Economic Analysis

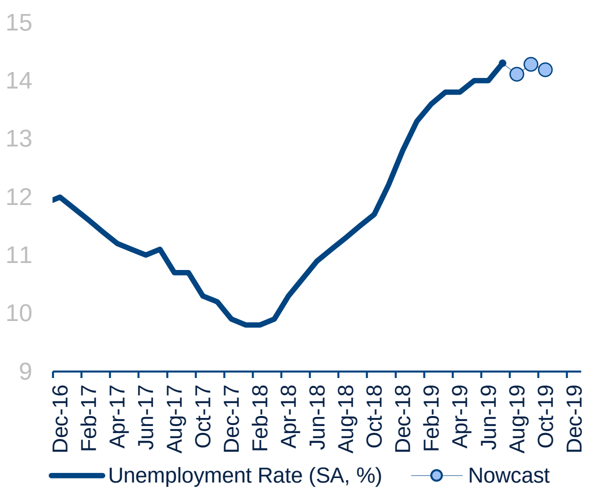
Nowcasting Turkish Unemployment Using Real Time Data From Google[1]

Fernando Bolívar / Álvaro Ortiz / Tomasa Rodrigo / Garanti BBVA Research Team

October 2019

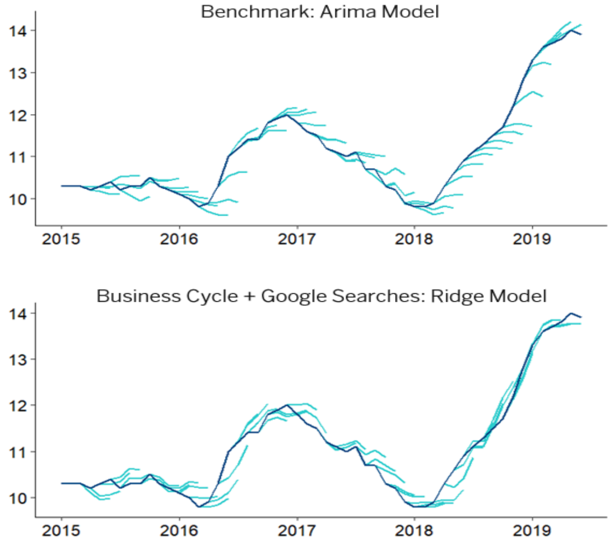
Turkish unemployment data is published with a significant delay (nearly three months), making it difficult to analyze the response of the labor market to business-cycle conditions. In order to assess the evolution of the unemployment in high frequency, we have developed a dynamic model that includes the own past of unemployment, the business cycle conditions and Google searches related to unemployment. We have tested the nowcasting performance of several models, including some potential explanatory variables among these groups, in order to check the nowcasting ability with regard to unemployment. We implement a Bayesian Model Analysis (BMA[2]) to select the set of variables to be included in the models from the potential candidates. Once the variables were selected, we tested the nowcasting pseudo-out-of-sample ability of several alternative models[3], and we found that Ridge Regression had the best performance from January 2007 to June 2019. The results show that both the Business Cycle and Google searches for jobs add extra information beyond the own dynamics of unemployment. Thus, the information content of Google Searches has been proved to be a relevant indicator for nowcasting unemployment. The results from the updated model foresee a stabilization of the unemployment rate after one year of increases.

Figure 1. **Unemployment rate: Data & Nowcasts**



Source: BBVA Research. ID Bloomberg: GBTRUNEM

Figure 2. **Out-of-sample errors benchmark & Google Models**



Source: BBVA Research

- Google Trends data in real time (daily) allows us to cover the gap in the timeline for the release of official data regarding the Turkish unemployment rate (almost three months), thereby improving our nowcasting accuracy.
- Bayesian moving average techniques show that the own dynamics of unemployment, economic activity indicators such as Industrial Production and Capacity Utilization and Google Searches related to unemployment, are important variables for nowcasting the Turkish unemployment rate.
- Nowcasting results[4] for August, September and October show that unemployment is stabilizing slightly and improving in line with the renewed strength of economic activity and Google searches.

