BCU

A predictive model of sovereign investment grade using machine learning and natural language processing

María Victoria Landaberry
Kenji Nakasone
Johann Pérez
María del Pilar Posada

Nº 005 - 2022

Documento de trabajo
ISSN 1688-7565

BCU

A predictive model of sovereign investment grade using machine learning and natural language processing[☆]

María Victoria Landaberry[a*], Kenji Nakasone[b], Johann Pérez[b], María del Pilar Posada[a]

a *Banco Central del Uruguay*
b *UTEC - Universidad Tecnológica*

Documento de trabajo del Banco Central del Uruguay 005-2022

Autorizado por: Jorge Ponce
Disponible en línea: 15/9/2022

Resumen

Las agencias calificadoras de riesgo como Moody's, Standard and Poor's y Fitch califican los activos soberanos basados en un análisis matemático de factores económicos, sociales y políticos conjuntamente con un análisis cualitativo de juicio de experto. De acuerdo a la calificación obtenida, los países pueden ser clasificados como aquellos que tienen grado inversor o cuentan con grado especulativo. Tener grado inversor es importante en la medida que reduce en costo de financiamiento y expande el conjunto de potenciales inversores en una economía. En este documento nos proponemos predecir si la deuda soberana de un país será calificada con grado inversor utilizando un conjunto de variables macroeconómicas y variables obtenidas a partir del análisis de texto de los reportes de Fitch entre 2000 y 2018 utilizando técnicas de procesamiento natural de lenguaje. Utilizamos una regresión logística y un conjunto de algoritmos de machine learning alternativos. De acuerdo a nuestros resultados, el índice de incertidumbre, construido a partir de los reportes de Fitch, es estadísticamente significativo para predecir el grado inversor. Al comparar los distintos algoritmos de machine learning, random forest es el que tiene mejor poder predictivo fuera de la muestra cuando la variable dependiente refiere al mismo año que las variables explicativas mientras que k-nearest neighbors tiene el mejor desempeño predictivo cuando las variables independientes refieren al año anterior en términos del f1-score y recall.

JEL: E22, E66, G24
Palabras clave: Riesgo soberano, agencias calificadoras, variables macroeconómicas, análisis de texto, procesamiento natural del lenguaje; machine learning

[☆] The views expressed herein are those of the authors and do not necessarily represent the opinion of the institutions with which they are affiliate.

[*] Corresponding author. E-mail: mlandaberry@bcu.gub.uy

1 Introduction

The investment grade status of countries is linked to the risk of default on their debt. Even though "the terms investment grade and speculative grade are market conventions and do not imply any recommendation or endorsement of a specific security for investment purpose" (Fitch, 2011), having an investor grade will reduce the cost of financing and expand the pool of potential investors in an economy. In many cases, companies and institutional investors are limited to investing only in countries that have an investor grade rating.

The three main credit rating agencies (Moody's, Standard and Poor's and Fitch) indicate that their sovereign risk ratings depend on the analysis of economic, social and political factors and use an alphanumeric code to rate risk. In this research we focus on the sovereign credit rating and sovereign investment grade status provided by Fitch. Sovereign credit ratings are focused on the risk of a sovereign government defaulting on its debt obligations. According to Fitch sovereign rating criteria there are a total of 22 rating scores (Fitch, 2020). Countries with a rating greater than or equal to BBB- by Fitch are considered to have investor grade, while countries with lower ratings are considered countries with speculative class assets, without investment grade (Table 1).

Table 1: Fitch Rating

Status	**Fitch Rating**
Investment grade	AAA , AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-
Speculative grade	BB+, BB, BB-, B+, B, B-, CCC+,CCC, CCC-,CC,C, RD/D

Source: (Fitch, 2011)

In this paper we want to predict whether a sovereign has investment grade status or not on a yearly basis, considering, as explanatory variables, a set of macroeconomic variables plus text analysis variables obtained from the reports issued by Fitch between 2000 and 2018. We consider all the reports issued by the rating agency related to the sovereign instead of using only credit rating action reports. We estimate a series of logistic regression as it is the model more used in previous literature related to predicting investment grade status or credit ratings. We then compare the results of this estimation with other machine learning algorithms as bayesian model average, k-nearest neighbor, support vector machine, classification and decision trees and random forest and discuss their predictive performance out of sample. Our main contribution to the previous literature related to predicting sovereign ratings is the use of explanatory variables obtained from the reports made by the credit rating agencies for the different countries using natural language processing techniques and simultaneously incorporating this information to the prediction of the sovereign investment grade using a variety of machine learning models.

Although the macroeconomic fundamentals are relevant to explain the loss or gain of the investment grade and the sovereign rating itself, there is an expert judgment component on the part of the rating agencies (qualitative overlay) that is not captured through the macroeconomic variables. According to the results obtained by Agarwal et al. (2015) adding the variables resulting from the natural language processing techniques to capture the qualitative overlay would improve the predictive performance of the machine learning models.

Also, we use Fitch's rating data, which is not used frequently in the literature[1]. In this way we contribute to the generalization of the discussion about the effect of sentiment analysis on the reports issued by rating agencies.

Through sentiment analysis, we extract text features from Fitch reports, in order to understand statistical significance and predictive power of these features, and the relevance of the fore mentioned reports. Whereas the constructed variables do not have an important impact over predictive power, we do find uncertainty index to be statistically significant.

Based on both macroeconomic and sentiment features, we compare alternative models to predict the investor grade of countries, finding slight improvements in the results of k-nearest neighbors (KNN) and random forests over the use of logistic regressions.

The remaining part of this work is organized as follows: In Section 2 we present a brief review of literature related to this topic. Section 3 describes the database used in this research project, the source of data and the variable selection criteria. In this section we focus on the exploratory analysis of the variables obtained from Fitch reports and the sentiment analysis techniques implemented. In Section 4 we present a description of the methodology used to build the prediction models. Section 5 presents the results and finally Section 6 the conclusions.

2 Literature Review

Most of the existing literature on investment grade explores the determinants of sovereign credit ratings, focusing the analysis on the economic, social and political variables that affect the rating status of a country. In particular, the assignment of Fitch's sovereign ratings reflects a combination of a Sovereign Rating Model[2] and a Qualitative Overlay (Fitch, 2020).

While the sovereign rating model is estimated using ordinary least squares over a set of economic and financial variables for all Fitch-rated sovereigns over 2000-2018 inclusive, its results represent a starting point to the final rating that a country obtains in each rating revision. Recognizing that no quantitative model can fully capture all the relevant influences on sovereign creditworthiness, Fitch employs a forward-looking qualitative overlay to adjust for factors not reflected or not fully reflected in the Sovereign Rating Model output for any individual rating (Fitch, 2020). According to the methodology description and previous literature about sovereign credit ratings determinants, final credit ratings are driven by a combination of hard and soft information (Slapnik and Loncarski, 2019).

Most of the previous literature is focused on the effects of hard information on sovereign credit ratings.Cantor and Packer (1996) using ordinary least squares to explain the numerical equivalents of Moody's and Standard and Poor's ratings found six factors that plays an important role in determining a country's credit rating: per capita income, GDP growth, inflation, external debt, level of economic development, and default history. Borraz et al. (2011) analyze sovereign credit ratings using an ordered logit regression for 53 countries between 2000 and 2010 to predict Moody's and

[1] Previous literature uses Moody's credit rating agency or/ and Standard and Poor's ratings and reports.

[2] For an extended discussion about variables considered in the Sovereign Rating Model see Section 3.

Standard and Poors sovereign ratings and a binary logit regression to predict investment grade status. According to their results GDP per capita, GDP growth, inflation, dollarization, government effectiveness, fiscal result over GDP, government debt over GDP, debt service, current account over GDP, default history and international reserves over GDP are statistically significant.Butler and Fauver (2006) use a sample of 86 counties to examine the cross- sectional determinants of sovereign credit ratings and find that the quality of a country's legal and political institutions plays a vital role in determining these ratings even after controlling by macroeconomic variables.

There is also a recent literature stream that tries to include to the rating models the soft information through the text analysis of the reports issued by the credit rating agencies. Agarwal et al. (2015) use the Naive Bayesian algorithm on the rating reports issued by Moody's in 62 countries for the period 2003–2013. They classify all sentences in each report as positive,negative, or neutral in tone and also, they classify each sentence into one of six content categories (i.e., macroeconomic, public and external finance, debt dynamics, financial sector, political and institutional, and others). They use this information and additional macroeconomic variables to predict the spread of the sovereign credit default. They find that the information contained in the reports is significant.

Slapnik and Loncarski (2019) use an ordered logistic regression with random effects for 98 countries in the period from 1996 to 2017, to explain the different sovereign credit ratings and incorporate, as dependent variables, a set of variables obtained from the sentiment analysis of Moody's credit action reports using a dictionary approach. They consider three sentiment variables and one subjectivity variable as proxies of the qualitative judgment of the rating agency. For the sentiment variables, each word in the last Moody's Credit Action reports issued in the year is classified using Loughran and McDonald (2011) dictionary in negative or positive. Then they consider negative sentiment as the ratio between the number of negative words in the text and the total number of words, positive sentiment is measured as the percentage of positive words in the text and net sentiment is then the difference between negative and positive sentiment. Finally, they add a subjectivity variable using the TextBlob package in Python which measures the degree of subjectivity in texts ranging between 0 and 1. According to this index, a report is more subjective if the sentences that compose it "refer to personal opinion, emotion or judgment whereas objective refers to factual information" (Slapnik and Loncarski, 2019). According to their results textual sentiment provides additional information not captured by traditional determinants of sovereign credit ratings if soft information proxies as governance and institutional quality are not taken into account.

Our main contribution to this literature can be defined in four aspects: First, we propose to perform a dictionary-based approach sentiment analysis of all the reports issued by the credit rating agency and not only credit rating action reports and we also add other sentiment indicators to the ones used by Slapnik and Loncarski (2019). Second, we compare the performance of the most traditional estimation techniques as binary logit with other supervised machine learning algorithms as bayesian model average, k-nearest neighbors, support vector machine, classification and decision trees and random forest. In third place we use Fitch sovereign rating and rating reports while most of the previous literature focus on Moody's and Standard and Poor's sovereign rating. Using Fitch's information can contribute to the generalization and discussion of the results found in previous literature for Moody's and Standard and Poor's rating agency. An extension of this work using other rating agencies sovereign ratings and reports can be done in the future.

Finally, instead of using the sovereign rating for each country by year as a dependent variable we try to explain through different models the investment grade status of a country by year (binary classification problem). This classification is important for countries that are dependent of foreign investment as having an investor grade will reduce the cost of financing and expand the pool of potential investors in an economy. In many cases, companies and institutional investors re limited to investing only in countries that have an investor grade rating. An extension including macroeconomic and text analysis variables to predict the rating itself can be addressed in future work.

3 Data

In this Section we provide a detailed description of the variables used in the models. In subsection 3.1 we describe the rating data used and the coverage in terms of countries considered. In subsection 3.2 we describe the macroeconomic variables and the criteria used to select the variables to be included in our models. Finally, in subsection 3.3 we describe the variables obtained from natural language processing techniques and we present some interesting results in terms of exploratory analysis of these variables.

