

Santander Customer Satisfaction Prediction

Frederic Marechal

Project Aim

- Identify dissatisfied customers early; using data from a Santander's anonymised dataset.

Why?

- To **reduce the churn rate** and **avoid reduction in P&L**, and potentially look for opportunity to **improve the upselling rate**.
- To **mitigate the reputational risk** - Research shows that dissatisfied customers tell more people about their experience than a happy customers (see <https://www.business.qld.gov.au/running-business/consumer-laws/customer-service/complaints/behaviour>).
- To **avoid further aggravation** – refunds, disputes, etc...

Problem Type

Supervised classification task.

Technologies

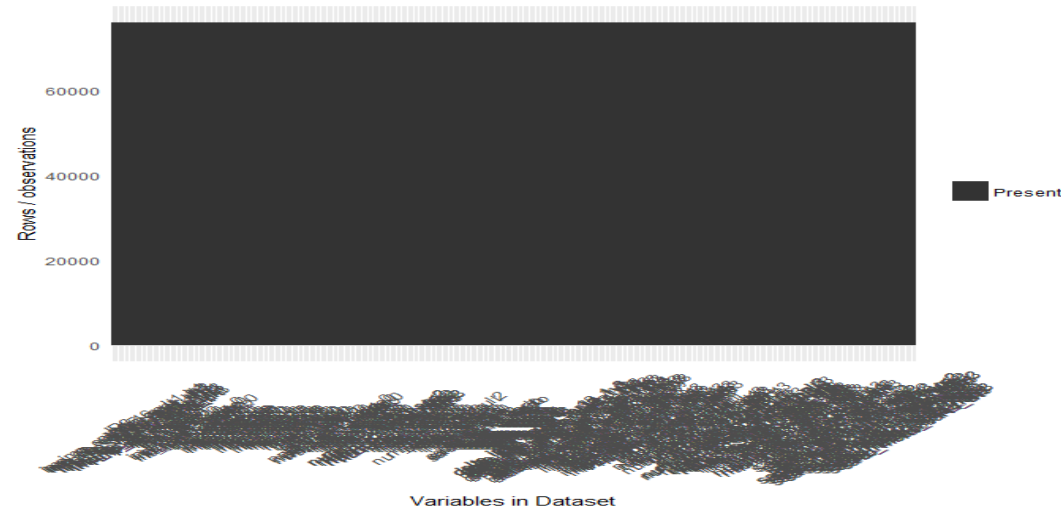
R Studio + Caret libraries and ML libraries

```
require(ggplot2) | require(tcltk) | require(caret) | require(randomForest)  
require(dplyr) | require(fBasics) | require(CORElearn) | require(e1071)  
require(reshape2) | require(caTools) | require(AppliedPredictiveModeling) | etc...
```

