Does gender make a difference? A labor demand analysis from an online job board

Working paper

Abstract

This paper seeks to analyze gender differences in job demand through data obtained from two websites that are widely used in Argentina both by people who demand employment and those who offer employment. Through a LASSO and L1 penalization logit analysis, the results suggest that there are differences in the words that are used in the description of job ads.

Introduction

In a spirit similar to this job Wu et al. 2017 attempted to identify tried to identify those words that best predict when a publication referred to a woman or a man. Through a logit-LASSO analysis he found that there were large differences between groups of words.

This work manages to contribute to diverse literatures. In the first place, gender preferences are explored from the labor demand side, this contribution is found within the literature on gender differences in the labor market. On the other hand, the work manages to account for certain mechanisms that could explain the wage gap.

Data

The data comes from webscrapping on zonajobs and boomeran sites. These sites are widely used in Argentina for job search.[1]

The webscrapping was carried out twice a month for the last N months. Through this process of obtaining data, a database was created containing the title of the publication, its description, the name of the company and other information.

Gender classification

Regarding the genre that the publication refers to, this variable was created through the title of the publication. In Spanish it is possible to differentiate, for example, when referring to a male or female business administrator, based on this possible distinction an algorithm was created that classifies publications.

Analysis

The objective of this work is to analyze whether there are differences between publications that are directed towards women compared to men or the general public. For this it was decided to use the LASSO and Logit methods with L1 penalty (henceforth LASSO-logit). The dependent variable will be whether a publication is directed towards men or towards women, and by finding the highest predictors it will be analyzed whether

[1] Por ejemplo, Bumeran.com tiene 6 millones de usuarios únicos al mes.

there are differences between predictors of publications directed towards men compared to publications directed towards women. Each method is detailed below.

LASSO

LASSO fulfills both a role of dimensionality reduction (by eliminating coefficients) and of improving predictions at the cost of producing biased estimates. In this case, as the dependent variable only takes values equal to 0 or 1, it is reasonable to use a logit model with L1 penalty as suggested in Friedman, Hastie, and Tibshirani 2001, p. 125.

$$\hat{\beta}_{lasso} = \min_{\beta} \sum_{n=1}^{N} (y_n - \beta x_n)^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

The LASSO method with logit maximizes the following equation and the intuition of the linear LASSO regarding the increase of lambda and increase of the penalty is maintained.

$$\max_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta^T x_i) - log(1 + e^{\beta_0 + \beta^T x_t}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

When using LASSO it is necessary to make a decision about the optimal λ, this will be selected through K-fold cross validation. In general, it is recommended to use 5-fold cross validation (Breiman and Spector 1992) or 10-fold cross validation (Kohavi et al. 1995). The Cross validation method will divide the sample into K parts and use all but one part to estimate the model and that remaining part will be used to test the model. This procedure will be carried out in such a way that all parties have fulfilled a training and testing role.

Results

Once the previous steps have been carried out, figure N shows the words ordered by magnitude of their associated coefficient, differentiating by sign (in the case of women, the coefficients will have a positive sign and otherwise a negative sign).

LASSO

The words that have the largest coefficients associated with them can be seen in Figure 1. Among these words we can find ... . This would indicate that when looking for women employers are interested in characteristics that have to do with flexible hours. On the other hand, the words that predict that an ad is directed towards men are associated with more important positions (Figure 2), for example the words benefits would indicate this.

Logit LASSO

If the same exercise is performed as before but this time using a logit model with L1 penalty (such as conventional LASSO), the results are similar. For example, you can see in Figure 3, the largest coefficients, that is, those that predict that the publication is