[1] We want to thank Joaquín Iglesias and Asuman Kemiksiz for their participation in the early stages of this project.

[2] An application of Bayesian inference to the problems of model selection, combined estimation and prediction. Further information here.

[3] We tested the out-of-sample performance of Linear Regression, Lasso, Ridge, Elastic Net, Principal Components, Spike and Slab and Bayesian GLM.

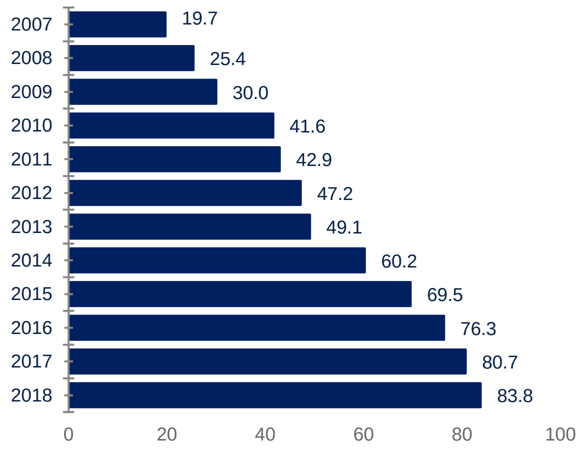
[4] This indicator can be tracked on monthly basis in Bloomberg database with the following ID: GBTRUNEM

Introduction

Internet usage has increased dramatically in the last two decades, soaring from 5.8% of the world population having access to the Internet in 2000 to almost 57.3% by mid-2019, according to Internet World Stats figures. The figure in Turkey is above this average, with 83.8% of the population having access to the Internet and 72.9% of the population using it (TurkStat, 2019). Figure 3 shows this increase over time from 2007 to 2018. This growth has been observed across all regions, ensuring distributed technological evolution and giving a balanced sample (Figure 4).

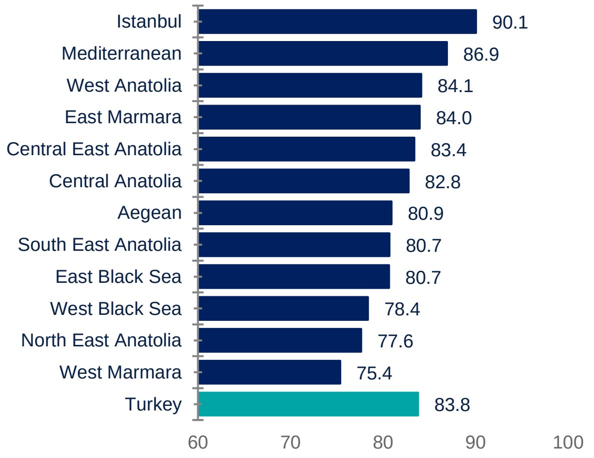
Since the Internet became an essential part of our daily lives, people are increasingly using it as the main tool to get information, search for opportunities and plan their lives. Their digital footprint from the Internet has generated powerful data in real time about populations' needs, plans, concerns and preferences, with a high potential for economic analysis. The analysis of labor-market evolution using this real time information is a clear example of this. People use the Internet to search for vacancies and job opportunities when they are unemployed. Here, we study if this new database, based on Internet queries regarding job searches, is able to explain labor-market evolution, offsetting the delay of almost three months in the release of official figures by TurkStat.

Figure 3. **Access to Internet in Turkey.** Percentage of population, annual evolution (2007-2018)



Source: TurkStat

Figure 4. **Access to Internet by region in Turkey. Percentage of population, by region (2018)**



Source: TurkStat

The unemployment rate in Turkey has special characteristics that make this variable an interesting case for nowcasting. First, unemployment is a key factor both in normal times but also to measures the real effect of the crisis. Second, it is an important variable for monitoring the resolution of banking problems and the potential for Non-Performing Loans. Third, the labor market normally reacts with a lag to economic conditions and, given the time lag in the release of both economic conditions, and even longer gaps on the labor market, assessing the evolution of the labor market will be a very important tool for economic analysis. Turkish economic activity is recovering very quickly, and knowing the response of the labor market in advance will complement our set of high-frequency indicators for assessing the health of the Turkish Economy.

