Models for Anti Money Laundering

FIE453-H20 Big Data with Applications to Finance

Final Project

Ruhui Ding, Candidate no. 453013

Agenda

- Project description
- Whether a company is suspicious
- Whether an individual transaction is suspicious
- Conclusion

Agenda

- Project description
- Whether a company is suspicious
- Whether an individual transaction is suspicious
- Conclusion

Datasets

- **Customer data.** All customers who were flagged by the bank’s old flagging system for suspicious transactions. With historical information about whether or not the customer ended up actually being reported to the authorities (after a manual inspection), along with information about the customer.

- **Transaction data.** All the transactions of the bank’s customers, with information about each transaction.

Objectives

Objectives

- Create machine learning models to calculate the probability of a corporate customer being involved in suspicious activities

- Create machine learning models to indicate whether an individual transaction is suspicious.

Models

- Supervised model
  - Linear logistic regression model
  - XGBoost model

- Unsupervised model
  - Anomaly detection model

Agenda

- Project description
- Whether a company is suspicious
- Whether an individual transaction is suspicious
- Conclusion

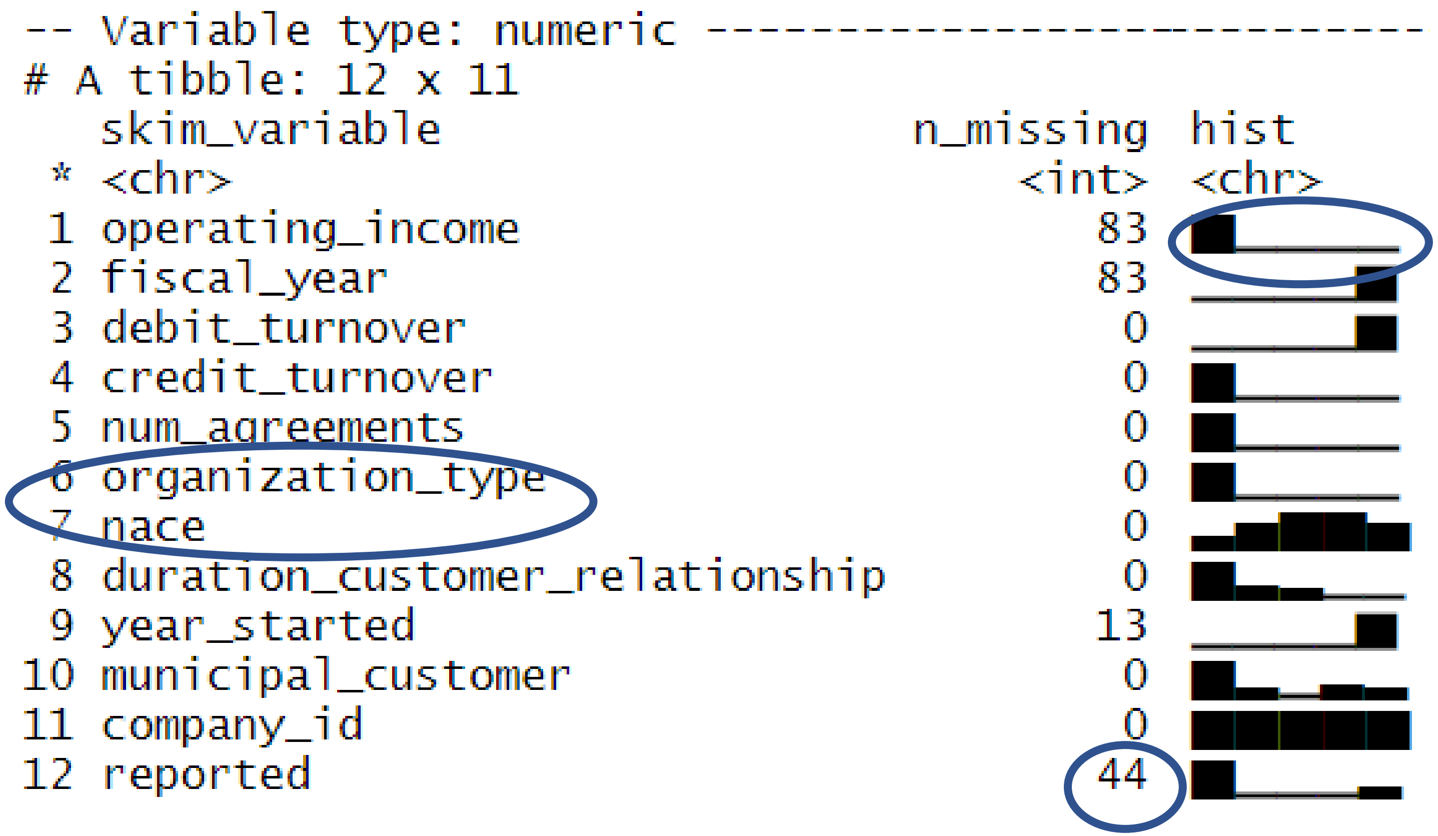
Whether a company is suspicious

Content

- Data visualization
- Linear logistic regression model
- XGBoost model
- Comparison of the two models

Data visualization

Data summary statistics

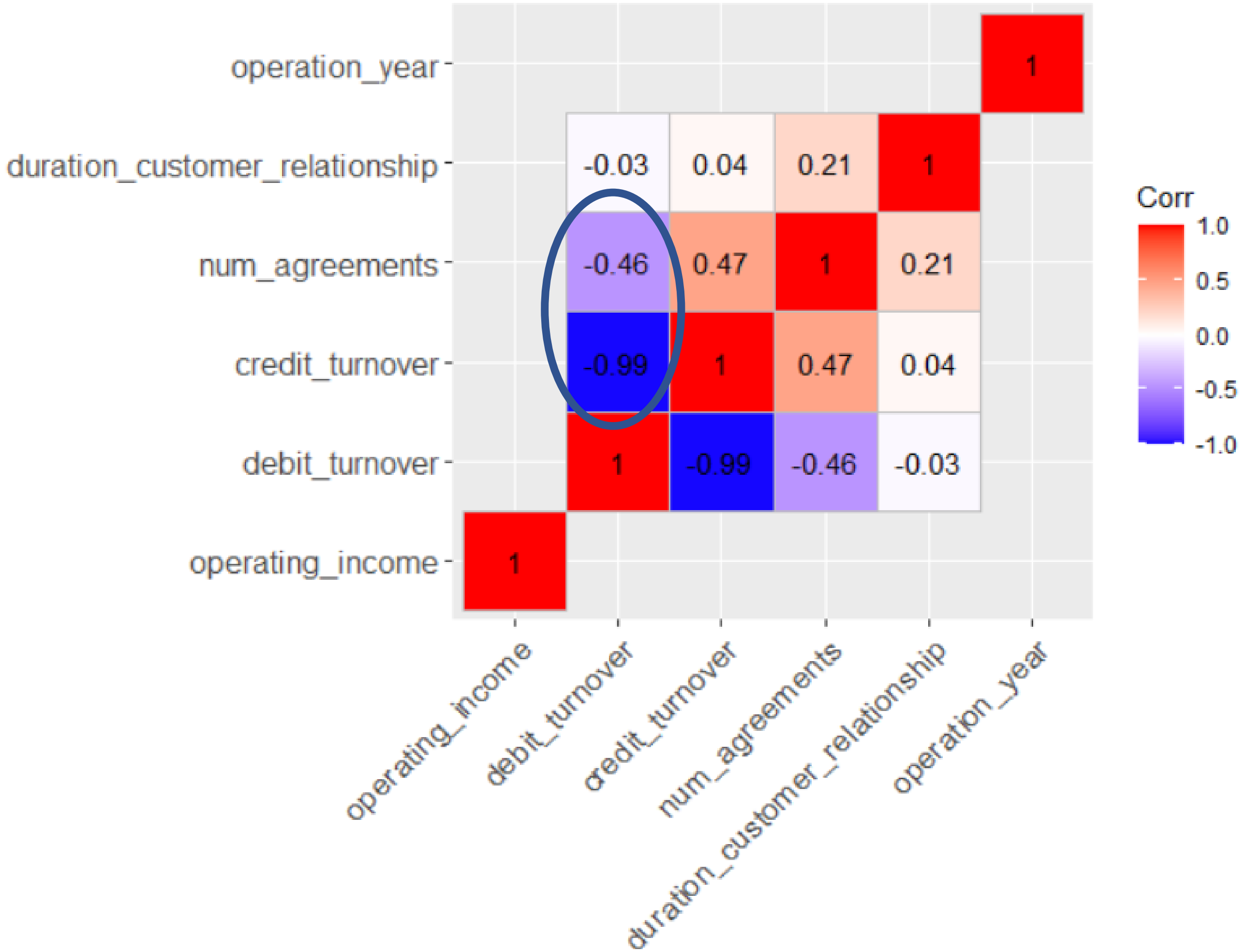


```
-- Variable type: numeric ----------------------------
# A tibble: 12 x 11
   skim_variable                  n_missing  hist
 * <chr>                              <int>  <chr>
 1 operating_income                      83  ▇▁▁▁▁
 2 fiscal_year                           83  ▁▁▁▁▇
 3 debit_turnover                         0  ▁▁▁▁▇
 4 credit_turnover                        0  ▇▁▁▁▁
 5 num_agreements                         0  ▇▁▁▁▁
 6 organization_type                      0  ▇▁▁▁▁
 7 nace                                   0  ▂▅▇▇▅
 8 duration_customer_relationship         0  ▇▂▂▁▁
 9 year_started                          13  ▁▁▁▁▇
10 municipal_customer                     0  ▇▂▁▂▂
11 company_id                             0  ▇▇▇▇▇
12 reported                              44  ▇▁▁▁▂
```