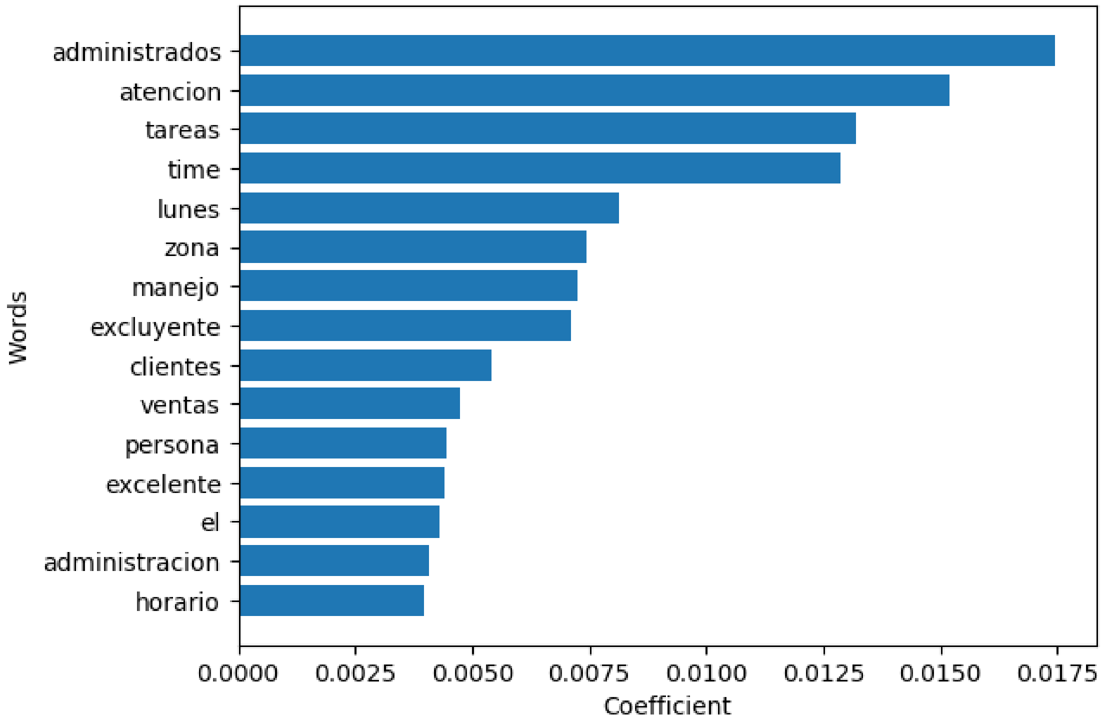

Figure 1: LASSO positive coefficients

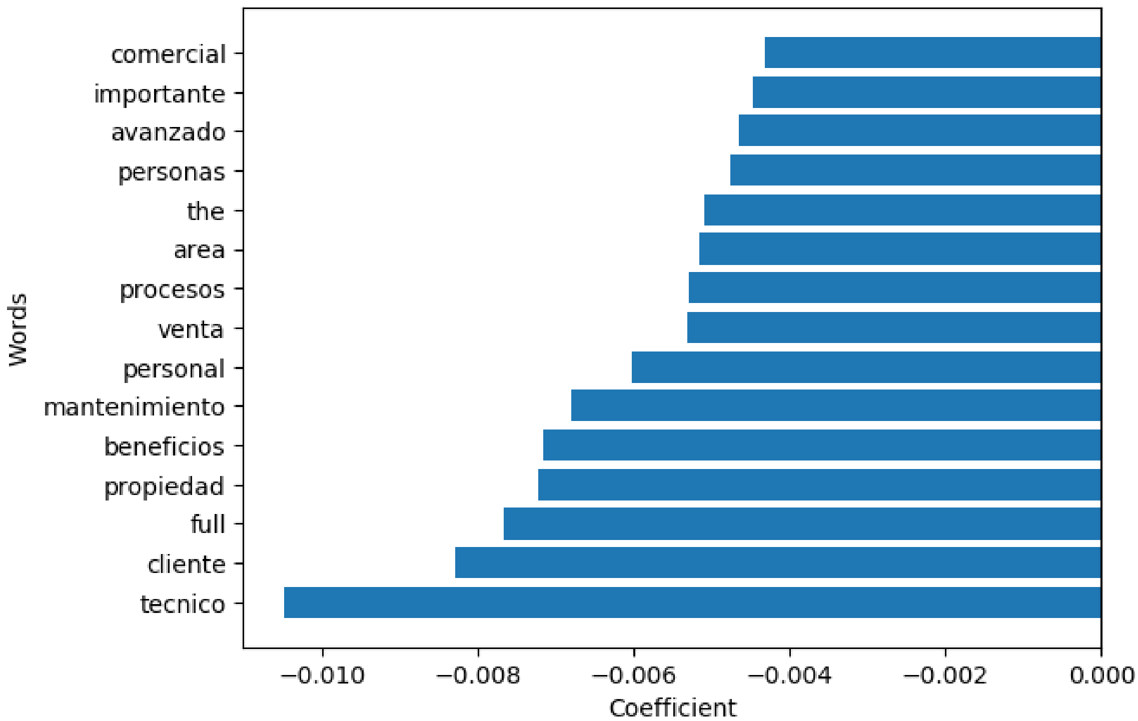

Figure 2: LASSO negative coefficients

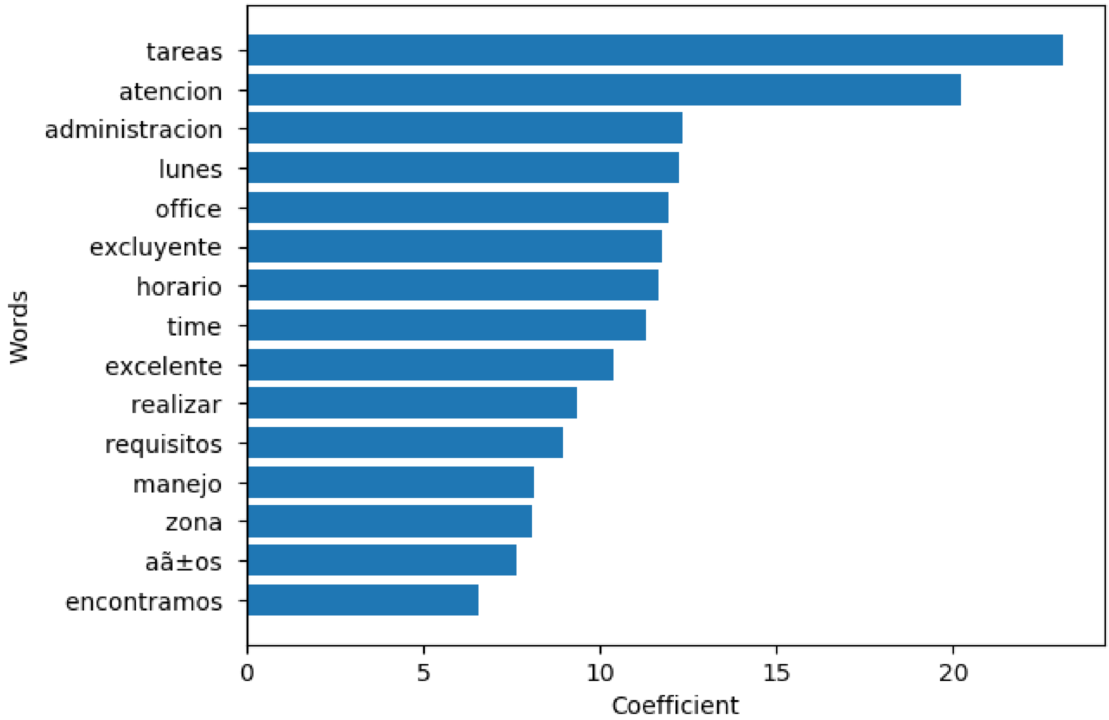

Figure 3: LASSO-Logit positive coefficients

directed towards women. On the other hand, in Figure 4, you can see the words that predict that the publication is directed towards men.