3.1 Rating data and coverage

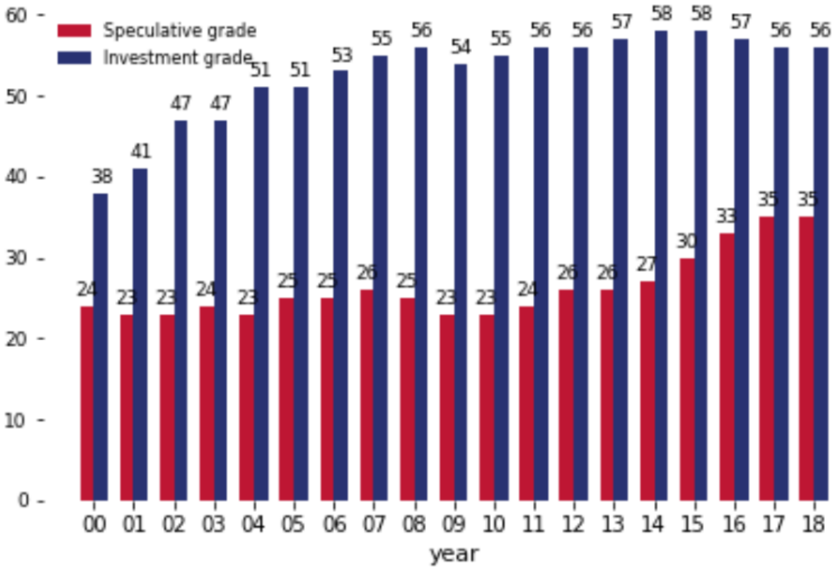
We use sovereign rating data issued by Fitch's rating agency over 91 countries between 2000 and 2018[3]. According to Fitch sovereign rating criteria there are a total of 22 rating scores (Fitch, 2020). Countries with a rating greater than or equal to BBB- by Fitch are considered investor grade, while countries with lower ratings are considered countries with speculative class assets, without investment grade (Table 1). We define a binary variable that takes the value equal to 1 if the country has investment grade on the year considered and 0 if the country does not have investment grade status. For those years where a country changes between investment and non-investment grade status or vice versa we consider that the country has non-investment grade status. This is our dependent variable in the binary classification problem [4]. From a total of 1502 observations 33% have speculative status (Figure 1). 27 countries in the sample have changed at least once their investment grade status. In total there are 39 events of investment grade status change (23 events of countries obtaining an investment grade rating and 16 events of countries losing their investment grade status).

In Figure 2 we present the countries included in our database and in the Annex I we provide a detailed table of the countries considered in our models, the years for which we have data for each country and information about changes in their investment status in the sample.

[3] Period considered in this work matches with the period used in Fitch's sovereign rating model (Fitch, 2020)

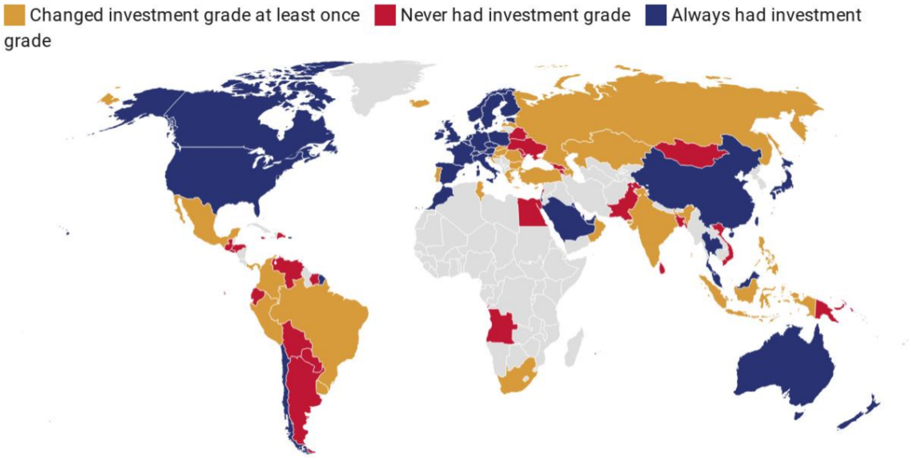
[4] If instead of using this definition we consider the rating at the end of the year or we assign that the country has investment grade the year the status change coefficients in the model and significance in the variable does not change, meaning that results are robust to these alternative definitions of the dependent variable

Figure 1: Countries with investment and speculative grade by year



Source: Author's calculation

Figure 2: Countries and investment grade status according to Fitch


Changed investment grade at least once
Never had investment grade
Always had investment grade

Source: Author's calculation

3.2 Macroeconomic variables

In our models we include macroeconomic variables that capture the quantitative information about the situation of each country in a particular year.

According to Fitch Sovereign rating criteria (Fitch, 2020) the assignment of sovereign ratings reflects a combination of a Sovereign Rating Model (SRM) and a Qualitative Overlay (QO). The

SRM is a starting point for assigning sovereign ratings and it is a multiple regression rating model that employs historical, current and forward-looking data for quantitative variables grouped into four analytical pillars: structural features, macroeconomic performance and public finance. In their model they estimate a numeric variable that is calibrated to each one of the classification categories, using ordinary least squares (OLS) for all Fitch-rated sovereigns over 2000-2018 inclusive. We consider the same period of time in our models.

They use a total of 18 variables in a centered three years averages (therefore including Fitch´s forecast for the current year) for the more dynamic variables, such as the current account and fiscal balances, to smooth the impact of volatility on the output. Variables used in Fitch SRM model and the corresponding analytical pillar are presented in Table 2.

We do not have access to all the variables used as independent variables in the OLS estimation in order to reproduce the results or incorporate all of them in our estimations. In particular we do not have the Fitch forecast for each year that is used in the centered average of the variables when considered and we do not have data about the real GDP growth volatility, the reserve currency flexibility, commodity dependence, official international reserves for non-reserve currency sovereigns of sovereign net foreign assets. In general terms, when possible we include the variables as considered in the SRM model. For three years centered average variables we consider only the value of the variable in the year we are interested in and finally in some cases where data is not available, but we get a similar variable we replace it by the one that is available.If variables have more than 10% of missing values, we do not consider them into the analysis[5].

In Table 2 we present the variables used in our models and the changes if any with respect to the definition used in Fitch's SRM model. In Table 3 we provide the summary statistics of these variables and in Annex II we include additional information about our preliminary exploratory data analysis.

[5] Money supply and public foreign currency debt variables are excluded according to these criteria

Table 2: Fitch explanatory variables versus our models

Variable	Description	Changes in the data we used if any /data source.
Structural features		
Composite governance indicator	Simple average percentile rank of world bank governance indicators: "rule of law", "government effectiveness", "control of corruption" and "voice and accountability", "regulatory quality", "political stability and absence of violence"	No change. The same variable is used. / World Bank data.
GDP per capita	Percentile rank of GDP per capita in the US dollars at market exchange rate	No change. The same variable is used. / Eurostat, AMECO,Official National Source, Moody's and authors calculations.
Share in world GDP	Natural logarithm of % share in world GDP in US dollars at market exchange rate	No change. The same variable is used. / Eurostat, AMECO,Official National Source, World Bank, Moody's and author's calculations.
Years since default of restructuring event	Non-linear function of the time since the last event: the indicator is zero if there has not been such event after 1980. For each year that elapse the impact on the model output declines	Dummy variable that takes 1 if the country in the last 21 years has entered in default at least once. / Data until 2010 was obtained from Asonuma and Trebesch (2016) After 2010 data was obtained from Cheng et al. (2018), Moody's (2020) and Standard and Poor's (2018).
Money supply	Natural logarithm of broad money (%GDP)	Not included in our models
Macroeconomic performance		
Real GDP growth volatility	Natural logarithm of an exponentially weighted standard deviation of historical annual percent change in the real GDP	Not included in our models.
Consumer price index	Three-year centered average of the annual % change in consumer price index (CPI), truncated between 2% and 50%.	Annual % change in consumer price index (CPI), truncated between 2% and 50%. / Eurostat, Official National Source, Moody's and author's calculations.
Real GDP growth	Three-year centered average of the annual % change in real GDP	Annual % change in real GDP /Eurostat, Official National Source , Moody's
Public Finance General government		
Gross general government debt	Three-year centered average of the Gross (general) government debt (% GDP)	Annual Gross (general) government debt (% GDP) /IMF, OECD, Eurostat, AMECO, Official National Source, Moody's
Interest payment	Three-year centered average of gross government interest payment (% general government revenues)	Annual gross government interest payment (% general government revenues) / IMF, OECD, Eurostat, AMECO, Official National Source, Moody's
General government fiscal balance	Three-year centered average of general government (budget) balance (% GDP)	Annual general government (budget) balance (% GDP) / IMF, OECD, Eurostat, AMECO, Official National Source,Moody's
Public foreign currency debt	Three-year centered average of public foreign currency-denominated (and indexed) debt (% of general government debt)	Not included in our models.
External Finance		
Reserve currency flexibility	Reserve currency flexibility based on the natural logarithm of the share of that country's currency in global foreign- exchange reserve portfolios (plus a technical constant), as reported by the IMF in its COFER database (updated quarterly with a four-month lag)	Not included in our models.
Commodity dependence	Non-manufactured merchandise exports as a share of current account receipts.	Not included in our models.
Official international reserves for non-reserve currency sovereigns	Year-end stock of international reserves (including gold) expressed as month's cover of current external payments. This variable is set to zero for all all sovereigns with a reserve currency flexibility score above zero.	International reserves (% GDP) / IMF, OECD, Eurostat, AMECO, Official National Source, Moody's
Sovereign net foreign assets	Three-year centered average of sovereigns net foreign assets (%GDP)	Not included in our models.
Current account balance plus net foreign direct investment	Three-year centered average of external interest services expressed as a share of CXR	Current account balance (% GDP) and Foreign direct investment(% GDP) are included but separately.
Additional variables	(not included in Fitch's SRM)	
Developed		Dummy that takes the value equal to 1 if the country is developed according to IMF classification and 0 otherwise.

3.3 Fitch's Reports text analysis variables

As mentioned in the previous subsection, Fitch's sovereigns' ratings model is the starting point to determine the final sovereign rating for a country. The rating agency combines the information provided by the model with a Qualitative Overlay that reflects some relevant factors that are not included in SRM because they are not quantifiable as geopolitical risk, or variables that cannot

Table 3: Macroeconomic variables summary statistics

Variable	Observations	Mean	Std. dv.	Min	Max
Composite governance indicator	1518	0.54	0.30	0	1
GDP per capita rank	1518	0.49	0.28	0	1
Share in world GDP	1518	-1.73	1.79	-6.50	5.47
Default variable	1518	0.17	0.37	0	1
Consumer price index	1518	4.89	5.76	2	50
Real GDP growth	1516	3.41	3.77	-16.2	34.5
Gross general government debt	1518	50.84	34.63	0	264.8
Interest payment	1518	8.71	9.35	0	92.8
General government fiscal balance	1518	-1.87	5.27	-32.1	48.5
Official international reserves	1518	17.67	19.79	0	126.25
Current account balance	1515	1.23	74.78	-802	436
Foreign direct investment	1507	1.89	11.61	-157.3	243.4
Developed	1518	0.34	0.48	0	1

Source: Author's calculations

be included in the model because they are not available for all the countries, data gaps, capturing non linearities or events that happens on a more frequent basis that the data (Fitch, 2020). We try to capture this qualitative assessment including in our models some variables obtained from the text analysis of all the reports issued by Fitch referring to sovereigns for each country by year. These reports are available from Fitch's website without a paid subscription. We have a total of 7979 reports covering the period 2000-2018 for the countries in our database (Figure 3)

We consider two different approaches to extract the information from the reports. The first approach consists in considering the number of documents issued by Fitch in the period 2000- 2018 and the second approach consists of using within the natural language processing techniques the analysis of the tone of reports issued by Fitch.

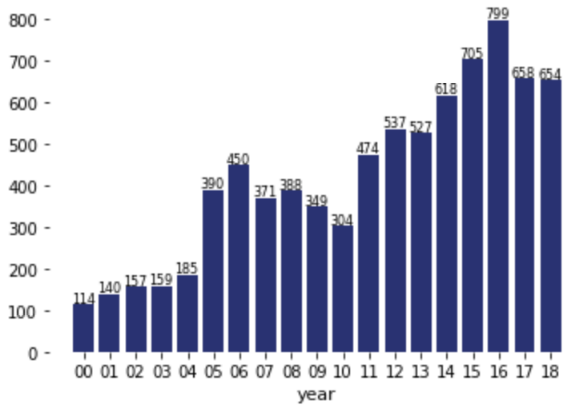
3.3.1 Quantity of documents issued

We present two alternative hypotheses about the number of reports issued according to the sovereign status of a country. The first hypothesis is that investment grade status countries obtain more attention than speculative grade status countries. The second hypothesis is that when a sovereign from a particular country is going to change their investment grade status it is more likely that more reports are issued related to that country.

