Predicting subscriptions with bank direct marketing data

Langyi Tian

Nov 2019

Project summary

- Mission: Predicting client subscription to a term product
- Data: Direct marketing campaign results
- Exploratory data analysis:

MCA with categorical variables, correlation matrix and PCA with numerics

- Data preprocessing:

Missing value imputation, one hot encoding, scaling, train/test split, under-sampling training set

- Model training and comparison:

Logistic regression, support vector machine, random forest

- Model selection:

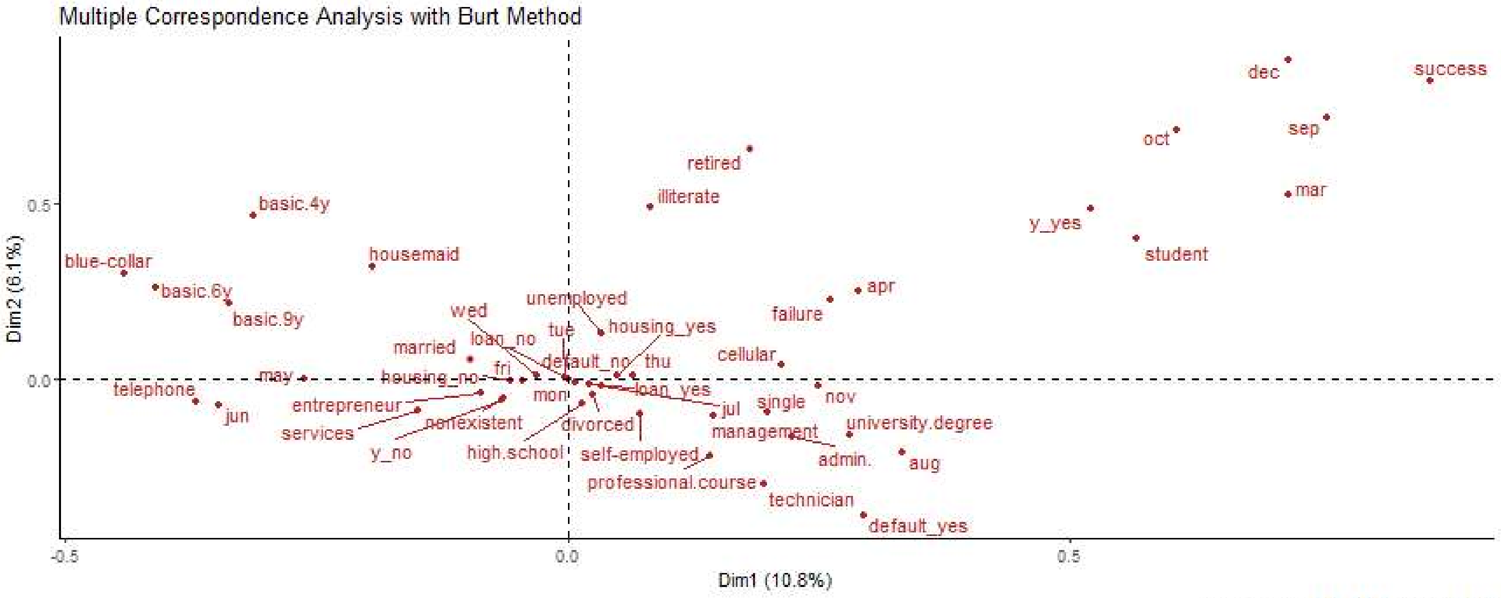
Random forest (82% accuracy without hyperparameter tuning)

- Future use cases:

