Data Mining and Text Mining - Course Project

Pythons on a Plane

Maddalena Andreoli, Riccardo Pressiani, Andrea Battistello,
Pietro DiMarco, Carmen Barletta

⟨*name.surname*⟩@mail.polimi.it

June 2017

1 Introduction

The task we were given was to devise a data mining model able to predict the risk of default for credit card users. In order to find the best possible model, we passed through three phases:

- Problem definition: through data exploration and researches in the credit cards domain
- Features aggregation and selection
- Model selection and validation

2 Problem definition

2.1 Data exploration

The first examination of the data was focused on trying to grasp the general characteristics of the dataset. 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. This are the general remarks we agreed upon by analysing the attributes:

- The higher `LIMIT_BAL`, the lower the probability of default is, which is reasonable.
- `SEX` is heavily skewed towards female, representing around 60% of the dataset. 0,9% of the values are missing. There is no particularly significant difference in limit balance between the two sexes, however females have a slightly lower probability of default than males (∼21% against ∼24% of males).
- There doesn't seem to be any significant correlation between the `EDUCATION` and `BIRTH_DATE` attributes and default probability.
- Single people have a slightly higher probability of not being insolvent than married ones.
- The attributes `PAY_<month>` are aligned with the common sense of the domain, i.e. that the less a customer pays back timely their due debt, the higher the chance that this client will be insolvent.

It was soon clear that we had to produce a prediction model that reduced as much as possible the number of false negative, to achieve a high recall.

2.2 Data preprocessing

First of all, we converted the attribute `BIRTH_DATE` to `AGE`, which is easier to process and analyze.
The train dataset contained a few (1323; only 39 in the test set) missing values, exclusively in the census attributes (`BIRTH_DATE`, `SEX`, `EDUCATION`, `MARRIAGE`). Most of them concerned the `AGE` attribute. Since there were no more than two attributes missing for the same person, we supposed that these were **Missing completely at random**, but we decided not to discard them. `AGE` missing values were filled, using a regression algorithm based on k-nearest neighbors (K Neighbors Regressor with k=5 from sklearn library), applied using people with no missing age as

the train set. For SEX attribute, missing values where predicted using the majority class principle, thus 'Female'; while EDUCATION and MARRIAGE where filled using the label 'other'. We also added three columns, for EDUCATION, MARRIAGE and SEX to mark whether the value was missing or not.
We transformed nominal attributes in their respective one hot encoding, with the exception of SEX that we transformed to a binary vector.

3 Features Aggregation and selection

The data cleaning and preprocessing phase was followed by the addition of new aggregated features that we considered relevant. Those added features are:

- LIMIT-MEAN_BILL = LIMIT_BAL - avg(BILL_AMT): represents the average availability every customer has with respect to their LIMIT_BAL.
- AVG_<x>_LAST_<y>, where <x> is either PAY, PAY_AMT or BILL_AMT, represents the average of the last y months for a given x. y is between 2 and 6.
- <x>_TREND, where <x> is either PAY, PAY_AMT or BILL_AMT. It represents the trend of the x attributes throughout the months.
- <x>_SKEW, where <x> is either PAY, PAY_AMT or BILL_AMT, shows the skeweness of the x attributes (meaning how unbalanced wrt a gaussian distribution they are).
- <x>_KURT, where <x> is either PAY, PAY_AMT or BILL_AMT, shows the Kurtosis coefficient for x.
- DELTA_<x> = BILL_AMT_(<x>) - PAY_AMT_(<x>+1): represents how much the bank is still due from a customer of the debit of the previous month.
- PROD = BILL_AMT_TREND * PAY_AMT_TREND
- FRAC_PAY_<x> = PAY_AMT_<x> / BILL_AMT_(<x>-1) indicates the fraction of debt that the customer has payed the following month.
- <x>_AR , <x>_I , <x>_MA and <x>_SIGMA: are the AR (Auto Regressive), I (Integrated) and MA (Moving Average) coefficients of an ARIMA process that describes <x> as a temporal series of samples. SIGMA is the variance (i.e. noiseness) of such process.
- TOTAL_PAY_AMT is the sum of all PAY_AMT_<month>.

Some features selection algorithms were used to understand which one were the most relevant for the prediction task. In particular we used the recursive features elimination algorithm cross validation (RFECV from sklearn library) with five folds and XGBoost as the estimator parameter and we discovered that there were 51 most relevant features. So we relayed on them to perform the final classification.

4 Model selection and validation

We performed model selection with a 10-fold cross validation grid search. As it turned out, choosing the right threshold is very important to improve the F1 measure, so we slightly modified the cross validation process to take into account this aspect. The threshold, then, is the average of the optimal thresholds that maximize the F1 measure of each validation fold inside an inner cross validation process. Such threshold is computed for each fold of the outer cross validation, so that we can be quite confident that the model is able to generalize well enough.
We tried out different models: Decision Trees, Logistic Regression, AdaBoost, Random Forests and XGBoost. We have obtained the best result with XGBoost (F1-measure of 0.547 as crossvalidation score and 0.538 with the holdout set), so we used that for prediction.
We also tried to perform stacking with various combinations of models, but its performance was not as good as the XGBoost classifier.