IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

Predicting foreign investors' behavior and flows projection in Indonesia government bonds market using machine learning[1]

Anggraini Widjanarti, Arinda Dwi Okfantia and Muhammad Abdul Jabbar,
Bank Indonesia

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

Predicting Foreign Investors' Behavior and Flows Projection in Indonesia Government Bonds Market Using Machine Learning

Anggraini Widjanarti[1], Arinda Dwi Okfantia[2], Muhammad Abdul Jabbar[3]

Abstract

Capital flows is one of the factors that greatly affect exchange rate stability in emerging markets. In Indonesia, the share of foreign investor ownership in government bonds market is large and increasing, signifying the importance of analyzing the behavior of foreign investors in government bonds market. This study aims to explain and predict the behavior and capital flows of individual foreign investors. We apply various machine learning techniques on daily data of government bonds transactions by foreign investors, combined with macroeconomic and market indicators. Investors are clustered using a data-driven algorithm based on their portfolio management, that is real money investors (long term) or traders (short term). For both investor groups as well as 35 investors with the largest share of government bond ownership, we build two additional sets of machine learning models: decision trees that explain each investor's behavior, and predictive models for the capital flows decision (net sale/net buy/hold) and the flows amount on a daily basis. The preliminary result show that the models have potential to support monetary operation strategy based on the direction of investors' decision.

Keywords: government securities, investors' behavior, flows projection, machine learning

JEL classification: C02, F34, E58

[1] Statistics Department – Bank Indonesia; E-mail: anggraini_widjanarti@bi.go.id

[2] Statistics Department – Bank Indonesia; E-mail: arinda_dwi@bi.go.id

[3] Statistics Department – Bank Indonesia; E-mail: muhammad_abdul@bi.go.id

Contents

1. Background

The movement of global capital flows has 2 main driving factors, namely: push and pull. Push factors are characterized by global macroeconomic conditions, central bank monetary policy, international financial market asset returns, and global liquidity. Meanwhile, pull factor capital flows from/to a country are determined by domestic macroeconomic conditions, perceived risk, and returns on domestic assets. In line with the conducive global push factor accompanied by the maintained Indonesian pull factor, the flow of foreign funds to Indonesia, especially to government bonds market (SBN) increased quite significantly in 2019. This condition aligns in many countries, which foreign holders make up the largest share of the investor base (Andritzky, 2012).

The increase in inflows to SBN on the one hand had a positive impact on the external balance in the context of deficit financing. The current account is getting wider. However, the increasing position of foreign investors in SBN has the potential to cause volatility in capital flows and exchange rates, which in turn disrupts economic stability (Agung & Darsono, 2012).

In line with high capital flows to Indonesia which will impact exchange rate volatility, it is necessary to understand the behavior of individual foreign investors in government bond market by classifying investors based on their portfolio management behavior, such as real money (long term) or trader (short term). By mapping foreign investors, we can see which groups of foreign investors are dominant and sensitive to financial market sentiment, so central bank can formulate more precise policy responses. In line with the increasing number and dynamic behavior of foreign investors in government bond market[1], it is necessary to calibrate the classification of investors with an enhanced methodology through data-driven techniques of Big Data Analytics. Big Data Analytics also have good potentials to predict the behavior of foreign individual investors in various scenarios of economic indicators or financial markets. The goal of this study is to develop methodology of Big Data Analytics that hopefully is able to strengthen the analysis of foreign investor behavior and to recalibrate foreign investor behavior classification.

2. Literature Review

2.1 Identification and Grouping of Foreign Investors

Identification of Unique Foreign Investors with Entity Resolution

One of the issue that we found when we explore the raw transactions that we have on our transaction database is that the investor name doesn't fully represents a single entity of foreign investor that we need. Before we can do any analytics with the transactions data, we have to figure out how to identify different names as single entity of the investor of interests that we want to analyze.

[1] In 2019, several global bond indexes such as Bloomberg Barclays (BBGA), FTSE Russell (WGBI), and JP Morgan (GBI –EM) increased China's weight in the benchmark index, and the Norges Pension Fund which lowered its portfolio in EM debt which had an impact on capital flows to Indonesia.

Entity resolution (ER), the problem of extracting, matching and resolving entity mentions in structured and unstructured data, is a long-standing challenge in database management, information retrieval, machine learning, natural language processing and statistics (Getoor & Machanavajjhala, 2013). Entity resolution is necessary when there is clear indication that the entity that we have on our dataset doesn't necessarily reflect the needs of the big data analytics goals. Entity resolution can be done using several methodology from natural language processing to clustering. Measuring text similarity to group similar names together is one of the methodology to do entity resolution for text based entities. We experiment with 4 string similarity metrics in our study:

Jaro-winkler distance

Jaro–Winkler distance is a string metric measuring an edit distance between two sequences. It is a variant proposed in 1990 by William E. Winkler of the Jaro distance metric. Jaro-Winkler is computed by measuring Jaro distance and apply length of common prefix and constant scaling factor.

Normalized Levenshtein Distance

Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other. It is named after the Soviet mathematician Vladimir Levenshtein, who considered this distance in 1965. Normalized Levenshtein Distance based on a study by Yujian & Bo in 2007 to make a normalized version of Levenshtein distance that can mathematically satisfy Triangle Inequality.

Weighted Levenshtein Distance

Weighted Levenshtein Distance or Damerau-Levenshtein Distance is a string metric for measuring the edit distance between two sequences. Informally, the Damerau–Levenshtein distance between two words is the minimum number of operations (consisting of insertions, deletions or substitutions of a single character, or transposition of two adjacent characters) required to change one word into the other. The Damerau–Levenshtein distance differs from the classical Levenshtein distance by including transpositions among its allowable operations in addition to the three classical single-character edit operations (insertions, deletions and substitutions) (Levenshtein, 1966).

Metric Longest Common Sub-sequences (MLCS)

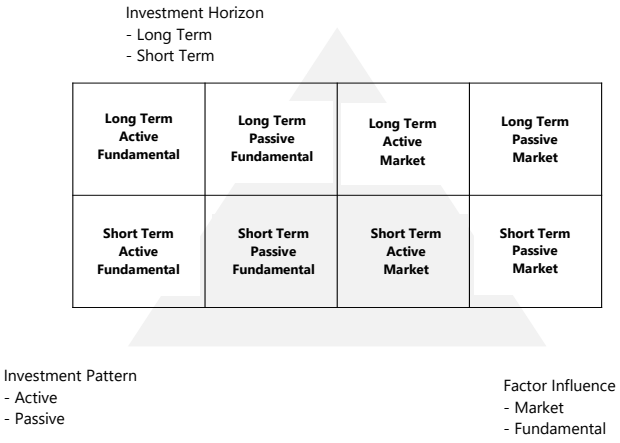
The longest common subsequence (LCS) problem is the problem of finding the longest subsequence common to all sequences in a set of sequences (often just two sequences). MLCS is a metric based on the LCS problem proposed by Bakkelund in 2009. MLCS as a metric have the properties of Positive Definiteness, Symmetry, and Triangle Inequality.

Grouping of Foreign Investors

Composite Index

The classification used previously divides investors into two groups, namely long term and short term investor group. The groupings are based on three different dimensions, namely: **investment horizon**, **transaction frequency**, and **transaction volume** (Indawan, Fitriani, Permata, & Karlina, 2013). In its use, foreign investors with short term classification are identified with investors who have a short investment horizon, high transaction frequency, high transaction volume, and tend to be influenced by market movement indicators (Hoffmann, Shefrin, & Pennings, 2010). Meanwhile, foreign investors with the long term classification are identified with investors who have a long investment horizon, low transaction frequency, low transaction volume, and tend to be influenced by economic fundamental indicators.