Data and Methodology

In order to know how people search for job opportunities in Google, we used Google Correlate to get the most correlated search topics (in Turkish) relating to unemployment since 2004. We obtained 75 related searches (see Appendix Table 1) and, once we had the individual queries, we were able to analyze their evolution since 2004 using Google Trends. This tool provides the data on a monthly basis by taking its search volume and dividing it by the total search volume in Turkey. The index is scaled between 0 and 100, where 100 points corresponds to the highest number of searches for the query under analysis[5].

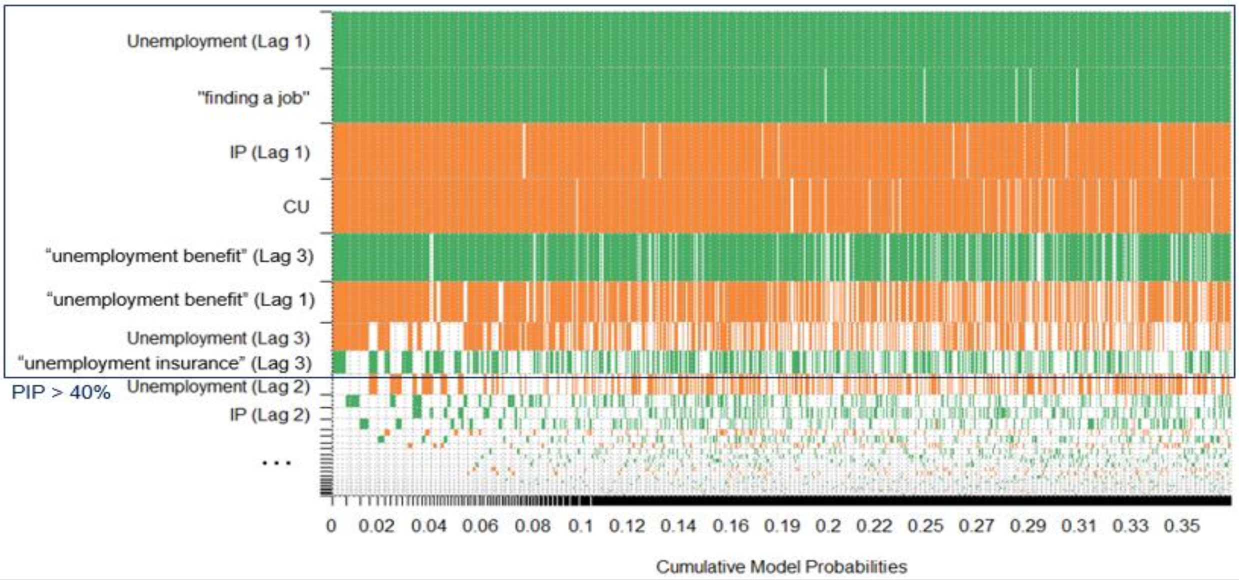
In order to develop a robust model, we also introduce a set of key macroeconomic variables normally used to monitor economic activity used in literature to explain unemployment. Among these, we checked Industrial Production[6] (IP), growth in electricity consumption and the capacity utilization (CU) ratio published monthly by the Central Bank of Turkey. We also tested other variables related to the labor market, such as the number of monthly applications for unemployment benefits (from the Turkish Employment Agency, ISKUR).

Our dependent variable is the seasonal adjusted Turkish unemployment rate published by TurkStat in its labor force survey statistics. The data is published near mid-month and reflects the three-month average (i.e. it includes the current month, the previous month and the coming month, in relation to the release). In order to maintain consistency, we also use the centered moving average for the economic activity variables. In the case of Google data, we opted to maintain the original structure, since this is previously processed information. Both our soft data and the official labor force statistics are monthly, going back to 2007, in line with official data. The latest available data on the unemployment rate was for June 2019, meaning that our current nowcast horizon is August, September and October 2019.

