Data Mining and Text Mining - Course Project

Pythons on a Plane

Maddalena Andreoli, Riccardo Pressiani, Andrea Battistello,
Pietro di Marco, Carmen Barletta
<name.surname>@mail.polimi.it

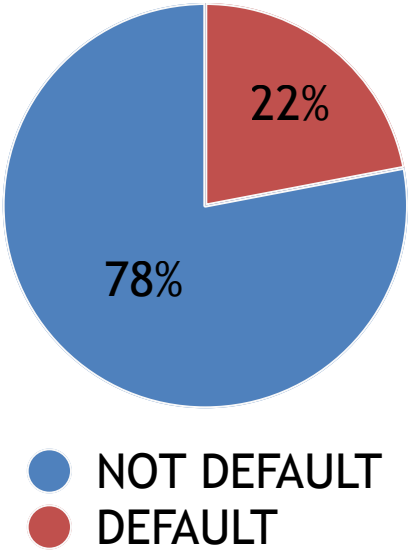
9/6/2017, DEIB Conference Room

Task

Devise a data mining model able to predict the risk of default for credit card users

Three phases:

- Data exploration
- Feature aggregation and selection
- Model selection and validation



According to data provided in the train set **only 22% of people were in default** in January, putting our problem in the **anomaly detection** category.

Other remarks:

- The higher **credit limit**, the lower the probability of default.
- **Females** have slightly lower probability of default than males.
- **Single people** have higher probability of not being insolvent than married ones
- The less a customer **pays back timely** their due debt, the higher the chance he/she will be insolvent

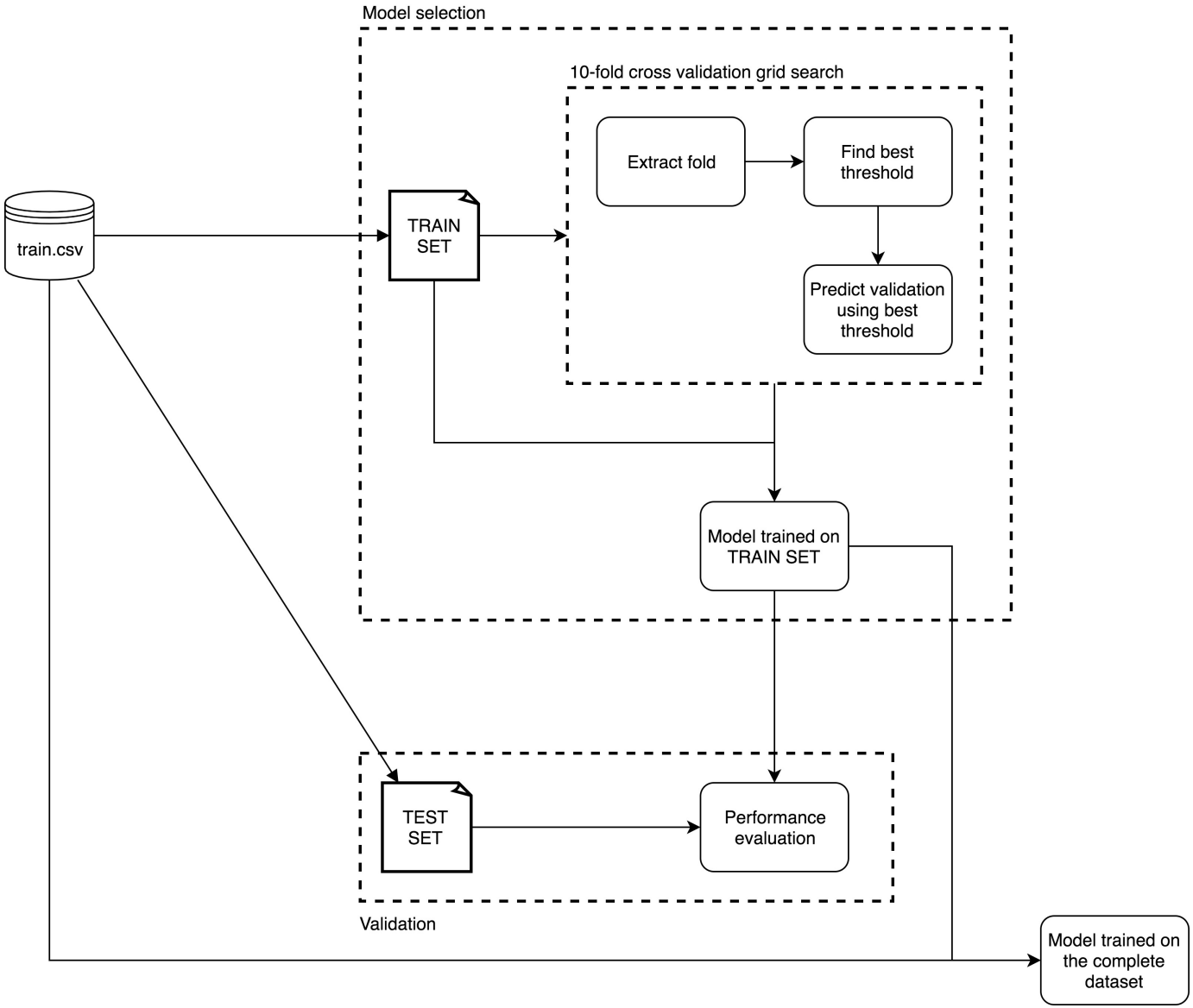
We added a **new set of aggregated feature** that we considered to be relevant:

- LIMIT - MEAN_BILL
- TOTAL_PAY_AMOUNT
- <x>_TREND
- <x>_SKEW
- <x>_KURT
- ARIMA process coefficients

We used the ***recursive features elimination algorithm cross validation*** (RFECV from sklearn library) with five folds and XGBoost as the estimator parameter.

We discovered that there were 51 most relevant features.

Model Selection and Validation



We have obtained the best performance with respect to the F1-measure with **eXtreme Gradient Boosting**

Crossvalidation score:	0.547
Holdout set score:	0.538

Thanks for the attention!

Maddalena Andreoli, Riccardo Pressiani, Andrea Battistiello,
Pietro di Marco, Carmen Barletta
<name.surname>@mail.polimi.it

9/6/2017, DEIB Conference Room