According to the densities presented in Annex II, countries with speculative status have fewer reports than investment grade status countries. Considering the type of report[6], countries with speculative status have less articles and similar press releases than investment grade status. This result is in line with our first hypothesis. Alternatively, we consider the deviation from the mean of reports issued by country and the deviation from the mean of reports issued by year. As presented in Annex III, countries with investment grade status have a left skewed density of reports deviation from the mean of reports issued by country but countries with speculative grade status seem to have more variance in the deviation of the mean of reports issued by year than investment grade

[6] Reports are classified in three categories: Articles, Press Release and FS Multimedia. The last type of reports with a small number of observations in the sample is excluded in the density plots analysis

Figure 3: Fitch's reports by year (total)



status. Considering these variables there is more evidence in favor of the first hypothesis that countries with investment grade status receive more attention in terms of reports issued by the rating agency. Summary statistics of text variables included in our models are presented in Table 4.

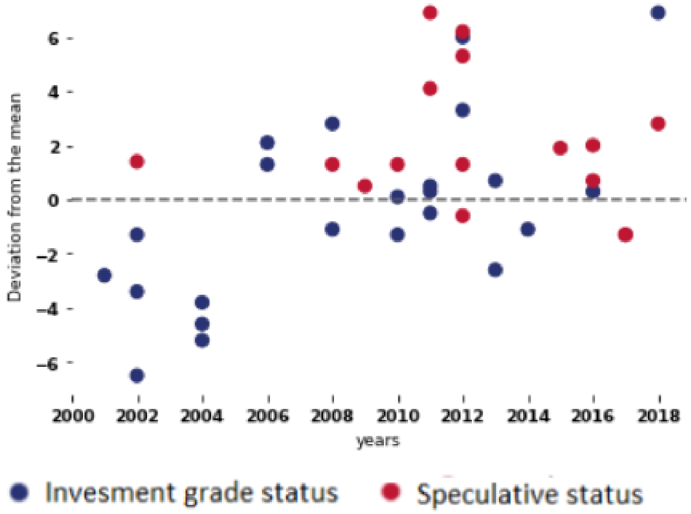
In our sample 27 countries have changed their investment grade status at least once over the period 2000-2018. The countries that changed their investment grade status and the years where these changes took place are presented in Annex I. In this section we explore if the total of document frequency changed in these countries in the previous or the same year where the investment grade status change takes place and also if the sentiment analysis indexes that we build reflect that an investment grade change will take place. We consider the difference between the total documents issued by Fitch the previous year and the same year when a sovereign obtained an investment grade status and when a sovereign loses its investment grade status with respect to the mean of reports issued for each country between 2000-2018[7] (Figure 4).

In general terms, countries that gain investment grade status do not have on average more reports than the mean for 2000-2018 the year the change is produced or the year before. For these countries on average there are 0.76 reports less than the mean for 2000-2018 the year the change is produced and 1.59 less than the mean for 2000-2018 the year before. If we consider the change to speculative status, on average the same year and the year before, countries in this situation have more reports issued than the mean observed for the period 2000-2018. The same year there is an average deviation from the mean over the period of 2.5 documents and the year before the average deviation is 2.9 documents. 80% of the countries that were rated with a speculative grade had more documents than the average the same year the rating changed, and the number increased to 86% the year before.

This result is evidence in favor of the second hypothesis that when a sovereign from a particular country is going to change their investment grade status it is more likely that more reports are issued related to that country. But there is no symmetry. Increases in reports occur when a country is going to lose their investment grade status, but the opposite does not hold true.

[7] In Annex IV we present the plots of reports issued and changes in the investment grade status for all the countries in the sample except Oman as it only has observations for two years

Figure 4: Deviation from the mean of reports issued for each country between 2000-2018 the year the status changed



Source:Author's calculation

3.3.2 Sentiment Analysis

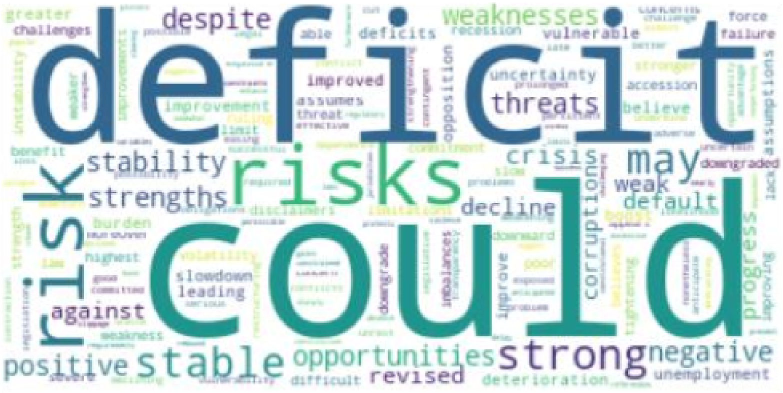
The second approach consists of using within the natural language processing techniques the analysis of the tone of reports issued by Fitch. For this purpose, we use the Loughran and McDonald (2011) sentiment dictionary with a total of 4150 words classified in six categories (negative, positive, uncertainty, constraining, litigious and superfluous) used to capture the tone in financial or business documents. After preprocessing Fitch´s sovereigns reports, including transforming all words in lowercase and dropping information that is not part of the main text as contact information or disclaimer footnote, we classify all the words in the reports in Negative, Positive or Uncertain according to Loughran and McDonald (2011) sentiment dictionary. In Figure 5 we present the most frequent words from the Loughran and McDonald (2011) sentiment dictionary across all Fitch documents considered in our work. While in Figure 6 we show the same frequency broken down into the positive, negative and uncertainty sentiments.

Then we construct four variables that reflect the tone of the documents issued in the year for each country. Following Moreno Bernal and González Pedraz (2020) we define the Net Negative Index for country i in year t (Equation 1). This index consists in the total of negative words minus the total of positive words in all the reports issued in year t for country i, over the sum of the total of words that are classified as positive or negative in year t for country i.

$$\text{Net negativity index}_{it} = \frac{\#\text{negative words}_{it} - \#\text{positive words}_{it}}{\#\text{negative words}_{it} + \#\text{positive words}_{it}} \tag{1}$$

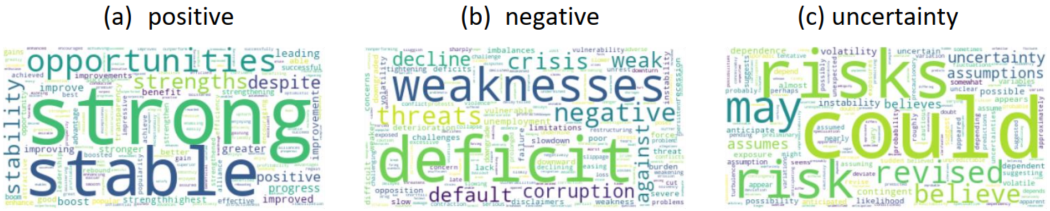
Net negativity index ranges from -1 to +1. The higher the index the higher the negativity in the tone of the reports issued by Fitch in year t for the country i. Alternatively, we define the Negativity index for country i in year t (Equation 2) that consists in the total of negative words

Figure 5: Loughran and Mc Donald (2011) word frequency across all Fitch documents



Source:Author's calculation

Figure 6: Loughran and Mc Donald (2011) word frequency across all Fitch documents broken down by sentiment


(a) positive
(b) negative
(c) uncertainty

Source:Author's calculation

over the sum of the total of words that are classified as positive, negative or reflecting uncertainty. As the Net negativity index, the higher the index the higher is the negative tone of the report issued by Fitch in year t for the country i. This index ranges from 0 to 1.

$$\text{Negativity index}_{it} = \frac{\#\text{negative words}_{it}}{\#\text{negative words}_{it} + \#\text{positive words}_{it} + \text{uncertainty words}_{it}} \tag{2}$$

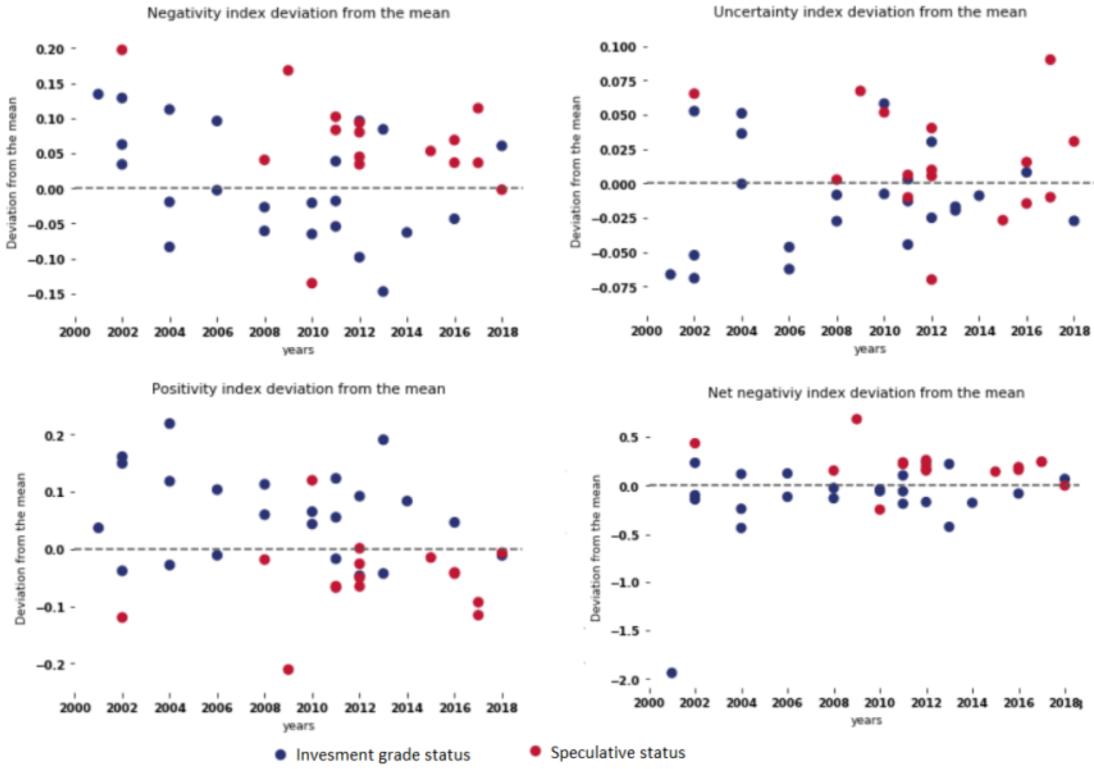
We also define a Positivity index and an Uncertainty index (Equation 3 and 4). The Positivity index consists in the total of positive words over the sum of the total of words that are classified as positive, negative or reflecting uncertainty in all the reports issued in year t for country i. Uncertainty index is the total of words reflecting uncertainty over the sum of the total of words that are classified as positive, negative or reflecting uncertainty in all the reports issued in year t for country i. Both indexes range from 0 to 1. Values of the Positivity index close to 1 reflect a more positive tone while values of the Uncertainty Index close to 1 reflect more uncertainty.

$$\text{Positivity index}_{it} = \frac{\#\text{positive words}_{it}}{\#\text{negative words}_{it} + \#\text{positive words}_{it} + \text{uncertainty words}_{it}} \tag{3}$$

$$\text{Uncertainty index}_{it} = \frac{\#\text{uncertainty words}_{it}}{\#\text{negative words}_{it} + \#\text{positive words}_{it} + \text{uncertainty words}_{it}} \tag{4}$$

We can do a similar analysis considering the variables that we previously defined related with the tone of the reports. We analyze the indexes for the 39 events of change of rating status from speculative grade to investment grade and vice versa (Figure 7).

Figure 7: Deviation from the year mean of Positivity index, Negativity index, Uncertainty index and Net negativity index for events of status between 2000-2018



As expected, if we analyze the indexes for the 23 events of gain of the investment grade status, we can verify that in 16 cases the Positivity index is superior to the mean of the year, and in 13 cases the Negativity index is inferior to the mean. On the other hand, in 14 of 16 events of loss of the investment grade, the Negativity index is superior to the mean of the year and the Positivity index is inferior in the same proportion. The performance of the Uncertainty index also behaves as expected, being superior to the mean in cases of loss the investment grade status and inferior in the cases of gain the status. Finally, Net negativity index deviation with respect to the mean of the year is higher for those countries that lose their investment grade status than for those that gain investment grade status. In Annex III we present the densities of these indexes for investment grade and speculative grade status and in Table 4 the summary statistics.