Discussion

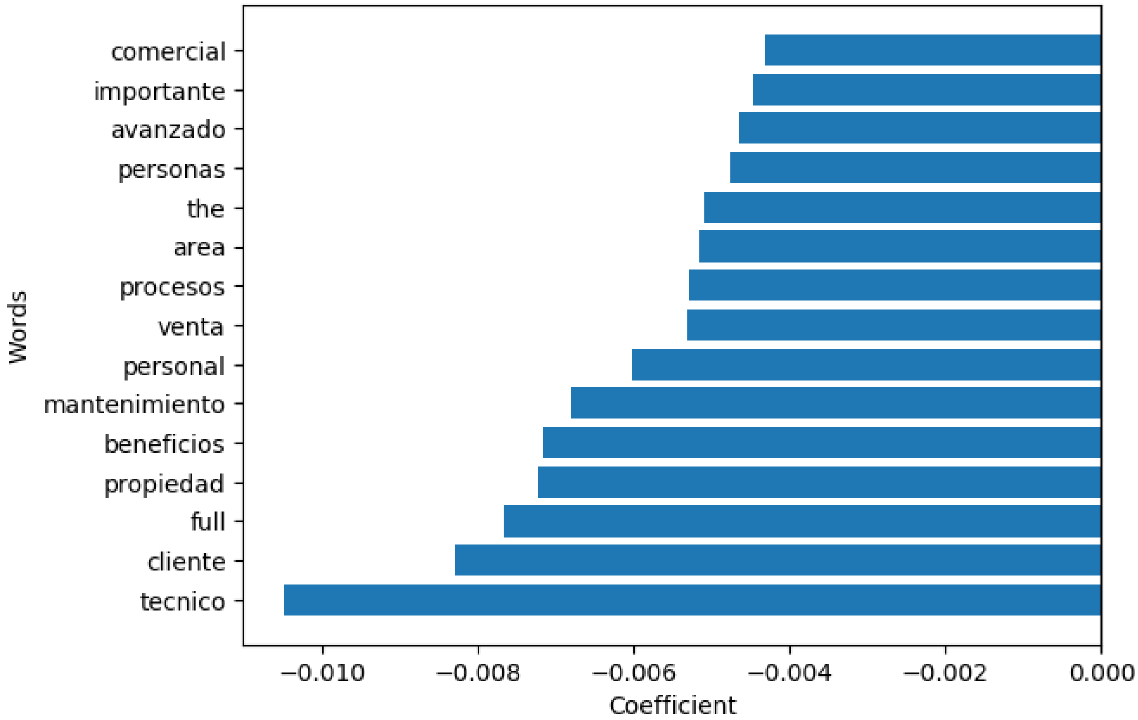

Figure 4: LASSO-Logit negative coefficients

References

Breiman, Leo and Philip Spector (1992). "Submodel selection and evaluation in regression. The X-random case". In: *International statistical review/revue internationale de Statistique*, pp. 291–319.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York.

