Comparative classification of online shoppers’ purchasing intention - Project status presentation

Gaëlle Jakubowski, Guillaume Billés, Christian Nestroy, Rodolfo Rubino

Aim of the project

❑ Predict whether the online shopping session is concluded with a transaction or not (binary classification)

❑ Inference about importance of analytical and context variables for online marketing purposes

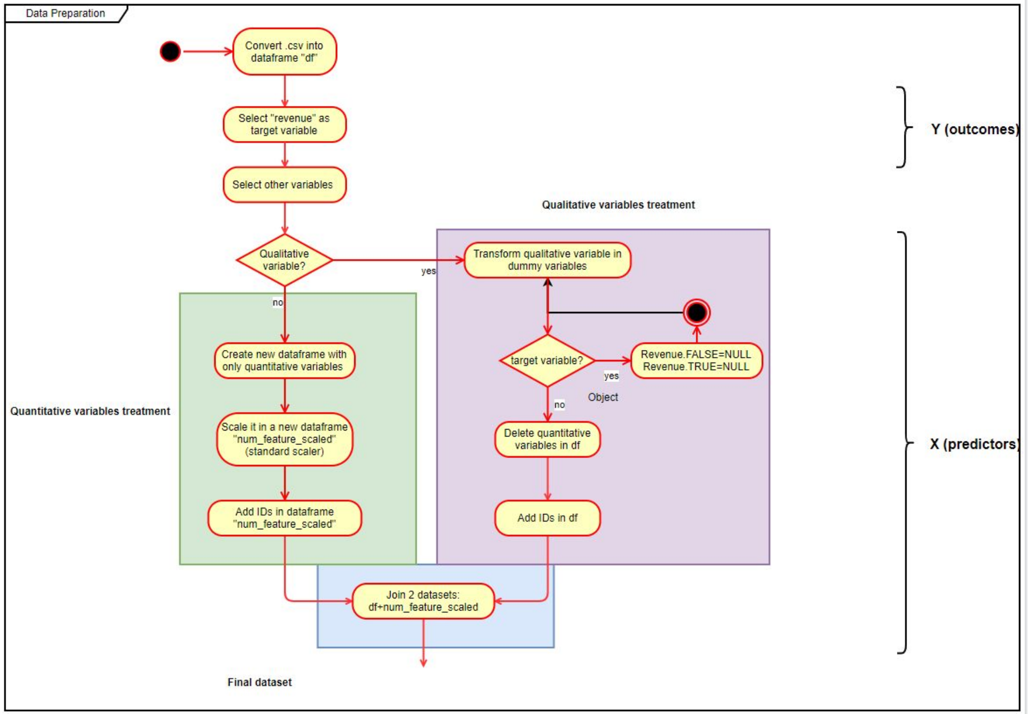
Dataset description

- “online shoppers’ intention”
- 12330 observations (online shopping sessions)
    - 10422 negatives, 1908 positives
- 18 variables (10 numerical, 7 categorical, 1 categorical target)
- source: UCI Machine Learning Repository
- format: structured, CSV

Data preprocessing

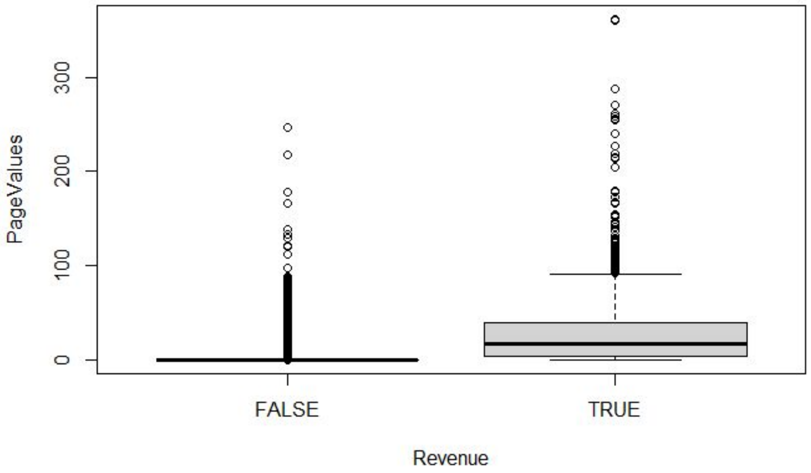
❑ one-hot encoding of categorical variables