Table 4: Text analysis variables summary statistics

Variable	Observations	Mean	Std. dv.	Min	Max
Total of reports	1518	5.29	4.32	0	52
Deviation from yearly mean reports	1518	0	3.78	-1.89	43.12
Deviation from country mean reports	1518	0	3.65	-10.70	41.32
Net negativity index	1518	0.23	0.22	-0.77	1
Negativity index	1518	0.45	0.14	0	1
Positivity index	1518	0.28	0.12	0	1
Uncertainty index	1518	0.21	0.09	0	1

Source: Author's calculation

4 Methodology

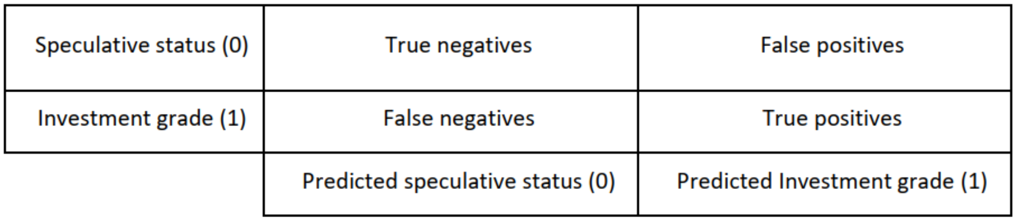
We split our database into a train and test sample with 75% and 25% of the data respectively following Müller and Guido (2016) rule of thumbs. Because our classification problem does not have a balanced number for each class label (33% of observations are speculative status and 77% are investment grade) we split the dataset using a stratified train-test split strategy. This method consists of split train and test sets preserving the same proportion of examples in each class, as observed in the original dataset (Brownlee, 2020).

Our main problem is to predict if a country has or not investment grade status on a yearly basis and as explanatory variables we consider macroeconomic and text analysis variables. This problem can be described as a binary classification problem.

4.1 Logistic regression and predictive performance

In the Binary classification problem, our benchmark machine learning algorithm is a logistic regression as this is the algorithm more used in the previous literature. We consider six different specifications for the logistic regression: In the first model we consider as independent variables only the macroeconomic variables specified in Table 2 of Section 3.2. In the second model we include from the variables obtained from the text analysis of Fitch's reports described in Section 3.3: the net negativity index, the uncertainty index and deviation from country mean report's and yearly mean reports. In the third specification of the model, we substitute the net negativity index with the positivity index. Finally, we repeat the first three models but considering the independent variables lagged one year, as we are interested in predicting the investment grade with information available before the rating is issued by the credit rating agency. We compare the results between these models in terms of the variables that are significant and a series of summary

Figure 8: Confusion Matrix



Speculative status (0)	True negatives	False positives
Investment grade (1)	False negatives	True positives
	Predicted speculative status (0)	Predicted Investment grade (1)

measures obtained from the confusion matrix to compare the predictive performance.

The confusion matrix (Figure 8) it's a two by two array where the rows correspond to the true classes in the test set and the columns correspond to the predicted classes by the model. Each entry in the matrix counts how often a sample that belongs to the class corresponds to the row, in our model investment grade or speculative status was classified as the class corresponding to the column.

Entries on the main diagonal, true negatives (TN) and true positives (TP) of the confusion matrix correspond to correct classifications, while false negatives (FN) and false positives (FP) represent how many samples of one class got mistakenly classified as another class. From the confusion matrix we obtain a series of summary measures about the predictive performance over the test sample of our models. These measures are precision, recall and f1-score (Müller and Guido, 2016).

The precision measure, also known as positive predictive value, is defined as the number of true positives divided by the number of true positives plus the number of false positives (Equation 5). If the model does not produce many false positives it has a high precision.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

The recall measure, also known as sensitivity or true positive rate, measures the proportion of positives that are correctly identified (Equation 6).

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

There is a trade-off between optimizing recall and optimizing precision[8]. f1-score summarizes both measures and is the harmonic mean of precision and recall (Equation 7). The highest possible value of the f1-score is 1, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero.

$$\text{f1 score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

The model with the highest f1-score will be considered as the model with the best predictive performance out of the sample. If two models have the same f1-score then, we select the one with

[8] We can obtain a perfect recall if the model predicts all samples to belong to the positive class. In this situation there will be no false negatives, and no true negatives either. However, this model will result in many false positives, and therefore the precision will be very low (Müller and Guido, 2016)

higher recall since in our model there is a high cost associated with a False Negative, i.e. misclassify a sovereign as non-investment grade status (Shung, 2018).

Even Though accuracy (correct prediction over total predictions) is one of the most used metrics to evaluate the performance of a classification predictive model we do not consider this measure for evaluating our models since our data set is unbalanced and accuracy does not perform correctly to evaluate models where the distribution of examples in the training dataset across the classes is not equal (Brownlee, 2020c).

In the logistic regression the outcome we obtain is the probability that a sample observation has investment grade status, i.e. is labeled 1. To assign to the observation the label, instead of using the default threshold of 0.5 we tune the optimal threshold to obtain the higher f1-score, because this metric is one of our predictive performance out of sample measure chosen to select the best model (Brownlee, 2020b).

If the probability of the sample observation obtained from the prediction of the logistic model is higher or equal than the threshold obtained by maximizing the f1-score, the sample observation is labeled as 1 and 0 otherwise. Finally, we consider alternative supervised machine learning algorithms (bayesian model average, support vector machine, k-nearest neighbors, classification and decision trees and random forest) and we compare the results in terms of their predictive performance with the logistic regression results. The alternative algorithms used are described in subsection 4.2 and the corresponding results are presented in subsection 5.2.

4.2 Supervised alternative machine learning algorithms

The goal of supervised machine learning algorithms for classification problems is to learn a function that, given a sample of data and desired outputs, best approximates the relationship between input and output observable in the data. When our goal is to obtain the label or the target instead of predicting a value, the algorithm solves a classification problem. In this section we briefly present a series of alternative supervised machine learning algorithms considered to compare our benchmark logistic regression. In subsection 5.2 we present the corresponding results, and we identify the best model in terms of predictive performance out of sample.

The first supervised alternative algorithm that we use is **Bayesian Model Average (BMA)**. This model is an application of bayesian inference to the model selection problem. Analyze all possible model specifications (possible space) instead of choosing a model specification on a discretionary or random basis giving up information from other models. In this methodology 2^k models are estimated, where k is the total of proposed regressors. We assign a uniform prior for independent variables and we obtain a model that is the weighted average of all possible models. Weights are posterior model inclusion probabilities obtained from the Bayes rule application (Amini and Parmeter, 2011).[9]. We obtain the out of sample prediction of the probability to have investment grade status and define the threshold maximizing the f1 -score as explained in Section 4.1.

The second supervised machine learning algorithm proposed is **linear support vector machine (SVM)**. SVM finds a separating line or a hyperplane (decision boundary); depending on the total number of features that best separates the two classes. It can also be used in a multi-label

[9] Implementation is done using the R statistical program and BMA package following Raftery and Volinsky (2005)

classification problem. The SVM margin is the distance between the hyperplane and the nearest data points and the algorithm finds the best decision boundary such that the margin is maximized (Loukas, 2020).

Algorithms that use distance measures are affected by the units of the features used. This is the case of SVM and K-nearest neighbors. As a consequence, before applying these algorithms it is necessary to preprocess the dataset in order to scale the data. In general, the scaling is performed in the range of $[0, 1]$ or $[-1, 1]$ in those variables that have a bigger range (Arora and Bhambhu, 2014). We choose a range between 0 and 1 to scale the variables before applying these algorithms. There are two techniques recommended to perform the scaling: normalization and standardization. Normalization uses minimum and maximum values to rescale data within the new range of 0 and 1 (Equation 8)[10].

$$\text{Normalized value} = \frac{x - xmin}{xmax - xmin} \tag{8}$$

Standardization assumes that the data follows a Gaussian distribution and rescales the values so that the mean is 0 and the standard deviation is 1 (Equation 9).[11]

$$\text{Standarized value} = \frac{x - \text{mean x}}{\text{standard deviation of x}} \tag{9}$$

Standardization can result in values that are both positive and negative centered around zero and hence outside the range [0,1] and normalization can be applied after standardization in order to guarantee that the rescaled data is in the desired range(Brownlee, 2020d).

To decide the method of rescaling we first standardize the variables that have a range bigger than [0,1] and then perform a Jarque-Bera test, based on the skewness and kurtosis, to determine if the standardized variable follows a Gaussian distribution. If according to the test the variable follows a Gaussian distribution, then we normalize the standardized variable to guarantee that the variable remains in the range between 0 and 1. Otherwise[12] we apply the normalization over the original variable. Results on Jarque-Bera test and more details about the data preprocessing are presented in Annex VII.

K-nearest neighbor algorithm (KNN) is a very simple machine learning algorithm. To make a prediction of a new data point, it finds the closest[13] k-data points in the training set and according to their label classifies the new data point. In the simplest version, the one that we use, we consider only one nearest neighbor and the prediction for the new data is the label of this nearest neighbor in the training set (Müller and Guido, 2016). When considering more neighbors to select the label for the test data the algorithm assigns the majority class among the k-nearest neighbors. As we use distances, we apply this algorithm over the dataset rescaled and we consider the simplest case of 1 neighbor. We also perform a grid search over possible values of k to maximize f1-score.

Classification and decision trees (CART) learn a sequence of if/else questions that gets

[10] Implementation is done using MinMaxScaler from scikit-learn in Python

[11] Implementation is done using StandardScaler from scikit-learn in Python.

[12] If we reject a Gaussian distribution according to the Jarque- Bera test

[13] We use the Euclidean distance to measure closeness.

us to the true classification the most quickly. When a feature is a dummy variable, the question is stated in terms of being 1 or 0 and when the feature is a continuous variable the question is “Is feature i larger than value a?” These questions are named tests. To build a tree, the algorithm searches over all possible tests and finds the one that is most informative about the target variable, that is, that best separates between the two classes[14]. After this first split the process continues recursively until each partition only contains a single class of points. Building a model with this criteria leads to a high overfitting of the data to the training set[15]. To avoid this overfitting, we can early stop the creation of the tree, “limiting the maximum depth of the tree, limiting the maximum number of leaves, or requiring a minimum number of points in a node to keep splitting it “(Müller and Guido, 2016). We limit the depth of the tree as strategy and choose the depth that provides us with a higher f1-score on the test set.

One of the advantages of this algorithm is that it is interpretable and we can obtain the feature importance summary statistic, a number between 0 and 1 for each feature, where 0 is when the feature is “not used at all” and 1 means “perfectly predicts the target” (Müller and Guido, 2016).

Random forest algorithm addresses the overfitting problem in classification and regression trees and it consists of aggregating a number of n random decision trees that are slightly different from one another. In our model we select the default of 100 trees. The algorithm will produce different trees through a bootstrap sample[16] of our train data. To make a prediction, the algorithm performs a prediction for each decision tree. Then a majority vote rule is applied and the sample observation will have the class that was prevalent in the n decision trees (Müller and Guido, 2016).

5 Results

5.1 Logistic regression model

In Table 5 we present the results for the logistic regression considering the dependent variable and the independent variables at the same period of time t. In Model 1 we present the results of using exclusively the macroeconomic variables. According to the results all variables, except general government fiscal balance, current account balance, foreign direct investment and GDP per capita are statistically significant and with the expected sign. Countries with a better governance index, bigger in terms of their share in the world GDP, with higher GDP annual rate of real growth, developed and with more international reserves in terms of GDP, are more likely to be countries with investment grade status. Countries with a past history of default, high levels of inflation, higher government debt over GDP and annual gross government interest payment as a percentage of general government revenues, have less probability to have investment grade status. In Model 2 and Model 3 we introduce some variables related to Fitch’s reports’ analysis.