Foreign Investor Classification — Figure 1



Investment Horizon

Classification of investors based on the investment horizon is done by using the transaction ratio calculation approach derived from the calculations of Lakonishok, Shleifer, & Vishny (1992).

First Equation: Transaction Ratio

$$Transaction\ Ratio(t) = \frac{|Rpbuy(t) - Rpsell(t)|}{Rpbuy\ (t) + Rpsell\ (t)}$$

Where:
Rpbuy(t) = Buy Value (Rp) SBN on t period
Rpsell(t) = Sell Value (Rp) SBN on t period

A high ratio (close to 1.00 ratio) can explain that investors have a tendency to carry out all or most of their transactions to increase (net buy) or reduce (net sell) ownership of a financial asset in period (t). Conversely, a low ratio (close to the 0.00 ratio) can explain that transactions made by investors have little or no effect on changes in ownership of a financial asset in period (t) (buy for resale and vice versa).

Threshold First Equation:

1 – Transaction Ratio > 0,5: Short Term Investor
1 – Transaction Ratio < 0,5: Long Term Investor

The problem with this calculation is that if the investor does not make any transactions during the observation period, the number is zero. Zero transaction ratio number within the threshold will be categorized as Short Term Investor [1 – Transaction Ratio (0)] = 1. In the future, it is necessary to revisit the equation and threshold to determine the classification of investors based on the investment horizon.

Transaction Frequency:

In addition to the investment horizon, investor classification is also determined from the frequency of transactions which is calculated using the average transaction day per year with the following formula.

Second Equation: Average Transaction Day Per Year

$$Average\ transaction\ day\ per\ year = \frac{\Sigma_{i=1}^{n}\ transaction\ day\ (t)}{\Sigma_{i=1}^{n} year\ (t)}$$

Where:
transaction day (t) = transaction day on t period
year (t) = sum of year on t period

Threshold Second Equation:
Average transaction day per year > 0,2: Short Term Investor
Average transaction day per year < 0,2: Long Term Investor

The threshold assumption of 0.2 from equation 2 is taken from the calculation of 48 days/240 days. A 48-day approach is assumed with 4 primary SUN auctions per month and 12 months per year. Investors are expected to adjust their SUN portfolio at least 48 days for 240 working days. If the number of transactions is more than 48 days during a year, it is assumed that the investor is a short term investor. The problem with this calculation is the assumption of 48 times in 1 year based on professional judgment.

Transaction Volume:

In this study, the need for investor portfolio rebalancing is also added with the following formula.

Third Equation: Transaction Volume

$$Transaction\ Volume = \frac{\Sigma_{i=1}^{n}\left\{\left(Rp\ Buy(t) + Rp\ Sell(t)\right)/Position(t)\right\}}{n}$$

Where:
Rpbuy (t) = Buy Value (Rp) SBN on t period
Rpsell (t) = Sell Value (Rp) SBN on t period
Position (t) = Ownership of SBN Position on t period
n = number of days

Threshold Third Equation:
Transaction Volume > 0,05: Short Term Investor
Transaction Volume < 0,05: Long Term Investor

The calculation assumption of equation 3 is the assumption of investors' need for asset rebalancing. The asset rebalancing threshold is 5% of the total portfolio. If the threshold is > 5%, it is assumed that the investor is short term in line with the frequent rebalancing of the portfolio to the total SUN portfolio in one transaction. The assumption of 0.5% rebalancing of the total portfolio is derived from professional judgment.

Composite Indicator for Investor's Classification (Short Term and Long Term):

Fourth Equation: Composite Index

$$Investor\ Type = \left[1 - \frac{|\sum_{t=1}^{T} Net\ Volume_t|}{\sum_{t=1}^{T}(Rp\ Buy + Rp\ Sell)_t}\right] x \left[\frac{\sum_{i=1}^{n} transaction\ day\ (t)}{\sum_{i=1}^{n} year\ (t)}\right] x \left[\frac{\sum_{i=1}^{n}\{(Rp\ Buy(t) + Rp\ Sell(t))/Position(t)\}}{n}\right] x100$$

Threshold Investor Type:
Investor Type > 0,05: Short Term Investor
Investor Type < 0,05: Long Term Investor

The total composite calculation of 3 indicators (horizon, frequency, and transaction volume) uses multiplication and is not weighted[2]. From the result of this composite index, the majority of individual investor composite values are close to zero so that the classification of investors is mostly long term. In the future, it is necessary to improve the threshold and calculation method to increase the number of short-term investors.

Clustering Methodology

We have the composite index as our benchmark to classify the investors. Next, in this study, we explore the use of clustering methodology to map investor behavior that is carried out without class information and annotations on which investors are short term and long term. Clustering can be done to group based on the proximity between the data according to various characteristics of the data.

Clustering is a machine learning (unsupervised learning) method that can group data points into groups based on the similarity and proximity of the data points. Clustering is usually used to see the group structure in the data without labelling or annotating the dataset. Das (2003) uses k-means clustering to classify hedge-fund investors based on their investment strategy and style. Validation for clustering can be done by using the Silhouette Coefficient[3] calculation to calculate the intra-cluster and inter-cluster distances and also using the average transaction frequency results from both short term and long term investor groups. In this study we use K-Means Clustering.

[2] The composite calculation of three indicators with multiplication actually produces a number that gets smaller and closer to zero as the value of each indicator is zero. The composite value that is getting closer to zero will result in a long term classification

[3] Silhouette coefficient is a measure that can be used to evaluate whether clustering results are good. The Silhouette Coefficient is calculated by using the proximity between points in one cluster and its cluster members and also calculating the proximity between points in one cluster and other cluster members. A well-separated cluster is a cluster with the distance between points in its adjacent cluster of members and the distance between the outer points of the cluster of members who are far apart.

K-Means Clustering

The k-means algorithm is an iterative algorithm that partitions the dataset into exclusive clusters with a predetermined number of clusters. This algorithm separates the cluster until it succeeds in finding a number of cluster centers that separate the clusters by the distance between the farthest cluster center and the distance between the closest cluster members with the closest distance until it becomes a single cluster containing all data points.

2.2 Behavior Modelling

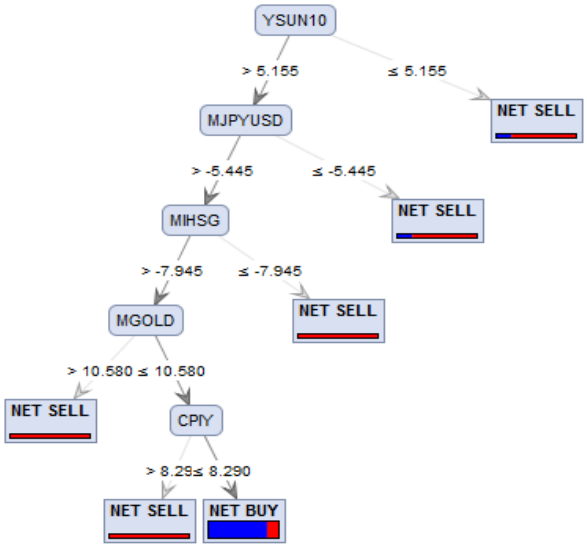
There are many fundamental and market factors that may influence investors in making decisions. To select the factors/variables that most influence investors in making buying/selling/hold decisions, and to avoid overfitting the model, a feature selection methodology is necessary. By using feature selection methodology, we are able to identify those attributes that best describe how investors decide to buy or sell their positions in an objective and statistically correct manner (Silva, Tabak, & Ferreira, 2019).

Our feature selection process gives us an objective way of identifying important variables that should be accounted to be put on forecasting model for flows projection. Every investor will have different important variables. With those important variables is the foundation for us to build model for flow projection.