- Missing values
- Wrong data types
- Highly skewed variables

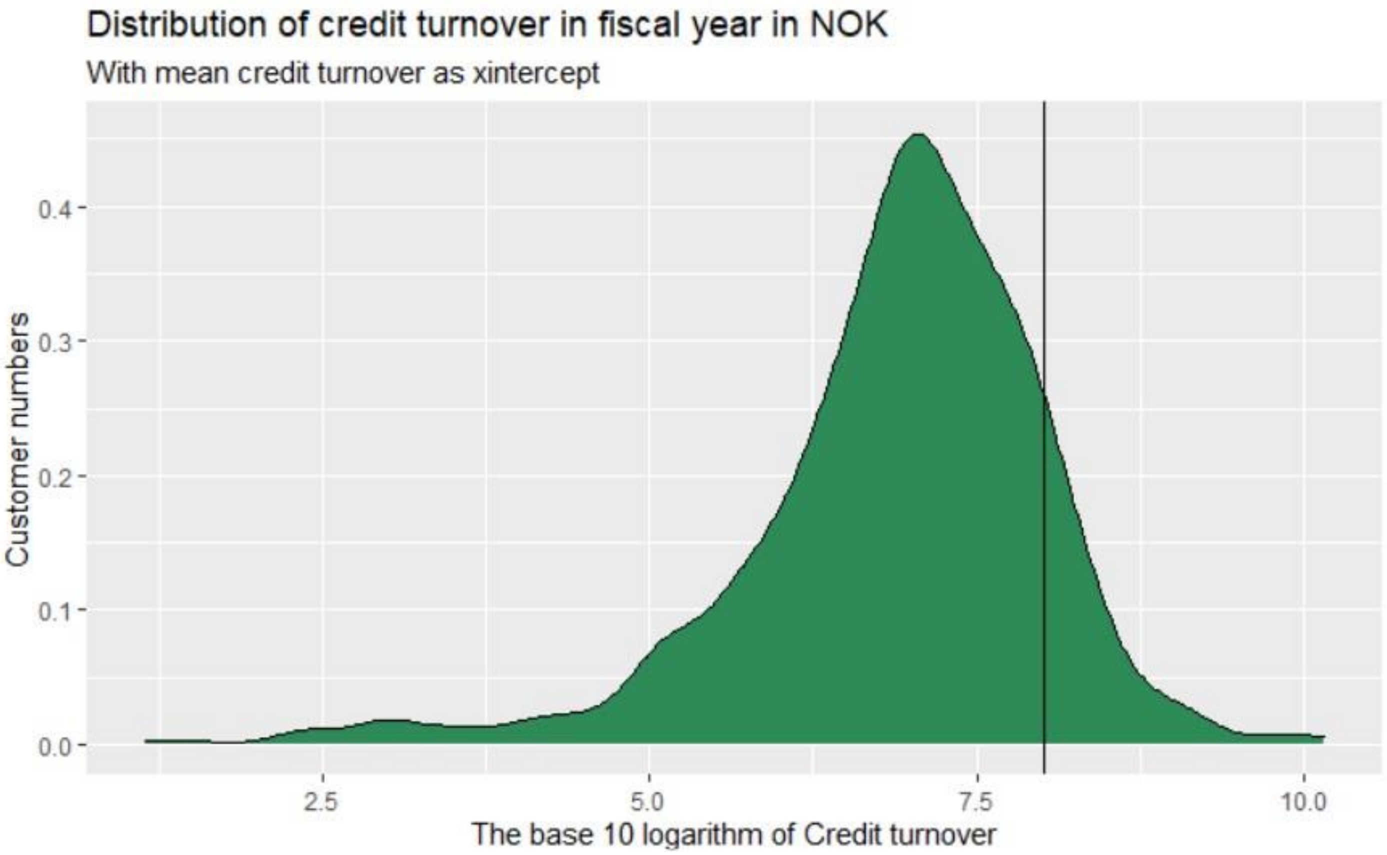
Data visualization

Correlation Plot and Correlation Coefficients



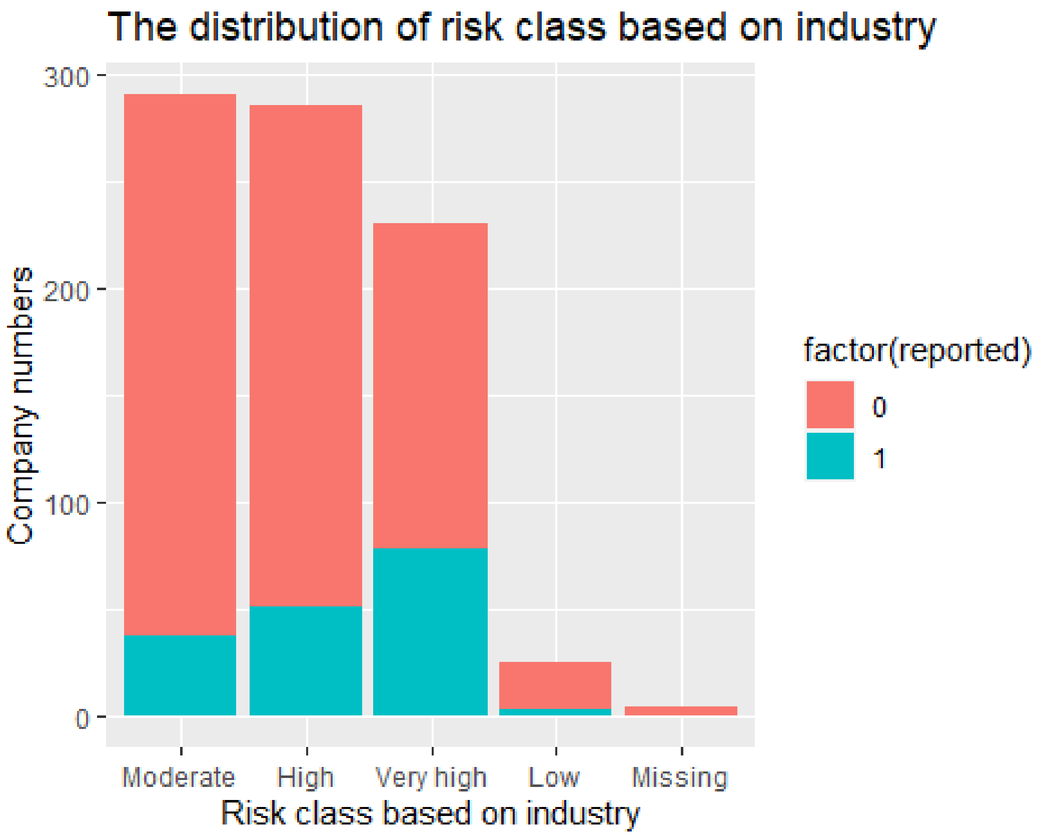
- Check the collinearity between independent variables.
- *debit_turnover* and *num_agreements* are correlated to *credit_turnover*.