The coefficients of the macroeconomic variables do not change substantially and maintain their level of significance in the three models. According to the results, when controlling with

[14] We consider the Gini index as the cost function for selecting the split in terms of features and values. A Gini index of 0 corresponds to a perfect separation

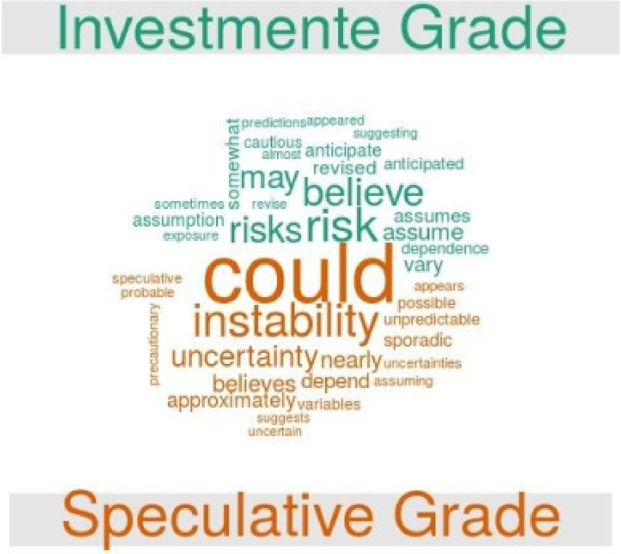
[15] Accuracy in the train set will be 100%

[16] It takes samples with replacement of the data n times and creates a data set that has the same number of observations for each tree, but some data points are missing and some of them are repeated

macroeconomic variables, the uncertainty index is statistically significant and an increase in the index is negatively correlated with having investment grade status. The Positivity index, net negativity index and the deviation in the reports with respect to the country mean or the year mean are not statistically significant.

Given the statistical significance of the uncertainty index, it is interesting to notice that amongst all documents, words like “could”, “risk(s)”, “revised”, “may”, “believe”, “uncertainty”, “assumption”, “volatile” seem to be the most frequent uncertainty related words. Regarding uncertainty associated words alone, and the difference of frequency between documents associated to investment grade versus speculative grade (Figure 9), we find that words like “could”, “instability”, “uncertainty” are more frequent in speculative grade documents than in investment grade. On the other hand “risk” and “risks” are more frequent in investment grade documents, which could sound counterintuitive at first, but since Loughran and McDonald (2011) sentiment dictionary only considers monograms, the “risk” word alone can vary greatly its meaning according to its surrounding context, which is indeed a limitation of this approach.

Figure 9: Loughran and Mc Donald (2011) uncertainty words frequency difference between investment and speculative grade



Countries with reports over the year that have more words reflecting uncertainty with respect to words with some classification (positive or negative) have less probability to have investment grade. This result suggests that there is information in the reports issued by the rating agency that contains additional information that the one present in the macroeconomic variables and as mentioned by Agarwal et al. (2015). Our result is different from the obtained by Slapnik and Loncarski (2019), since even controlling for governance and institutional quality, we find a variable obtained from the text sentiment analysis that is statistically significant. In terms of the predictive performance out of sample the three models achieve the same f1-score (0.95) and recall measures. In Table 6 we present the result of the same models but considering the independent variables lagged one period. These models allow us to predict if a country will have investment grade with information available before the rating is issued by the credit rating agency. In the Model 4, the same macroeconomic variables that were statistically significant in the previous models remain significant and with the same sign when considering the first lag. When including the lag of the variables obtained from the text analysis of the reports issued by Fitch, the current account balance lagged one period over the GDP is statistically significant and positively correlated with having investment grade status. In Model 5 and 6, the uncertainty lagged one year is statistically significant and negatively correlated with having investment grade status. In Model 6, the positivity index lagged one year is also significant and positively correlated with having investment grade the next year. A higher proportion of uncertain words in Fitch's reports decrease the probability of having investment grade status the next year and the opposite effect occurs the higher the positivity index. The f1- score is the same for all the models but recall is higher in models 5 and 6.

Table 5: Logistic regression model results: Dependent variable Investment grade

Independent variable	Model 1	Model 2	Model 3
Constant	3.375***	4.330***	4.405***
	(0.689)	(0.888)	(0.922)
Composite governance indicator	8.797 ***	8.965***	8.866***
	(0.858)	(0.910)	(0.895)
GDP per capita rank	-0.5479	-0.477	-0.577
	(0.443)	(0.451)	(0.447)
Share in world GDP	1.207***	1.256***	1.285***
	(0.110)	(0.129)	(0.130)
Default	-0.929***	-0.987**	-0.949****
	(0.304)	(0.312)	(0.312)
Consumer price index	-0.248***	-0.267***	-0.259***
	(0.046)	(0.048)	(0.048)
Real GDP growth	0.085 *	0.100 **	0.082**
	(0.040)	(0.042)	(0.042)
Gross general government debt	-0.054***	-0.051 ***	-0.051***
	(0.008)	(0.008)	(0.008)
Interest payment	-0.113***	-0.126***	-0.133***
	(0.029)	(0.030)	(0.031)
General government fiscal balance	-0.002	0.013	-0.094
	(0.040)	(0.040)	(0.040)
Official international reserves	0.046***	0.047***	0.043***
	(0.017)	(0.016)	(0.015)
Current account balance	0.014	0.021	0.018
	(0.022)	(0.022)	(0.022)
Foreign direct investment	-0.014	-0.008	-0.0129
	(0.021)	(0.022)	(0.021)
Developed	1.271**	1.023*	0.993*
	(0.553)	(0.574)	(0.569)
Deviation from yearly mean reports		-0.012	-0.221
		(0.050)	(0.052)
Deviation from country mean reports		0.007	0.032*
		(0.053)	(0.054)
Net negativity index		0.982	
		(0.719)	
Positivity index			1.006
			(1.364)
Uncertainty index		-4.836***	-4.349**
		(1.778)	(1.748)
Threshold	0.418549	0.431570	0.433043
Pseudo R-squared	0.702	0.708	0.707
Precision	0.92	0.92	0.92
Recall	0.98	0.98	0.98
f1-score	0.95	0.95	0.95

standard errors between parentheses, N= 1126

***, **, * represent statistical significance at 1%, 5% and 10% respectively

Table 6: Logistic regression model results: Dependent variable Investment grade and independent variables in t-1

Independent variable	Model 4	Model 5	Model 6
Constant	2.891*** (0.750)	4.471*** (0.933)	3.579*** (0.986)
Composite governance indicator t-1	9.943 *** (0.951)	10.138*** (0.985)	10.088 *** (1.027)
GDP per capita rank t-1	- 0.093 (0.483)	-0.025 (0.496)	0.030 (0.496)
Share in world GDP t-1	1.239*** (0.122)	1.368*** (0.153)	1.380*** (0.155)
Default t-1	-0.931*** (0.318)	-0.944*** (0.325)	-0.929*** (0.329)
Consumer price index t-1	-0.3224*** (0.053)	-0.346*** (0.055)	-0.338*** (0.055)
Real GDP growth t-1	0.099*** (0.036)	0.090*** (0.038)	0.081 (0.039)
Gross general government debt t-1	-0.053*** (0.009)	-0.049*** (0.009)	-0.050*** (0.009)
Interest payment t-1	-0.108*** (0.030)	-0.140*** (0.034)	-0.145*** (0.033)
General government fiscal balance t-1	0.055 (0.046)	0.046 (0.046)	0.034 (0.047)
Official international reserves t-1	0.069*** (0.017)	0.065*** (0.017)	0.066*** (0.017)
Current account balance t-1	0.038 (0.024)	0.045* (0.024)	0.044* (0.024)
Foreign direct investment t-1	0.023 (0.029)	0.029 (0.025)	0.025 (0.028)
Developed t-1	1.140* (0.585)	0.845 (0.597)	0.984* (0.595)
Deviation from yearly mean reports t-1		-0.017 (0.059)	-0.041 (0.061)
Deviation from country mean report t-1		-0.034 (0.060)	-0.004 (0.062)
Net negativity index t-1		-0.493 (0.754)	
Positivity index t-1			2.976** (1.485)
Uncertainty index t-1		-4.496** (1.782)	-4.740 * (1.836)
Threshold	0.518952	0.369267	0.356977
Pseudo R-squared	0.717	0.726	0.728
Precision	0.94	0.91	0.91
Recall	0.96	0.98	0.98
f1-score	0.95	0.95	0.95

standard errors between parentheses, N=1059
***, **, * represent statistical significance at 1%, 5% and 10% respectively

5.2 Alternative supervised machine learning algorithms

Alternatively, to the logistic regression model presented in subsection 5.1., we implemented bayesian model average, support vector machine, k-nearest neighbor, classification and decision trees and random forest algorithms to the binary classification problem and compare the predictive performance out of the sample of these models. To compare the results, we present as a benchmark the best logistic regression in terms of f1-score and recall.

Using **Bayesian model average** a total of 23 logistic models were selected and averaged considering the dependent and independent variables referred to the same period of time and a total of 43 models were selected and average considering the independent variables lagged one year. In the Annex VI we present the results in terms of the expected value of the independent variables included in the models.

In the model where independent variables and investment grade status are referred to the same year the uncertainty index is included in 17 of the 23 models and negatively correlated with having investment grade status reflecting that it is a relevant variable to be included. The BMA predicted values, using a threshold that maximizes f1-score[17], achieves a lower f1-score than our benchmark logistic models (Table 7).

In the model where independent variables are lagged one year the uncertainty index is selected in 30 out of the 43 models estimated and it is negatively correlated with having investment grade status the following year. In this specification positivity index is included in 22 models out of the 43 models estimated. This index is positively correlated with having investment grade status the following year. In terms of the out of sample predictive performance, using a threshold that maximizes f1-score[18], achieves the same f1-score than our benchmark logistic models (Model 5 and 6) but with a slightly lower Recall metric (Table 8)

The second alternative algorithm is support vector machine (SVM). Before applying the algorithm, we normalize the variables in the dataset as described in section 4.1.1 and Annex VII. Comparing the predictive performance, SVM algorithm achieves a lower f1-score and recall than the benchmark when the model is referred to the same year or independent variables are lagged one year (Table 7 and Table 8).

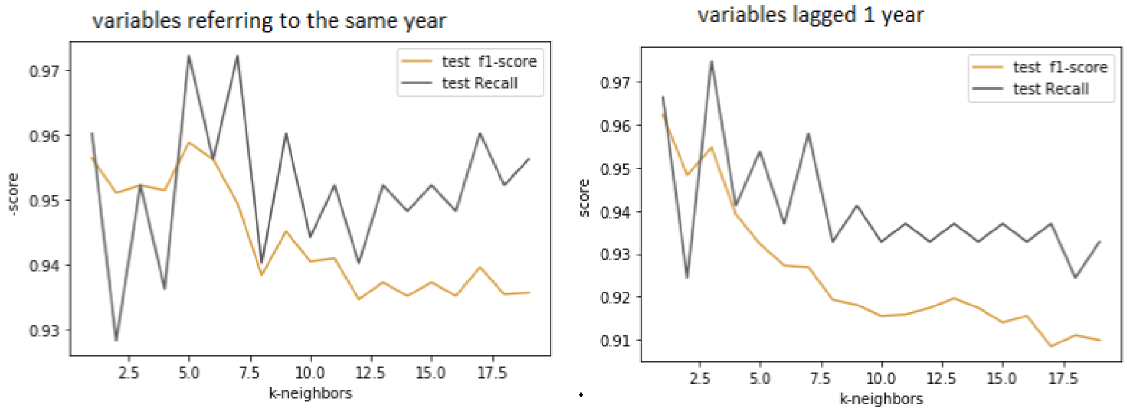
K-nearest neighbors algorithm (KNN) considering k=1, only the nearest neighbor, performs better in terms of the f1-score and precision than the benchmark when the model is referred to the same year or independent variables are lagged one year (Table 8). We perform a grid search to determine the number of neighbors to be considered that maximizes the f1-score and the recall. For the first group of models, with independent variables referred to the same year, considering a total of 5 neighbors maximizes the f1-score obtained with a higher recall than the model considering only 1 neighbor. When considering the models with one lagged variable, the best predictive performance in terms of f1-score is obtained by considering only 1 neighbor and the best predictive performance in terms of recall is obtained considering a total of 3 neighbors (Figure 10 and Table 8).