Decision Tree

In this study, we use data-driven machine learning methods to do feature selection. We use decision tree algorithms for this part of the methodology. One of the great features of decision tree algorithms is that they inherently estimate a suitability of features for separation of objects representing different classes (Grabczewski, & Jankowski, 2005). The Decision Tree algorithm will sort the variables based on the largest information gain and will eliminate variables that have no effect on the investor decision (information gain is close to 0). We use 119 variables that consist of market and fundamental variables as an input for this process.

Decision Trees Illustration — Figure 2



We also apply Random forest and XGBoost algorithm as more advanced algorithms of decision trees algorithms that have feature importance calculation that we can use to measure the variable importance.

Random Forest

Random forest is a machine learning technique that can be used to solve regression and classification problems. Due to the random exploration of features, Random Forest lends itself to feature selection well and the measure of feature importance adopted here is the average information gain achieved during forest construction (Rogers, & Gunn, 2006). It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the overfitting of datasets and increases precision.

XGBoost

XGBoost is a scalable tree boosting system that is widely used by data scientists and provides state-of-the-art results on many problems (Chen & Guestrin, 2016) .XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework that has some notable features:

- Clever penalization of trees
- A proportional shrinking of leaf nodes
- Newton Boosting
- Extra randomization parameter
- Implementation on single, distributed systems and out-of-core computation
- Automatic Feature selection

2.3 Flows Projection

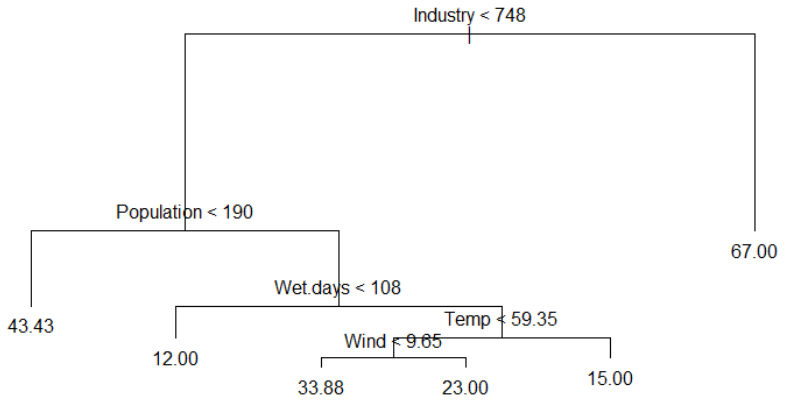
To build model for flow projections, we use Machine Learning and Big Data Analytics techniques can be used to generate capital flows projections with better precision. Here are some algorithms that can be used to perform time series projections:

Regression Tree

A regression tree is similarly a tree-structured solution in which a constant or a relatively simple regression model is fitted to the data in each partition (Loh, 2014). Regression tree is a class of decision tree algorithms that make decision trees made from observations of various variables with the final result in the form of predictive continuous numbers. Regression tree is created through a binary recursive partitioning process which divides data iteratively into partitions and branches based on the value of the existing data set.

Regression Trees Illustration

Figure 3

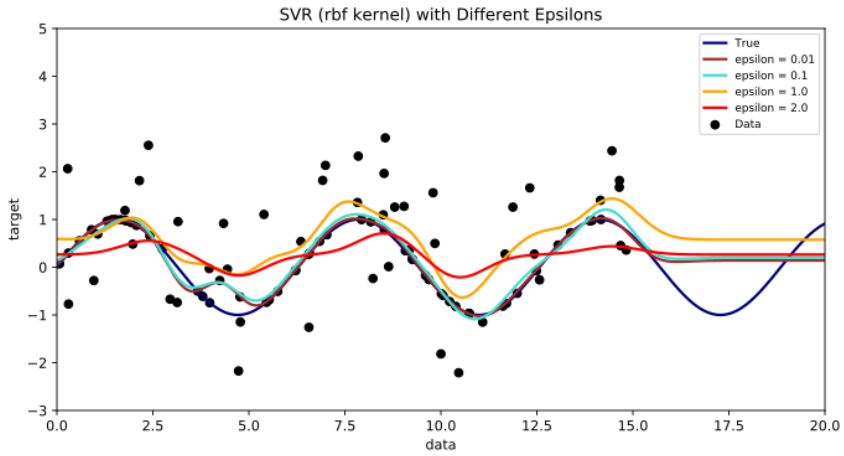


Support Vector Regression

Support vector regression is a variation of the support vector machine algorithm that creates a hyperplane function that can classify data in the function scope. Support-vector network implements the following idea: it maps the input vectors into some high dimensional feature space Z through some non-linear mapping chosen a priori (Vapnik & Cortes, 1995). Support vector regression can produce functions with non-linear constraints. Support Vector Regression also uses the parameter as a threshold value of how far the prediction result is from the original predicted value. The result of the support vector regression is in the form of a hyperplane function that gives a range where data values can be in the next data period and data projections for the next period.

Support Vector Regression Illustration

Figure 4

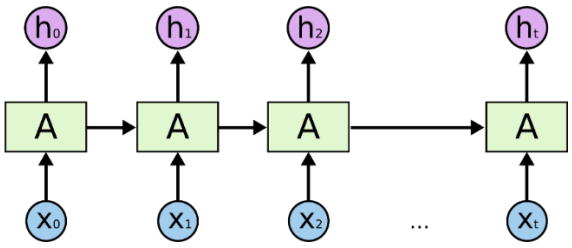


Long-Short Term Memory (LSTM)

LSTM is part of the Deep Learning algorithm class that can perform pattern and sequence recognition well. LSTM is a novel recurrent network architecture in conjunction with an appropriate gradient based learning algorithm (Hochreiter & Schmidhuber, 1997). LSTM can be used to study existing patterns in data, including patterns of changes in a value in time series data such as capital flows, stock prices,

and others. LSTM as part of the Deep Learning algorithm contains activation functions that can recognize patterns from data points with high accuracy even though it requires large amounts of data and high computational complexity. The LSTM artificial neural network as illustrated below is given input in the form of time series values at a time, then learns the pattern of values based on the values in the previous sequences of time.

LSTM Illustration Figure 5



2.4 Model Interpretation

Machine learning models are often considered as a black box. To understand the complexity of machine learning models, we need to apply model interpretability methodology to verify whether the model is in line with what our goal is.

In this study, we will use Local Interpretable Model-agnostic Explanations (LIME) as a method for interpreting machine learning models by using a simple model approach at a point of observation. Instead of training a global surrogate model, LIME focuses on training local surrogate models to explain individual predictions (Ribeiro, 2016). LIME learns what happens to predictions when given variations of data into a machine learning model. Then LIME generates a new dataset consisting of the correct and error samples according to the machine learning model results. On this new dataset, LIME then trains an interpretable model, which is weighted based on the proximity of the sample instance to the desired instance.

3. Methodology

3.1 Data

In this study, the daily net value of buy and sell transactions that indicate foreign investors' decision in the government bond market is used as the target variable. Specifically, the net transaction is flagged as “net buy” if it is positive (>0), “net sell” if it is negative (<0), and “hold” if it is equal to zero.

This study considers several market indicators collected from foreign exchange market, money market, bond market, commodity market, and stock market.