Variable Selection

To avoid noise and co-llinearity, and given the amount of available data, we use Bayesian variable selection techniques as the Bayesian Model Averaging (BMA). The BMA tool provides a consistent method to account for model uncertainty, and it is particularly useful in exercises with a large number of potential regressors and a relatively limited number of observations. Thus, we combine the priors (uniform or non-informative priors) with the data to obtain the posterior probability of any testing variable to be included in the model[7].

Figure 5. **Model inclusion based on best 3000 models**



Source: BBVA Research

[5] For further information about how trends data is adjusted, see here.
[6] Since this value is also published with a two-month delay, we include internally predicted values.
[7] In order to capture potential time lag effects, we include lagged variables up to three months, including the own dependent variable. Therefore, we consider 79 variables (75 google queries and four macroeconomic indices) and the lagged unemployment, and perform the variable-selection exercise with 319 possible regressors. 3000 different linear models have been tested to analyze the convergence between their likelihoods to obtain the average coefficients, giving the total posterior inclusion probability (PIP) of each variable, i.e. the sum of the posterior model probability (PMP) for all models in which a covariate was included.

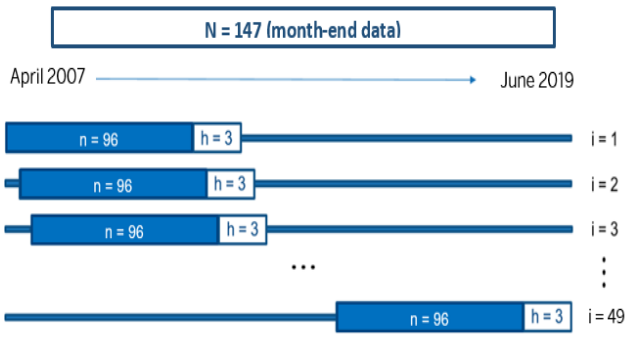
The results are shown in Figure 5. The vertical axis shows the predictors ranked by their posterior inclusion probability (PIP) and the horizontal axis shows the best models, scaled by their posterior model probabilities (PMP). Each column in the resulting matrix corresponds to one of the 'top' models obtained, ordered from left to right according to the accumulated PMP. In terms of layout, green color corresponds to a positive coefficient, orange to a negative coefficient, and white to non-inclusion (a zero coefficient).

Finally, we select the variables with PIP higher than 40%, taking into account the most significant lag in the event that a variable appears twice. The dependent variable itself lagged one month (with a positive coefficient) consistent with the specification of a dynamic model and business cycle conditions as capacity utilization (CU) and industrial production (IP) (both with negative coefficients consistent with the theory). In the case of Google searches, "finding a Job", "unemployment benefit" and "unemployment insurance" (lagged three months) are the Google queries with the highest inclusion probabilities for explaining the evolution of unemployment. The three-month lag in the Google query shows that people normally look for unemployment even before becoming unemployed. The rest of the business-cycle indicators, like electricity or the number of monthly applications for unemployment benefits, do not seem to add extra information to the previously selected ones. Once we have obtained the most relevant variables through BMA, we analyze their statistics in the same way as the existing correlations with the dependent variable (Appendix Table 2).

Model Selection

Once the variables have been pre-selected, we proceed to develop a robust nowcasting model for the unemployment rate in Turkey. To this end, we examine the out-of-sample predictive performance of several linear models to test whether the Google queries add some nowcasting accuracy to the models. To do this, we compare a naïve model (i.e Arima) as a benchmark with several models used to work with massive data.