Classification and decision tree algorithm (CART) when considering the dependent and inde-

[17] BMA threshold equal to 0.347484

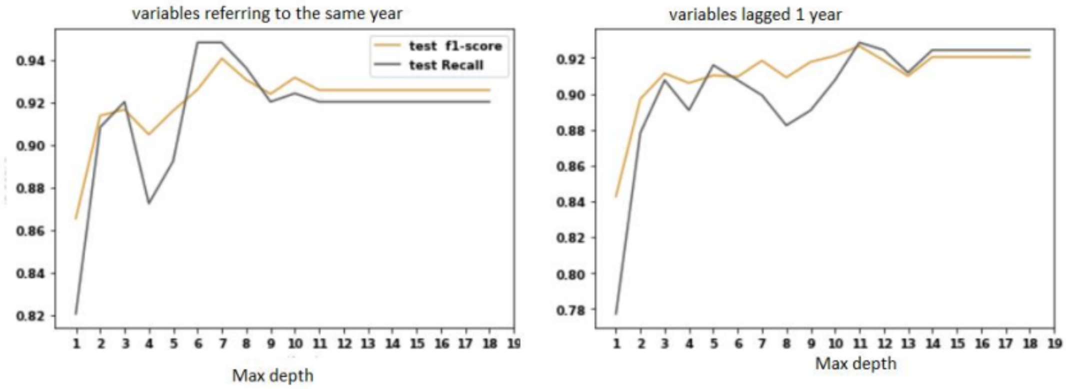
[18] BMA with independent variables lagged one year equal to 0.380729

Figure 10: Grid search of the optimal number of neighbors



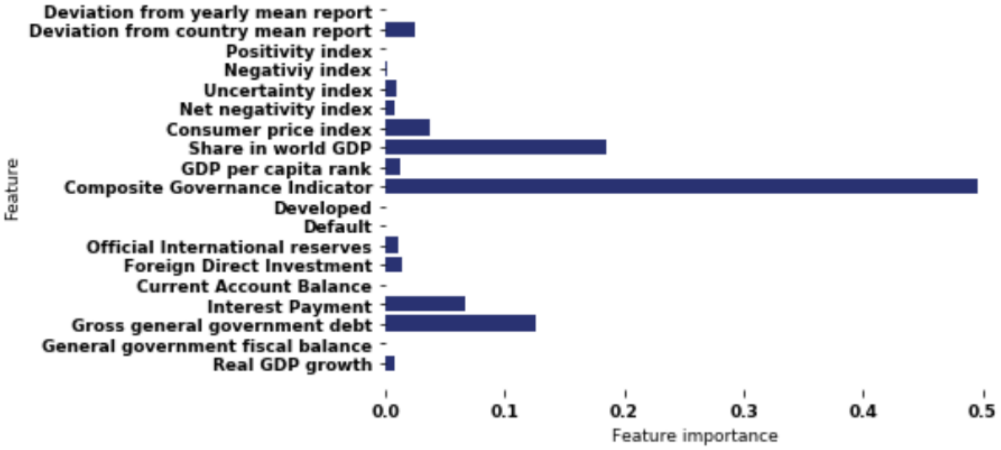
pendent variables referred to the same year achieves an f1- score of 0.93, a recall of 0.92 and a precision of 0.93. To avoid overfitting, we limit the depth of the decision tree algorithm doing grid search and choosing the max depth that maximizes the f1-score (Figure 11). With a maximum depth of 7, f1-score increases to 0.94, recall to 0.95 and precision to 0.93. When considering the independent variables lagged one year CART algorithm achieves an f1- score of 0.92, a recall of 0.92 and a precision of 0.92. Using a maximum depth 11, resulting for a grid search, f1-score, recall and precision increases to 0.93.

Figure 11: Grid search for maximum depth in CART



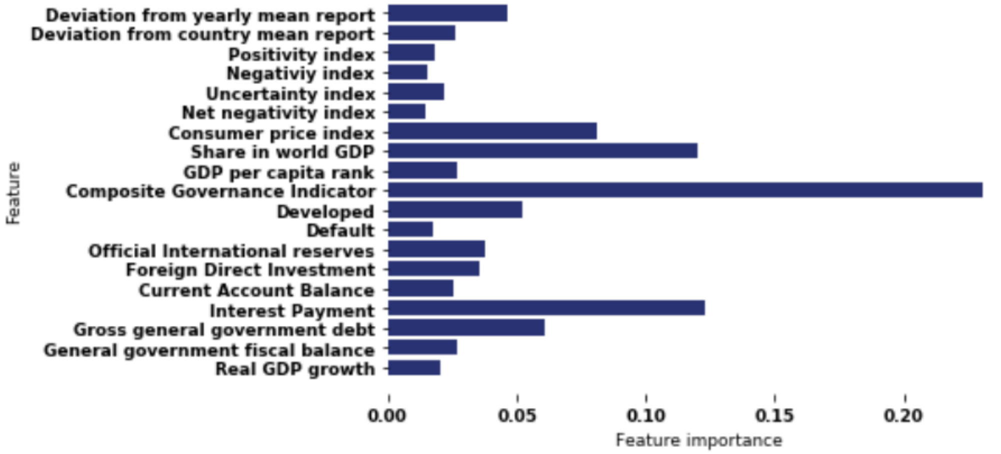
Variables that are more relevant for the classification problem according to this algorithm are the composite governance indicator, followed by the share in world GDP, gross general government debt and interest payment (with feature importance of 0.5, 0.19, 0.13 and 0.12 respectively). Respect to the variables obtained from Fitch's report text analysis the deviation from countries mean report, the uncertainty and net negativity index have feature importance metrics different from zero (Figure 12).

Figure 12: Feature importance captured by the CART algorithm



Random forest algorithm, with the default of 100 trees achieves the higher f1-score and recall, of 0.97 and 0.99 respectively and it is the best model in terms of predictive performance on the test set. When considering the independent variables lagged one period, the algorithm reaches an f1-score of 0.96 and a recall of 0.97. In this model the predictive performance is as good as the K-nearest neighbor algorithm with K=1. Feature importance in the random forest algorithm gives non-zero importance to many more features than the CART algorithm (Figure 13). As random forest aggregate features importance over the trees considered, the result is “more reliable than the ones provided by a single tree”(Müller and Guido, 2016).

Figure 13: Feature importance captured by the Random Forest algorithm



The composite governance indicator is the most important feature, followed by the interest payment and the share in the world GDP variables. From the text analysis variables, deviation from the yearly mean report followed by the deviation from the country mean report and the uncertainty index are the most relevant features. If we aggregate the feature importance for the text analysis variables, they represent a 14% of the total. This result suggests that including these variables is relevant to predict investment grade status of sovereigns.

The best model in terms of predictive performance out of sample considering the f1-score and recall when considering the dependent and independent variables referred to the same year is Random Forest (Table 7). In this model, text sentiment analysis variables are relevant as their feature importance summary metrics are non-zero.

Table 7: Predictive performance of supervised algorithms

	Benchmark	**BMA**	**SVM**	**KNN(k=1)**	**KNN(k=5)**	**CART**	**Random Forest**
Precision	0.92	0.91	0.92	0.95	0.95	0.93	0.95
Recall	0.98	0.98	0.96	0.96	0.97	0.95	0.99
F1-score	0.95	0.94	0.94	0.96	0.96	0.94	0.97

When considering the independent variables lagged one-year, Random Forest and K-nearest neighbors with K=1 have the higher performance in terms of f1-score (Table 8). The predictive performance from using the independent variable lagged one period or the contemporaneous value is almost the same.

Table 8: Predictive performance of supervised algorithms with independent variables lagged one year

	Benchmark	**BMA**	**SVM**	**KNN(k=1)**	**KNN(k=5)**	**CART**	**Random Forest**
Precision	0.91	0.92	0.91	0.95	0.94	0.93	0.95
Recall	0.98	0.97	0.95	0.97	0.98	0.93	0.97
F1-score	0.95	0.95	0.93	0.96	0.95	0.93	0.96

6 Final comments

We successfully use text sentiment analysis in Fitchs' reports to generate new text variables to use them along with macroeconomic variables to apply machine learning algorithms in order to understand statistical significance and predictive power of these features. We applied logistic regression as our base algorithm. Considering the dependent and independent variables at the same period of time we found that when controlling with macroeconomic variables the uncertainty index created is statistically significant. This suggests that there is additional information in the reports that the one present in the macroeconomic variables as mentioned by Agarwal et al. (2015) However, our results differ from the obtained by Slapnik and Loncarski (2019) since even controlling for governance and institutional quality, we find a variable obtained from the text sentiment analysis that is statistically significant.

Regarding the predictive performance there is no improvement when included text variables (Table 5). However, if we consider the independent variables lagged one period we see a slight improvement in the predictive performance (Table 6). This allows us to predict if a country will have investment grade with information available before the rating is issued by the credit rating agency.

We implement five alternative supervised machine learning algorithms and found that with respect to our base algorithm two have a slightly better performance: K-Nearest Neighbor algorithm and Random Forest. To conclude the low incidence of the text feature analyzed from the reports can be expected since the incidence of the Qualitative Overlay (QO) on the rating is probably less than the Sovereign Rating Model (SRM) and the macroeconomic variables.

As a final comment this paper can be extended by using this research to predict if a sovereign will maintain, upgrade or downgrade its rating and also to predict the rating issued by Fitch. This is a multi-label classification problem and is outside the scope of this work. Future work can include the estimation of the models presented in this work to other rating agencies as Moody's and Standard and Poor's.

References

Agarwal, S., Chen, G., V.and Sim, and Zhang, W. (2015). The information value of sovereign credit rating reports.

Amini, S., and Parmeter, C. (2011). Bayesian model averaging in r. *Journal of Economic and Social Measurement*, *36(4)*.

Arora, M., and Bhambhu, L. (2014). Role of scaling in data classification using svm. *InternationalJournal of Advanced Research in Computer Science and Software Engineering*.

Borraz, F., Fried, A., and Gianelli, D. (2011). Análisis de las calificaciones de riesgo soberano: El caso uruguayo. *Revista de Economía del BCU*.

Brownlee, J. (2020). A gentle introduction to threshold-moving for imbalanced classification.

Brownlee, J. (2020b). Failure of classification accuracy for imbalanced class distributions.

Brownlee, J. (2020c). Train-test split for evaluating machine learning algorithms.

Brownlee, J. (2020d). How to use standardscaler and minmaxscaler transforms in python.

Butler, A. W., and Fauver, L. (2006). Institutional environment and sovereign credit ratings. *Financial Managment*, *35(3)*, 53-79.

Cantor, R., and Packer, F. (1996). Determinants and impact of sovereign credit ratings. *Economic policy review*, *2(2)*.

Fitch. (2011). Rating definitions. special report.

Fitch. (2020). Sovereign rating criteria. rating criteria.

Loughran, T., and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 35-65.

Loukas, S. (2020). Support vector machines (svm) clearly explained: A python tutorial for classification problems with 3d plots.

Moreno Bernal, I., and González Pedraz, C. (2020). Análisis de sentimiento del "informe de estabilidad financiera". *Documentos de trabajo. Banco de España*, *N 2011*.

Müller, A. C., and Guido, S. (2016). Introduction to machine learning with python: a guide for data. *O'Reilly Media, Inc.*.

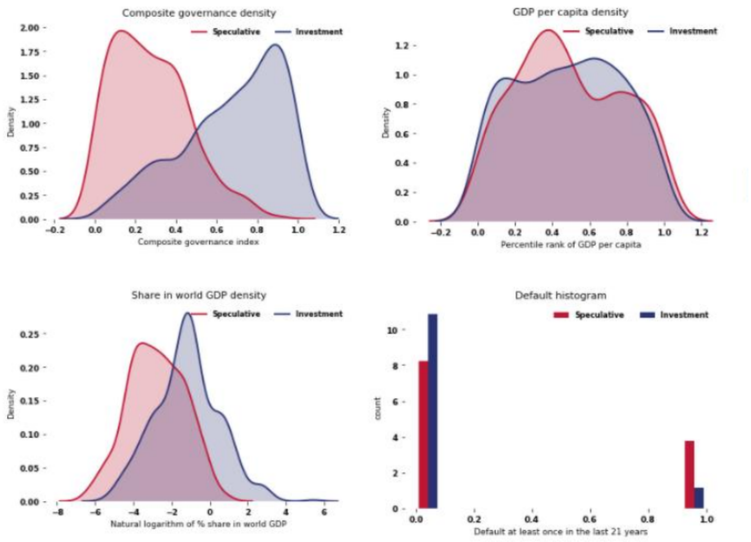
Raftery, I. S., A. E.and Painter, and Volinsky, C. T. (2005). Bma: an r package for bayesian model averaging. *The Newsletter of the R Project*, *5(2)*.