Market and Fundamental Indicators Example

Table 1

FX Market	Money Market	Bond Market	Commodity Market	Domestic Fundamental
1. Indonesia Rupiah Exchange Rate	1. Overnight Interbank Rate	1. Yield SUN Secondary Market	1. Oil WTI	1. BI Policy Rate
2. NDF 1 M onth Exchange Rate	2. LIBOR USD	2. IDMA	2. Gold Spot Price	2. GDP
3. USD Index	3. LIBOR-OIS USD 1M Spread	3. HSBC Asia Local Bond Index	Stock Market	3. Inflation Rate
4. EURUSD	4. Bloomberg Financial Conditi on Index US	4. HSBC Asia Dollar Bond Index	1. DJI Index	4. Trade Balance
5. USDJPY	5. Bloomberg Financial Conditi on Index EU	5. JP Morgan Indonesia Total Bond Return Index	2. Stoxx 600 Index	5. CA Balance
6. ADXY	6. Bloomberg Financial Conditi on Index Asia	6. Yield UST-SUN 2Y	3. Nikkei Index	6. Fiscal Budget
7. REER		7. Yield Spread Bund PIIS	4. MSCI Asia Index	7. Foreign Reserve
		8. CDS PIIS	5. MSCI EM Index	Global Fundamental
		9. Yield Spread 10Y-2Y	6. IHSG	1. FFR
		10. CDS 5Y Indonesia		2. ECB Rate
				3. BOJ Rate

Moreover, the domestic and global fundamental indicators such as central bank policy rate, GDP, inflation rate, etc., also used to comprehend the effect of macroeconomic factors on investment decisions. The data used in this research was obtained from Bank Indonesia – Scripless Securities Settlement System (BI-SSSS) and Bloomberg from January 2016 to May 2020. BI-SSSS is a database that serves as both securities depository in the form of electronic registry and provider of securities settlement services and is directly connected between participants, operators, and the Bank Indonesia - Real Time Gross Settlement System (BI-RTGS System).

The data are split into two: training dataset (January 2016 - December 2019) and testing dataset (January - May 2020). In summary, this study uses 119 features, including 108 market indicators and 11 fundamental indicators as the independent variable.

We filter the transactions to include only the investor of interests which are:

1. Top 30 investors based on their ownership of government bonds in the market.
2. 5 investors in Indonesia investment focus group.

These investors accounted for 65% of the foreign investor's ownership in Indonesia government bonds market in December 2019.

We apply lag to the market and fundamental indicators from 1 day to 14 previous business days to factor the influence of previous day's indicators to the transaction settlement that captured in our database. We also feature engineer the variables value to get the periodical changes of the variables. The periodical changes that we feature engineered are as follows:

- Year to date (ytd)
- Quarter to date (qtd)
- Month to date (qtd)
- Daily changes (dΔ).

3.2 Entity Resolution

There is an issue where single investor entity can have multiple Single Identification Identifier (SID) in the database. Therefore, we need to group SID with similar names as single entity.

SID Investor Names Examples — Table 2

SID Investor Name	Investor Name
CSTDBK1 INVESTOR-A BOND FUND	INVESTOR-A
CSTDBK2 INVESTOR-A STRTEGY PLUS	INVESTOR-A
CSTDBK3 INVESTOR-A GL-MUL BND FUND	INVESTOR-A
CSTDBK2 INVESTOR-B LCL DEBT INDEX PRTF	INVESTOR-B
CSTDBK3 INVESTOR-B SBP	INVESTOR-B
INVESTOR-B GLBL ALLOC FND	INVESTOR-B

We develop entity resolution models using string similarity metrics to group investor names from multiple SIDs into a single entity. First, we study the pattern of names from a sample of the investor names in the database. Then we separate the raw names into the custodian bank names, investor names, and other texts. Then we label the raw investor names to the investor real names and produce a small database of investor names labelled with their real names. Finally, we do a horse race of 4 string similarity metrics to match the raw investor names to labelled investor names that we have in the database and use the best model as the entity resolution model to group similar SID names as single investor entities. The string similarity metrics that we experiment with are as follows:

1. Jaro-Winkler Distance
2. Normalized Levenshtein Distance
3. Weighted LevenShtein Distance
4. Metric Longest Common Subsequence (MLCS)

3.3 Grouping of Foreign Investors

One of our goals is to create a grouping of investors into long term (LT) investor or short term (ST) investor based on its yearly activities and behavior in Indonesia's government bonds market. Using the components of the composite index mentioned

in 2.1.1 which are Investment Horizon, Transaction Frequency and Transaction Volume, then using the entity resolution result, and K-means clustering, we create clusters of investors based on their behavior and activities in the government bond market. We tried 2-6 number of clusters in our clustering methodology. Then, we calculate silhouette coefficients to measure the clusters quality with different number of clusters.

3.4 Behavior Modelling

For each individual investor we use decision tree algorithms feature importance to filter the important variables to be used for the investor decision model and projection flows model. This is done to avoid overfitting, minimize noises and redundant data, and improve the performance of the investor decision and flows projection model. We use decision dree algorithms feature importance which are Decision Tree, Random Forest and XGBoost models and information gain measures to calculate the important features.

We experiment with inherent lags in the model from 1 to 5 days to accommodate possible delays that may occurred from the time the investor get their information to the time of the transaction settled in the database. We evaluate the model using F1 scores with the variation of inherent lags and algorithms for each of the individual investors. We didn't use the model for prediction, but to help decide which of the lags and variables that produced the best F1 score. The lags and variables produced by the best model then used again as one of the input for investor decision prediction and flows projection model that used wider varieties of algorithms, pre-processing, and experiment.

3.5 Investor Decision Prediction and Flows Projection

We develop investor decision prediction model using classification machine learning algorithms to predict daily individual investor decision of whether the investor will have a net buy, net sell, or hold decision in the corresponding day. The algorithms that we use are Logistic Regression, SVM, KNN, Decision Tree, Random Forest, XGBoost, and LSTM. Before modelling we apply pre-processing techniques such as PCA, lag adjustment, and variables adjustment to see which option produced the best result from using all of the variables, only the important variables or using PCA transformed variables. The model then evaluated using F1 scores. The model that produces the best F1 scores for each individual investors then used to help the flows projection model.

The flows projection model experiment design is similar to the investor decision model with different algorithms and different dependent variable. We use regression machine learning models such as Logistic Regression, KNN, Regression Tree, SVR, LSTM, and XGBoost. In the flows projection model, we use daily transaction nominal capital flows of each investor as the dependent variable. We evaluate the flows projection model using Mean Average Error (MAE) regression error metric.

We use the investor decision model to help the flows projection model. If the investor decision model prediction is a hold, then the flows projection model will calculate the flows as 0 for the day. If it's net buy then the flows projection model will project positive values, and vice versa for net sell decision.

3.6 Model Interpretation

In order to conduct further analysis, it is very important to interpret the model to find out what influences investors in making decisions. However, the machine learning models are mostly black box models which tend to be either difficult or impossible to interpret.

We use LIME to interpret our machine learning model decision. We apply LIME on several random date on the testing period to see what variables affect the model to predict the direction of investor decisions. We chose top 10 (ten) most influential (highest weight) variables that drive the prediction of the model. We apply LIME to all the best models for each investor.

4. Result & Analysis

4.1 Entity Resolution

The models that have been trained in the previous steps need to be evaluated in order to measure their accuracy in predicting each target class. We use F1-score as the metric for entity resolution and prediction model evaluation, in order to get a balanced classification model with the optimal balance of recall and precision.

The results of evaluation for best entity resolution models using 4 string similarity metrics and a threshold from 0.90 to 1.00 shown in Table 2. The best model is obtained using the Jaro-Winkler metrics with threshold of 0.97, which produce F1 score of 87%. We also evaluate using 500 randomly taken out of sample data, and produced good accuracy results of 89%. Therefore, we believe that the model is robust enough to be implemented in all data in January 2014 - May 2020 period. From 4.215 unique SIDs and unique investor names in the Indonesia Government bond transaction data on period January 2014 – December 2020, we get 1.846 unique investors from the entity resolution results.