Data visualization



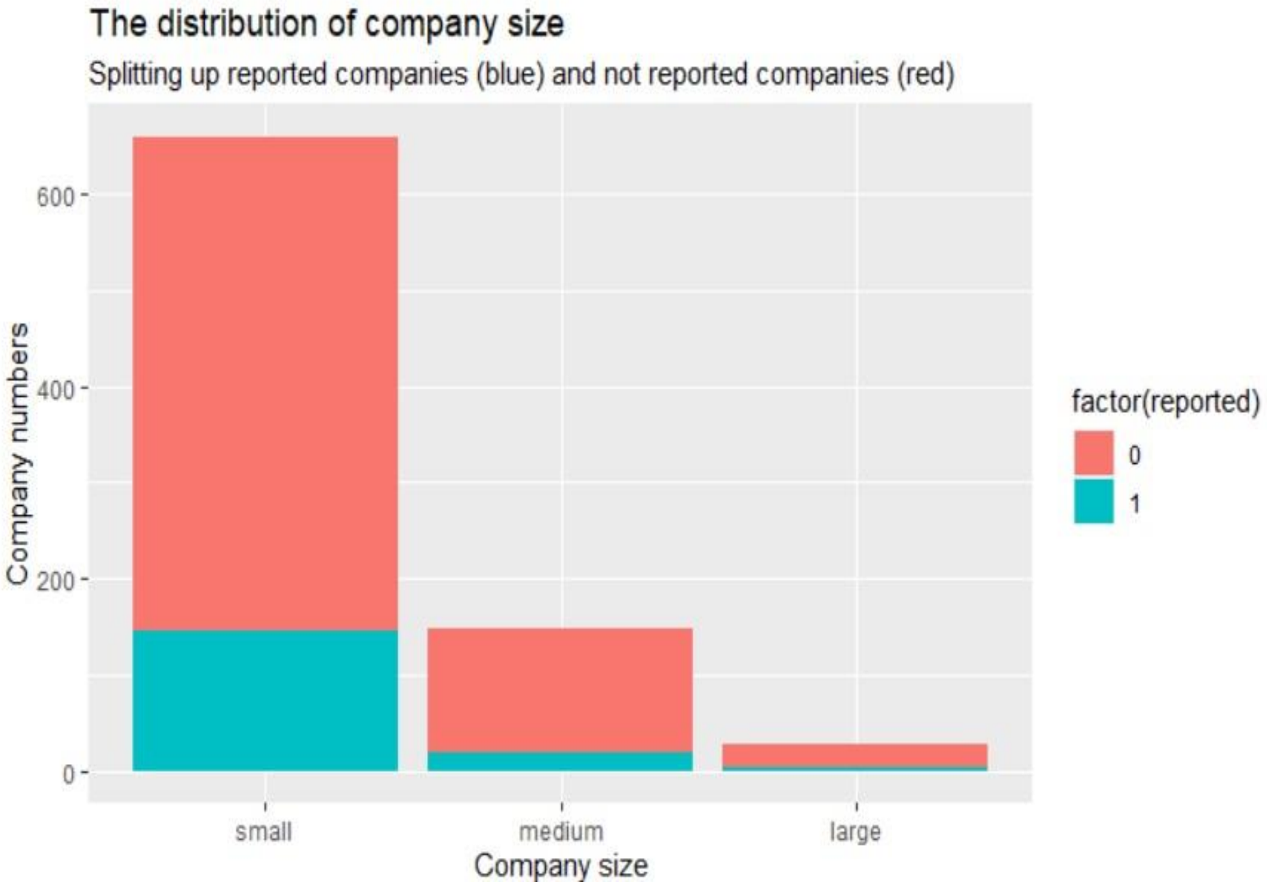
Distribution of credit turnover in fiscal year in NOK

With mean credit turnover as xintercept



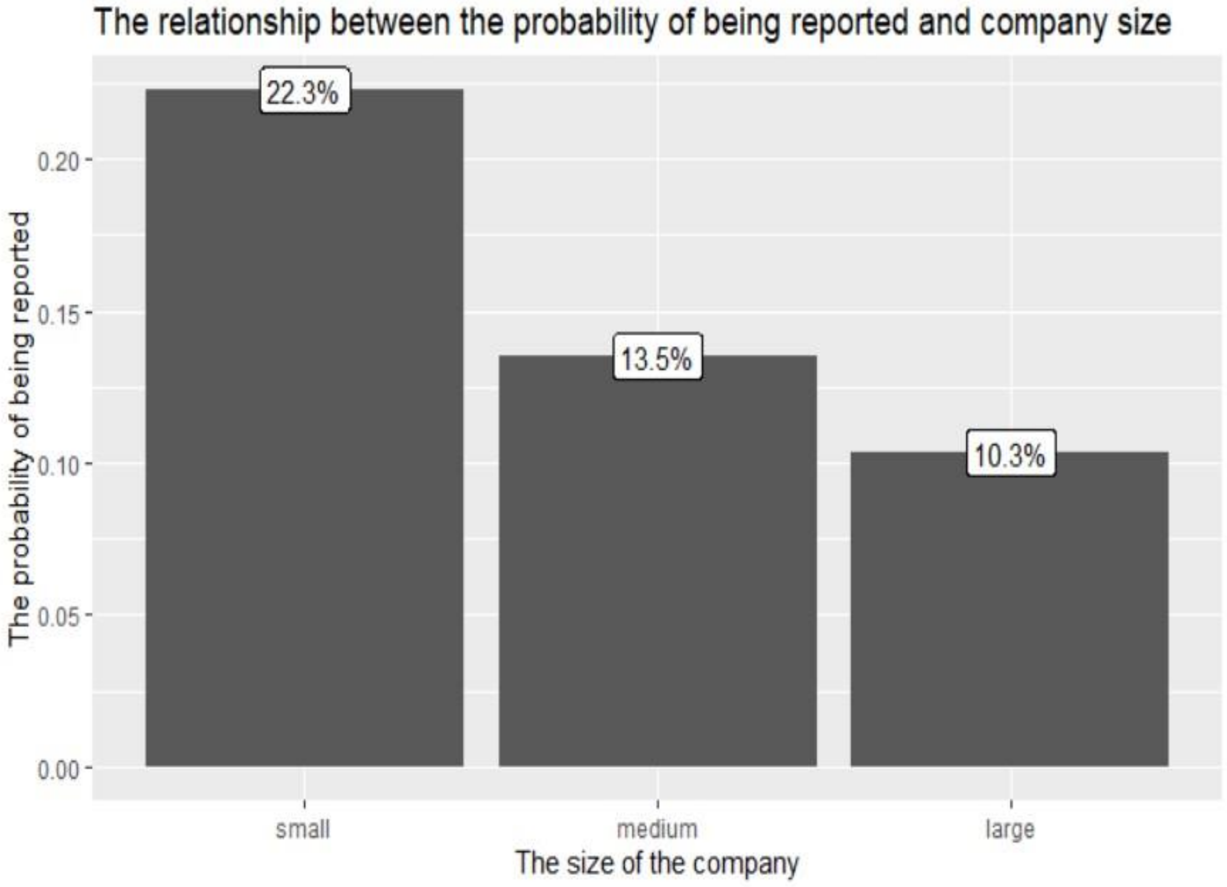
The distribution of risk class based on industry

Data visualization



The distribution of company size

Splitting up reported companies (blue) and not reported companies (red)



The relationship between the probability of being reported and company size

Linear logistic regression model

Variable selection and transformations

Not choose

- *company_id*
- *debit_turnover*
- *num_agreements*
- *fiscal_year*
- *year_started*
- *country_customer*
- *organization_type*
- *Bankrupt*
- *nace*

Choose

- *language_form*
- *num_accounts*
- *risk_industry*
- *company_size*
- *reported*