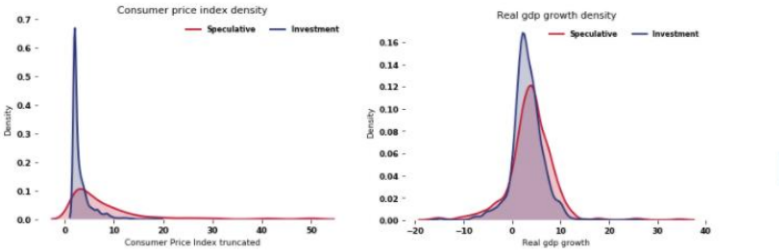
Shung, K. (2018). Accuracy, precision, recall or f1?

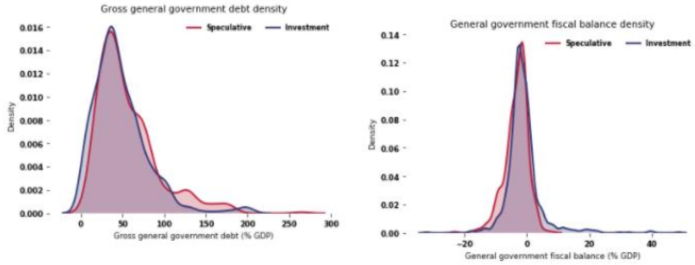
Slapnik, and Loncarski, I. (2019). Understanding sovereign credit ratings: Text-based evidence from the credit rating reports. *Capital Markets: Market Efficiency eJournal*.

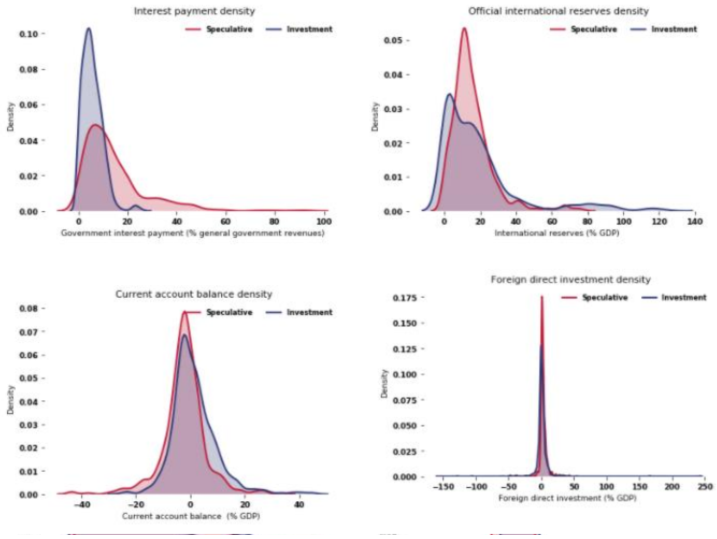
Annex

I. Macroeconomic variables exploratory analysis

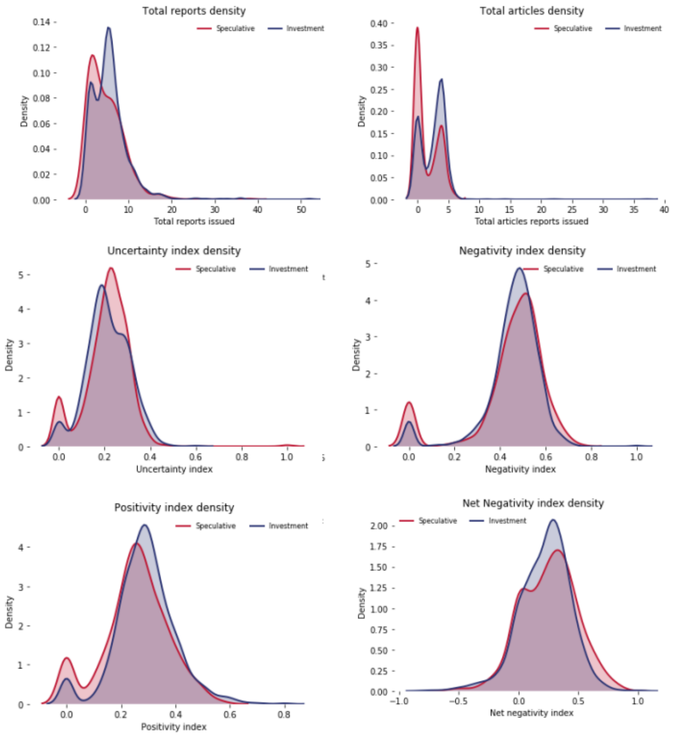








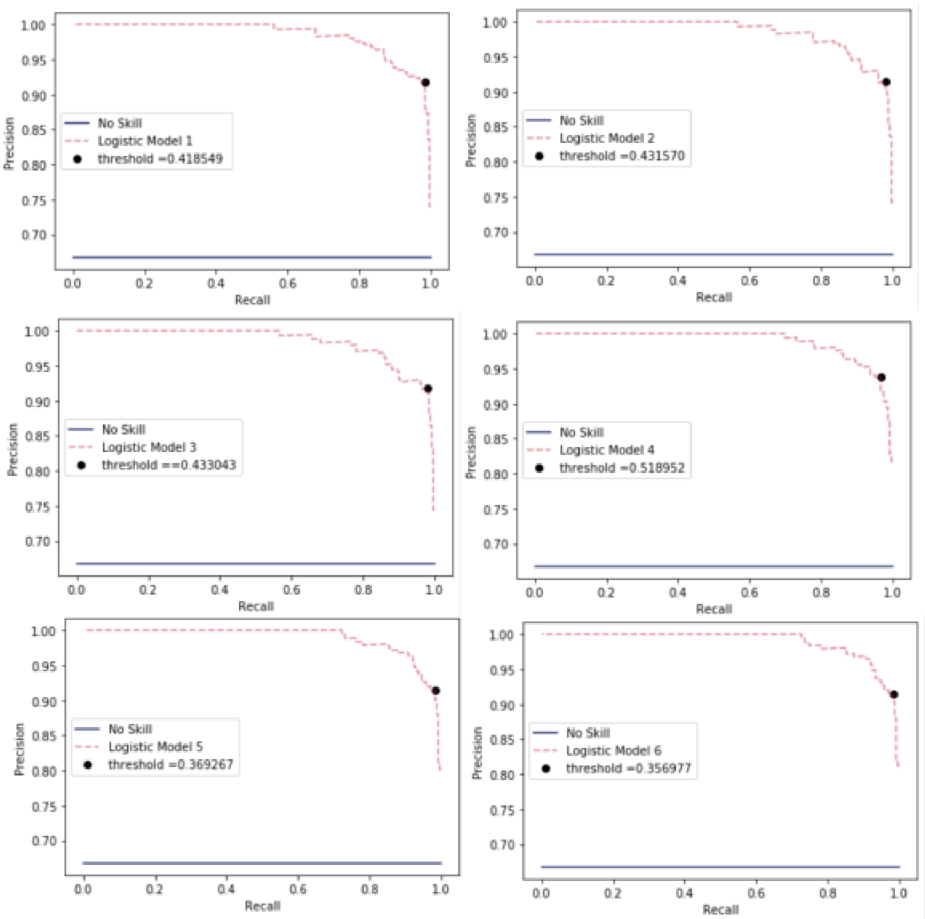
II.Text variables exploratory analysis



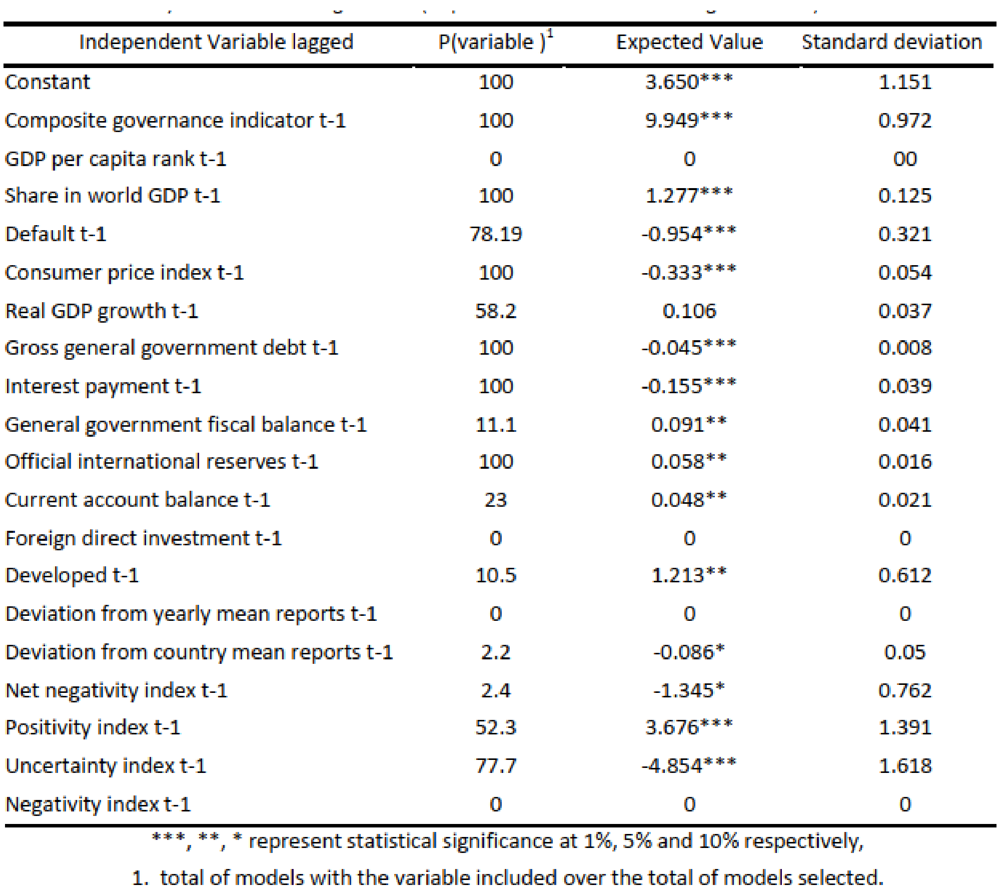
II. Countries and change in their investment grade status

