International, Credit Ratings

Predicting Bond Ratings with Machine Learning and Logistic Regression

A machine learning program in Python is used to predict which countries currently have investment grade debt

Mongolia has been downgraded three times by Moody's Investors Service this year, bringing the country's credit rating to Caa1.

By ALEX MONAHAN

Updated Oct. 11, 2016 12:29 p.m. ET

When a bond issuing country wants to raise additional capital, the country will issue new securities in the debt market. Although bonds can have various structures (i.e. “10 year,” “30 year,” “callable,” “puttable,” “zero-coupon”), the credit rating is an essential aspect of any debt security. Credit ratings help an investor determine the risk of a particular security, with a low credit rating indicating increased risk and a higher probability of default. Thus, as the credit rating of a country decreases, the issuing country typically needs to offer investors a higher yield.

Bond ratings can be divided into two main segments: investment grade debt and non-investment grade (or “junk”) debt. The distinction between investment grade debt and non-investment grade debt is important, as many investors prefer to only purchase investment grade debt, since these securities are viewed as “safer” assets. In this snapshot report, machine learning and logistic regression are utilized together to predict whether or not a country currently has an investment grade level credit rating.

Logistic regression was the appropriate analysis tool in this report because the dependent variable (whether or not a country has an investment grade level credit rating) is dichotomous, or binary. In addition, logistic regression is frequently used to analyze financial data. This form of analysis rose to popularity in financial fields when researchers used logistic regression to study the non-investment grade bond market, classifying companies as either defaulters or non-defaulters.

In this analysis, I solely used credit ratings from Moody’s Investors Service to determine whether a particular country currently has investment grade debt. Investment grade debt is defined as securities with a Moody’s ratings of Baa3 or above. Since the logistic regression model requires that the dependent variable can have only two possible outcomes, countries were separated into two different categories: investment grade debt or non-investment grade debt. Thus, for the i^{th} country in the dataset (115 total countries were analyzed), Y_i is defined such that $Y_i = 0$ if the country currently has investment grade debt and $Y_i = 1$ if the country currently has non-investment grade debt.

Using a variety of sources, including Federal Reserve Economic Data (FRED) and the World Bank Data Catalog, I also collected seven attributes for each country to be used as covariates in the logistic regression model. These attributes were the most recent full-year GDP growth rate (attribute 1, x_{i1}), the current unemployment rate in the country (attribute 2, x_{i2}), the most recent full-year inflation rate (attribute 3, x_{i3}), the current central bank interest rate (attribute 4, x_{i4}), a binary response of whether or not the country is running a current account deficit or surplus (attribute 5, x_{i5}), a binary response of whether or not the country is a net oil exporter (attribute 6, x_{i6}), and the current ratio of government debt to GDP (attribute 7, x_{i7}). For the i^{th} country, each of the attributes was encoded numerically as an individual covariate ($x_{i1}, \ldots, x_{i7}$). As an example, x_{i4} represents the interest rate of the i^{th} country in the dataset.

Following the steps to complete logistic regression, each country (from i = 1…115) was assumed to be an independent random variable with a distribution of Bernoulli(p_i), where each p_i was calculated with the specific covariates for the i^{th} country. For each country, the log odds of p_i, or the logit, was modeled as a linear combination of the seven covariates described above. In other words, for the i^{th} country, the data was modeled with the following formula, where $x_{i1} \ldots x_{ip}$ represent the different covariates.

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

The coefficients β_1 … β_7 are the unknown parameters we are estimating in logistic regression, where each β_j (for j=1…7) represents the increase or decrease in the logit if the jth covariate is increased by one. For each country, the equation above can be rewritten as follows:

$$\mathbb{P}[Y_i = 1] = p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}}}$$

As stated previously, in logistic regression, the covariates (x_{ij}) are treated as known, fixed quantities. For example, if Turkey were the 17th country in the dataset, then $\log(p_{17}/(1-p_{17})) = \beta_0 + \beta_1 x_{17,1} + \beta_2 x_{17,2} + \beta_3 x_{17,3} + \beta_4 x_{17,4} + \beta_5 x_{17,5} + \beta_6 x_{17,6} + \beta_2 x_{17,7}$, where $x_{17,1}$ = the most recent full year GDP growth rate in Turkey, $x_{17,2}$ = the most recent Turkish unemployment rate, $x_{17,3}$ = the most recent full year inflation rate in Turkey, $x_{17,4}$ = the current Turkish central bank interest rate, $x_{17,5}$ = whether Turkey currently has a current account surplus or deficit (0 = deficit, 1 = surplus), $x_{17,6}$= whether Turkey is a net oil importer or exporter (0 = net importer, 1 = net exporter), and $x_{17,7}$ = the current debt to GDP ratio of Turkey.