Figure 6. **Data slicing summary (cross-validation)**
N = total number of months, n = fixed window (months), h = horizon (months), i = iterations (folds)



Source: BBVA Research

To select the best nowcasting model, we implement cross validation from April 2007[8] to June 2019 (i.e. we check the out-of-sample error by repeating and calculating the arithmetic mean obtained from the evaluation measures of different partitions). Given that the unemployment rate is published with a three-month delay, we calculate the cross validation of each model in 49 fixed folds of eight years (96 months) for the following three periods (see Figure 6). In addition, a hyper-parameter optimization exercise has been developed for the models (grid search), as well as prior standardization of the data. The metric used to evaluate the prediction accuracy is the Root Mean Square Error (RMSE), averaging the recursive forecasts of each model.

Results

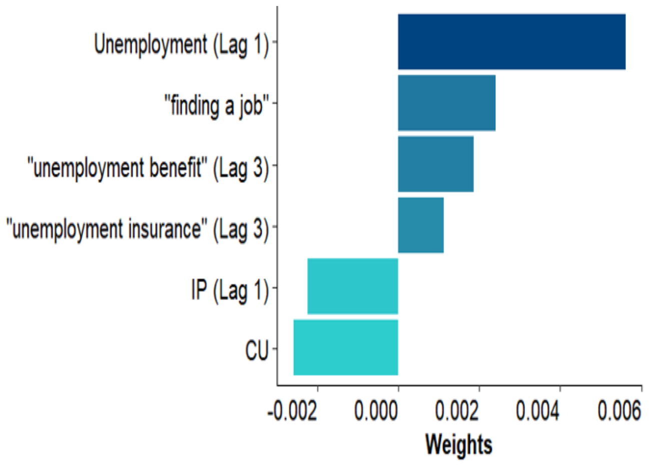
According to the error exercise (results in Table 3), the Google information content is validated in all the models tested, as they outperform a simple dynamic ARIMA model in our horizon of analysis. The best performing model is Ridge Regression, although the results for Elastic Net results are very close. Figure 2 displays both the unemployment ratio and the out-of-sample recursive predictions since 2015 for both the benchmark and the more accurate model (Ridge Regression)[9]. These graphs show how the prediction traces (light blue lines) fit the series

[8] We lost the first three months of 2007 due to lagged variables.
[9] More information about the benchmark (ARIMA) and the best performance models (Ridge Regression and Elastic Net Regression) are found in the appendix.

(dark blue lines). The greater the deviation from the predicted values, the greater the prediction error in the long term. The Ridge model including Google beats ARIMA.

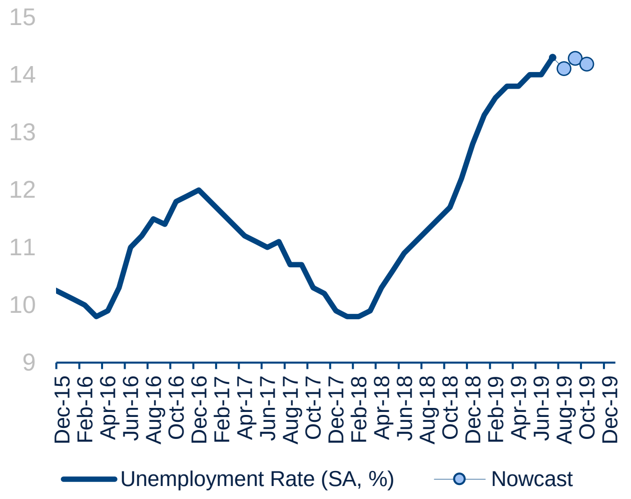
One additional advantage of top regressions obtained is that they allow us to understand the variables' significance, as assigned by the method. We present the estimated absolute coefficients different from zero in decreasing order in Figure 8. As expected, the high inertia of unemployment dependent variable (favored by the use of the centered average by TurkStat), makes that the one-month-lagged unemployment variable has the highest weight in the model. This variable is followed by one of our proxies of economic activity, the capacity utilization ratio (with negative sign, as expected). These are followed by the Google query "finding a job", the first lag of industrial production (with negative sign) and the Google queries "unemployment benefit" and "unemployment insurance". The Google queries are positively related to the dependent variable, that is, the greater the number of searches, the higher the unemployment rate. Besides, the dynamic impact of the Google Search is not uniform, and while the "finding a job" has a contemporary effect, the influence of searches for "unemployment benefit" and "unemployment insurance" has a three-month delay.

