Scripts Description

1. Fraud Detection

Very first approach to predict if a transaction is a fraud or not. Full\_train- / Full\_test\_data sets were built using joins with train\_- / test\_- id sets. Then the “meaningful” variables (mngful\_variables) have been gathered separately. So were the numeric variables (train\_numeric) – they have been standardized with StandardScaler as well. Log-transformation of the dependent variable (TransactionAmt) has been implemented to assure normal distribution.

Finally, Logistic Regression has been chosen to make first prediction attempt. Its accuracy has been evaluated via:

* Confusion matrix,
* Sensitivity,
* Specificity,
* ROC Curve,
* AuC (Area Under Curve)

1. Future Sales

In the beginning a triple loop (through years, quarters and blocks) has been run to fill the MQ\_Dict dictionary. It will be used to map each month/year to respective quarter (column: date\_quarter) (to check the season impact on sales). The data has been aggregated to build a model that could be later tested on the test set. Test set has been assigned with date\_quarter = 4 and date\_block\_num = 34.

There were 4 regressors used to calculate the predictions:

* Dummy Regressor,
* Linear Regressor,
* Random Forest Regressor,
* ANN Regressor,

The ANN has been visualized using ann\_visualizer library. Later, RMSE has been calculated to compare all of the models. Basic attempt on Grid Search has been taken (with number of estimators). For exploratory analysis an object *single\_values* has been prepared. Uncommeting the XKCD-plot section leads to outputting nicely visualized time series for shop&item pair specified as arguments.