String Similarity Metrics Evaluation — Table 3

String similarity Metrics	Threshold	F1
Jaro Winkler	0.97	**87%**
Normalized Leveinshtein	0.96	79%
Weighted Leveinshtein	0.96	82%
MLCS	0.85	73%

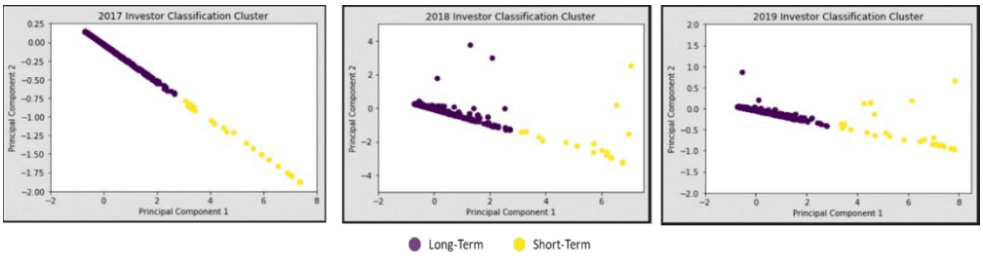
Note: Blue-shaded cells denote the best result for each model

4.2 Grouping of Foreign Investors

The clustering model is able to group investors well into short term (ST) investors and long term (LT) investors, with a Silhouette Coefficient of 0.89. These results are also in line with the grouping of investors with the expert judgement done by the Monetary Management Department using the Composite Index. Clustering result that we use as the grouping is the 2019 investor transactions data, with 1.075 investors grouped as long term investors and 35 investors grouped as short term investors.

Grouping of Foreign Investor Cluster Results

Figure 6



4.3 Behavior Modelling

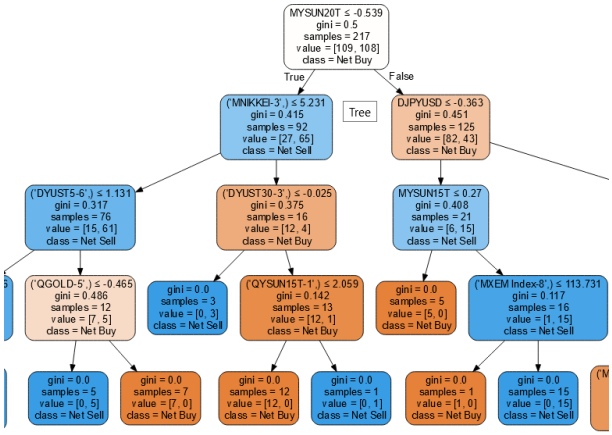
We use behavior analysis models to find the indicators that influence the decision of each investor of interests, and each the investor groups (short term and long term). Our findings are as follows:

1. We find that for both of the investors group, Indonesian Bonds Yields in different maturity are considered important with short term investors group considers shorter maturity of government bonds.
2. JAKCONS which is Jakarta Consumer Goods stock index is important for both of the investors group.
3. Short term investors group are affected by more daily and high frequency indices while long term investors group are highly affected by a fundamental indicator (Indonesia YoY Core Inflation).

Furthermore, the behavior model result will be used as selected important features for decision prediction and flows projection model.

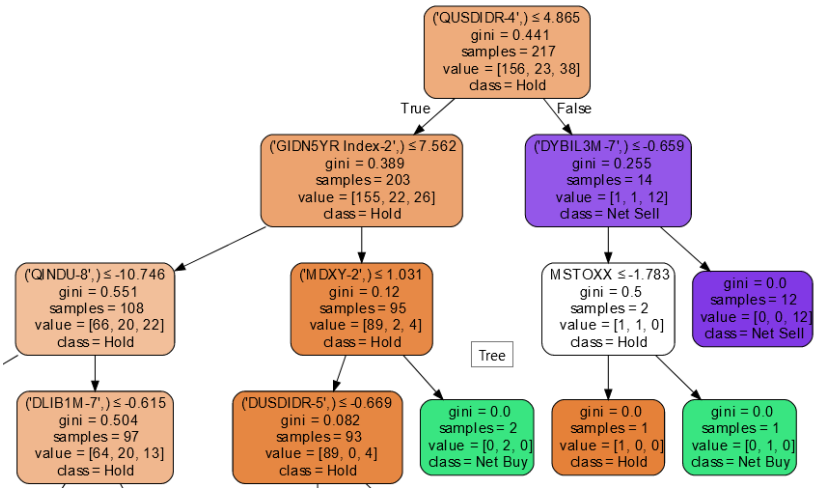
Behavior Model Single Decision Tree Short Term Investor Example

Figure 7



Behavior Model Single Decision Tree Longer Term Investor Example

Figure 8



Behavior Model Top 10 Important Variables for Investors Group

Table 4

Short Term Investors		Long Term Investors	
Ticker	Description	Ticker	Description
IDR Currency	Rupiah Currency	JASFXBAT Index	Indonesia Stock Capital Flows
HSITR Index	HSBC Asia Dollar Bond Index	FSSTI Index	Straits Times Index
SENSEX Index	S&P Bombay Stock Exchange	MGIY10Y Index	Malaysian 10-year Government Bond Yield
USGGBE10 Index	US Breakeven 10 Y	INR Currency	Indian Rupee Currency
GIDN12YR Index	Gov. Bonds Generic Yield 12 Years	GIDN30YR Index	Yield SUN Generic 30 Year
JAKINFR Index	Jakarta Infrastructure, Utilization and Transportation Stock Index	JAKCONS Index	Jakarta Consumer Goods Stock Index
IDPPON Index	PUAB o/n	IDIFCRIY index	Indonesia Yoy Core Inflation
SXXP Index	STOXX Europe 600 Index	IHNI1M Currency	Implied NDF 1M
JAKCONS Index	Jakarta Consumer Goods Stock Index	IBPRTRI Index	IBPA Government Bond Index
MXEM Index	MSCI Emerging Market Index	THB Currency	Thailand Baht Currency

Behavior Model Top 10 Important Variables for Individual Investors Example

Table 5

Example Short Term Investor		Example Long Term Investor	
Ticker	Description	Ticker	Description
EPUCNUSD Index	US Economic Policy Uncertainty Index	GIDN1YR Index	Generic Gov. Bonds Yield 5 Year
GIDN1YR Index	Generic IDN Gov. Bonds Yield 1 Year	IHN+1M Index	IDR NDF 1 Month
ADXY Index	Asian Dollar Index	USGG30YR Index	Generic US Treasury Yield 30 Years
GIDN15YR Index	Generic IDN Gov. Bonds Yield 15 Years	GIDN10YR Index	Generic IDN Gov. Bonds Yield 10 Years
GIDN5YR Index	Generic IDN Gov. Bonds Yield 5 Years	USGG3M Index	Generic US Treasury Bills Yield 3 Month
EURUSD Currency	Euro Currency	USSOA Index	USD Overnight Indexed Swap 1M
BFCIEU Index	Bloomberg Financial Condition Index EU	GIDN5YR Index	Generic IDN Gov. Bonds Yield 5 Years
GIDN10YR Index	Generic IDN Gov. Bonds Yield 10 Years	IDPPON Index	PUAB o/n
USGG3YR Index	Generic US Treasury Bills Yield 3 Years	JCI Index	IDX Stock Composite Index
DXY Index	Dollar Index	EURUSD Currency	Euro Currency

4.4 Investor Behavior Prediction and Projection Flows Model

In this study we develop 35 behavior prediction models for each investor. In addition, we also develop 1 model for the LT investor group and 1 model for the ST investor group. Based on the result, our investor behavior prediction models are able to predict the decision (buy, sell, or hold) made by 18 of 35 investors and the two investor groups with satisfying result (>60% F1 Score). The model for LT investors group has F1 score of 77%, while for ST investors group has F1 score of 64%. While for the individual investor prediction models, the best model has F1-score of 86%, while the lowest F1-score model get F1-score of 47%.

