Team Jago
Data Mining 2017 project

Sklearn, KNIME, H2O

Team members:
Crippa Mattia -- 10397252
Tran Khanh Huy Paolo -- 10401830
Pirovano Alberto Mario -- 10396610
Vetere Alessandro -- 10425802

Sklearn approach

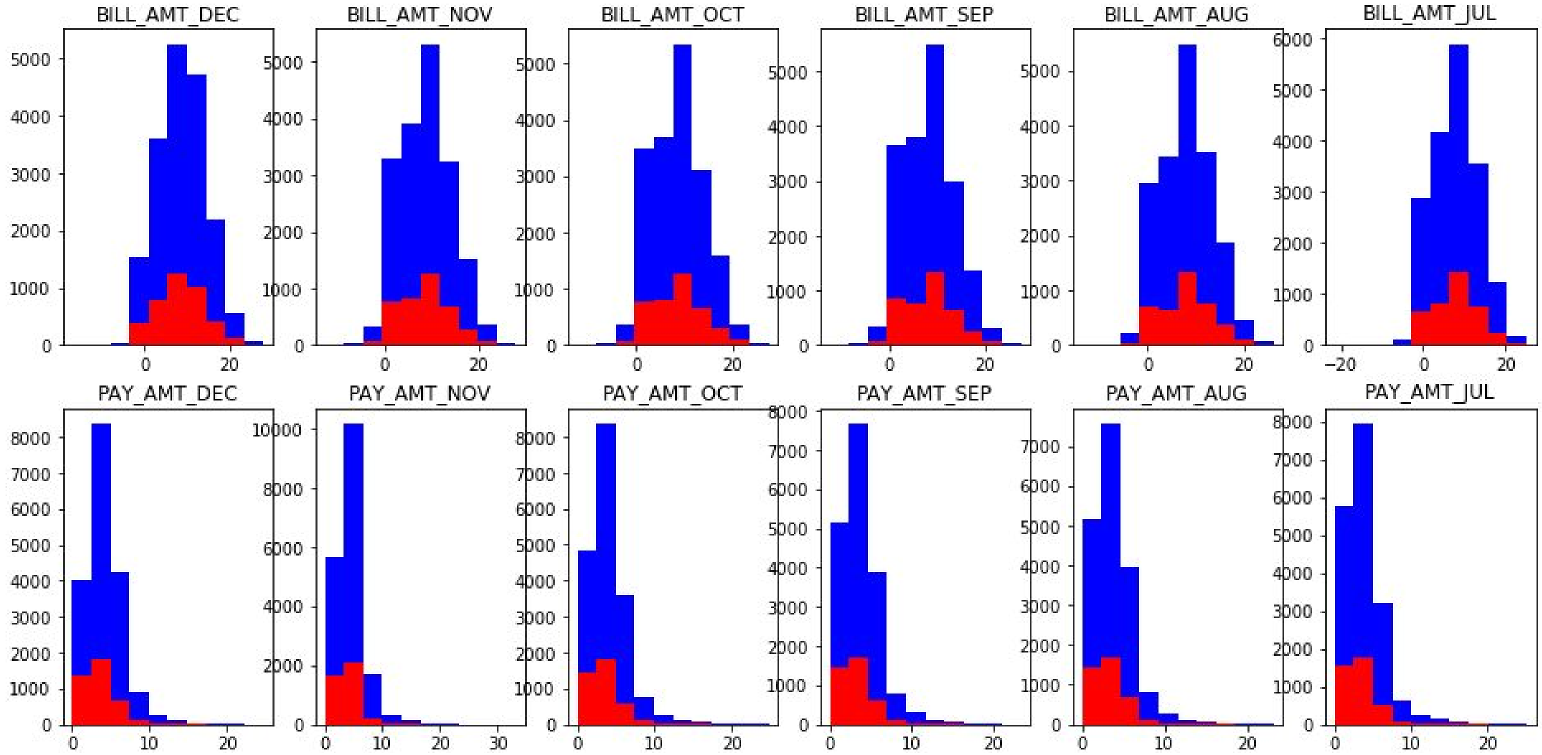
Performance assessment:

- Stratified train test split with 67% train set and 33% test set
- To evaluate algorithms we used stratified 10 fold cross validation
- Stratification was necessary to face unbalanceness of classes