1. Customer portraits supporting target marketing
2. Predictive information product for campaign operations

Some months and students linked with more subscription

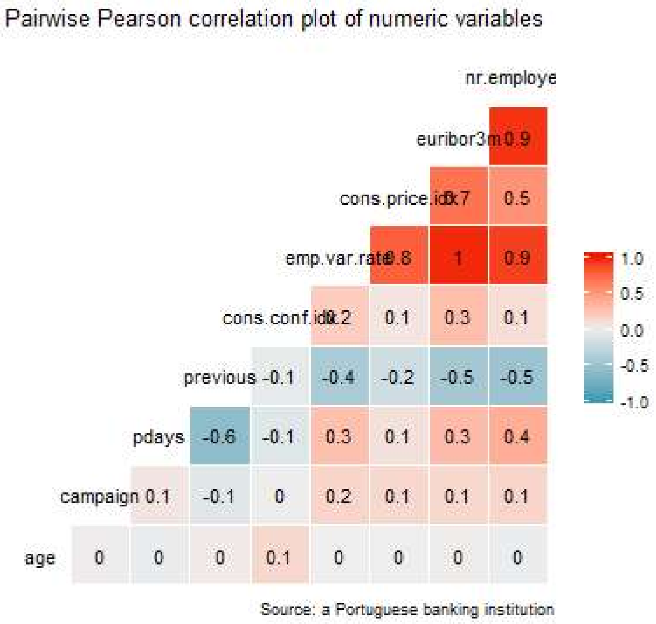
- Totally missing rows and duplicate rows were removed first
- A MCA with all categorical variables


Multiple Correspondence Analysis with Burt Method

Source: a Portuguese banking institution

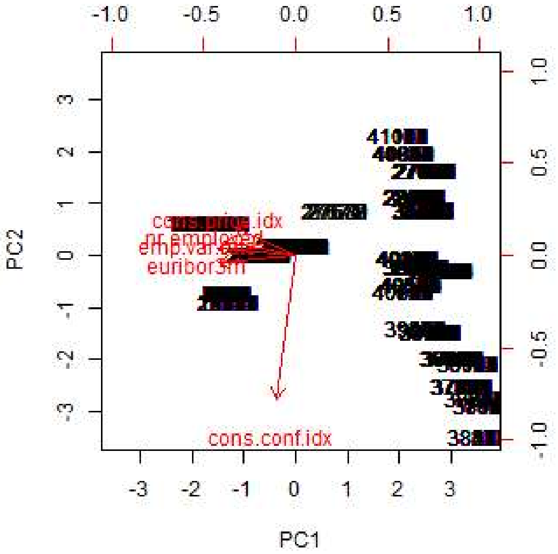
- Overall, 4639 out of 41179 (11%) subscribed
- On top right, more calls at March (51%), Sep (44%), Oct (44%), Dec (49%) led to subscription
- 31% student subscribed, among the top in job categories