Investor Behavior Prediction Result

Table 6

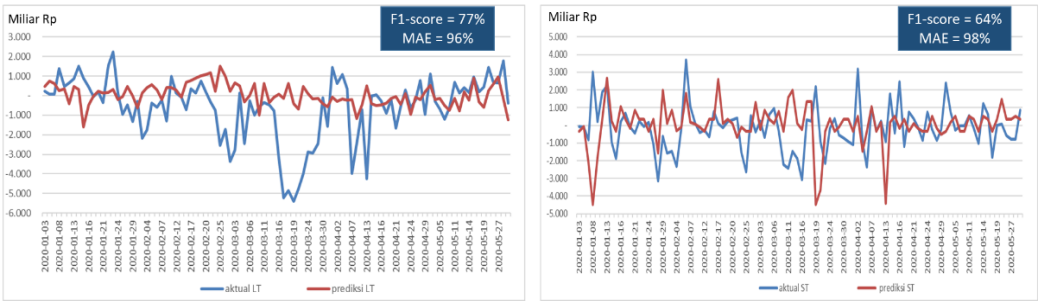
Investor	Algorithms	Historical data (days)	Using PCA	F1 Score
LT investors group	XGBoost	4	Yes	77%
ST investors group	XGBoost	6	Yes	64%
Best individual investor	Regression Tree	2	Yes	86%
Lowest individual investor	XGBoost	9	Yes	47%

Note: The model is each individual investor model with the highest f1 score.

However, the flows projection model's ability to predict the amount of flows still needs to be improved, considering the error (mean absolute error/MAE) is still quite

large. The average MAE for the 37 investors is 93% of the average investor transaction, with the smallest MAE at 50% and the largest MAE being 101% of the average investor transaction.

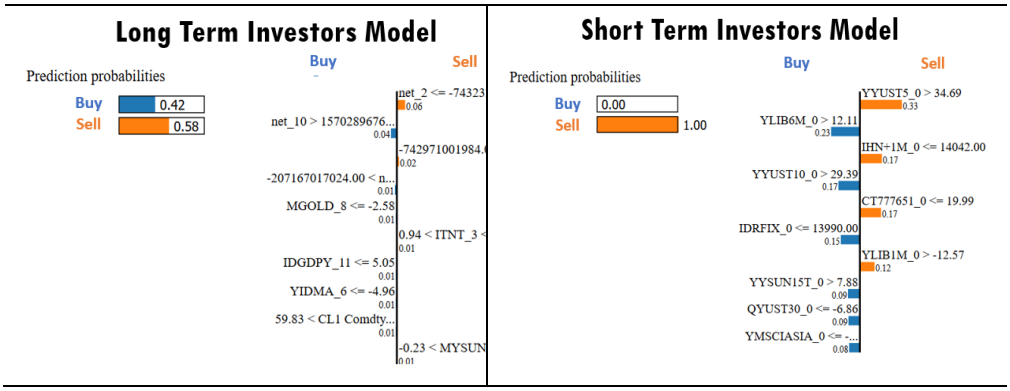
Flows Projection of Long Term and Short Term Result Plot — Figure 9



4.5 Model Interpretation

Based on LIME result on 3rd January 2020, we found out that the prediction model of long-term (LT) investors on that day is more influenced by the decision (net buy) of the investors on previous period. While for the prediction model of short-term (ST) investors, we found out that on the same day the model is more influenced by market variables, such as US Treasury yields and USD LIBOR.

LIME Top 10 Important Variables for Investors Group — Figure 10



5. Conclusion & Future Work

5.1 Conclusion

Firstly, to deal with the issue of the same investor entities that has many different SIDs in the database of government bond transaction that we have, we develop a methodology using string similarity metrics to match similar investor names as single entities. The result is able to match random out-of-sample investor names very well with accuracy of 89%.

Then, we develop clustering model to complete the analysis of the grouping foreign investors into real money investors (long term / LT) or traders (short term / ST), based on their portfolio management behavior with a data driven approach using machine learning. The cluster result has high silhouette coefficient in grouping the investors with similar activities in the government bonds market and matched the grouping using composite index done by Monetary Management Department.

Lastly, we develop prediction model for each investor decisions using market and fundamental data, namely 119 variables and their feature engineered and lag adjusted form with a total of approximately 2.000 variables. We use decision tree algorithms to do feature selection and filter the most influential variables for each individual investor. We then develop investor decision (buy, hold, or sell) prediction model and flows projection regression model using machine learning algorithms. The results are individual investor prediction decision models with 18 investors and 2 investor group's prediction decision models that are able to produce satisfying prediction power (>60% F1 Score). As for the flows projection model, the result is not yet good and still need to be improved.

5.2 Future Work

There are several improvements in the methodology that can be applied for future works.

- Improve the accuracy of the investor decision and flows projection models for all of the individual investors

For some investors, the prediction model that we have is not accurate enough, so it needs to be improved. Improvements can be done by adding longer data (perhaps from 2013) as well as adding other market or fundamentals variables that are considered influential in determining investor decisions.

Furthermore, we can try to do 2-stage classification, considering that for some investors the frequency of holding is much higher than the frequency of buying or selling. The first stage is to predict whether investors will hold or not hold. Furthermore, if the prediction results is not "hold": then stage 2 predictions will predict whether investors will net buy/net sell.

- Develop investor investment prediction and flows projection model that can predict well during abnormal flows period (COVID 19 Pandemic)

During the pandemic period, around 2020 - 2021, the pattern of transactions in the government bonds market, both nominally and transaction frequency, is drastically different from the previous period. Investor behavior was also presumably different from the previous periods. Therefore, we need to include data from COVID-

19 pandemic period into our training data. As an alternative, it also worth to try to separate model in the abnormal period and the normal period. However, this could be quite challenging since the data for the abnormal COVID-19 period is limited to only 1 year of data.

- Develop model automation and dashboard for daily visualization of government bonds daily data and prediction

To use the prediction model to predict the foreign investors' decisions on daily basis, it is necessary to automate the process and disseminate the prediction results. We can try to automate the prediction process of the models and visualize it in a dashboard so that it can be used to analyze foreign investors' behavior in government bond market to support decision making related to monetary operations strategy in timelier manner.

- Develop model for foreign investors in stock and currency market

To complete the analysis of capital flows in Indonesia, the foreign investor behavior prediction models using machine learning can be applied to foreign investor in stock and currency market.

References

Agung, J., & Darsono. (2012). Post-Global Crisis Capital Inflows to Indonesia: Challenges and Policy Responses. SEACEN.

Andritzky, J. R. (2012). Government Bonds and Their Investors: What Are the Facts and Do They Matter? *IMF Working Paper WP/12/158*.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). Das, N. (2003). Hedge Fund Classification using K-means Clustering Method. *Computing in Economics and Finance 2003, 284*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. Hoffmann, A. O. I., Shefrin, H., & Pennings, J. M. E. (2010). Behavioral Portfolio Analysis of Individual Investors. *SSRN Electronic Journal*.

Indawan, F., Fitriani, S., Permata, M. I., & Karlina, I. (2013). Capital Flows in Indonesia: The Behavior, The Role, and Its Optimality Uses for The Economy. *Bulletin of Monetary, Economics and Banking*, 23-54.

Lakonishok, J., Shleifer, A., & Vishny, R. W. (1992). The Impact of Institutional Trading on Stock Prices. *Journal of Financial Economics, 32*, 23-43.

Levenshtein, V. I. (1966, February). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, No. 8, pp. 707-710).

Loh W.-Y. (2014). Classification and Regression Tree Methods. Wiley StatsRef: Statistics Reference Online.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should I trust you?: Explaining the predictions of any classifier." 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

Rogers, J., & Gunn, S. (2005, February). Identifying feature relevance using a random forest. In *International Statistical and Optimization Perspectives Workshop*" *Subspace, Latent Structure and Feature Selection*" (pp. 173-184). Springer, Berlin, Heidelberg. Silva, T. C., Tabak, B. M., & Ferreira, I. M. (2019). Modeling Investor Behavior Using Machine Learning: Mean-Reversion and Momentum Trading Strategies. *Complexity*, 1-14.