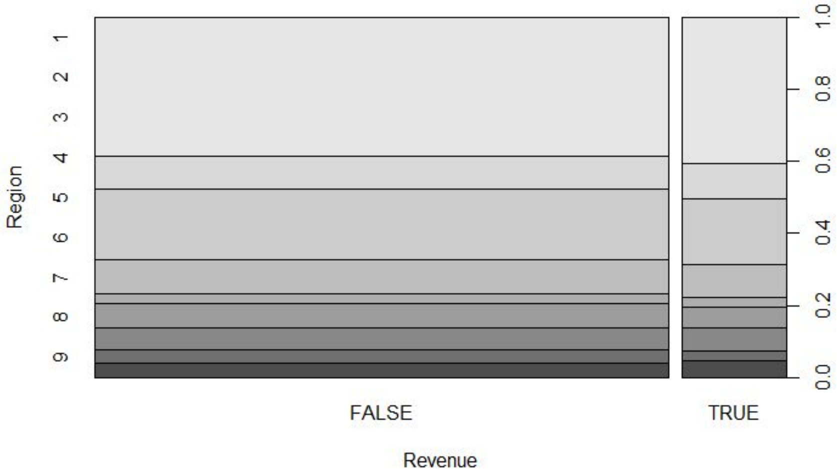
❑ scaling of numerical variables to standard normal distribution


Data Preparation
Convert .csv into dataframe "df"
Select "revenue" as target variable
Select other variables
Qualitative variable?
yes
no
Qualitative variables treatment
Transform qualitative variable in dummy variables
target variable?
yes
Revenue.FALSE=NULL
Revenue.TRUE=NULL
Object
no
Delete quantitative variables in df
Add IDs in df
Quantitative variables treatment
Create new dataframe with only quantitative variables
Scale it in a new dataframe "num_feature_scaled" (standard scaler)
Add IDs in dataframe "num_feature_scaled"
Join 2 datasets: df+num_feature_scaled
Final dataset
Y (outcomes)
X (predictors)

Inference about variable importance

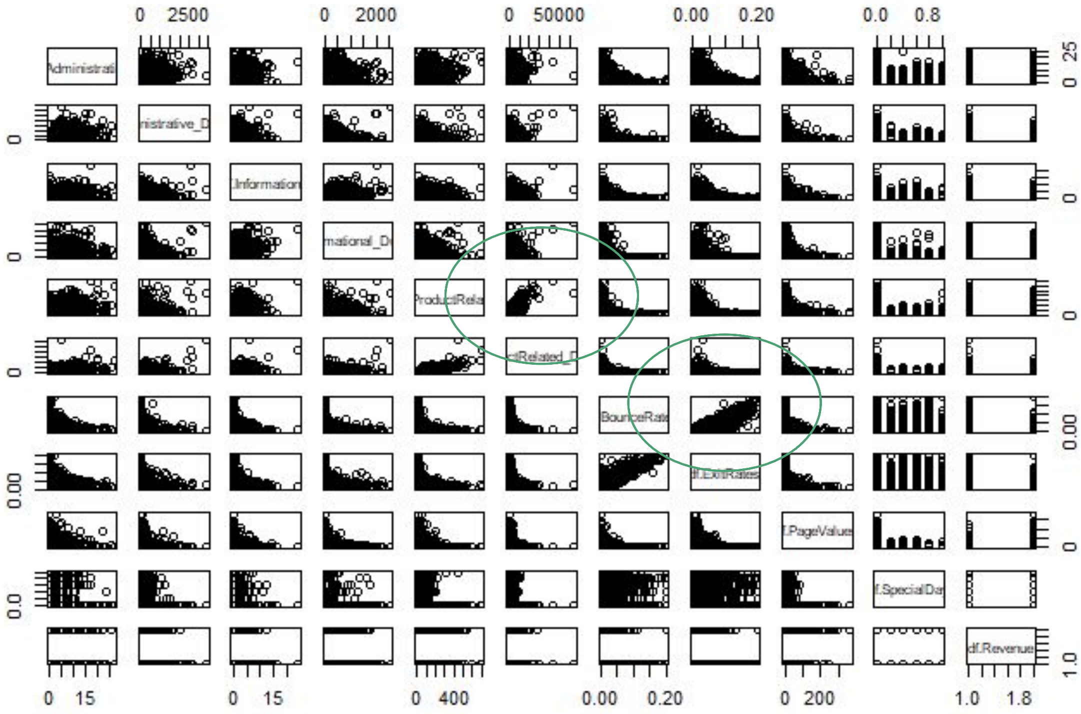
- influence on categorical target investigated by chi-square test for categorical variables and by ANOVA for numerical variables
- most important variables: PageValues and ExitRates (Google Analytics metrics)