Data Analysis & Visualisation

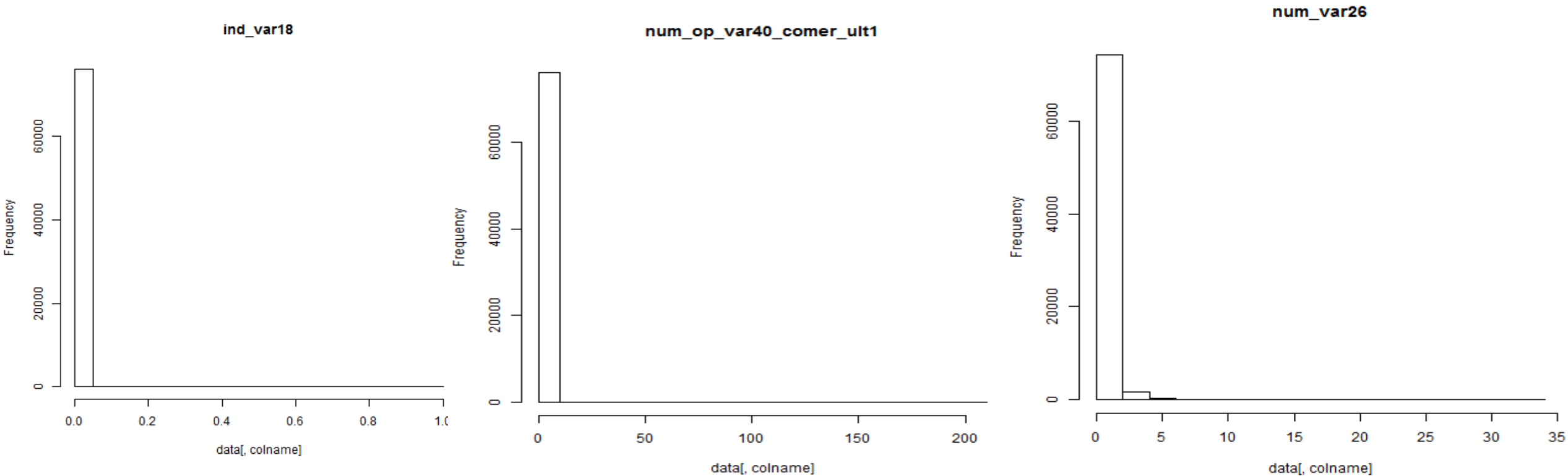
A. Initial exploration

- Large number of variables -> 369 of interest (ID column removed from analysis...)
- 76,020 rows for the training set and 75,818 for test set
- All columns data types are numeric either discrete (integer) or continuous (double)
- Reasonably clean data (no missing data)