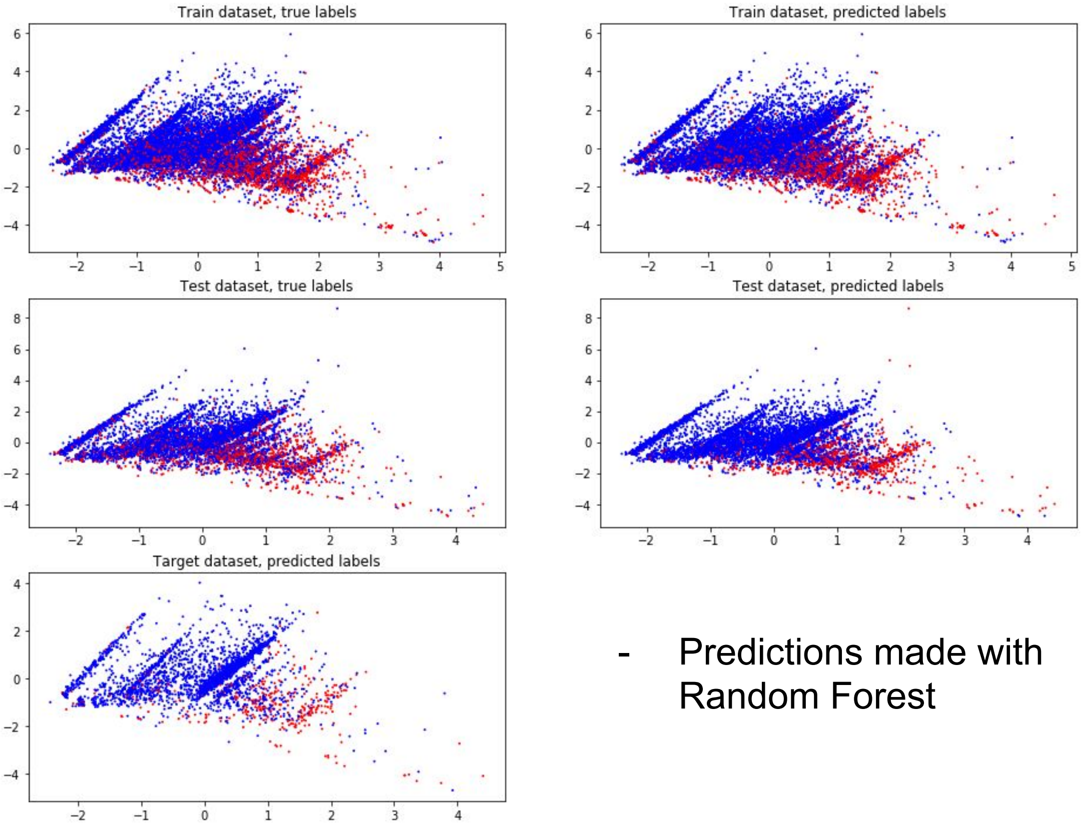
Preprocessing:

- Some variables had a too skewed distribution and to fix this we used the cubic root transformation to reshape them
- We used RobustScaler for normalizing and both OHE and scoring for categorical variables
- We applied KMeans to generate 4 clusters as additional columns

Sklearn approach: After cubic root



Sklearn approach: Visualize data using PCA



Train dataset, true labels

Train dataset, predicted labels

Test dataset, true labels

Test dataset, predicted labels

Target dataset, predicted labels

- Predictions made with Random Forest

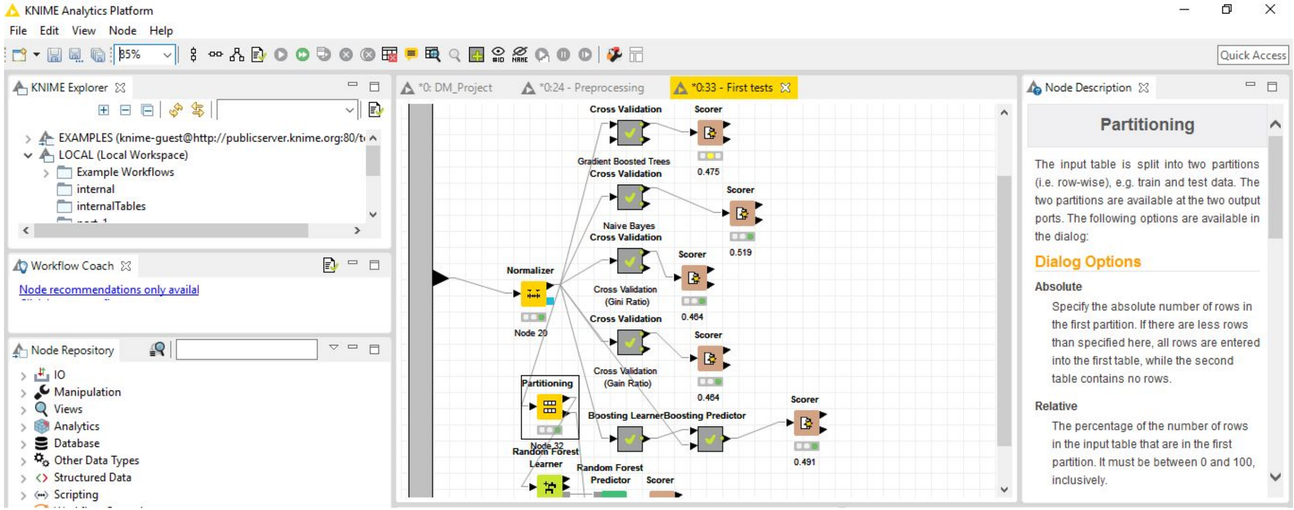
Sklearn approach: Models and Scores

Model	Threshold	F1 Cross Validation	F1 Test
Decision Tree	0.25	0.524 ± 0.024	0.525
Gaussian Naive Bayes	0.48	0.522 ± 0.016	0.528
Random Forest	0.24	0.545 ± 0.022	0.547
K Neighbors Classifier	0.24	0.531 ± 0.021	0.526
Multi Layer Perceptron	0.29	0.541 ± 0.022	0.542
Logistic Regression	0.25	0.525 ± 0.025	0.519
XGBoost	0.27	0.547 ± 0.023	0.547
Linear Discriminant Analysis	0.21	0.524 ± 0.026	0.516
Quadratic Discriminant Analysis	0.30	0.530 ± 0.022	0.524
Soft Voting Ensemble (RF, XGB, MLP)	0.24	0.548 ± 0.022	0.549