Figure 8. **Total Unemployment Coefficients** (Top models average)



Source: BBVA Research

Figure 9. **Seasonally Adjusted Unemployment Rate Nowcast**



Source: BBVA Research. ID Bloomberg: GBTRUNEM

The model allows to track the recent events of the labor market. In line with our GDP nowcasting model, which is consistent with a rapid acceleration of the economic recovery, the unemployment nowcasting model results point to a stabilization in the upward trend of unemployment observed since the beginning of 2018 (Figure 9). The out-of-sample results for the next three months indicate an inter-monthly variation of -1.36%, 1.24% and -0.66% respectively (using Ridge and Elastic Net Regression), signaling a recovery of unemployment in line with the rest of the macroeconomic indicators (ID Bloomberg to track the indicator on monthly basis: GBTRUNEM).

Conclusions

High Frequency and Big Data information have become a key factor for having timely information for economic analysis nowadays. In this research note we have shown how the Google internet searches can be used in combination with the labor market dynamics and the business-cycle conditions to nowcast unemployment in Turkey. The out-of-sample results signal that Ridge and Elastic Net regressions are the best performers as they provide extra information valuable for unemployment nowcasting. The model anticipates a stabilization of the unemployment rate following the upward trend observed during the last year.

References

Gülenay M. and Sengül G. (2012): “Nowcasting Unemployment Rate in Turkey: Let's Ask Google” Central Bank of the Republic of Turkey Working Papers.

Hoerl A. and *Kennard* R. (1988): “Ridge Regression.” *Encyclopedia of Statistical Sciences* 8:129-136. New York: Wiley.

Kucharcukova O. B. and Bruha J. (2016): “Nowcasting the Czech Trade Balance.” CNB WP 11/2016.

Schams A. (2019): “Bias, Variance, and Regularization in Linear Regression: Lasso, Ridge, and Elastic Net — Differences and uses” https://towardsdatascience.com/bias-variance-and-regularization-in-linear-regression-lasso-ridge-and-elastic-net-8bf81991d0c5

The Caret package (short for Classification And REgression Training), https://topepo.github.io/caret/index.html

Tuhkuri J. (2015): “Big Data: Do Google Searches Predict Unemployment?”

Zeugnet S. and Feldkircher M. (2015): “Bayesian Model Averaging Employing Fixed and Flexible Priors: The BMS Package for R”. Journal of Statistical Software, Volume 68, Issue 4.

Zou H. and Hastie T. (2005): “Regularization and variable selection via the elastic net.” Journal of the Royal Statistical Society 67(2): 301-320.

Appendix

Table 1. **Terms downloaded from Google Trends**

Query	Turkish	English	Query	Turkish	English
Query1	garanti iş ilanları	Garanti vacancies	Query39	iş başvurusu	job application
Query2	iş ilanları	vacancies	Query40	iş başvuruları	job applications
Query3	kariyer	kariyer (website)	Query41	insan kaynakları	human resources
Query4	örnek cv	sample CV	Query42	iş ilanları istanbul	vacancies Istanbul
Query5	özgeçmiş	resume	Query43	iş ilanları istanbul anadolu	vacancies Istanbul Anatolia
Query6	banka iş ilanları	bank vacancies	Query44	iş ilanları istanbul avrupa	vacancies Istanbul Asia
Query7	iş bankası iş ilanları	işbank vacancies	Query45	avrupa yakası iş ilanları	European side vacancies
Query8	işkur	İşkur (Labor Agency in Turkey)	Query46	acil iş arıyorum	I'm looking for an urgent job
Query9	işkur iş ilanları	İşkur vacancies	Query47	ek iş ilanları	side job vacancies
Query10	yenibiriş	yenibiriş (website)	Query48	izmir iş arayanlar	izmir job searchers
Query11	cv örneği	Sample cv	Query59	eleman arayanlar	staff searchers
Query12	eleman net	eleman net (website)	Query50	eleman arayanlar istanbul	staff searchers Istanbul
Query13	mezun iş	graduate job	Query51	eleman ilanları	staff vacancies
Query14	işsizlik maaşı	unemployment benefit	Query52	vasıfsız eleman ilanları	unskilled worker vacancies
Query15	işsizlik maaşı başvuru	unemployment benefit application	Query53	cv indir	download cv
Query16	kariyer net	kariyer net (website)	Query54	cv örnekleri indir	download cv samples
Query17	kıdem tazminatı	severance payment	Query55	boş cv indir	download empty cv
Query18	tazminat	compensation	Query56	boş cv örneği	empty cv sample
Query19	cv	cv	Query57	cv örneği indir	download cv samples

