Landing .Jobs TAIKAI

Landing.Jobs Data Challenge

Project: “should you ask for a raise?”

Author: Wilson Capitão

Abstract

This project was developed as part of the *Landing.Jobs Data Challenge*. The goal was to build a model that could serve as a decision-making helper for every full-time tech professional, living in Portugal, that wanted to ask for a raise above 30k€ (Gross Annual Income). Reading the Landing.Jobs report, we can clearly understand that female professionals are underpaid comparing to male professionals and that served as a trigger to build an ethical model – for this reason, Gender, Nationality and Age were removed. Due to the nature of several features, I faced the “sparse matrix” problem.

The data used was entirely from the dataset given to every participant with 3371 anonymized samples – from this data, 2931 were full-time professionals and only those were considered. I’ve made a split 80%-20% for test, validation datasets. The data was transformed to be processed and I’ve tried to understand what was the model that produced the best accuracy and the best explanation possible. The final step was to optimize and tune the model.

The best model with less stochasticity and best interpretability was Logistic Regression (LR). After data transformation and feature reduction through the coefficient’s importance, the model was able to perform an accuracy of 80.75%. Feature reduction with PCA had a slight better result (0.6% better) but lacked interpretability. I noted that manually doing a bagging with predictions from a Logistic Regression, XGBoost and Support Vector Classification boosted the accuracy of the final model (with LR) in 2.3% (final accuracy of 82.62%) using a cross validation with 10 splits.

The final conclusion is that no matter what model was being used, in this case, removing gender, nationality and age, the amount of “working experience” from the professional, is always very important. In fact, in the final model it’s the most important feature with a coefficient 2,6 times bigger than the second coefficient. The only skills in the top 20 more important features, by importance, are: Education Level, English Level and knowledge on Perl Language. This means that all the other features are about the company, where is the professional living and the role. I find this model, with an accuracy of 82.62%, a good tool to give confidence for any tech professional, living in Portugal, that is thinking about to ask for a rise above 30k€, no matter the gender, age or nationality, to support their decision.