BANK INDONESIA

PREDICTING FOREIGN INVESTORS' BEHAVIOR AND FLOWS PROJECTION IN INDONESIA GOVERNMENT BONDS MARKET USING MACHINE LEARNING

Anggraini Widjanarti, Muhammad Abdul Jabbar, Arinda Dwi Okfantia
Statistics Department – Bank Indonesia
Email: anggraini_widjanarti@bi.go.id, muhammad_abdul@bi.go.id, arinda_dwi@bi.go.id

OUTLINE

BACKGROUND AND GOALS

FRAMEWORK

METHODOLOGY

RESULTS AND ANALYSIS

CONCLUSION AND FUTURE WORKS

BACKGROUND AND GOALS

Analytical needs to **monitor individual foreign investors activities** in the government bonds market that potentially create currency volatility.

Increasing foreign investor ownership in Indonesia government bonds market

Utilization of data sources that BI has e.g Government Bonds transactions, fundamental and market indicators that can be used to predict foreign investor behavior

"Predict foreign investor behavior on the Government Bonds Market **by using various scenarios of macroeconomic and *market indicators* and machine learning methods that can produce a good level of accuracy.** "

FRAMEWORK

This study uses Machine Learning and Text Mining to predict the capital flows of foreign investors in the government bonds market. We use granular government bonds transactions data, fundamental and financial market indicators. The results are foreign investor behavior clusters, important variables that influence investor decision, and foreign investors flows projection.

Behaviour analysis **of foreign investor in government bonds market includes:**

- **Classification of investor group cluster(*short term/long term*),**
- **Identification of important variables that influence individual investor decision**
- **Foreign investor decision prediction and flows projection.**

REFERENCES:

- Hoffman, A. et al (2010). Behavioral Portfolio Analysis of Individual Investors
- Bontempi, G. et al. (2009). Machine Learning Strategies for Time Series Forecasting.
- Agung, J. and Darsono. (2012). Post-Global Crisis Capital Inflows to Indonesia: Challenges and Policy Responses.
- Mody, A. et al. (2001). Modelling Fundamentals for Forecasting Capital Flows to Emerging Markets.

COLLECT

PROCESS

ANALYTICS

DISSEMINATE

DATA CAPTURING

SI - BISSSS

FOREIGN GOVERNMENT BONDS TRANSACTION SETTLEMENT

BLOOMBERG

FUNDAMENTAL ECONOMY & *MARKET DATA(AUTOMATIC RETRIEVAL)*

ANALYTICS

PRE-PROCESSING

A

- *Data Cleansing*
- *Entity Resolution – String Similarity*

INVESTOR TYPE MAPPING & VALIDATION

B

- *Composite Index Formula*
- *K-Means & Hierarchical Clustering*

INDIVIDUAL INVESTOR BEHAVIOR ANALYSIS

C

Decision Tree on market & fundamental variables

FLOWS PROJECTION, VALIDATION, & ANALYSIS

D

- *Machine Learning : SVR & Regression Tree*
- *Deep Learning : Long Short Term Memory (LSTM)*

RESULT

INDIVIDUAL INVESTOR BEHAVIOR & CLASSIFICATION

INVESTOR DECISION PREDICTION AND CAPITAL FLOWS PROJECTION

METHODOLOGY - DATA SOURCE AND PRE-PROCESSING

Bloomberg indicators data is pre-processed to have the *lag adjusted* version of the data, *and feature engineered* to produce dtd, mtd, qtd, and ytd changes. *The SID (Single Investor Identification)* are processed using text mining so that we can group each SID to its *approximate investor name*.

BI-SSSS (BI Scripless Security Settlement System) Data

2016 – 2020 Raw Government Bonds Settlements

Entity Resolution

Calculate Investor Net Flows

2016 – 2020 Investor of Interests* Indonesia's Bond Net Settlements

Bloomberg Economic Indicators

119 *Market and Fundamental Variables*

Lag Adjustment

Feature Engineering

Lag Adjusted and Feature Engineered Variables

*Top 30 foreign investors of Indonesian Government Bonds + 5 foreign investment forums investors

METHODOLOGY - ENTITY RESOLUTION

Entity Resolution is done to obtain list of SID and and group foreign investors name into unique entities. Entity resolution is necessary because some investors have multiple SID in the government bonds transactional data.

Entity resolution model is developed using ***string similarity algorithm***, which is a class of algorithm that measures similarity between texts.

ALGORITHM	THRESHOLD	F1-SCORE
Jaro-Winkler	**0,97**	**0,87**
Normalized Leveinshtein	**0,96**	**0,79**
Weighted Leveinshtein	**0,96**	**0,82**
MLCS	**0,85**	**0,73**

*Names listed above are masked

METHODOLOGY - INVESTOR TYPE CLUSTERING

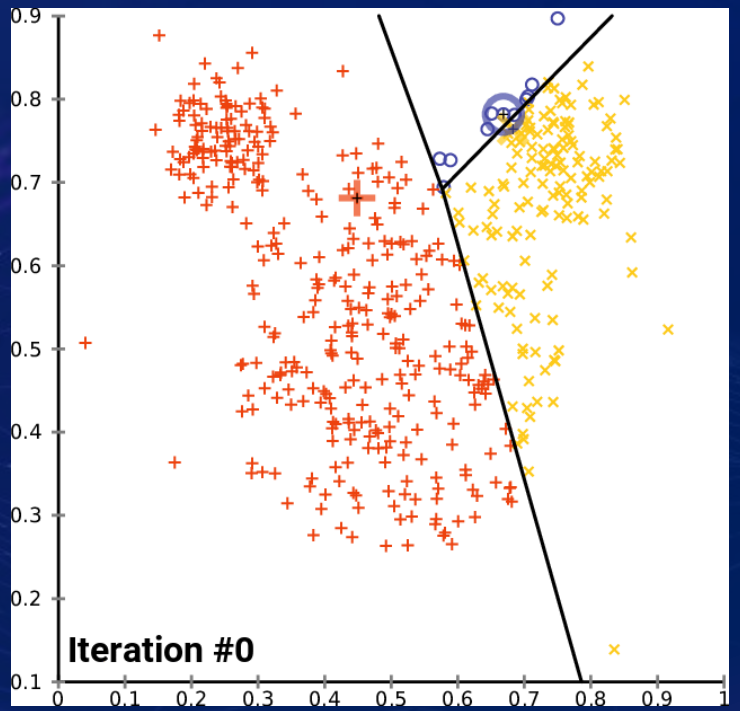
We measured *3 indicators to group investors*. Then, we use *K-means clustering* using the indicators to form foreign investor clusters as group of investors based on their yearly behavior in the government bonds market. The data-driven clustering result aligns with the expert judgement done by the Monetary Management Department and have a *high silhouette coefficient* as evaluation metrics.

Investment **Horizon**

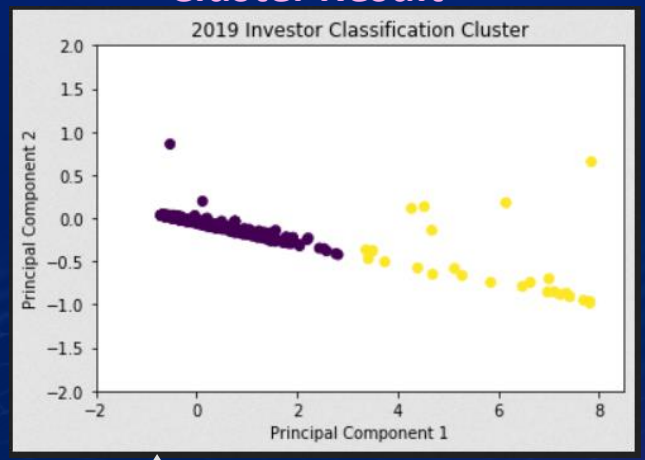
Transaction **Frequency**

Rebalancing **Ratio**

K-Means Clustering



Cluster Result



Investor Long Term

Investor Short Term

Silhouette Coefficient*:

0.892

Number of investors ST/LT:

35/1075

METHODOLOGY - BEHAVIOR MODELLING

We built *Behavior Modelling* to get better understanding of the indicators that *influence the decision* of each investor of interests, and each the investor group. The behavior model result will be used as selected *important feature for decision prediction and flows projection model*.