Query20	cv örnekleri	cv samples	Query58	cv hazırlama	cv preparation
Query21	eleman aranıyor	looking for a staff member	Query59	cv formu indir	download cv form
Query22	iş	job	Query60	işten çıkarıldım	I was laid off
Query23	iş arama	job search	Query61	tazminat hesaplama	compensation calculation
Query24	iş arayanlar	job searchers	Query62	ssk tazminat hesaplama	Social Sec. Adm. compensation calc.
Query25	iş arıyorum	i'm looking for a job	Query63	işten çıkış	leaving a job
Query26	iş bulma	finding a job	Query64	sgk işten çıkış	Social Security Ins. leaving a job
Query27	iş ilanı	vacancy	Query65	işten çıkış bildirgesi	leaving a job declaration
Query28	işçi bulma kurumu	labor agency	Query66	işten çıkış bildirgesi sgk	leaving a job declaration sgk
Query29	işsiz	unemployed	Query67	işten çıkarılma tazminatı	getting laid off compensation
Query30	işsizlik sigortası	unemployment insurance	Query68	tazminat hesaplama netten brüte	compensation calc. net to gross
Query31	kariyer.net	kariyer.net(website)	Query69	yenibiriş iş ilanları	yenibiriş (website) vacancies
Query32	kariyernet	kariyetnet(website)	Query70	kariyer iş ilanları	kariyer (website) vacancies
Query33	personel alımı	recruitment of personnel	Query71	hürriyet iş ilanları	hürriyet vacancies
Query34	ek iş	side job	Query72	hürriyet seri iş ilanları	hürriyet seri vacancies
Query35	elemanonline	elemanonline (website)	Query73	hürriyet seri ilanlar iş ilanları	hürriyet seri ads vacancies
Query36	evden iş	work from home	Query74	sabah iş ilanları	sabah vacancies
Query37	part time	part time	Query75	posta iş ilanları	posta vacancies
Query38	secretcv	secretcv (website)			

Source: BBVA Research

Table 2. **Descriptive Statistics and correlations**

Variables	num	mean	sd	min	max	correlation
Unemployment (*)	1	0.104	0.015	0.080	0.140	1.000
Unemployment (Lag 1) (*)	2	0.104	0.015	0.080	0.140	0.984
"finding a job"	3	45.546	12.589	25.000	83.000	0.539
IP (Lag 1) (*)	4	0.049	0.079	-0.212	0.238	-0.639
CU (%)	5	76.287	3.791	62.000	83.300	-0.613
"unemployment benefit" (Lag 3)	6	27.302	18.751	3.000	100.000	0.466
"unemployment insurance" (Lag 3)	7	24.611	25.328	3.000	100.000	-0.023

(*) percentage per 1

Source: BBVA Research

Table 3. **Relative RMSE results**

Model	RMSE
ARIMA (Benchmark)	1
Ridge Regression	0.722
Elastic Net Regression	0.723
Linear Regression	0.735
Bayesian GLM	0.735
Lasso Regression	0.736
Spike and Slab Regression	0.745
PC Regression	0.924

Source: BBVA Research

ARIMA (Benchmark)