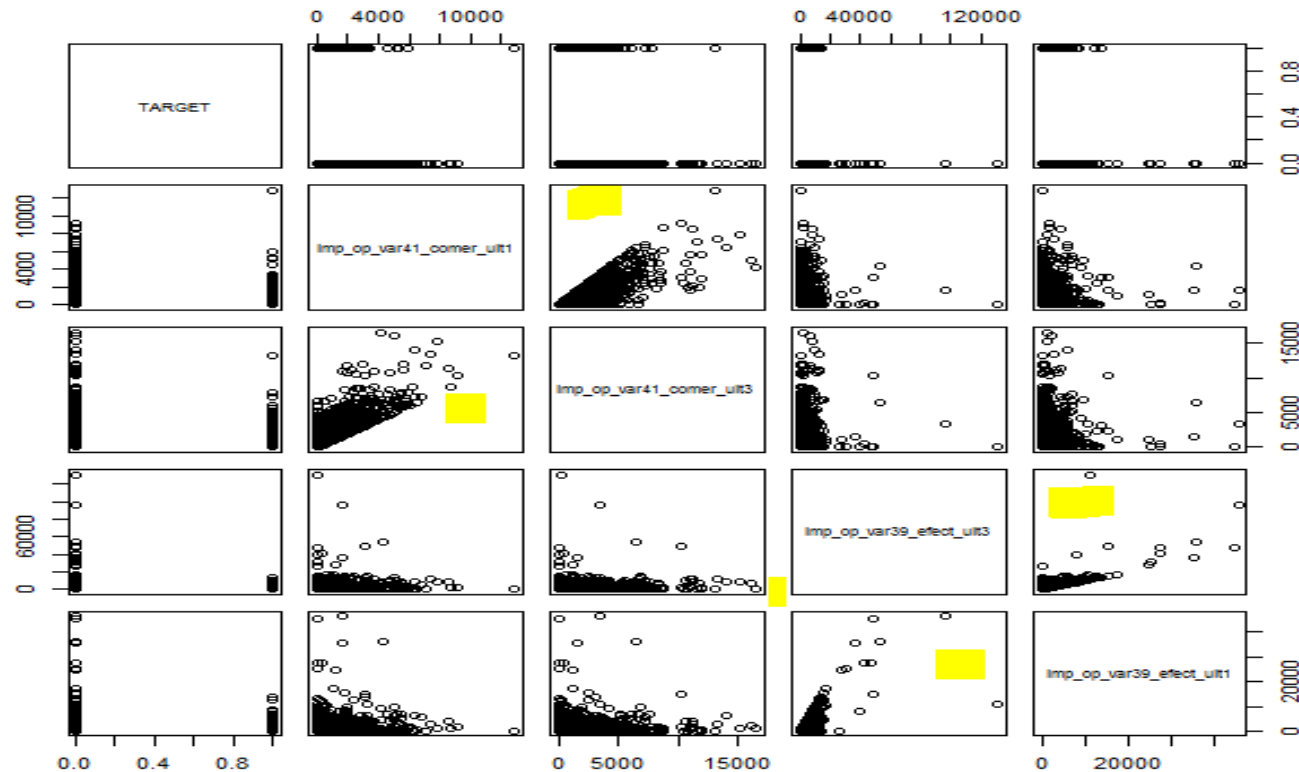
Data Analysis & Visualisation

- Presence of zero or near zero variance columns (just a few examples)