Data Collection

- Daily Settlement Data BI-SSSS 2016-2020
- Daily Bloomberg Data 2016-2020
- Exchange Rate and Financial Pressure Indices

Pre-processing

- *Data Cleansing*
- *Data Consolidation*
- *Data Pre-process*

Modelling

- Lag, Year and Shock Tuning
- Hyperparameter Tuning
- Random Forest & XGBoost Model Fit and Feature Importance

Model Result

Feature Importance Model

METHODOLOGY - FLOWS PROJECTION AND INVESTOR DECISION PREDICTION

Flows projection and investor behavior prediction experiment is done by using machine learning algorithms from *Logistic Regression to Deep Learning using LSTM*. The model that produce the best result is used to predict the investor investment decision and project the net flows of each individual investor daily.

Behavior Model Result

- Important variables
- Optimum lag

Features:

- Behavior Model Important Variables
- Up to 14 day lagged fundamental and market variables.
- Exchange Rate and Financial Market Pressure Indices

Pre-processing

1. Data Collection
2. Data Lag Adjustment
3. PCA Dimension Reduction

Modelling

Investor Behavior Prediction Algorithm:

- Logistic Regression
- Decision Tree
- XGBoost
- SVM
- KNN
- LSTM

Net Flows Projection Algorithm:

- Regression Tree
- XGBoost
- SVR
- LSTM
- ZIPR
- KNN

Hyperparameter Tuning

Model Evaluation

- Investor Behavior Prediction Model Evaluation using F1-score
- Flows Projection Model Evalution using MAE

BEHAVIOR MODELLING – LONG TERM AND SHORT TERM INVESTOR GROUP RESULT

- For both of the investors group, *Indonesian Bonds Yields in different maturity are considered important* with short term investors group considers shorter maturity of Government Bonds. JAKCONS Stock Index also important for both the groups.
- *Short term investors group* are affected by *more daily frequency indices*, while *long term investors group* are affected by a fundamental indicator which is the *Indonesia Yoy Core Inflation indicator* (IDIFCRIY Index)

Short Term Investors

Ticker	Description
IDR Currency	Rupiah Currency
HSITR Index	HSBC Asia Dollar Bond Index
SENSEX Index	S&P Bombay Stock Exchange
USSGGBE10 Index	US Breakeven 10 Y
GIDN12YR Index	Yield SUN Generic 12 Year
JAKINFR Index	Jakarta Infrastructure, Utilization and Transportation Stock Index
IDPPON Index	PUAB o/n
SXXP Index	STOXX Europe 600 Index
JAKCONS Index	Jakarta Consumer Goods Stock Index
MXEM Index	MSCI Emerging Market Index

Long Term Investors

Ticker	Description
JASXFBAT Index	Indonesia Stock Capital Flows
FSSTI Index	Straits Times Index STI
MGIY10Y Index	Malaysian 10-year Government Bond Yield
INR Currency	Indian Rupee Currency
GIDN30YR Index	Yield SUN Generic 30 Year
JAKCONS Index	Jakarta Consumer Goods Stock Index
IDIFCRIY index	Indonesia Yoy Core Inflation
IHNI1M Currency	Implied NDF 1M
IBPRTRI Index	IBPA Government Bond Index
THB Currency	Thailand Baht Currency

FLOWS PROJECTION AND INVESTOR BEHAVIOR PREDICTION – RESULT SUMMARY

The investor groups investor behavior prediction models are able to predict the decision made by 18 of 35 investor and the two investor groups with satisfying result (>60% F1 Score). As for the flows projection model the result is not good enough yet and still have to be improved in future works

Training Data
Jan 2016 – Des 2019

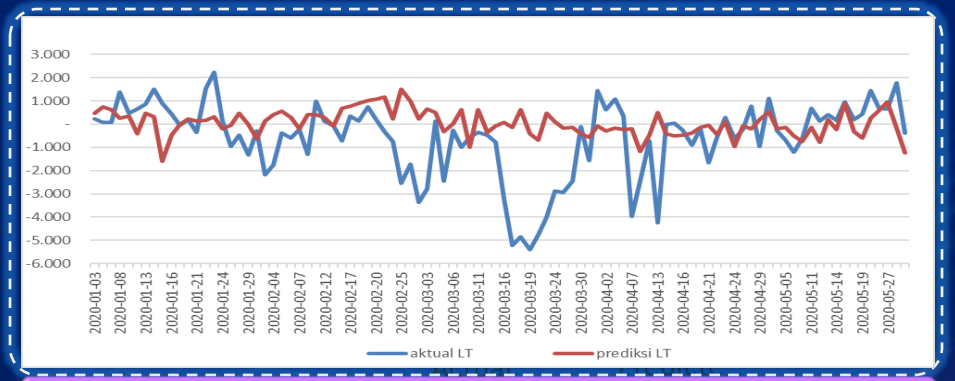
Testing Data
Jan 2020 – May 2020

Max F1-Score
86%

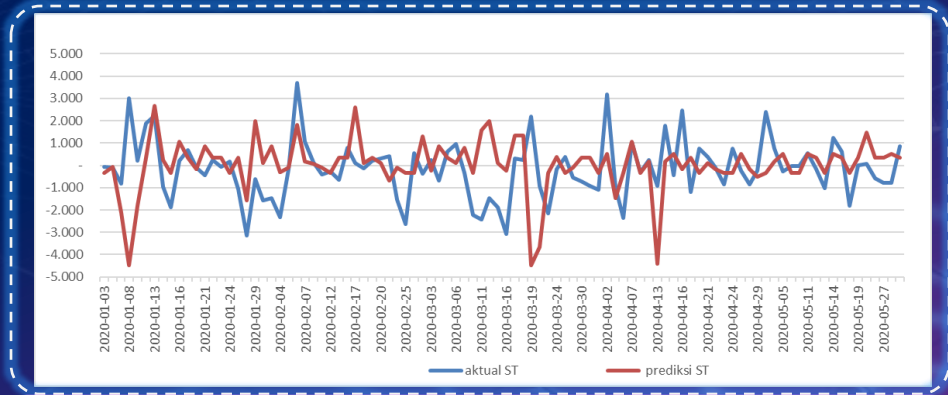
Min F1-Score
47%

Investor Model with F1 Score > 60%

20 / 37



Long term Investor Flows Projection



Short term Investor Flows Projection

CONCLUSION AND FUTURE WORKS

CONCLUSION

With the result we are confident that machine learning methodology has been able to identify single investor based on its name similarity using string similarity method, cluster investor group using yearly behavior and predict investor decision on Government Bonds Transaction (buy, sell or hold). But there is still a lot of works to improve the prediction accuracy of the flows projection model.

FUTURE WORKS

1. *Improve the accuracy* of the investor decision and flows projection models for all of the individual investors.
2. Develop *investor investment prediction and flows projection model* that can predict well during *abnormal flows* period (COVID-19 Pandemic).
3. Develop *model automation* and *dashboard* for daily visualization of government bonds daily data and prediction.
4. Develop model for *foreign investors in stock and currency market*.

BANK INDONESIA

"The future belongs to those who prepare for it today."

Malcolm X