Log transformation

- *operating_income*
- *duration_customer_relationship*
- *credit_turnover*

Lump levels

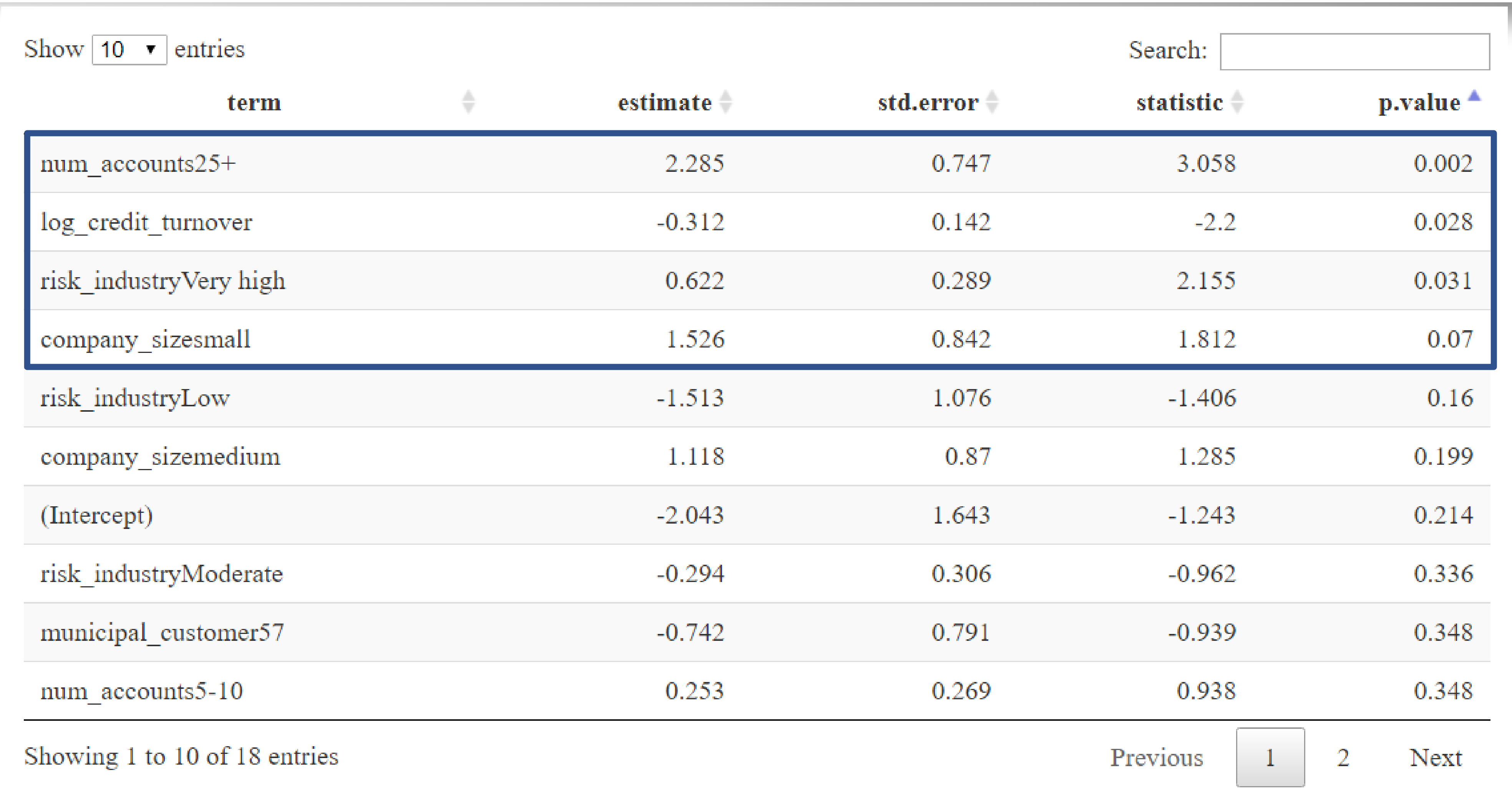
- *municipal_customer*

New variables

- *operation_year*

Linear logistic regression model

The findings from the model

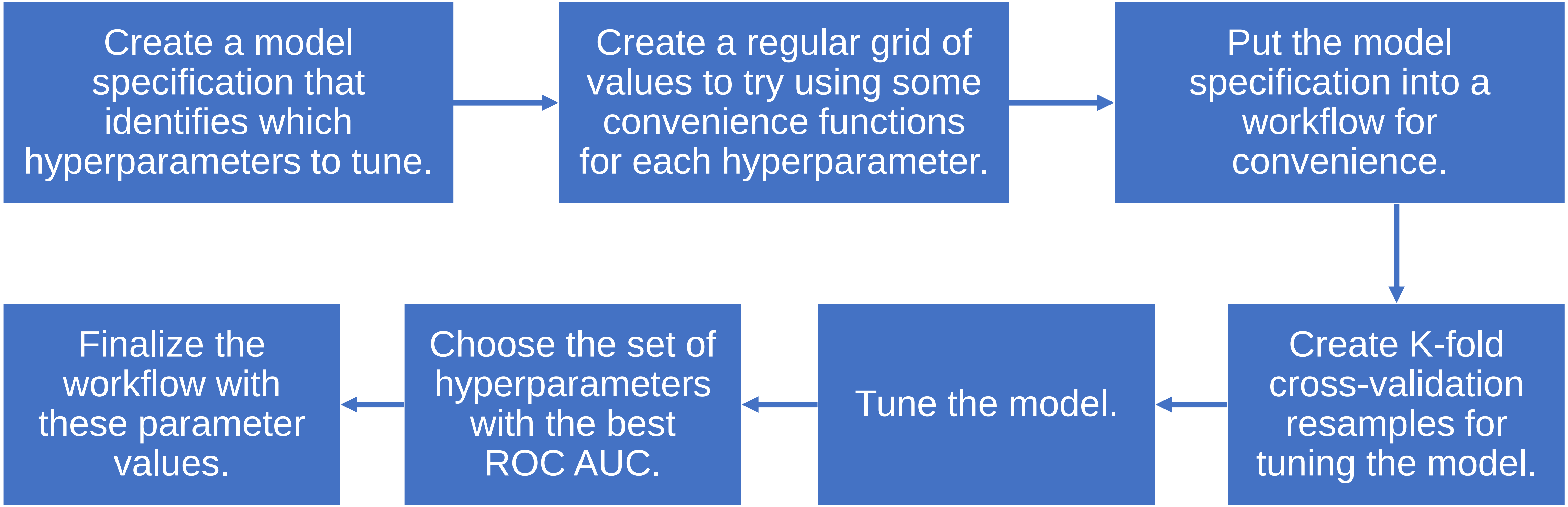


term	estimate	std.error	statistic	p.value
num_accounts25+	2.285	0.747	3.058	0.002
log_credit_turnover	-0.312	0.142	-2.2	0.028
risk_industryVery high	0.622	0.289	2.155	0.031
company_sizesmall	1.526	0.842	1.812	0.07
risk_industryLow	-1.513	1.076	-1.406	0.16
company_sizemedium	1.118	0.87	1.285	0.199
(Intercept)	-2.043	1.643	-1.243	0.214
risk_industryModerate	-0.294	0.306	-0.962	0.336
municipal_customer57	-0.742	0.791	-0.939	0.348
num_accounts5-10	0.253	0.269	0.938	0.348

Showing 1 to 10 of 18 entries

XGboost

Tuning hyperparameters



Create a model specification that identifies which hyperparameters to tune.

Create a regular grid of values to try using some convenience functions for each hyperparameter.

Put the model specification into a workflow for convenience.

Create K-fold cross-validation resamples for tuning the model.

Tune the model.

Choose the set of hyperparameters with the best ROC AUC.

Finalize the workflow with these parameter values.