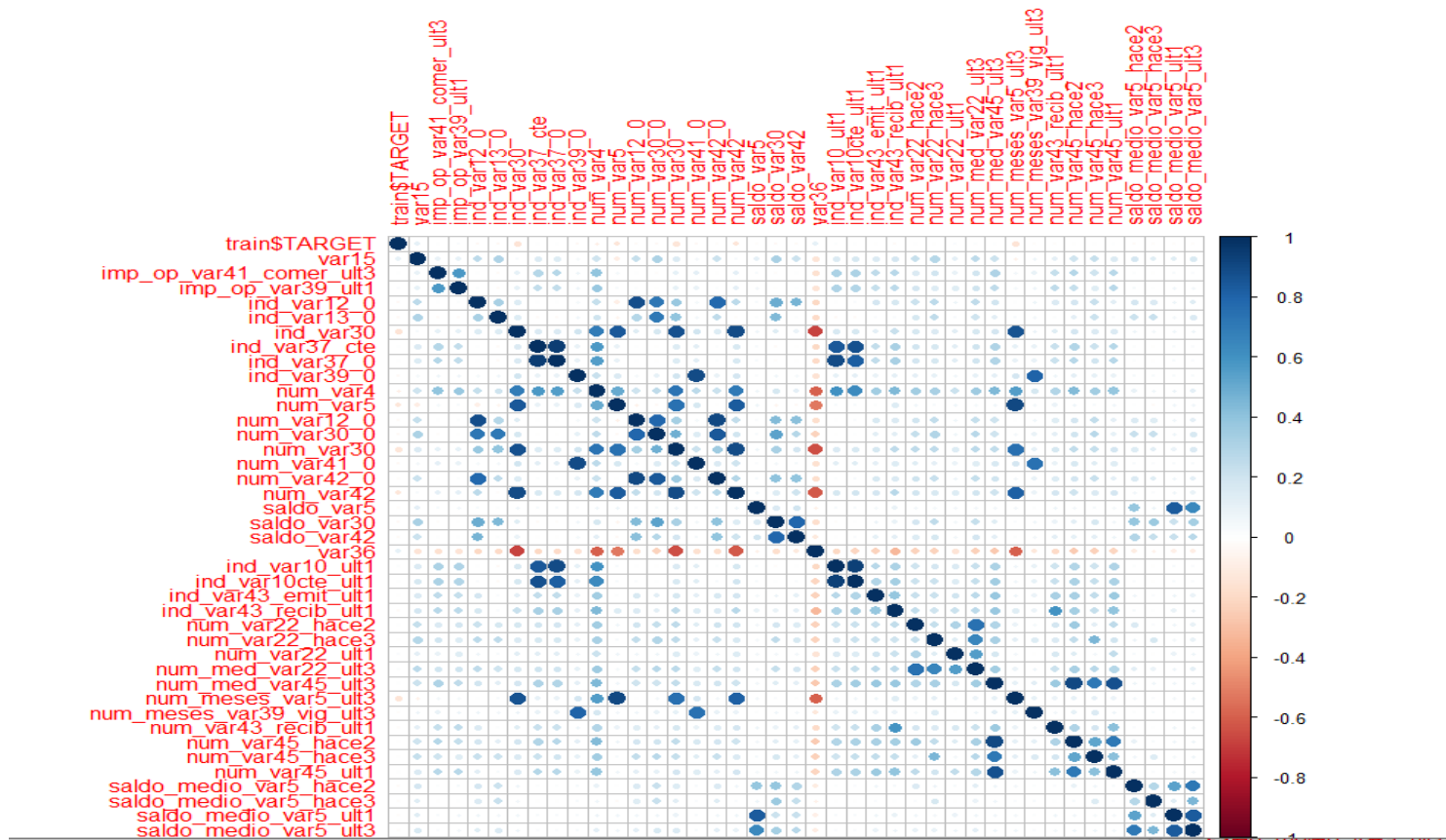
Data Analysis & Visualisation

- No presence of correlation between the explanatory and response variable, the correlation range is $[-0.1498, 0.1029]$
- Presence of linear dependencies within explanatory variables (just a few examples).



Data Analysis & Visualisation

- Potential presence of correlation between predictors (just a few examples).