Next, the likelihood function of the logistic regression model was calculated as follows, where each Y_i is modeled as a typical Bernoulli variable. Again, we are solving for the β terms in this analysis, as the parameters $\beta_1 \ldots \beta_7$ are the unknown values.

$$\mathrm{lik}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = \prod_{i=1}^{n}(1-p_i)\left(\frac{p_i}{1-p_i}\right)^{Y_i}$$

Then, introducing a logarithm yields the log likelihood equation, as seen below.

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} Y_i \log \frac{p_i}{1-p_i} + \log(1-p_i) = \sum_{i=1}^{n}\left(Y_i \sum_{j=0}^{p} \beta_j x_{ij} - \log\left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}}\right)\right)$$

To compute the Maximum Likelihood Estimator (MLE) for each coefficient $\beta_1 \ldots \beta_7$, we calculate seven separate partial derivatives (with respect to $\beta_1 \ldots \beta_7$). As seen below, setting each equation equal to zero yields seven equations with seven unknown values, and each MLE $\beta_1 \ldots \beta_7$ can be determined using the numerical Newton-Raphson method. As background on the Maximum Likelihood Estimator, intuitively, the MLEs of this analysis are the values of the parameters $\beta_1 \ldots \beta_7$ that make the observed data of this analysis "most probable" or "most likely."

$$0 = \frac{\partial l}{\partial \beta_m} = \sum_{i=1}^{n} x_{im} \left(Y_i - \frac{e^{\sum_{j=0}^{p} \beta_j x_{ij}}}{1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}}} \right)$$

In this report, Python was the programming language utilized to calculate the Maximum Likelihood Estimators for $\beta_1 \ldots \beta_7$ and to implement the machine learning predictive analysis (with import sklearn). As in most machine learning analyses, the classification algorithm was trained with the majority of the data, and then the algorithm was back-tested and used to analyze the remaining portion of the dataset. In this report, 80% of countries in the dataset (approximately 92 countries) were used to train the algorithm, and then the algorithm was tested on the remaining 20% of the countries.

The model was 81% accurate on the tested data. The AUC (area under the curve) of the logistic model was 0.77. The MLEs $\beta_1 \ldots \beta_7$ are given below.

$$\hat{\beta}_1 = 1.1078$$
$$\hat{\beta}_2 = 0.1718$$
$$\hat{\beta}_3 = 2.1045$$
$$\hat{\beta}_4 = 1.3454$$
$$\hat{\beta}_5 = -0.3868$$
$$\hat{\beta}_6 = -0.2447$$
$$\hat{\beta}_7 = 0.3812$$

As an example of what the above MLEs signify, for a one percent increase in year on year GDP growth of a given country (xi1), the logit of the probability of a country having investment grade debt (p_i) increases by 1.1078.

Finally, for each of the seven attributes (GDP growth rate, unemployment rate, etc.), 'R' was used to complete a Generalized Likelihood Ratio Test (GLRT) for the following null and alternative hypotheses: H_0: $\beta_j = 0$,

H_1: $\beta_j \neq 0$. Using a significance level of $\alpha = 0.05$, the GLRT rejected any null hypothesis with $p < 0.05$, where, as usual, p is the probability of a Type 1 Error. As seen in the hypotheses above, the GLRT is used to study whether a given covariate has a statistically significant effect on whether or not a country currently has investment grade debt.

Three of the seven covariates yielded Generalized Likelihood Ratio Tests with $p < 0.05$ (GDP growth rate, interest rate, and whether or not the country has a current account deficit or surplus). Thus, according to the data, only these three covariates have a statistically significant effect on whether or not a given country currently has investment grade debt.

As a note, logistic regression assumes no outliers are present in the data, and logistic regression also assumes none of the predictors have high correlations with other predictors. In addition, logistic regression assumes each country is distributed independently with a Bernoulli(p_i) distribution, where the logit of p_i is modeled as a linear combination of the seven covariates described earlier in this report. These assumptions of the logistic regression model are not trivial. As an example, the European Central Bank (ECB) has the ability to implement policies that affect numerous countries analyzed in this report. If the ECB implements a new policy that requires countries to lower their debt to GDP ratios, for instance, then the credit ratings of various countries in the European Union may improve, as securities issued by these countries could be viewed as safer investments. Thus, the credit ratings of countries may not be truly independent. Still, logistic regression is often used to analyze binary classification problems, even if the model does not yield an extremely accurate fit to the data. In these cases, the MLEs $\beta_1 \ldots \beta_p$ for $j = 1 \ldots p$ covariates represent the “closest” logistic regression model that could be fit to the true distribution of $Y_1 \ldots Y_n$.

Click here to download the data used in this snapshot financial report.