Auto-Regressive Integrated Moving Average models are, in theory, the most general kind of models for predicting a "stationary" time series by examining the differences between values of the series itself rather than through the actual values. An ARIMA model can be viewed as a "filter" that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts. The ARIMA prediction equation for a stationary time series is a linear (i.e., regressive type) equation in which predictors consist of dependent variable delays and/or forecast error delays, as shown below:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

Where μ is the constant, ϕ is the AR coefficient at lag k, θ_k is the MA coefficient at lag k, and $e_{t-k} = y_{t-k} - \hat{y}_{t-k}$ is the forecast error that was made at period t–k. In addition, p represents the number of autoregressive terms and q the number of lagged forecast errors. Notice that the MA terms in the model (the lags of the errors) are conventionally written with a negative sign rather than a positive sign.

Ridge Regression

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity, i.e. the least squares estimates are impartial, but their variations are large, so they may be far from the real value. This method aims to reduce standard errors by adding an additional term to the loss function, in this case a bias degree to the regression estimates. This regularization term will decrease the coefficients values, but is unable to force a coefficient to exactly 0 unlike Lasso Regression. This makes the use of Ridge Regression limited with respect to feature selection. However, when the number of predictors (p) is greater than the number of observations (n), it is able to select more than n relevant predictors if necessary. See the corresponding equation below:

$$L(\lambda, \beta) = \left|y - X'\beta\right|^2 + \lambda|\beta|_1$$

The lambda value is constant and arbitrary so it is advisable to try several to obtain models with different levels of regularization, with lambda = 0 corresponding to OLS and lambda approaching infinity corresponding to a constant function.

Elastic Net Regression

Elastic Net Regression (Zou and Hastie 2005) is a statistical learning linear method that offers selection and shrinkage of variables by adding a linear combination of ridge and lasso type penalties to the loss function (Hoerl and Kennard 1988; Tibshirani 1996):

$$\mathrm{L}(\lambda, \alpha, \beta) = |\mathrm{y} - \mathrm{X}'\beta|^2 + \lambda(\alpha|\beta|_2 + (1 - \alpha)|\beta|_1)$$

Where α and λ are the hyperparameters that give the mixing rate of the penalization type and the sensitivity to the constraint respectively. This is equivalent to carrying out the standard optimization process for linear regression, constrained to a combination of the first and second norms' areas. Despite these estimators being biased, they allow for automatic variable selection, as well as having an empirically demonstrated better prediction performance both in the literature (Zou and Hastie 2005; Kucharcukova and Bruha 2016).

DISCLAIMER

This document has been prepared by BBVA Research Department. It is provided for information purposes only and expresses data, opinions or estimations regarding the date of issue of the report, prepared by BBVA or obtained from or based on sources we consider to be reliable, and have not been independently verified by BBVA. Therefore, BBVA offers no warranty, either express or implicit, regarding its accuracy, integrity or correctness.

Any estimations this document may contain have been undertaken according to generally accepted methodologies and should be considered as forecasts or projections. Results obtained in the past, either positive or negative, are no guarantee of future performance.

This document and its contents are subject to changes without prior notice depending on variables such as the economic context or market fluctuations. BBVA is not responsible for updating these contents or for giving notice of such changes.

BBVA accepts no liability for any loss, direct or indirect, that may result from the use of this document or its contents.

This document and its contents do not constitute an offer, invitation or solicitation to purchase, divest or enter into any interest in financial assets or instruments. Neither shall this document nor its contents form the basis of any contract, commitment or decision of any kind.

With regard to investment in financial assets related to economic variables this document may cover, readers should be aware that under no circumstances should they base their investment decisions on the information contained in this document. Those persons or entities offering investment products to these potential investors are legally required to provide the information needed for them to take an appropriate investment decision.

The content of this document is protected by intellectual property laws. Reproduction, transformation, distribution, public communication, making available, extraction, reuse, forwarding or use of any nature by any means or process is prohibited, except in cases where it is legally permitted or expressly authorised by BBVA.