SEB Assessment

Michał Chmura

1 June 2023

1 Problem description

A dataset "alerts.xlsx" was used. It consists of 10177 observations with 14 features:

FIELD	DESCRIPTION
intID	Alert ID
AlertType	Alert type
AlertState	Alert state (1st line)
DateCreated	Date when alert was generated
DateClosed	Date when 1st line closed the alert
CaseOpen	Date when 2nd line started the investigation
CaseClosed	Date when 2d line closed the investigation
CaseReported	Date when SAR (Suspicious Activity Report) was sent to FIU
CaseState	Case State (2nd line)
PEP	PEP/not PEP
CusRiskCategory	Customer Risk Category
Type	Customer type: private banking or LC&FI
IndustryCode	Industry Code

Also, one additional undescribed feature was given - it was the only one without duplicates, so it might be some sort of ID (and has been renamed as such).

2 EDA

After standardizing all null values to NAs, it was detemined that some of the columns needs imputation. IndustryCode was imputed with numeric 0, CusRiskCategory with "Not specified", PEP with "N", CaseState with "Not Opened". More detailed comments can be found in the R file included.

After that, a separate dataset was created - $data_cleaned$. Columns related to time were dropped, as there is no consistent way to impute all of them, so they would be useless for modeling. They were kept in the original dataset, as they migth provide some additional insight upon further inspection.

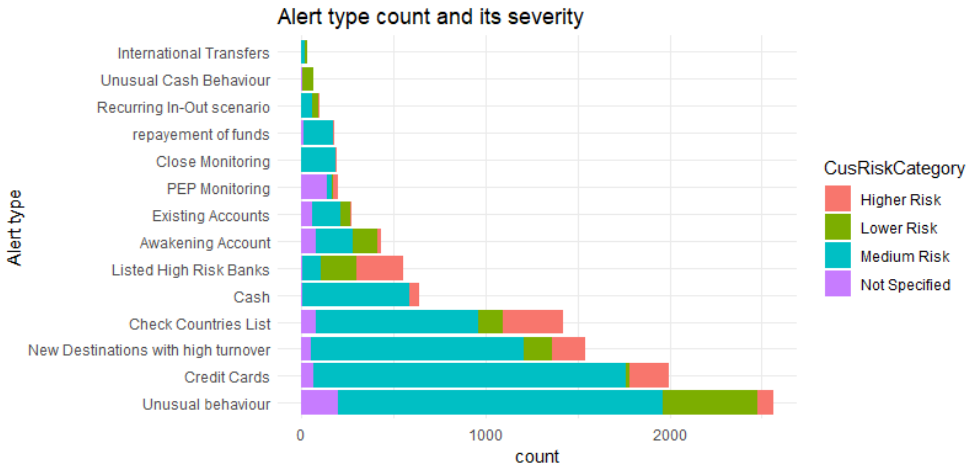
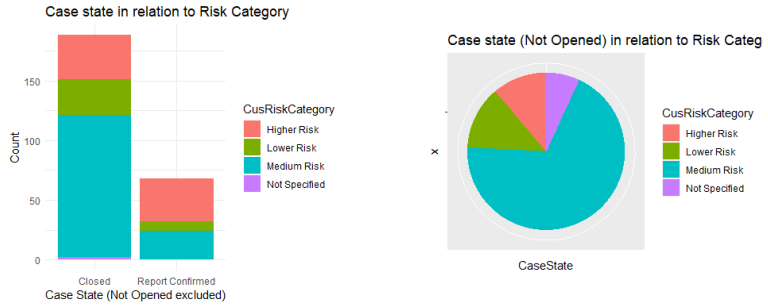


Figure 1: Number of alerts by alert time

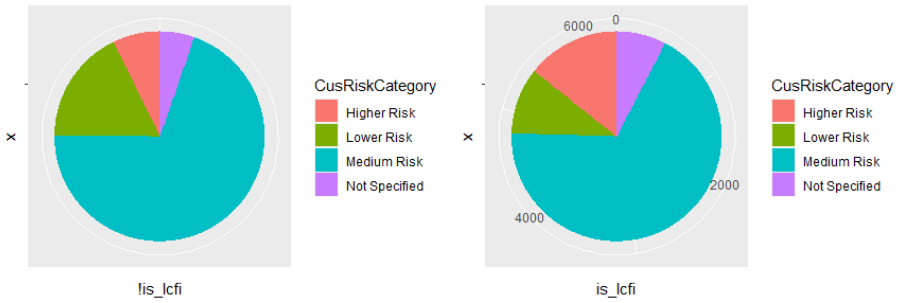


(a) Linear, $C = 1$ (b) Linear, $C = 0.04$

Figure 2: Case state and risk type

2.1 Visualizations

From the Figure 1 we can learn that credit cards usage and unusual behaviour are the most common reasons for alert. Most of them are medium level risk. We will also investigate the distribution of risk level in the next plots. Next two figures show case state in relation to specific risk categories. "Not Opened" category is showed on a separate piechart, as it has much more entries that the other two, so the barchard would become unreadable. We can read from those charts, that if the case was reported, it was more likely for it to be higher risk. Distributions of closed and not opened are similar to each other.



(a) Linear, $C = 1$ (b) Linear, $C = 0.04$

Figure 3: Client type and risk type

Next we will analyze the distribution of risk levels depending on type of the customer - private banking or corporate:

We can see that corporate clients are less likely to be at lower risk - those cases might just be employees making small mistakes, but more likely to be at high risk - as e.g. being targeted with a cyberattack

3 Encoding

If we want to use statistical models on the data, encoding of categorical text data is in order. PEP, Type and CaseState can be one-hot encoded, and will be so - converted to is_lcfi,is_PEP, caseState_Closed,caseState_RepCon. Alert type can be also encoded in such a way, though as it would double the number of columns in the dataset, it was not done here. Same goes for the column AlertState, though keeping it in text form might also form train and test set for the model - as we have clear distinction which cases are done and which are not. IndustryType might be encoded with the help of information included in $additional\_info$ file, but I did not do it, as it is already a categorical variable, so for the sake of interpretability it is left as it was. Finally, CusRiskCategory was hierarchically encoded, with levels from 0 as "Not Specified" to 3 as "Higher Risk".

	vars <dbl>	n <dbl>	mean <dbl>
ID*	1	10177	5089.00
intID	2	10177	31246.26
AlertType*	3	10177	7.49
AlertState*	4	10177	3.99
IndustryC...	5	10177	11022.40
is_lcfi	6	10177	0.62
is_PEP	7	10177	0.01
caseState...	8	10177	0.02
caseState...	9	10177	0.01
enc_CusR...	10	10177	1.85

Figure 4: Basic summary table

3.1 Summary table

Basic summary table was created - please disregard IDs, text variables and IndustryType. What we can learn from it, is that PEPs are just 0.01 of the dataset, and corporate clients are 62% off all clients whom the reports concern. Most of the cases are currently "Not Opened".

4 Additional analysis and insights

- Proper ID structure needs to be established
- Random forest classifier or logistic regression can be implemented to determine whether the case needs to be escalated/researched more
- If there would be real world data available with descriptions e.g. transcripts from chats regarding fraud, sentiment analysis can be done
- Thresholds for risk size are identical for private and corporate clients, it might be better to evaluate them separately
- MLLib library might be useful, as one can try to implement TM model which utilizes Spark to make real-time predictions
- Data processing pipeline should be established