Exploratory Data Analysis

We are given a dataset of shape 2572x55. Each column belongs to one of the following type group: ID, Numeric, Categorical, Date Ordinal, Binary. We now briefly explore the type groups of features.  
  
The first one is ID. Among the 4 features belonging to this group, we find that country\_id and product\_id are constant. Thus, we discard them in preprocessing phase. The remaining two features, application\_id and customer\_id, contain 2571 unique value and we can consider them as unique key for the rows.

The second one is Numeric.

The third one is Categorical. We encode these features in the preprocessing phase to be fed to the model. We store the dictionary storing the encoding at *“data/categorical\_dictionary.pkl”.* The feature Variable\_5 and we drop it in the preprocessing phase. Variable\_45 describe the sex. Seven entries in the dataset have the variable sex set to “?”. As we ignore the source of data, we don’t know whether “?” is a valid vlaue (e.g. it comes from a form filled by users who don’t want to declare their sex). However, as the frequency of this value is very low, we consider it as an outlier, and we discard it in the preprocessing phase.

The fourth one is Date. In the notebook called 1\_exploratory\_data\_analysis, we show the plots of the distribution of the date features. The features due\_date, first\_status\_day\_date, paid\_date, arrived\_date start to increase in June 2015. Their frequency augment until May 2016 and then they drop. As the feature first\_status\_time\_of\_day includes only one day, the information is contained in hour, minutes and seconds.