June 2nd, 2019

R Group Project – Tasks, Deliverables and Plan of Action

Classification – whether client will subscribe to term deposit or not.

Event / encoding – 1 if Yes, 0 if No

Deliverables

1. Printable Report (Federico, Anwita, Scott + Hatem)

Analysis and explanation

1. shinydashboard (Gerald, Louis, Alberto, Hatem + Anwita)

User end / client interface

1. CSV with predictions

Evaluation criteria – Sensitivity / Recall

Structure

1. Report (RMarkdown HTML)
   1. Problem statement / executive summary
   2. Data import, dataset quality checks – variable types, missing values, etc.
   3. Exploratory Data Analysis – Correlations, Skewness, Bivariate and multivariate tables

Ideas for visualisations and other EDA:

* + 1. Histograms for numerics and bar plots for categorical (Already done, can they be interactive? ***DONE – to be added to Shiny app***
    2. Correlation matrix ***DONE – to be added to Shiny app***
    3. One bar plot with the target variable distribution (yes and no) which can be succeeded by explanation for doing under / over sampling. ***DONE – to be added to Shiny app***
    4. Heatmap – avg age by job and target variable (x axis yes/no, y axis jobs and colour variable avg age) ***DONE – to be added to Shiny app***
    5. Scatter plot – x axis previous, y axis pdays, colour variable would be poutcome. Can be converted to bubble chart if previous is taken as size parameter. ***DONE – to be added to Shiny app (or not ?)***

***Note: The interactive graphs are commented out of the RMD because it makes it too heavy and long, will only use interactive in Shiny.***

*You can add more (especially cross tables) if you can think of some multivariate relationships relevant to the problem. I will also keep updating this.*

* 1. Recode target variable – yes to 1 and no to 0. Convert to factor.
  2. Baseline modeling

No preprocessing. Binary logistic. caret. Train dataset split into 10-folds for cross val.

Plot variable importance (optional, could be replaced by a table with variables, coefficients and p-values) and confusion matrix.

Baseline notes:

Age as continuous not significant

Job retired, services, unemployed and unknown not significant (with admin as base level)

Marital status single not significant (with divorced as base level)

Default yes not significant

Pdays not significant

* 1. Iterations:

1. Rectify the sampling problem.

Approach: Undersampling and oversampling. (Already done)

Baseline with resampling.

1. Simple stepwise to see if baseline improves?

1. Next steps will be on resampled data.

Feature selection, feature creation

**Demographics**, **Transactional**, **Campaign-specific**

1. **Age** is a continuous variable. Not skewed but clear splits of age group. Can be binned. (Already done)
2. **Job** – 12 levels. Can be collapsed into fewer levels (Already done)
3. **Marital** **status** – Fairly distributed, does not need any change
4. **Education** – unknown has very few observations and so does primary. If levels are not significant in the baseline model then they can be clubbed together.
5. **Default** – Highly disproportionate, but should be kept unchanged.
6. **Balance** – Highly skewed with negative values and majority 0. 3052 negative values and 2799 0s). Can this be converted to categorical in order to account for the negative values? @Federico
7. **Housing** – No transformation needed
8. **Loan** – Slightly disproportionate, but should be kept unchanged
9. **Contact** – Majority of obs belong to cellular, but should be kept unchanged
10. **Day** – Fairly well distributed. But needs to be treated as a factor and not numeric. (@Federico thoughts?). If factor we can think of binning it as first week / first 10 days of the month, etc.
11. **Month** – Either unchanged or convert to quarters.
12. **Duration** – Highly skewed. *Has some 0. It has obvious correlation with response and should be avoided in the model?* (0 duration = no term deposit)
13. **Campaign** – Highly skewed
14. **Pdays** – Most obs are 0.
15. **Previous** – Highly skewed with most obs as 0.
16. **Poutcome** – Most obs are unknown but should be kept unchanged.

@Federico – could there be a possibility of new variable by combining poutcomes, previous and pdays?

1. Advanced modeling (After variables are finalized or for every iteration? @Federico)

Options – Random Forest, SVM, Gradient Boosting

1. PCA for variable reduction???? @Federico
   1. Best model and interpretation of results
   2. Testing model performance on test dataset
   3. Concluding remarks
2. shinydashboard

Three main tabs:

* 1. Data at a glance
  2. Hyperparameter tuning
  3. Plugging in new data and generating probabilities