No clear cut -> but Useful exercise in “getting to know the data”. ☺

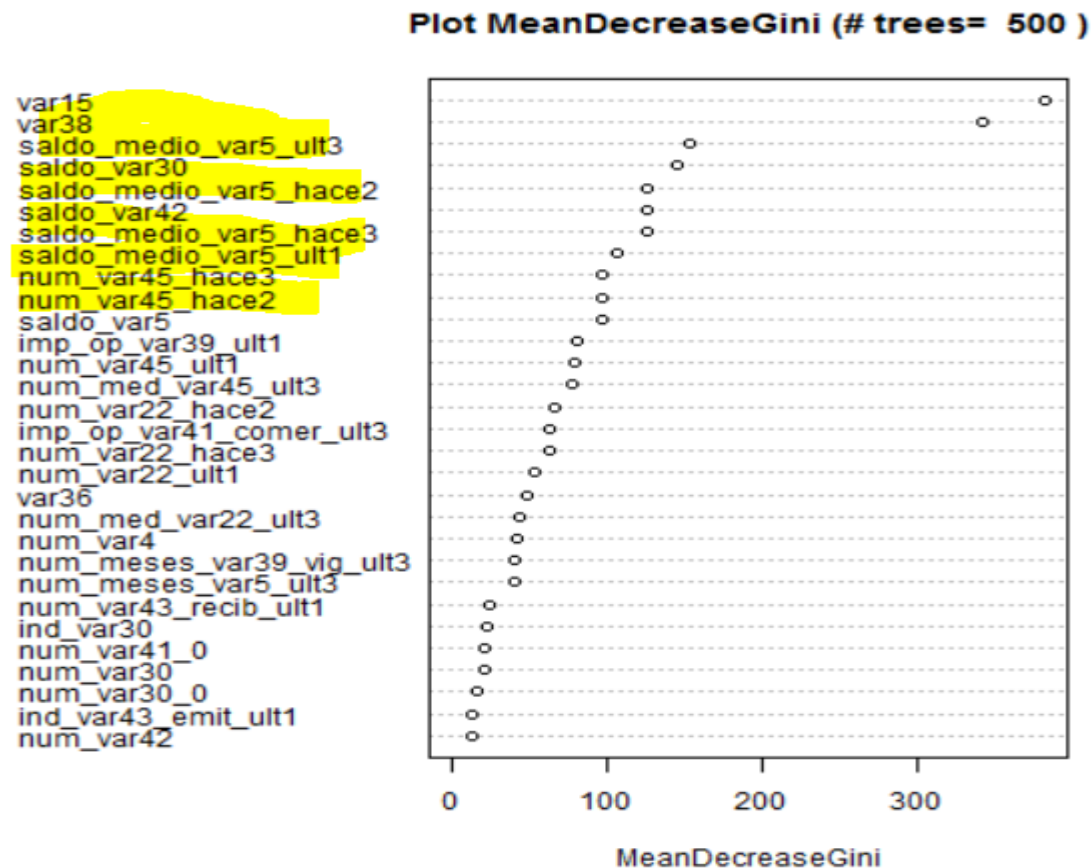
Data Pre-processing

- **Missing Data**
 - No action required
- **Class Imbalance**
 - 96% of 0s and 4% of 1s in the response variable (TARGET)
 - Use of SMOTE to rebalance... Experiment shows:
 - an over sampling of 100% of the minority class, coupled with
 - an under sampling of 200% of the majority class, with
 - Nearest neighbours set to 5On training data.. 0s = 53% and 1s = 47%
- **Near Zero Variance Columns**
 - Do not add explanatory power to the model
 - 310 columns removed
- **Linear Dependencies**
