Team MVA-2016

Data-Science Game 2016

Antoine de Maleprade
Sebastien Fischmann
Mastane Achab
Alexandre Garcia

Feature engineering

Best Random-Forest importance scores

0. feature : rowNumberForUser -- 0.474538396883

1. feature : NameOfPolicyProduct_NS -- 0.0328159814026

2. feature : PolicyHolderNoClaimDiscountYears -- 0.0307465654476

3. feature : DaysSinceCarPurchase -- 0.0279611458202

4. feature : CoverIsNoClaimDiscountSelected -- 0.0278649601994

5. feature : FirstDriverAge -- 0.0254410066302

6. feature : lda_feature_3 -- 0.0229693189225

7. feature : lda_feature_2 -- 0.0216706319752

8. feature : lda_feature_1 -- 0.0210556092396

9. feature : lda_feature_4 -- 0.0198186350213

10. feature : VoluntaryExcess -- 0.0177385677466

⟸ Number of times a user appears in the dataset

Other features :
- Number of unique values for different columns per user
- Binary features indicating whether the user Id has entered his most frequent car and most frequent driver.
- Gender Feature with DriverLicenceIdNumber

⟸ LDA Features : we consider the categorical values in each row as « words »

lda_feature_i encodes the projection of the row values on the i^{th} topic learnt

Regression step

- Xgboost only
  - 1st submission : 5 models trained on 5 folds with separated users and the predicted probabilities are then averaged.
  - 2nd submission : 4 most independent model chosen among 11 with the Pearson Correlation coefficient.

- What we tried (but didn't use):
  - Undersampling the 0 class -> better take the unbalanced repartition into account in the gradient-boosting parameters
  - Random Forest -> Too slow to train for this week-end