Prediction of Salary based on Social Media Data

Introduction

The aim of this project is to find a way to predict the salary of individuals in UK based on the data accessible via APIs provided by social media such as Facebook and LinkedIn. The idea is first to predict salaries using job offer description (with features like “title”, “skills” ...) and then to link the data obtained via the social media to the job description. The job offer descriptions dataset was built by scrapping job adds on the recruitment website reed.co.uk.

The products that will be generated are a predictive model of salary based on job offer description and a way to connect information provided by social media to features of this job offer. The several stages of the research will be:

* Clean of the dataset containing features for job description and job salary
* Build the best predictive model
* Analysis of the data that can be obtained from Social Media
* Find a way to link Social Media data to job description data

The research question for this project is:

How can I predict the salary of an individual using Social Media data?

The principal beneficiaries of this work would be:

* Insurance companies: the customers will be able to fill quicker online questionnaires using a connection via social media such as LinkedIn. It will also help them to figure out the market worth of different kind of positions.
* Bank and Insurance companies which can improve their current income prediction models.
* Recruitment websites such as Glassdoor, Reed, Indeed …

Critical context

The use of salary prediction

Few salary prediction models have been built to serve different purposes. The first to undertake this kind of prediction was Adzuna, a company that predict the salary of an individual using the data contained in his CV. This application allows users to have an estimation of the salary they can expect when applying for jobs. Job prediction also helps employers to figure out and have an overview of the market worth [1]. In [2], salary prediction was also used to figure out the salary of graduate students. The objective of this study was to motivate students to work harder, knowing that a well-paid job may awaits them after they graduate. This can help recruitment website to improve the experience of users searching jobs, and help employers and job seekers to have a better understanding of the market worth.

Salary is also considered as an important factor to determine success in life [3]. It is then a relevant factor used in Bank, Insurance and Pension industry to calculate several types of risks, for instance: financial risk for insurances and credit risk for banks. The value of individuals salary also helps pension actuaries to establish pension plans for their customers [4]. It also helps these institutions to target which product should be proposed to specific individuals. However, it is difficult for these financial institutions to figure out the incomes of their customers due to the fact that these ones rarely share this information [4]. Hence, these financial institutions build models to predict incomes of their customers using the data they have on him, in the objective of optimizing their decision-making systems.

Previous works

To build their system, Adzuna has begun by launching a Kaggle competition in which participants had to build a regression model to predict job salaries. The dataset used was based on job adds in UK taken from recruitment websites [5]. The 6 features of the dataset provide details on the jobs like “Full description”, “Location” or “Title”. Some participants have shared their code online with GitHub. All of them have used Random Forest algorithm to obtain their best accuracy.

P. Khongchai and P. Songmuang [6] have built a system to help student to predict their salary when they will graduate. The system was built using profiles of former students as training set, using several independent features like “Gender”, “Faculty” and “Program”. The predictive variable was a categorical variable with four levels as classes, each one being an interval of the salary. The authors then compared results with different methods and the best accuracy was found with Random Forest.

Liu et al. [7] have built a Bayesian Regression model to figure out the influence of factors like gender, race education, gifted or non-gifted student on yearly income in US. The dataset used was built from a survey on 4 years where “subjects were nationally representatively sampled eighth-grade students”.

Kibekbaev and Duman [8] have compare a mix of 16 linear and non-linear regression techniques on real-life dataset to predict bank customers incomes.

Social Media, a source of new information

Social Media offers a huge amount of data about users, their behavior, their societal interactions [9]. All this data is an incredible source of information about individuals and society in general which proved to have a great predictive power in various areas such as disease outbreaks, product sales, stock exchange prices and elections outcomes ([10], [11], [12]). Hence, social media can be an efficient indicator of real-world performance and individual’s behavior. Several kinds of information on an individual can be easily extracted from Social Media: his employer from LinkedIn, his age from Facebook, his gender, his likes on Facebook and follows on LinkedIn…Moreover, Social Media are very popular and an incredible number of people are active members (140 million on Twitter, 2 Billion on Facebook). However, for now, no study was focused on how to predict incomes with Social Media Data.

Unsupervised Learning

The data that can be found on Social Media and on job adds is very sparse and messy, especially when it is text data. We will then need to use unsupervised learning algorithms in order to build another representation of the data. Hence, unsupervised learning is used to find patterns in unstructured data by clustering it or by reducing the dimension [13].

LSA for dimension reduction with text data

Clustering for dimension reduction

Regression algorithms

Linear regression

Random Forest

Gradient Boosting

Neural Network

Approaches: Methods & Tools for Analysis and Evaluation

Description of dataset

Cleansing of Dataset

Implementation of Regression Model

Evaluation and comparisons

Analysis of Social Media Data

Linking Social Media Data with dataset

References

[1] N. Yasmin and K. Kavinilavurajan, “Salary Prediction using Big Data”, *International Journal for Scientific Research & Development* Vol. 4, Issue 01, 2016

[2] P. Khongchai and P. Songmuang, “Improving Students’ Motivation to Study using

Salary Prediction System”, *13th International Joint Conference on Computer Science and Software Engineering (JCSSE),* 2006

[3] S. Banerjee, “Why banks and other financial institutions predict income of their customers”. Available: <https://www.projectguru.in/publications/why-banks-and-other-financial-institutions-predict-income-of-their-customers/> [Accessed 01 July 2017]

[4] C. M. Bone and O. S. Mitchell, “Building Better Retirement Income Models” *North American Actuarial Journal*, *1*(1), 10–11, 1997

[5] Kaggle, <https://www.kaggle.com/c/job-salary-prediction> [Accessed 01 July 2017]

[6] P. Khongchai and P. Songmuang, “Random Forest for Salary Prediction System to

Improve Students’ Motivation”, *12th International Conference on Signal-Image Technology & Internet-Based Systems,* 2016

[7] H. Liu,Y-L. Kuo, Y. He and J. Liu, Yi, “Bayesian Regression Model For Predicting Income”, 2006

[8] A. Kibekbaev and E. Duman, “Benchmark regression algorithms for income prediction modeling”, Information Systems,60 40-52, 2016

[9] H. Schoen, D. Gayo-Avello, P. T. Metaxas, E. Mustafaraj, M. Strohmaier, P. Gloor, "The power of prediction with social media", Internet Research, Vol. 23 Iss: 5, pp.528- 543. 10.1108/IntR-06-2013-0115, 2013

[10] H. He, A. Subramanian, S. Choi, P. K. Varshney and T. Damarla, “Social media data assisted inference with application to stock prediction”, [*Signals, Systems and Computers, 2015 49th Asilomar Conference on*](http://0-ieeexplore.ieee.org.wam.city.ac.uk/xpl/mostRecentIssue.jsp?punumber=7405775)*,* [10.1109/ACSSC.2015.7421462](https://0-doi-org.wam.city.ac.uk/10.1109/ACSSC.2015.7421462)*,* 2016

[11] [S. Asur, and](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Sitaram%20Asur.QT.&newsearch=true)  [B. A. Huberman](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Bernardo%20A.%20Huberman.QT.&newsearch=true), “Predicting the Future with Social Media”, *Web Intelligence and Intelligent Agent Technology* (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, 2010

[12] K. Evangelos, T. Efthimios, T. Konstantinos, "Understanding the predictive power of social media", Internet Research, Vol. 23 Issue: 5, pp.544-559, 2013

[13] Z. Ghahramani, "Unsupervised learning." *Advanced lectures on machine learning*. Springer Berlin Heidelberg, 72-112, 2004