 - Use the *findLinearCombos()* R function to find and remove linear dependencies recursively
 - 3 columns deleted
- **Correlated predictors** (high multicollinearity increases variance/make prediction sensitive to minor change in model)
 - Deletion of correlated predictors (> 90% of correlation)
 - 9 columns deleted

From 369 to 41 columns....

Feature Selection

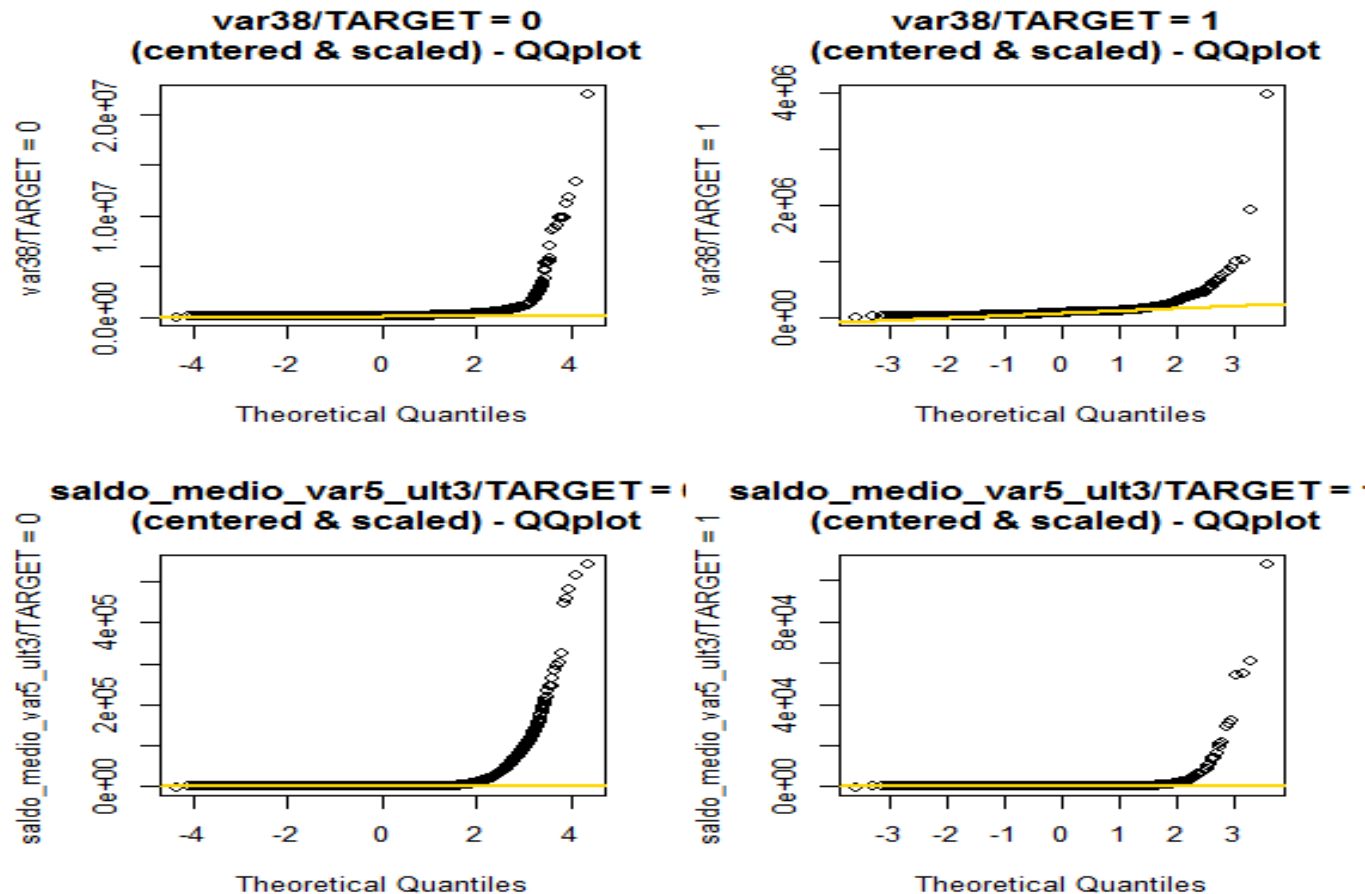
- Wrapper method: Variable Importance + Backward feature selection
- Aim: evaluate the performance of the feature selection for each a set of machine learning algorithm.
Plus point: Predictors are evaluated together, instead of individually
Minus point: Impact on computer resources + risk of over-fitting