Country	**Years**	**Change in investment grade status (SG= Speculative Grade, IG= Investment Grade)**	**For countries with changes in their investment grade status**	
			Year in which the country obtained an investment grade rating	Year in which the country obtained an speculative grade rating
Angola	2010-2018	No/SG		
Argentina	2000-2018	No/SG		
Armenia	2014-2018	No/SG		
Australia	2000-2018	No/IG		
Austria	2000-2018	No/IG		
Azerbaijan	2000-2018	Yes	2010	2016
Bahrain	2000-2018	Yes		2016
Bangladesh	2014-2018	No/SG		
Belarus	2016-2018	No/SG		
Belgium	2002-2018	No/IG		
Bolivia	2004-2018	No/SG		
Brazil	2000-2018	Yes	2008	2015
Bulgaria	2000-2018	Yes	2004	
Canada	2000-2018	No/IG		
Chile	2000-2018	No/IG		
China	2000-2018	No/IG		
Colombia	2000-2018	Yes	2011	
Croatia	2000-2018	Yes	2001	2012
Cyprus	2000-2018	Yes		2012
Czech RepubliC	2000-2018	No/IG		
Denmark	2000-2018	No/IG		
Dominican Republic	2003-2018	No/SG		
Ecuador	2003-2018	No/SG		
Egypt	2002-2018	No/SG		
El Salvador	2000-2018	No/SG		
Estonia	2000-2018	No/IG		
Finland	2000-2018	No/IG		
France	2000-2018	No/IG		
Georgia	2007-2018	No/SG		
Germany	2000-2018	No/IG		
Greece	2000-2018	Yes		2011
Guatemala	2006-2018	No/SG		
Honduras	2015-2018	No/SG		
Hong Kong	2000-2018	No/IG		
Hungary	2000-2018	Yes	2016	2012
Iceland	2000-2018	Yes	2012	2010
India	2000-2018	Yes	2006	
Indonesia	2000-2018	Yes	2012	
Ireland	2000-2018	No/IG		
Israel	2000-2018	No/IG		
Italy	2000-2018	No/IG		
Jamaica	2006-2018	No/SG		
Japan	2000-2018	No/IG		
Kazakhstan	2000-2018	Yes	2004	
Kuwait	2000-2018	No/IG		
Latvia	2000-2018	Yes	2011	2009
Lebanon	2000-2018	No/SG		
Luxembourg	2002-2018	No/IG		
Malaysia	2000-2018	No/IG		
Malta	2000-2018	No/IG		
México	2000-2018	Yes	2002	
Moldova	2000-2018	No/SG		
Mongolia	2005-2018	No/SG		
Morocco	2007-2018	No/IG		
Netherland	2000-2018	No/IG		
New Zealand	2002-2018	No/IG		
Norway	2000-2018	No/IG		
Omán	2017-2018	Yes		2018
Pakistan	2015-2018	No/SG		
Panamá	2000-2018	Yes	2010	
Papua New Guinea	2000-2018	No/SG		
Paraguay	2013-2018	No/SG		
Perú	2000-2018	Yes	2008	
Philippines	2000-2018	Yes	2014	
Poland	2000-2018	No/IG		
Portugal	2000-2018	Yes	2018	2011
Qatar	2015-2018	No/IG		
Romania	2000-2018	Yes	2006 and 2011	2008
Russia	2000-2018	Yes	2004	
Saudi Arabia	2004-2018	No/IG		
Singapore	2000-2018	No/IG		
Slovakia	2000-2018	Yes	2002	
Slovenia	2000-2018	No/IG		
South Africa	2000-2018	Yes	2002	2017
Spain	2000-2018	No/IG		
Sri Lanka	2007-2018	No/SG		
Suriname	2004-2018	No/SG		
Sweden	2000-2018	No/IG		
Switzerland	2000-2018	No/IG		
Taiwan	2000-2018	No/IG		
Thailand	2000-2018	No/IG		
Tunisia	2000-2018	Yes		2012
Turkey	2000-2018	Yes	2013	2017
Ukraine	2001-2018	No/SG		
United Arab Emirates	2007-2018	No/IG		
United Kingdom	2000-2018	No/IG		
United States	2000-2018	No/IG		
Uruguay	2000-2018	Yes	2013	2002
Venezuela	2000-2018	No/SG		
Vietnam	2016-2018	No/SG		

III.Precision-Recall curves logistic regression models



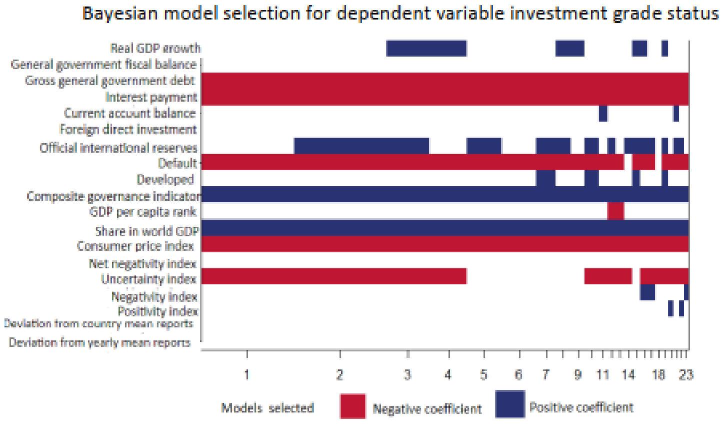
IV.Bayesian model average results (Dependent variable Investment grade status



Independent Variable lagged	P(variable)[1]	Expected Value	Standard deviation
Constant	100	3.650***	1.151
Composite governance indicator t-1	100	9.949***	0.972
GDP per capita rank t-1	0	0	00
Share in world GDP t-1	100	1.277***	0.125
Default t-1	78.19	-0.954***	0.321
Consumer price index t-1	100	-0.333***	0.054
Real GDP growth t-1	58.2	0.106	0.037
Gross general government debt t-1	100	-0.045***	0.008
Interest payment t-1	100	-0.155***	0.039
General government fiscal balance t-1	11.1	0.091**	0.041
Official international reserves t-1	100	0.058**	0.016
Current account balance t-1	23	0.048**	0.021
Foreign direct investment t-1	0	0	0
Developed t-1	10.5	1.213**	0.612
Deviation from yearly mean reports t-1	0	0	0
Deviation from country mean reports t-1	2.2	-0.086*	0.05
Net negativity index t-1	2.4	-1.345*	0.762
Positivity index t-1	52.3	3.676***	1.391
Uncertainty index t-1	77.7	-4.854***	1.618
Negativity index t-1	0	0	0

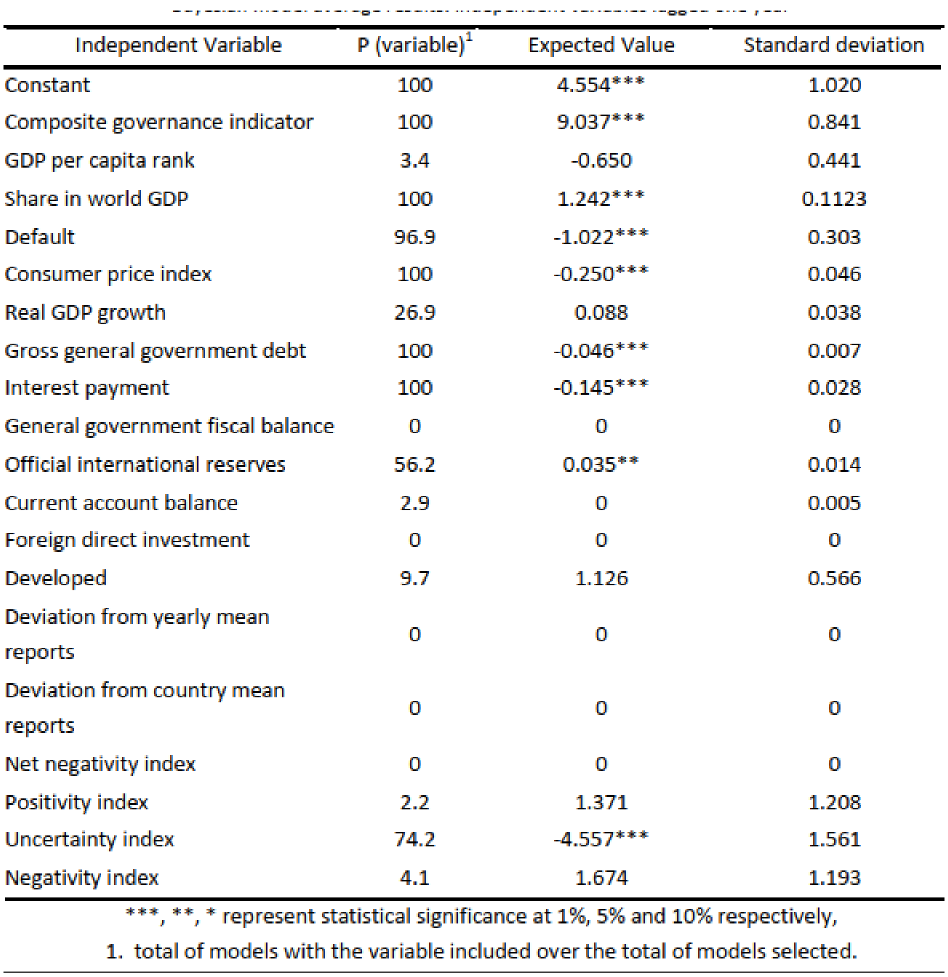
***, **, * represent statistical significance at 1%, 5% and 10% respectively,

1. total of models with the variable included over the total of models selected.



Bayesian model selection for dependent variable investment grade status

V.Bayesian model average results (Dependent variable Investment grade status and independent variables lagged one period

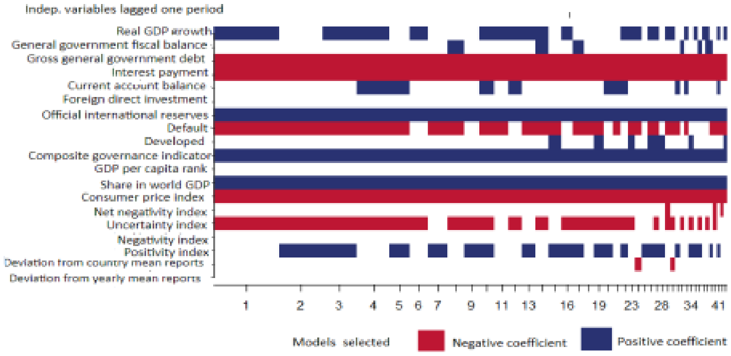


Independent Variable	P (variable)[1]	Expected Value	Standard deviation
Constant	100	4.554***	1.020
Composite governance indicator	100	9.037***	0.841
GDP per capita rank	3.4	-0.650	0.441
Share in world GDP	100	1.242***	0.1123
Default	96.9	-1.022***	0.303
Consumer price index	100	-0.250***	0.046
Real GDP growth	26.9	0.088	0.038
Gross general government debt	100	-0.046***	0.007
Interest payment	100	-0.145***	0.028
General government fiscal balance	0	0	0
Official international reserves	56.2	0.035**	0.014
Current account balance	2.9	0	0.005
Foreign direct investment	0	0	0
Developed	9.7	1.126	0.566
Deviation from yearly mean reports	0	0	0
Deviation from country mean reports	0	0	0
Net negativity index	0	0	0
Positivity index	2.2	1.371	1.208
Uncertainty index	74.2	-4.557***	1.561
Negativity index	4.1	1.674	1.193

***, **, * represent statistical significance at 1%, 5% and 10% respectively,

1. total of models with the variable included over the total of models selected.

Bayesian model selection for dependent variable investment grade status and independent variables lagged one period.



VI.Jarque- Bera test and results on the scaling process

The Jarque-Bera test is a goodness-of-fit test to verify whether a data sample has the skewness and kurtosis of a normal distribution.

Null hypothesis: The data belongs to a normal distribution.

JB: $\frac{n}{6} \times \ (S^2 \ + 0.25 \ \times \ (\ K - 3)^2) \sim \chi^2{}_2$ where S is the symmetry, K the kurtosis in the sample and n is the total of observations. If the p-value is less than 0.05 we reject the null hypothesis and use normalization over the original data-set. If the p-value is bigger than 0.05 we fail to reject the null hypothesis and we apply standardization over the original data and then normalization over the standardized data. In Table X we present the JB statistic, the p- value, the transformation according to the result of the test and the mean and standard deviation after it.

Jarque-Bera test results and variable statistics after transformation

Variable	JB	p-value	Mean	Std. dv.
Share in world GDP	6.42	0.04	0.400	0.150
Consumer price index	43785	0.0	0.06	0.121
Real GDP growth	3562	0.0	0.387	0.075
Gross general government debt	1707	0.0	0.191	0.130
Interest payment	20455	0.0	0.093	0.098
General government fiscal balance	21210	0.0	0.375	0.065
Official international reserves	4976	0.0	0.141	0.157
Current account balance	1606	0.0	0.492	0.094
Foreign direct investment	2235509	0.0	0.397	0.029
Deviation from yearly mean reports	22546	0.0	0.206	0.070
Deviation from country mean reports	31783	0.0	0.171	0.728
Net negativity index	87	0.0	0.568	0.123

Composite governance indicator, GDP per capita rank, default, developed, Positivity index and Uncertainty index ranges between 0 and 1. For these variables no transformation is implemented.

According to the results, we reject the null hypothesis for all the variables considered and we implement normalization as a rescaling strategy over the original variables.