Sigma_\beta \mathsf{x}$

Our case

The model response variable Y the logistic transformation of $\hat{\eta}$

$$g(t) = \frac{1}{1 + e^{-t}}$$
$$Y = g(\hat{\eta})$$

We do not have a closed form for

- the distribution
- the expected value
- the variance

But we can still compute them numerically using sample statistics.

Our case

The ensemble can output just two numbers, hence:

$$R \sim \mathrm{Be}(p_R|\boldsymbol{x})$$

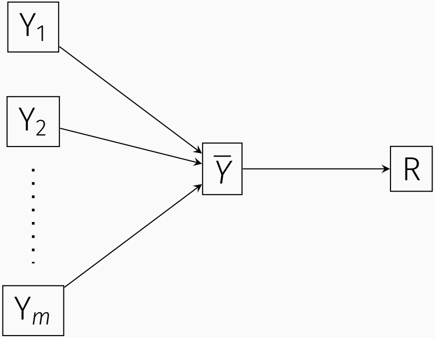
The question now is: *How do we make inference on p_R?*
Two main strategies:

- Soft voting
- Hard voting (most popular)

Soft voting

We make inference on p_R using as estimator $\hat{p}_R = \overline{Y}$. Under the independence assumption, the sample variance is disitrbuted normally:

$$\hat{p}_R \sim \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{m}\middle| x\right)$$



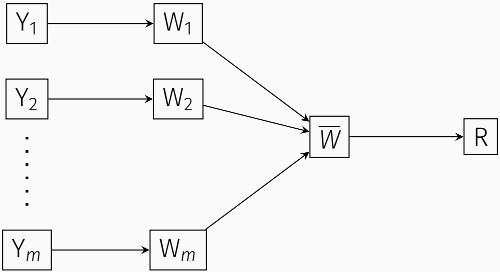
Hard voting

We first define $W = \begin{cases} 1 \text{ if } Y > \frac{1}{2} \\ 0 \text{ if } Y \leq \frac{1}{2} \end{cases}$,

so $W \sim \text{Be}(\tilde{p}|\boldsymbol{x})$, where $\tilde{p} = \mathbb{P}\left(Y > \frac{1}{2}\right) = \Phi\left(\frac{\mu_\eta}{\sigma_\eta}\right)$. Therefore we indirectly make inference on p_R by computing the sample mean over $\underline{W}$.

Under the independence assumption, we have that:

$$\hat{p}_R \sim \mathcal{N}\left(\tilde{p}, \frac{\tilde{p}(1-\tilde{p})}{m}\middle|\boldsymbol{x}\right)$$



Estimators comparison

Soft voting

$$\hat{p}_R \sim \mathcal{N}\left(\mu_Y, \frac{\sigma_Y^2}{m}\right)$$

- Unbiased
- Lower variance
- Lower or equal MSE

Hard voting

$$\hat{p}_R \sim \mathcal{N}\left(\tilde{p}, \frac{\tilde{p}(1-\tilde{p})}{m}\right)$$

- Biased
- Higher variance
- Higher or equal MSE

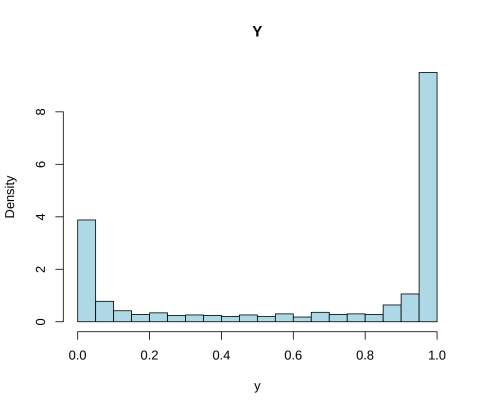
Self question

At this point I have a question for myself

Why does hard voting works?

Why hard voting works?

Consider the plot as an example of a possible distribution of Y:



Estimator	Mean	SS
$\hat{p}_S$	0.651	0.173852
$\hat{p}_H$	0.655	0.225975

Why hard voting works?

Expected value

- The better are the single models fitted to the data, the more concentrated would be the distribution around 0 and 1. If we consider the extreme case of Y following a Bernoulli distribution, then it follows that:

$$\overline{Y} = \mathbb{P}\left(Y > \frac{1}{2}\middle| x\right) = \mathbb{P}\left(Y = 1 | x\right)$$

  Which you can esteem by using the sample ratio $\frac{\#1}{\#1+\#0}$, which again, is equivalent to the sample mean.
- Another case is when the distribution of Y is simmetric.

Performance on CARAVAN

The validation set is 30% of the ROSE balanced set, while the test set is unbalanced

	Accuracy	Balanced Accuracy	F1 Score
Hard Voting - Validation	64.08	64.08	71.71
Soft Voting - Validation	**69.06**	**69.06**	**72.84**
Hard Voting - Test	**88.83**	61.19	**93.97**
Soft Voting - Test	80.70	**63.96**	88.99

Thank You!

Questions or Comments?