XGBoost model

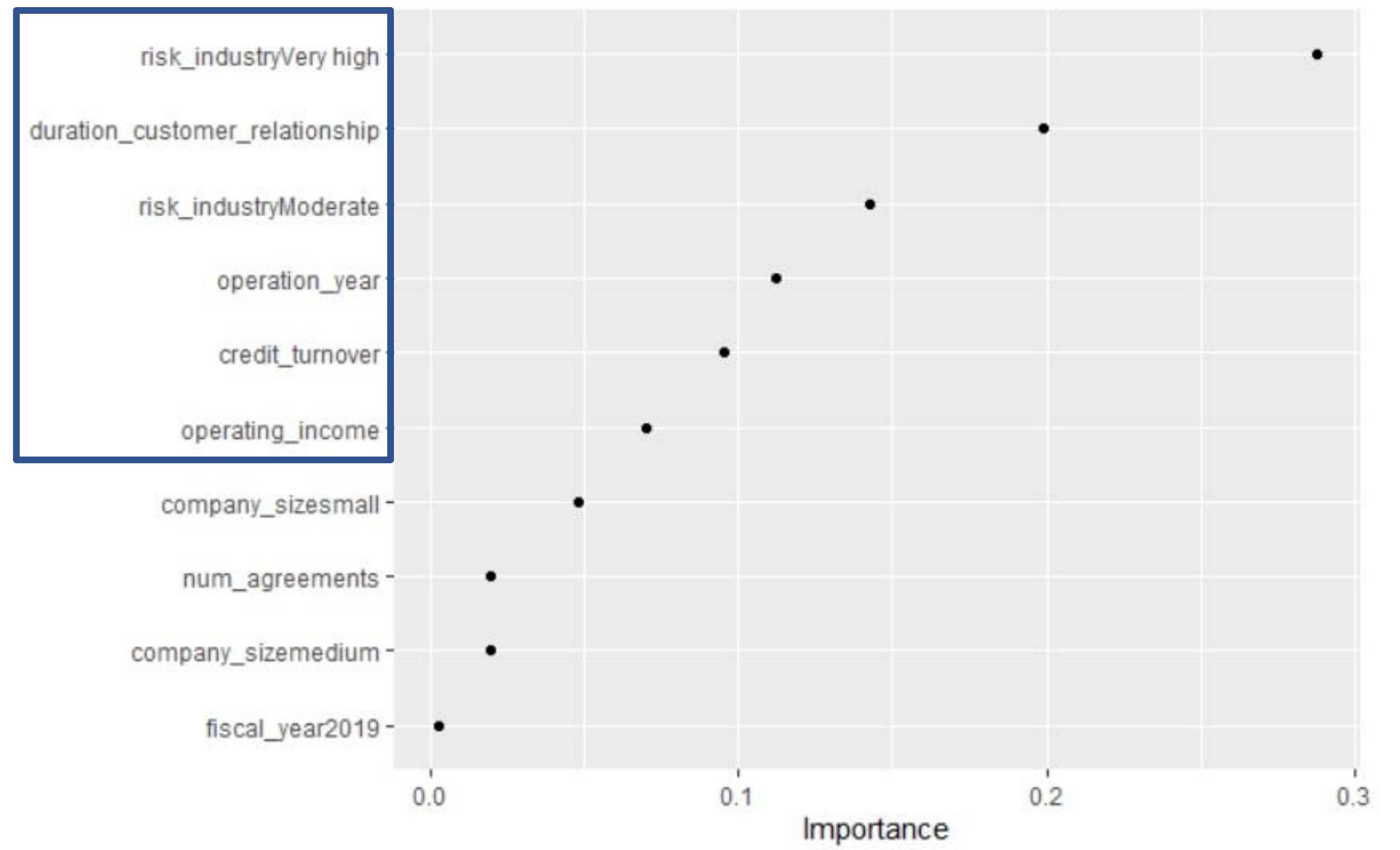
The set of hyperparameters with the best ROC AUC

Mtry	Min_n	Tree_depth	Learn_rate	Loss_reduction	Sample_size
13	19	13	0.0002354	0.0000008827	0.8666

- The number of predictors that are randomly sampled at each split is 13.
- The minimum number of data points in a node that is required for the node to be split further is 19.
- The maximum depth of the tree is 13.
- The rate at which the boosting algorithm adapts from iteration-to-iteration is 0.0002354.
- The reduction in the loss function required to split further is 0.0000008827.
- The amount of data exposed to the fitting routine is 0.8666.

XGboost

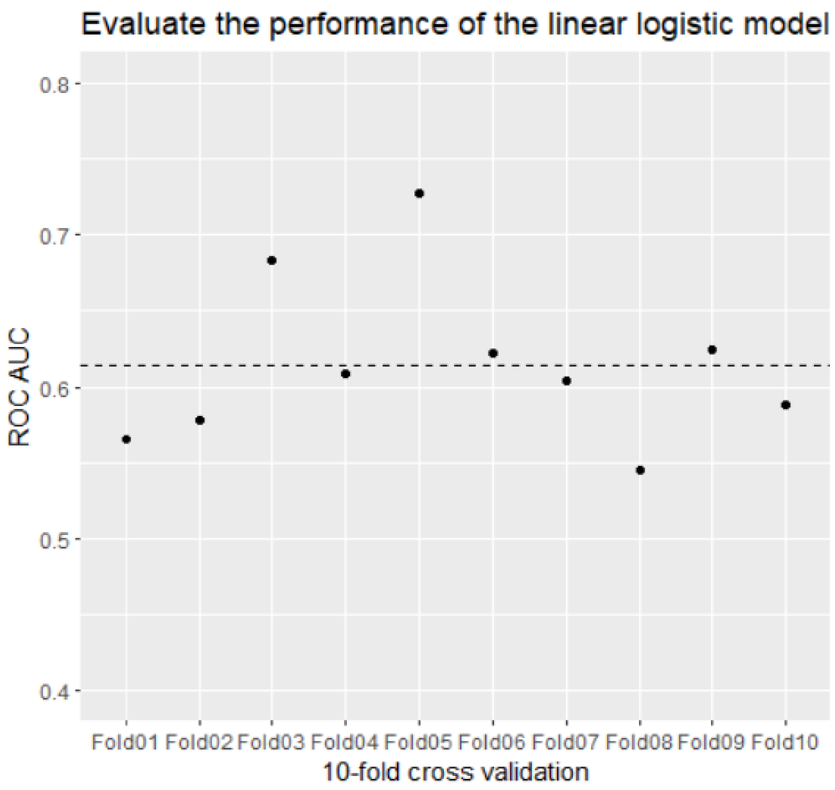
The most important variables

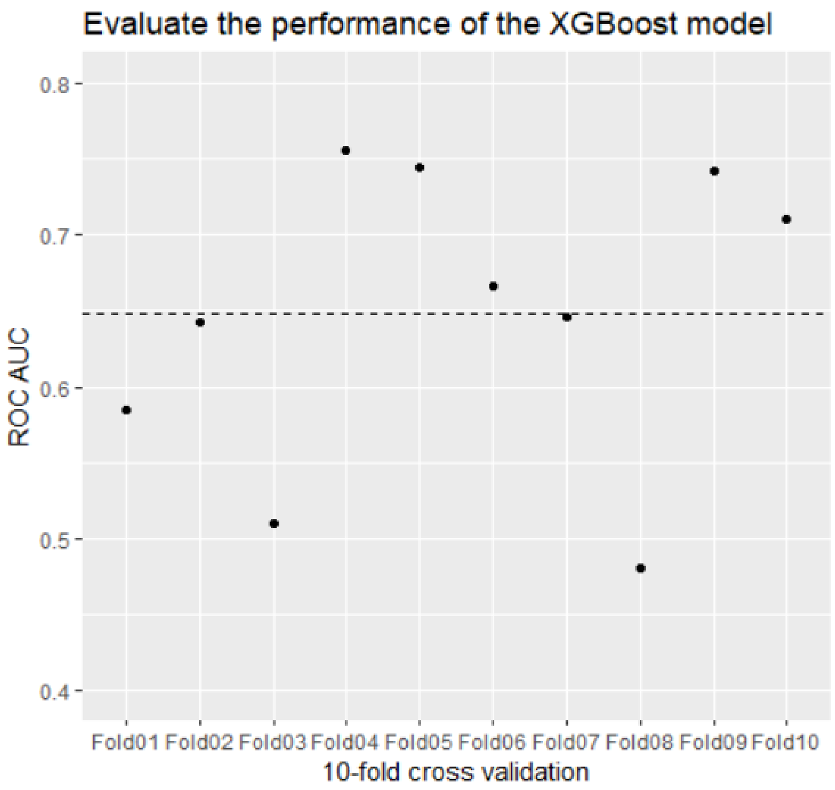


Comparison of the two models

Evaluate the models using 10-fold cross validation

Model	Metric	Mean
XGBoost	accuracy	0.7930
XGBoost	roc_auc	0.6484
glm	accuracy	0.7842
glm	roc_auc	0.6147


Evaluate the performance of the linear logistic model


Evaluate the performance of the XGBoost model

Comparison of the two models

Advantages and disadvantages

XGBoost model

Advantages

- Less data processing.
- Good model performance.
- With a proper tuning process, less likely to overfit.

Disadvantages

- Tuning process takes a long time.
- The interpretation is not very informative.
- Prone to overfit if hyperparameters are not tuned properly.

Linear logistic regression model

Advantages

- No need to tune hyperparameters.
- Fast to implement.
- More informative interpretation.

Disadvantages

- Need more data processing.
- Poorer model performance.

