Process the data for analysis and pre-modeling

The client files were extracted with SAS to a CSV. The original information from the company's DDBB was later anonymized. Population and other information related to customer location was taken from the INE webside.

DataFrame (notice this dataframe uses the ";" delimeter). Chunksize was also used to upload the file.

Path is used to return a normalized absolutized version of the pathname.

There were originally 400k instances and 51 features. Non residents were filtered and deleted to avoid noise in the analysys.

The dataframe used contains 395.937 instances and 44 features. 388.932 customers without a personal loan and 7.005 with a personal loan are used as the dependent variable.

21 of the 44 features have, at least, one null value and non duplicates were found.

Categorical and numerical features were treated separatelly. "sample" used for every iteration to make sure it is done correctly.

1. Categorical feature preparation

- Create a Dataframe "categorical_file" only with categorical features.
- Fill in NaN accordingly to prepare to encode categorical features.
- Replace special values of these categorical features.
- Change all categorical variables into dummies(get_dummies).

1.2 Numerical feature preparation

- Recuperate the numerical features previouslly dropped. Create a dataframe
- Replace commas representing decimal places with periodos for all columns with 'amount' as part of the name.
- Replace columns with Nan with 0.
- Change object dtype to float.

1.3 Merge categorical and numerical datasets

- Merge both files, dropping feature 'got loan' from one of them.
- Use POP to remove the item at the given position in the list. Avoid feature duplication.
- Change all column types to numerics and downcast them.
- Basic Dataframe statistics.

Although the initial information included in the dataset is real, in order to present this work and to comply with current regulations, all data has been anonymized.