Inference about variable importance

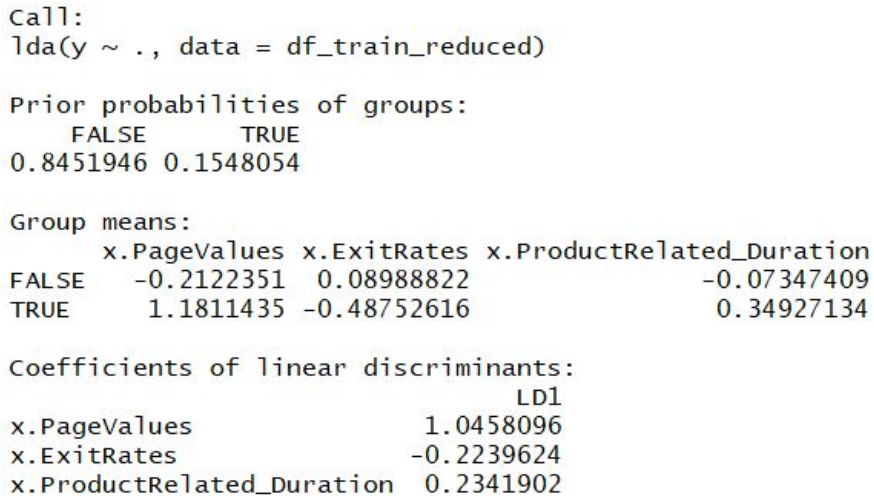
- ExitRates/BounceRates + ProductRelated/ ProductRelated_Duration strongly correlated
- only use one variable of these redundant pairs



Prediction results

- ❑ Use of accuracy because class variable not too imbalanced
- ❑ Simpler methods can compete with more sophisticated ones

ML method	Accuracy
Random Forest	0.8828
Support Vector Machine	0.8946
Logistic Regression	0.8799
Logistic Regression with LASSO	0.8881
k-Nearest Neighbor	0.8966
Linear Discriminant Analysis	0.8751



```
Call:
lda(y ~ ., data = df_train_reduced)

Prior probabilities of groups:
    FALSE      TRUE 
0.8451946 0.1548054 

Group means:
      x.PageValues x.ExitRates x.ProductRelated_Duration
FALSE   -0.2122351  0.08988822               -0.07347409
TRUE     1.1811435 -0.48752616                0.34927134

Coefficients of linear discriminants:
                                 LD1
x.PageValues               1.0458096
x.ExitRates               -0.2239624
x.ProductRelated_Duration  0.2341902
```

Outlook

❑ Further hyperparameter tuning

❑ Feature selection through wrapper methods (stepwise forward selection) and embedded methods (regularization, LASSO)

❑ Trying out (nested) cross-validation