Dashboard Project: Milestone IV

Data Warehousing and Business Intelligence - Computing Lab

Ivan Vallejo Vall, Saurav Poudel & Daniel Velásquez Vergara

November 28, 2016

1 Analytics

1.1 Regression

One of the objectives of the project is to identify key macroeconomic indicators strongly related to telecom development. For this exercise, the indicator "Percentage of Internet users" was chosen as the measure of telecom development for each country. In order to select the most influential macro indicators, we carried out the following procedure:

- The World Bank (WB) publishes information for 1440 indicators and 211 countries. We exctracted the data for the period 2000-2015. A first filter was carried out based on the data availability for a group of representative countries from all regions of the globe: if data for a given indicator were not available for these countries, the indicator was discarded. By means of this procedure, 745 variables were discarded. Out of the remaining variables, 54 indicators were selected based on a qualitative judgement of which of them could have a stronger relation with the percentage of Internet users.
- In the next step, we ranked the 54 preliminary variables considering the number of years and countries for which we had non-null observations. Based on this criterion, we selected 30 variables. These are the indicators for which the number of non-null observations is greater than 70% of the total number of observations.
- In order to avoid multicollinearity, we clustered the 30 variables chosen in the previuos step based on the absolute value of their correlations. The cluster analysis allowed us to make a final selection excluding strongly correlated variables. At this point, we had between 15 and 20 macroeconomic indicators.
- The final step of the variable selection process consists in the implementation of a regularized regresion model on the standardized data. We have implemented two regression models: Lasso regression and bayesian model selection. Both schemes allow us to identify the most relevant macroeconomic indicators.

1.1.1 Outliers

Another goal of the analysis is to detect outliers, which in this context does not have a negative connotation but rather indicates the countries that are under- or outperforming. That is, an outlier represents a country that, in a particular year, exhibits poor socioeconomic performance but high telecom improvement or viceversa. We identify outliers when observe a relevant deviation between the model prediction and the actual observation.

1.2 Classification

A second objective of the analysis was to fill in the data gaps for those countries with no data available on telecommunication performance.

In particular, the list of Least Connected Countries (LCCs) published by ITU was considered. Countries in the LCC list were coded as ones and countries with data but not considered LCCs were coded as 0. The classification algorithm fitted a generalized linear model (of the type binomial) on these data, and used the results to predict the LCC status for 44 countries with unknown LCC status in 2014 (latest year with data available). The procedure was as follows:

- The 54 variables pre-selected for the regression analysis from the WB, Google's data on broadband prices and ITU's data on Internet users, mobile subscriptions and fixed-telephone subscriptions were extracted for the years 2008, 2010, 2011, 2012, 2013 and 2014 (i.e. years with a published LCC list).
- The bottleneck in this regression exercise is the lack of data for the countries under consideration. Therefore, the variables selected in the previous step were ordered according to their availability for the countries to be classified, considering that the inclusion of a variable in the model implies that the countries where the variable is missing for the target year (2014) will be excluded from the model. Twelve indicators were selected, thus striking a balance between precision in the classification and countries covered (22 out of 44).
- All observations of countries with known LCC status were used to fit the GLM model. Time was not considered as a variable and this required a prior transformation of the data. Indeed, the LCC status of a country is determined based on the relative performance of all other countries in the same year, i.e. the lowest quartile is classified as LCC. Therefore, the variables used to fit the model were adjusted according to the value of the lowest quartile in each year. For example, 1st_Quartile_Indicator1_2008 = k; Indicator1_2010 = Indicator1_2010 * k / 1st_Quartile_Indicator1_2010; Indicator1_2011 = Indicator1_2011 * k / 1st_Quartile_Indicator1_2011 etc.
- The fitted GLM model was used to predict the LCC status of 22 countries with unknown value in 2014.

2 Code and results

All the scrips related to the analysis are available at the folder analysis in the project repository https://github.com/veldanie/dwbi_project.

Some outputs of the analysis are also available in the project repository, including the results of the variable selection for the classification algorithm (here) and a chart identifying the outliers in the regression model (here).