Correlated macroeconomic indicators



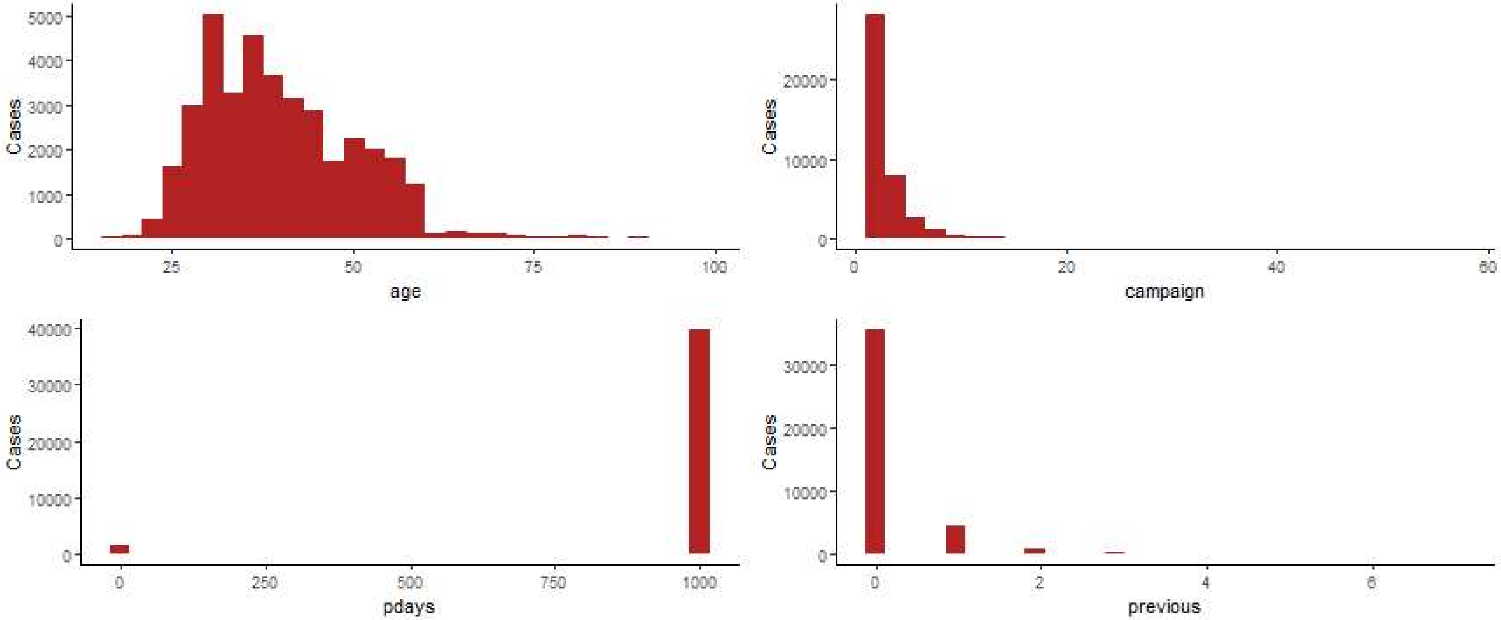
- Macroecononmics indicators are highly correlated, PCA can be a solution

Correlated macroeconomic indicators: PCA as a solution



- Consumer confidence index relatively independent in PCA

Need to normalize and drop pdays



- pdays contains little information, dropped in subsequent analysis
- Other variables right skewed, will be normalized later on

Transformations and missing value imputation

- One hot encoding all categorical variables
- Transform binary variables into dummies
- Impute missing ages and CPI with average values (~5400 values replaced)

Train/test split and preprocessing

- 0.8 train/test random split, training set contains 32,943 obs
- After split, centering and scaling both samples to keep the training set "independent"
- Further under-sampling training set to have a half-half balance between "no" and "yes" in y
- Final training set contains 7,424 obs

Model 1: Logistic regressions

- Baseline, generalist model

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6180  361
##          1 1129  566
##
##                Accuracy : 0.8191
##                  95% CI : (0.8106, 0.8273)
##     No Information Rate : 0.8874
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.335
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8455
##             Specificity : 0.6106
##          Pos Pred Value : 0.9448
##          Neg Pred Value : 0.3339
##              Prevalence : 0.8874
```

Model 2: Support vector machine

- Less overfitting, moderate dimensionality

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6285  389
##          1 1024  538
##
##                Accuracy : 0.8284
##                  95% CI : (0.8201, 0.8365)
##     No Information Rate : 0.8874
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3389
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8599
##             Specificity : 0.5804
##          Pos Pred Value : 0.9417
##          Neg Pred Value : 0.3444
##              Prevalence : 0.8874
```