+ Backward selection (only for LDA, not for SVM or RF/Bagging)

Data Shape

- Kolmogorov–Smirnov (*ks*): establish the presence of a divergence from the normal distribution.
- The Kolmogorov–Smirnov test *Null* hypothesis (H_0) indicates the data distribution seems to follow a normal distribution. H_0 is rejected when the Kolmogorov Smirnov test returns a p-value less than 0.05.
- Here all the feature selected columns do not seem to follow a normal distribution



Model Selection and Assumptions

These models were selected because they all belong to the supervised classification model family + compatible with numerical data.




- LDA

- Assumption: Explanatory variables to follow a normal distribution (H_0) + No scaling required.
- Advantages: no parameter tuning => lesser computer power required and fast time to completion.
- Drawbacks: potential to overfit the data (no parameter tuning)

- RF/Bagging/Boosting (ensemble methods)

- Assumption: No specific assumption relating to the data distribution
- Advantages: Overfitting reduction + graphical representation of trees provide an intuitive way of interpreting results
- Drawbacks: require more compute power and are have 'slower' time to completion (the higher the number of tree to slower)

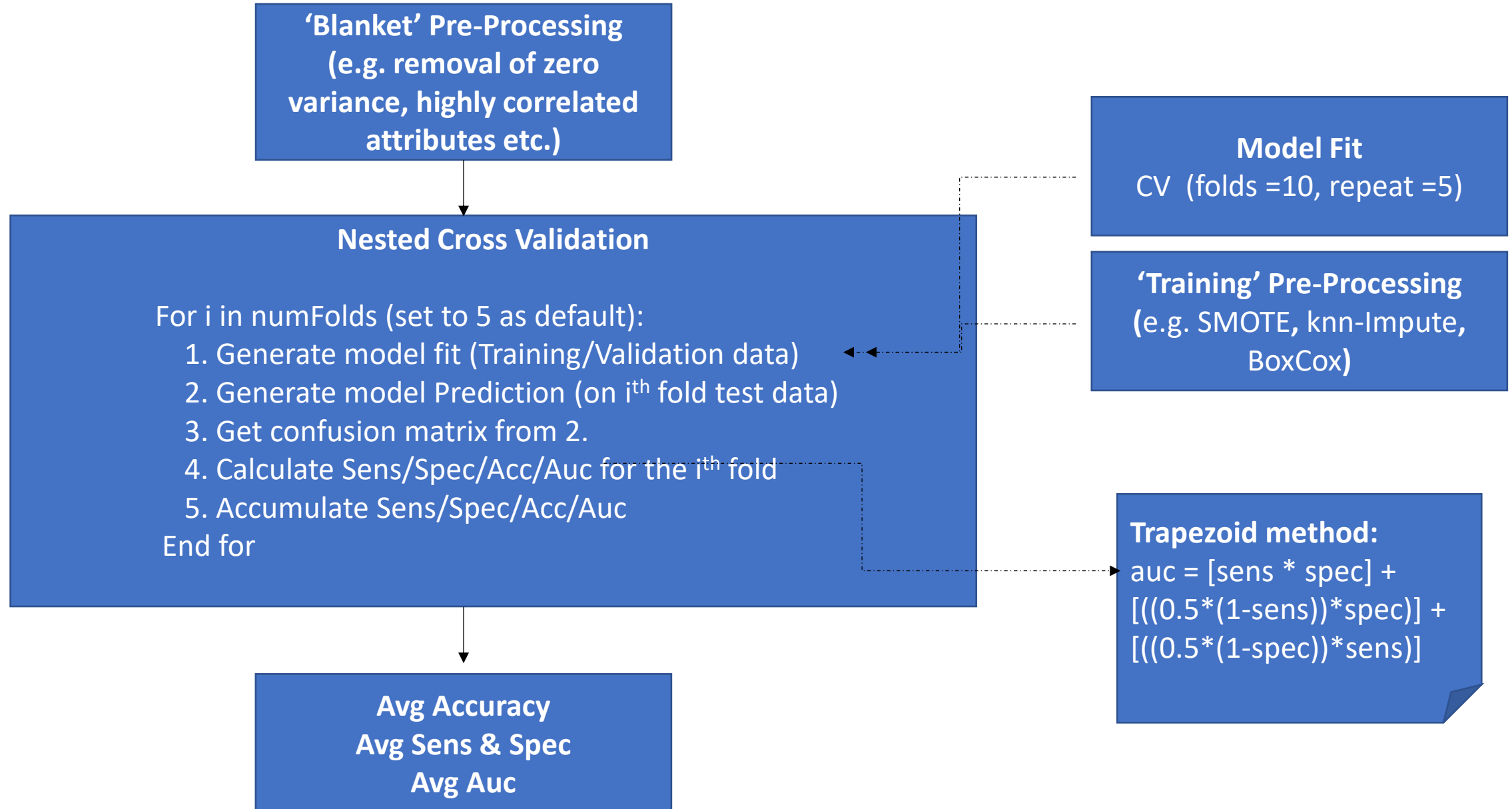
- SVM

- 
- Same assumptions/advantages and drawbacks than ensemble method. The higher the cost (linear kernel) & gamma (non linear kernel), the slower time to completion and higher memory foot print.
 - Furthermore it requires explanatory variable scaling.

Model Parameters

Model Name	Hyper-parameters
LDA	None
Random Forest	Number of trees = 500, number of splits = p (Bagging) & \sqrt{p} and $p/2$ for RF
Boosting	Number of trees = 500 and shrinkage = 0.01
SVM (Linear)	Cost = 0.0001, 0.0005, 0.001, 0.1, 1, 5, 10, 100

Methodology



Performance Measures

The Confusion Matrix

Total Population	Prediction 1	Prediction 0
Expected 1	True Positive (TP)	False Negative (FN)
Expected 0	False Positive (FP)	True Negative (TN)

