EDA report

Donald Duck

19/7/2900

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: lattice
##
## Attaching package: 'mice'
## The following object is masked from 'package:tidyr':
##
##     complete
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

Objectives of the analyses

Reported below are the objectives and result of Exploratory Data Analysis performed on internal portfolio data. the analysis was performed as a preliminary step to subsequent modeling activities. the report is structured into the following main parts:

- description of data treatments applied to the raw portfolio data
- description of each analysis performed and the obtained results
- final conclusions obtained from the performed analyses

data treatments applied

To gain a sufficient level of confidence on the soundess of accuracy of employed data, data quality checks were performed on the raw dataset obtained from legacy systems. data quality checks were performed to verify:

- presence of anomalous values
- presence of incoherent data
- presence of outliers

mentioned analyses showed the presence of both outliers and incoherent values. smell test activities were also performed to understand reasons of the observed data quality deficiencies. as a result of performed data quality assessments, the following treatments were applied:

- removal of *strategic address* attribute, showing an 88 records with missing values, equal to the 100% of total records
- censoring of *PD* attribute, showing two records with PD values higher than 1
- imputation of missing values on the *LGD* attribute, which showed a 11 missing values

- introduction of a interquartilic range treatment on the *defaulted_obligors* attribute, to change outliers values to the value of lower and upper bound following the Tukey's rule for outliers identification.

Following the introduction of described treatments, the final dataset employed for the analyses resulted composed of 12 attributes and 108 records.

performed analyses

described below are the rationales and result of the analyses performed on data obtained from the applied data treatments.

average values

as a first step some summary statistics were calculated. this preliminary analysis was performed to evaluate the general level of coherence within the observed data and also considering previous knowledge of involved portfolios. it was for instance computed the average value of probability of default and loss given default by any given portfolio. this was compared with the expected level of riskiness represented by each of the involved portfolios. Reproduced below is a table showing average values of PD and LGD by portfolio.

```
portfolio_data_comple_cens %>%
  group_by(Portfolio) %>%
  summarise(mean_LGD = mean(LGD), mean_PD = mean(PD)) %>% kable()
```

Portfolio	mean_LGD	mean_PD
RS1	0.3887819	0.0353888
RS10	0.2973043	0.0443230
RS11	0.2888401	0.0254973
RS12	0.0000000	0.0505667
RS13	0.1559120	0.0091716
RS14	0.4222288	0.0648266
RS15	0.3735040	0.0196156
RS16	0.0000000	0.6666667
RS17	0.0000000	1.1666667
RS18	0.8006509	0.1060031
RS2	0.0450739	0.0338803
RS3	0.2816686	0.0195141
RS4	0.0142338	0.1863905
RS5	0.2611357	0.0756213
RS6	0.0322339	0.1151104
RS7	0.4214273	0.0143236
RS8	0.1520841	0.0139800
RS9	0.1743367	0.0177692

A general positive evaluation was obtained from the analyses, since level of PD and LGD appeared aligned with what expected.

linear correlation

pearson coefficients were calculated among each numerical variable. results are reproduced below:

```
portfolio_data_comple_cens %>%
  select(non_defaulted_obligors,
         defaulted_obligors,
         exposure_defaulted,
         exposure_non_defaulted,
         PD,
         LGD) -> numeric_portfolio
cor(numeric_portfolio,method ="pearson") %>% kable()
```

	non_defaulted_obligors	defaulted_obligors	exposure_defaulted	exposure_non_defaulted
non_defaulted_obligors	1.0000000	0.6724627	0.7176267	0.7121169
defaulted_obligors	0.6724627	1.0000000	0.7498228	0.6599929
exposure_defaulted	0.7176267	0.7498228	1.0000000	0.7284908
exposure_non_defaulted	0.7121169	0.6599929	0.7284908	1.0000000
PD	-0.1045893	-0.0971819	-0.0989288	-0.0853805
LGD	0.0584066	0.0860841	0.0696989	0.0602789

also here a general coherence was observed.

conclusions

following analyses reported above, the following relevant conclusions was obtained:

- that the quality of our data resulted being not satisfactory for some of the attributes and that this should be further investigated for subsequent uses of this dataset
- that with the data we have is possibile to fit linear regression models since we haveverifiede the assumption related to both sample adequacy and residuals.

we can therefore conlcude that the performed analyses allowed to reach the objectives that were stated within the *objectives of the analyses* paragraph