Kohavi, Ron et al. (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Ijcai*. Vol. 14. 2. Montreal, Canada, pp. 1137–1145.

Wu, Alice et al. (2017). *Gender stereotype in academia: Evidence from economics job market rumors forum*. Tech. rep.

Appendix

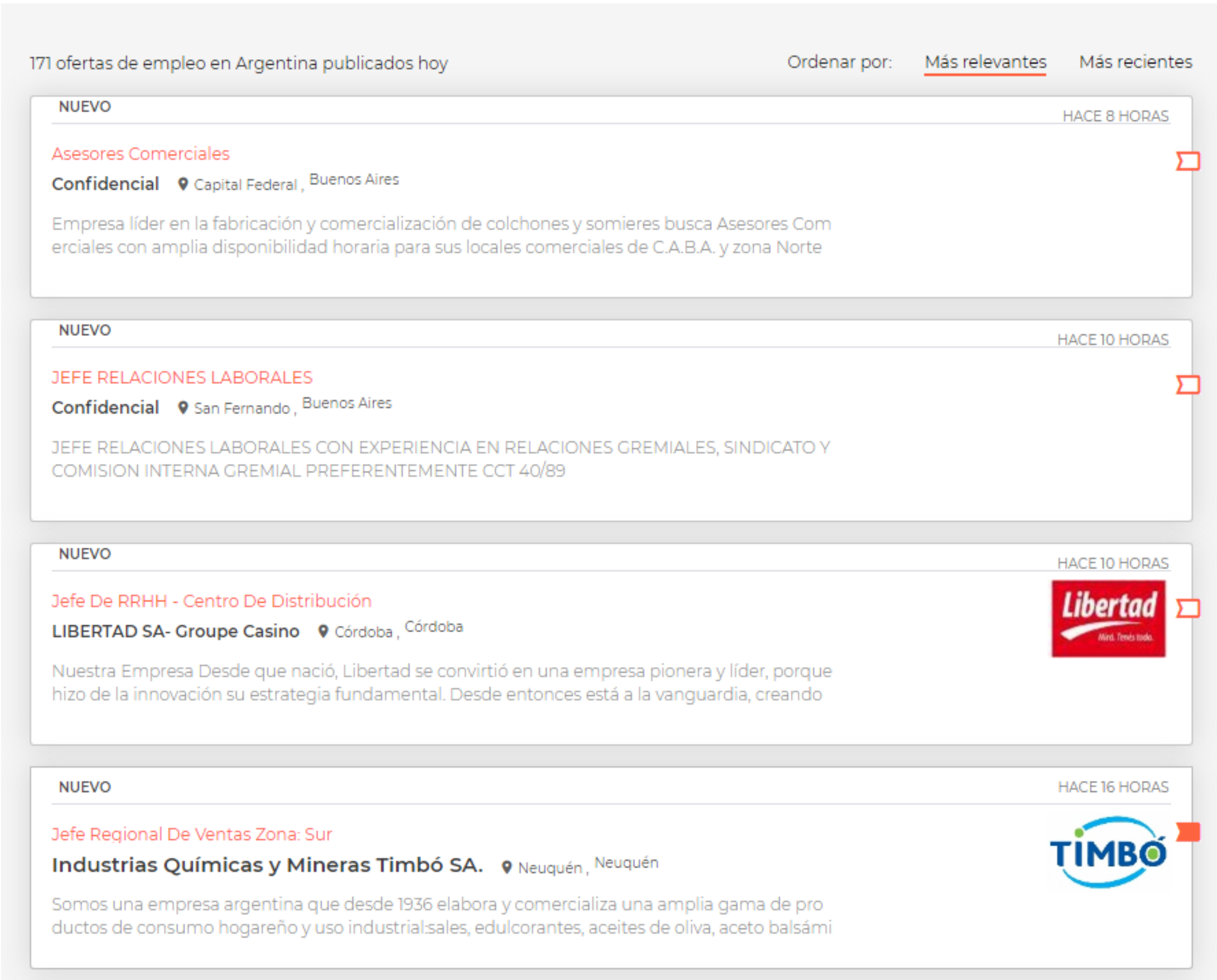

Figure 5: Some publications of the job board: ZonaJobs