KNIME approach: Models

The models we tried were:

- Gradient Boosted Trees
- Naive Bayes
- Decision Trees with Gini Index
- Decision Trees with Gain Ratio
- Boosting learner with Naive Bayes and with Decision Trees
- Random Forest



Partitioning

The input table is split into two partitions (i.e. row-wise), e.g. train and test data. The two partitions are available at the two output ports. The following options are available in the dialog:

Dialog Options

Absolute

Specify the absolute number of rows in the first partition. If there are less rows than specified here, all rows are entered into the first table, while the second table contains no rows.

Relative

The percentage of the number of rows in the input table that are in the first partition. It must be between 0 and 100, inclusively.

Screenshot of the KNIME workflow

Deep learning approach: Autoencoders

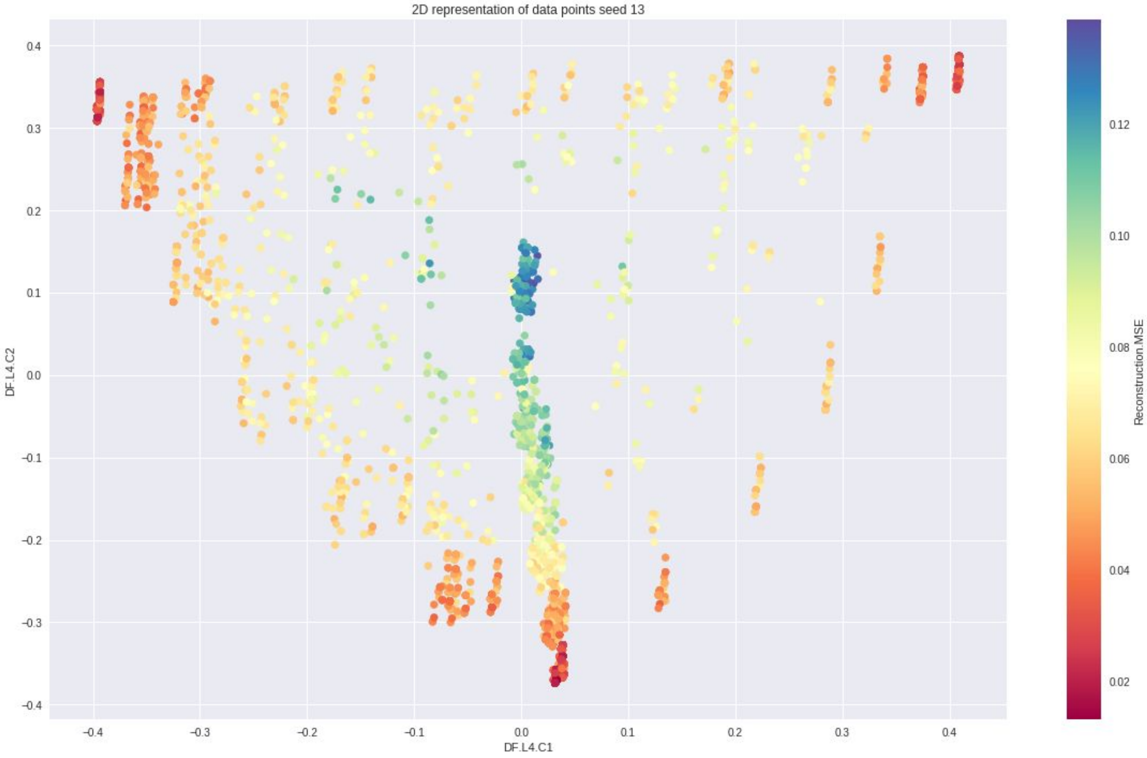
- The training set is made of samples from the majority class.
- The training phase is done in an unsupervised way with the objective of minimizing the reconstruction MSE
- At inference phase we check the reconstruction MSE for each test sample

We expect:

- reconstructionMSE(test_sample) = 0 if "test_sample" belongs to majority class
- reconstructionMSE(test_sample) > 0 if "test_sample" belongs to minority class

Practically we chose a reconstructionMSE_threshold of 0.065 for deciding how to classify data points and we achieved f1 = 0.47.

Deep learning approach: 2D hidden representation



We can see the test data points plotted in the 2 hidden dimensions of the 3rd hidden layer.

The color assigned to each point allow us to visualize the reconstruction MSE for each test point.