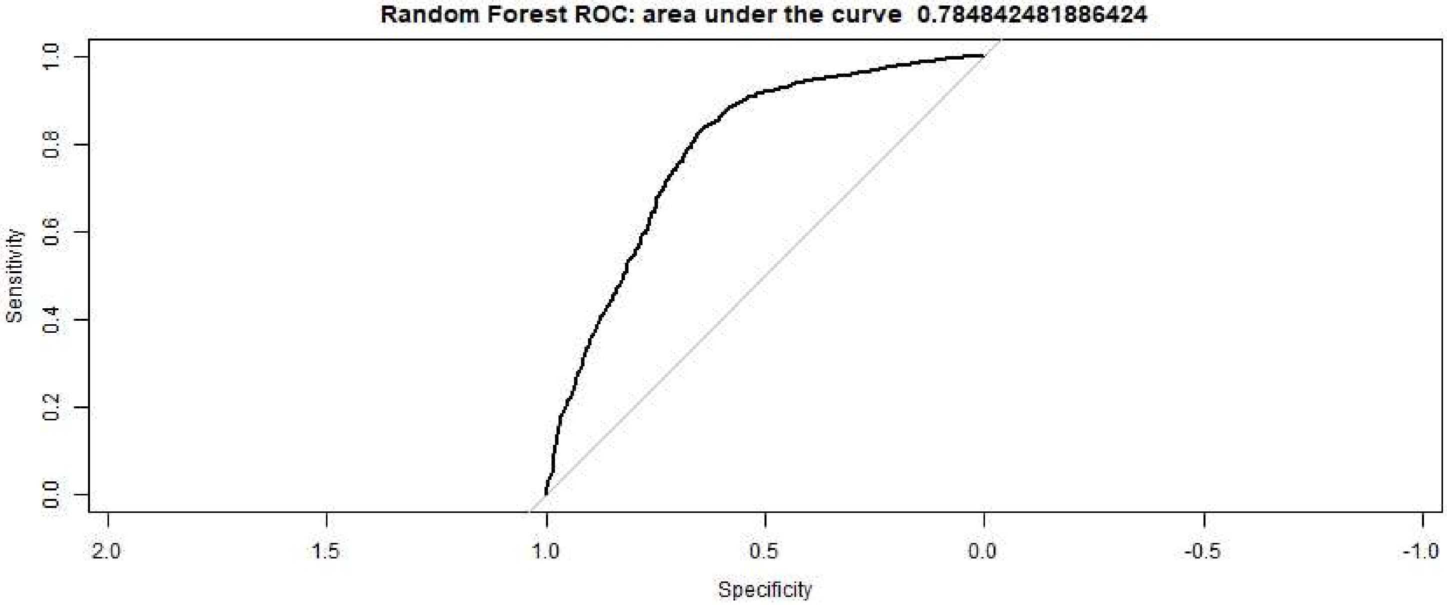
Model 3: Random forest

- Good for data mixed with categories and numerical, less overfitting

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6228  361
##          1 1081  566
##
##                Accuracy : 0.8249
##                  95% CI : (0.8165, 0.8331)
##     No Information Rate : 0.8874
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3455
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8521
##             Specificity : 0.6106
##          Pos Pred Value : 0.9452
##          Neg Pred Value : 0.3437
##              Prevalence : 0.8874
```

Model choice: random forest

- Comparative accuracy (82%) compared with logit (81%) and SVM (83%) even without tuning, good ROC
- Fewer false negatives: help to minimize huge loss cases where potential subscribers are not identified
- Good interpretability with the ability to return variable importance and visualize sample trees



- Next step: hyperparameter tuning and possibly compare with other tree algorithms (such as GBM)

Future use cases

- For marketers: customer portraits supporting target marketing effort

Knowing who to sell to helps business evolve from mass marketing to target marketing. A report on customer acquisision strategy regarding which kind of customers will have better likelihood to subscribe can be proposed. Customer persona can be generated from some models (such as the coeffecients of the logistics regression, or some sample tree structures of random forest). Delivered to marketers, it would support practices that approach target customer group better. For example, for a TV advertisement, it might be designed and personalized for students since they have a higher likelihood to subscribe among others.

- For sales: predictive information product during campaign event

In direct marketing campaigns, call agents are in charge of outbound sales calls, who usually have to scroll through random and endless customer profiles. An end-user product (such as a dashboard or report connecting to the CRM system) providing a metric of "likelihood to subscribe" based on the CRM information would help them to prioritize the contact task (i.e. contact highly likely customers first), make the work more fun for them, and enhance the convertion rate of the entire campaign.

In addition, the impact of campaign-specific information (date of campaign, outcome of previous contacts) in the model can be extracted to a report to help sales manager plan the next event better.