Agenda

- Project description
- Whether a company is suspicious
- Whether an individual transaction is suspicious
- Conclusion

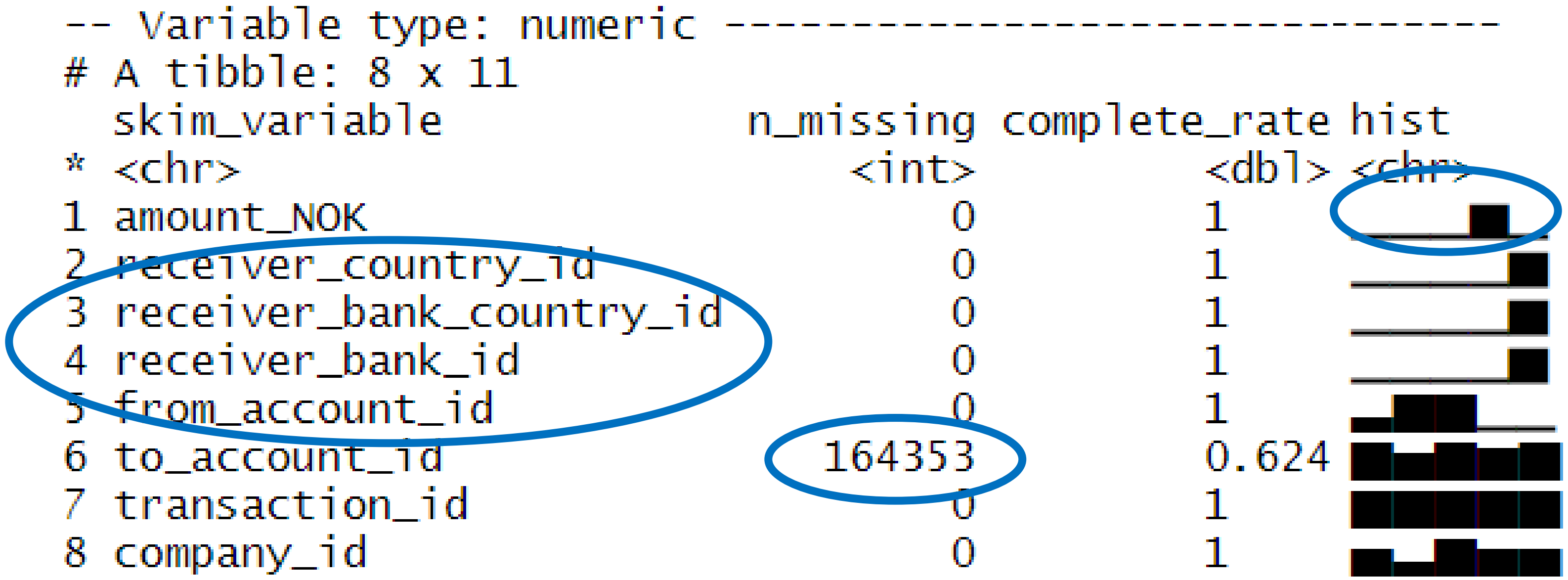
Whether an individual transaction is suspicious

Content

- Anomaly detection model
- Aggregated anomaly score
- Combine linear logistic regression with aggregated anomaly score
- Handle large amounts of data

Anomaly detection model

Data summary statistics



```
-- Variable type: numeric --------------------------------
# A tibble: 8 x 11
  skim_variable            n_missing complete_rate hist
* <chr>                        <int>         <dbl> <chr>
1 amount_NOK                       0         1     ▁▁▁▇▁
2 receiver_country_id              0         1     ▁▁▁▁▇
3 receiver_bank_country_id         0         1     ▁▁▁▁▇
4 receiver_bank_id                 0         1     ▁▁▁▁▇
5 from_account_id                  0         1     ▂▇▇▁▁
6 to_account_id               164353         0.624 ▇▆▇▆▇
7 transaction_id                   0         1     ▇▇▇▇▇
8 company_id                       0         1     ▃▁▇▃▃
```

- Missing values
- Wrong data types
- Highly skewed variables

Anomaly detection model

Variable selection and transformations

Not choose

- *transaction_date*
- *to_account_id*
- *company_id*
- *transaction_id*

Choose

- *currency*
- *transaction_type*
- *operation_year*
- *overfoering_egne_konti* (internal_transaction)

Lump levels

- *text_code*
- *receiver_country_id*
- *receiver_bank_country_id*
- *receiver_bank*
- *from_account_id*

New variables

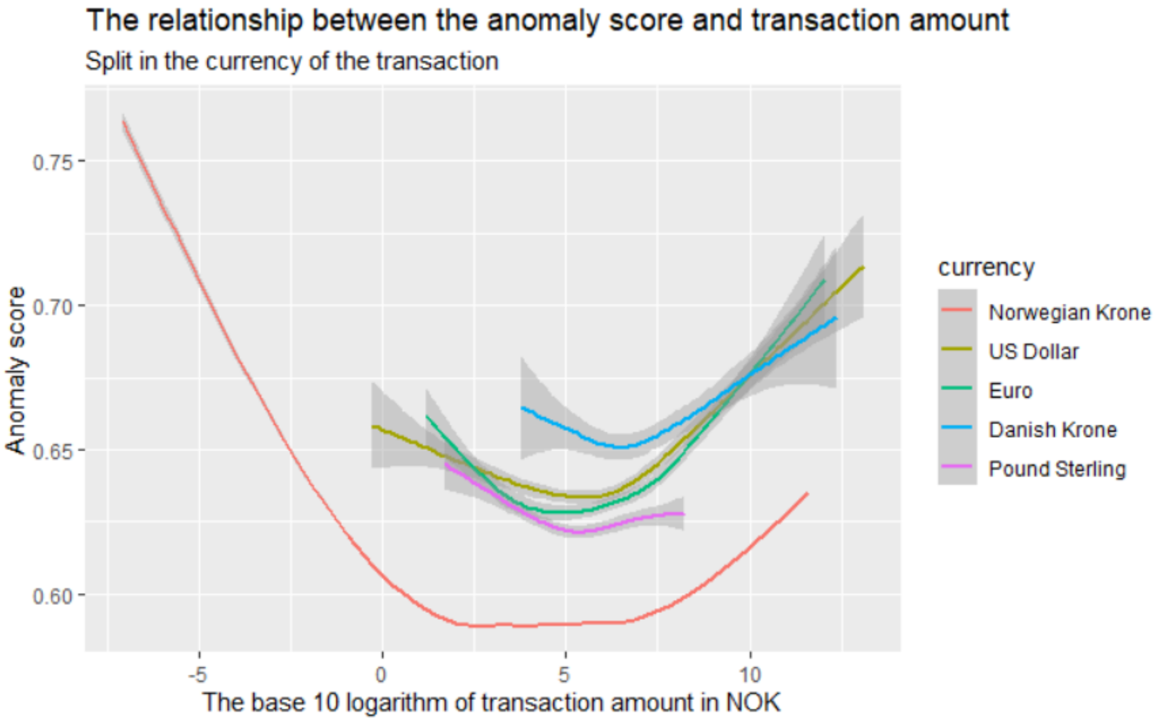
- *month*
- *day*
- *weekday*
- *relative_size*