Accuracy Rate

- Measure of the statistical bias. Accuracy is between 0 (maximum bias) and 1 (no bias).
- This provides a measure of the overall quality of the prediction.
- $\text{Accuracy} = (TP+TN)/(TP+FN+FP+TN)$

Sensitivity

- Measures the proportion of positives that are correctly identified (i.e. how good is a test at detecting the positives).
- $\text{Sensitivity} = TP/(TP+FN)$

Specificity

- Measures the proportion of negatives that are correctly identified.
- It is calculated as $TN/(TN+FP)$

AUC

- The Area Under the Curve (AUC) of a Receiver Operator Characteristic (ROC) is defined as the sensitivity plotted against (1-Specificity).
- The balance between the specificity and sensitivity is an indication of model performance.
- The AUC ranges between 0 and 1 (perfect classification). It measures the area under the ROC.

Results & Conclusion



Model Name	AUC	Sensitivity	Specificity	Accuracy	Time to Completion
Boosting	82.8%	67.7%	82.8%	82.2%	1h
Random Forest	80.9%	67.1%	81.5%	80.9%	1h
LDA *	66.0%	57.9%	74.1%	73.3%	15mins
LDA ** (CoxBox)	50%	0%	100%	95.6%	15mins
SMV*** (no class rebalancing)	37.5%	50%	50%	41.7%	1day

**In this experiment, Boosting is the preferred model.
It has the highest AUC, with the highest levels of sensitivity and specificity.**

* backward selection best feature selection: factor(TARGET) ~ var15+var38+saldo_medio_var5_ult3+saldo_var30+saldo_medio_var5_hace2+saldo_var42+saldo_medio_var5_hace3+saldo_medio_var5_ult1+num_var45_hace3"

** backward selection best feature selection: factor(TARGET) ~ var15

*** There was an issue with using SMOTE with SVM. The results have been published but do not represent the true prediction of the algorithm with class rebalancing.

Recommendation & Improvements

ML improvements

- Increase the hyper-parameter space, e.g. RF could use 1000, 2000, etc. trees instead of the 500 in this experiment.
- Increase the supervised classification ML offering: e.g. Naive Bayes, Regression, Lasso, Ridge, MLP NN, SVM with polynomial/radial kernel, QDA, etc.
- Fix the issue relating to SMOTE (i.e. class rebalancing) and the SVM model.
- Use another feature selection mechanism (e.g. Filter based) and compare against the results obtained with Wrapper feature selection used in this experiment.

Code improvements

- The code is quite manual when it comes to perform backward selection (need to make it more flexible)
- The code should be written so that it follows a more OO architecture (increase reuse) + could be design as services (Loader Service, Feature Selection Service, ML Service, etc.)
- R parallel class should be investigated to improve performance.
- Apply save points (e.g. after blanket pre-processing), to facilitate recovery and improve performance when re-run is required.

Q&A



Extras

AUC:

Method provided by <http://www.kdd.org/kdd-cup/view/kdd-cup-2009/Tasks>

We define the sensitivity (also called true positive rate or hit rate) and the specificity (true negative rate) as:

Sensitivity = recall = hit rate = true pos rate = tp/pos

Specificity = true neg rate = tn/neg

1- Specificity = False pos rate = false alarm

where $pos = tp + fn$ is the total number of positive examples and $neg = tn + fp$ the total number of negative examples

The results will be evaluated with the so-called Area Under Curve (AUC). It corresponds to the area under the curve obtained by plotting sensitivity against specificity by varying a threshold on the prediction values to determine the classification result. The AUC is related to the area under the lift curve and the Gini index used in marketing ($Gini = 2 \cdot AUC - 1$). The AUC is calculated using the trapezoid method. In the case when binary scores are supplied for the classification instead of discriminant values, the curve is given by $\{(0,1), (tn/(tn+fp), tp/(tp+fn)), (1,0)\}$ and the AUC is just the Balanced Accuracy BAC.

ROC:

https://en.wikipedia.org/wiki/Receiver_operating_characteristic