Log transformation

- *amount_NOK*
- *relative_size*

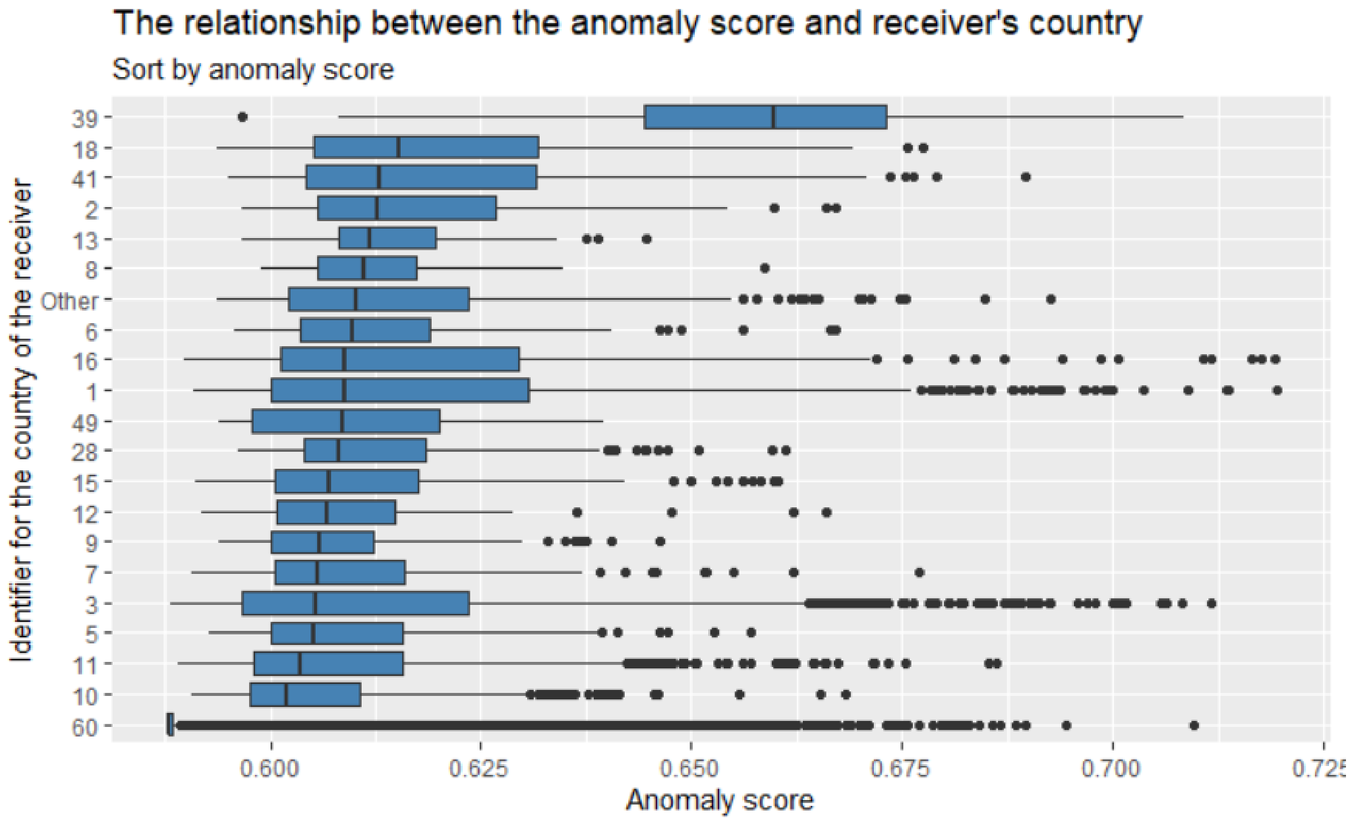
Anomaly detection model

Visualize some of the relationships



The relationship between the anomaly score and transaction amount

Split in the currency of the transaction

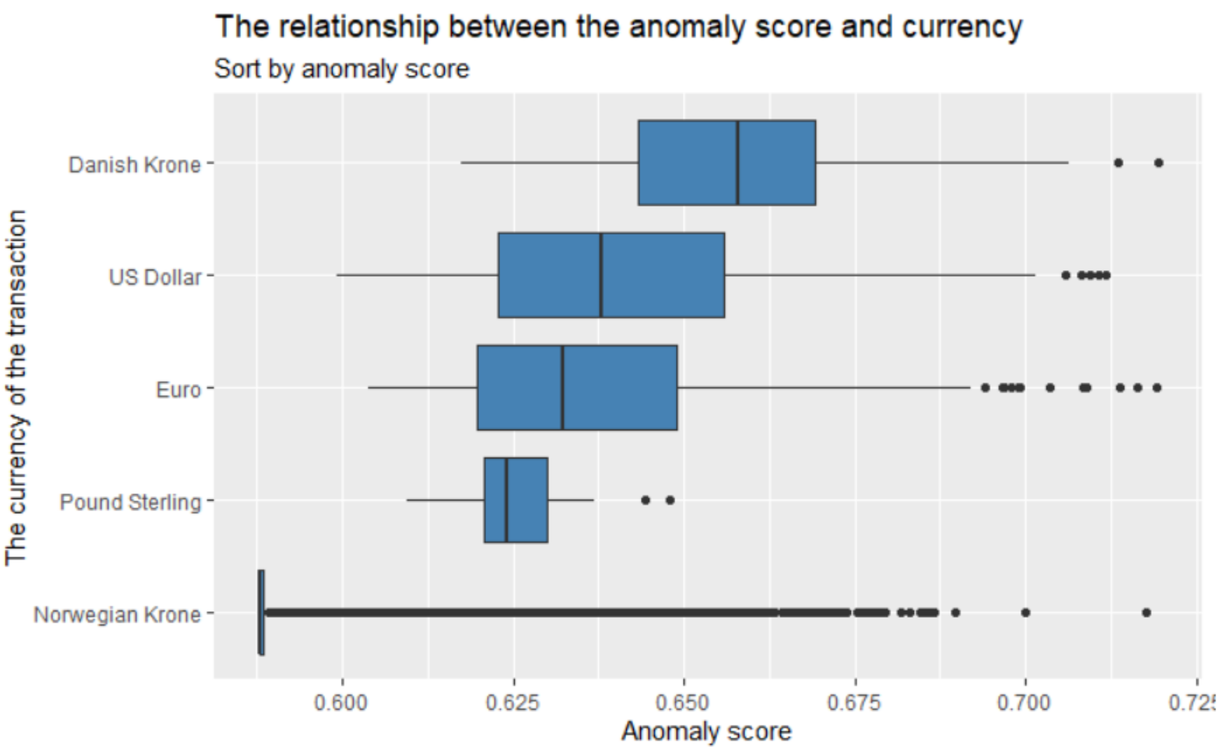


The relationship between the anomaly score and receiver's country

Sort by anomaly score

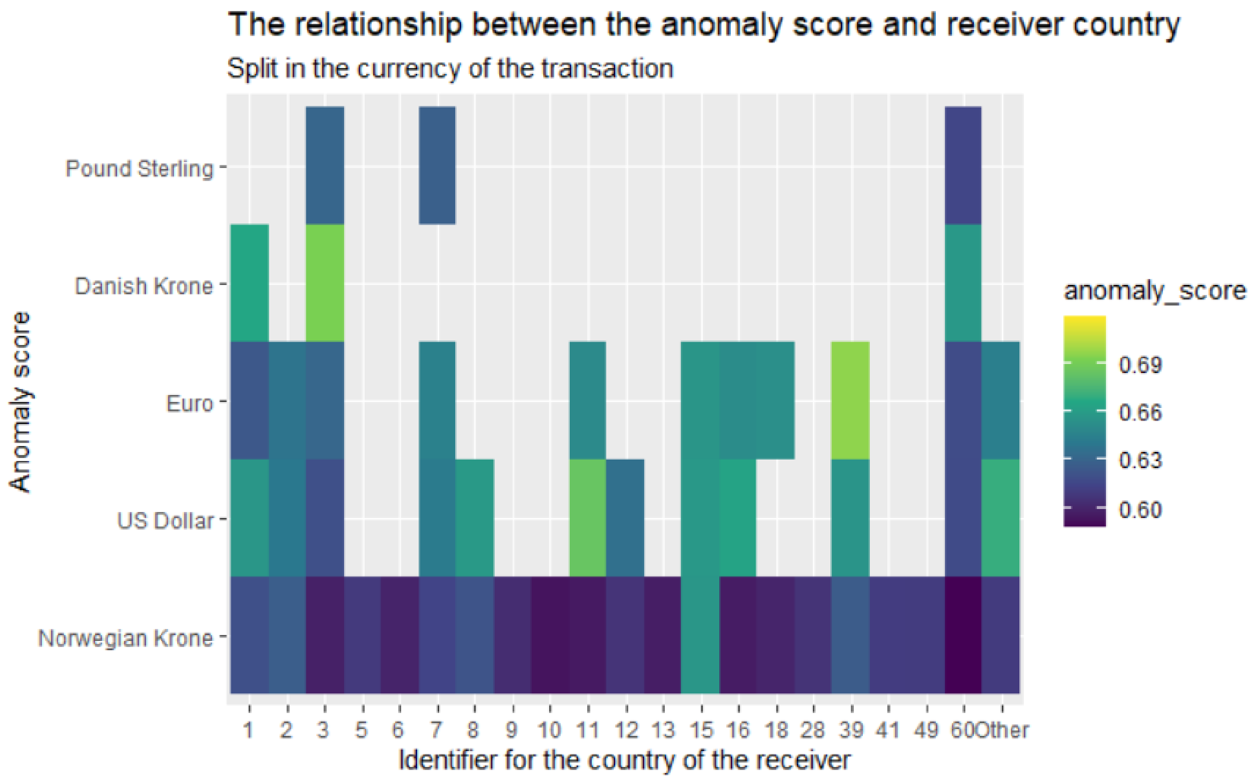
Anomaly detection model

Visualize some of the relationships



The relationship between the anomaly score and currency

Sort by anomaly score

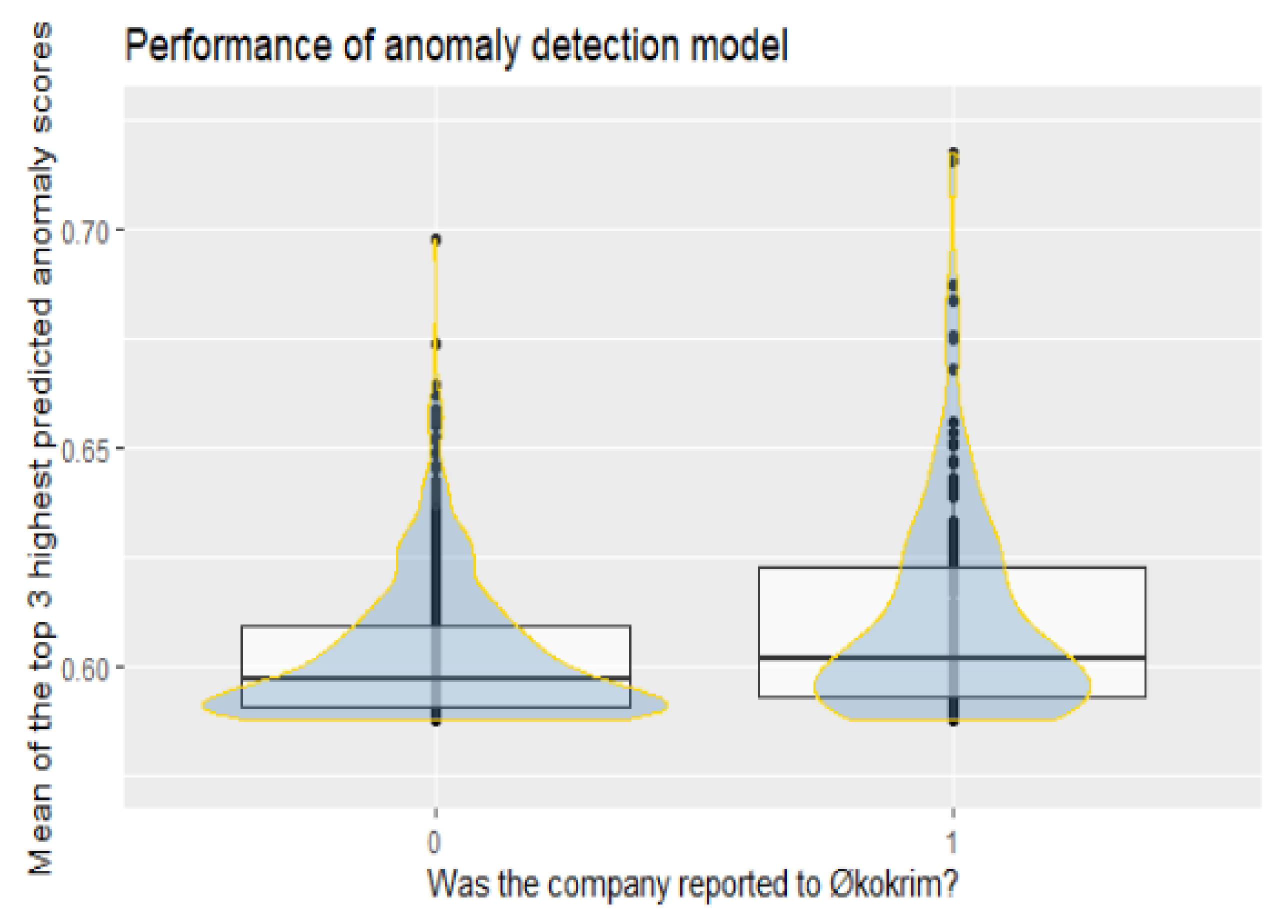


The relationship between the anomaly score and receiver country

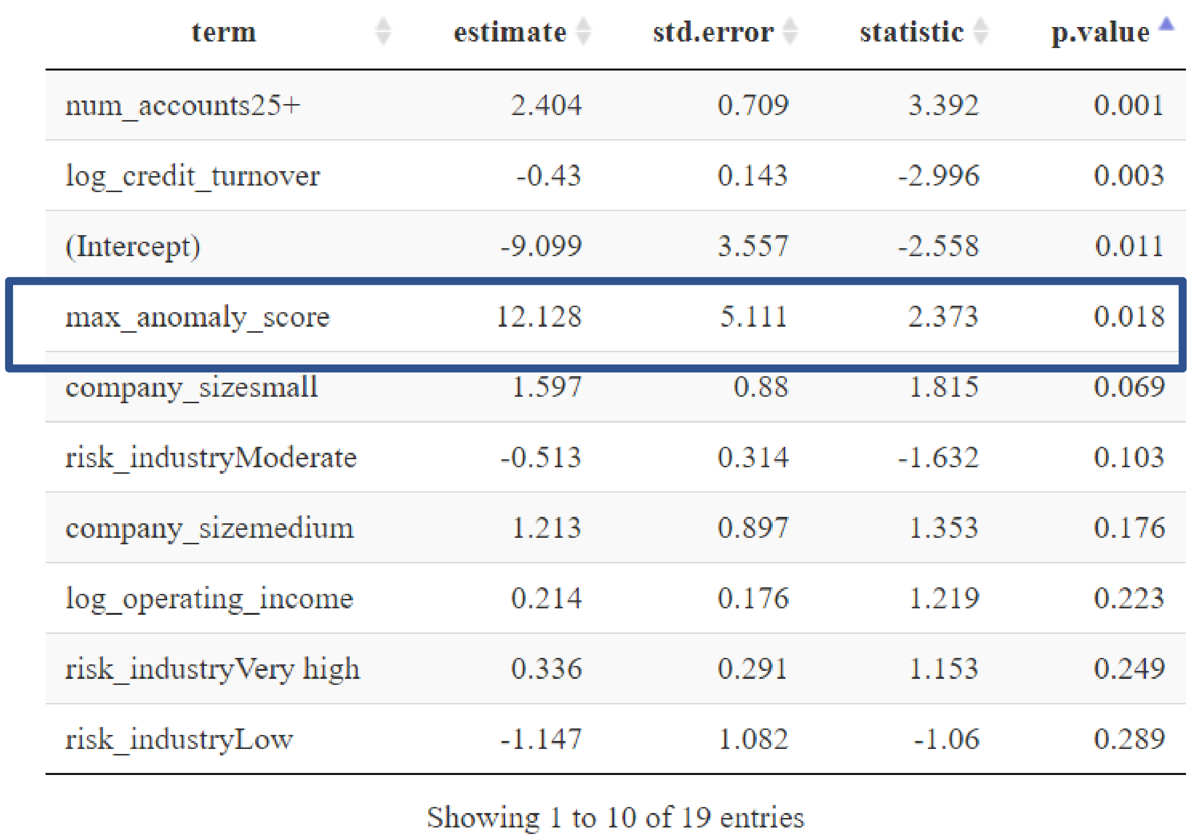
Split in the currency of the transaction

Aggregated anomaly score

Company ID	Top 3 anomaly score	Maximum anomaly score
1	0.614	0.620
2	0.592	0.593
3	0.622	0.627
4	0.647	0.657
5	0.588	0.588
6	0.608	0.610
7	0.602	0.609
8	0.602	0.608
With 687 more rows		



Combine linear logistic regression with aggregated anomaly score



term	estimate	std.error	statistic	p.value
num_accounts25+	2.404	0.709	3.392	0.001
log_credit_turnover	-0.43	0.143	-2.996	0.003
(Intercept)	-9.099	3.557	-2.558	0.011
max_anomaly_score	12.128	5.111	2.373	0.018
company_sizesmall	1.597	0.88	1.815	0.069
risk_industryModerate	-0.513	0.314	-1.632	0.103
company_sizemedium	1.213	0.897	1.353	0.176
log_operating_income	0.214	0.176	1.219	0.223
risk_industryVery high	0.336	0.291	1.153	0.249
risk_industryLow	-1.147	1.082	-1.06	0.289

Showing 1 to 10 of 19 entries

- Adding the aggregated anomaly score variable to the original linear logistic regression model, the ROC AUC on testing set improves from 0.6694 to 0.7178.

Handle large amounts of data

- After reading data to R studio, select the columns and rows that are necessary for the task and save the useful part of the data as RData file.
- When working with the data, take a random sample of the data and use the sample to work through the problem before fitting a final model on all the data.
- If possible, upgrade the computer with more memory; if not, use cloud service like Amazon Web Services.

Agenda

- Project description
- Whether a company is suspicious
- Whether an individual transaction is suspicious
- Conclusion

Conclusion

- Both linear logistic regression model and XGBoost model have good performance in predicting the probability of a company being report. The XGBoost is slightly more accurate.
- The anomaly detection model can separate actual reported cases from non-reported cases historically.
- Combining supervised model and unsupervised model improves the model performance of the supervised model.