An Interactive Data Analytics System for Employment Acceptance Recommendation

Opeyemi Fadairo, David Olatunbosun, Premsagar Ravula, and Manpreet Kaur

Abstract – Making decisions to accept or reject an employment offer can be quite difficult at times. There are various factors that should be considered in the process of decision-making. Therefore, an interactive system developed using data analytics and predictive analytics can forecast the degree of satisfaction at a prospective job. Users will be needed to provide the necessary information such as company name, job title, location and others in order for the system to make a prediction which will allow the user to gain a better insight for making their decision.

Introduction

As recent graduates in the data science field, we are looking to join organizations where we can contribute and achieve the set goals and mission stated. Also, as humans with the desire for self-fulfillment we would like to work for organizations that offer great career growth, conducive or enjoyable work atmosphere and outstanding benefits. Typically, when individuals have multiple job offers they consider various factors such as company recognition, company reviews, location, job position, career advancement, living expenses, and benefits in order to choose the right employment offer. It is not uncommon to find accepting or rejecting an employment opportunity tricky with a possibility of the outcome not aligning with our expectations.

The purpose of this interactive system is to make data and predictive analytics instrumental in forecasting job satisfaction range. Also, making it possible for users to visualize multiple employment offers and make decisions based off numbers and figures. For the pilot phase of development, we only considered the company reviews, housing expenses, and average salary to compare various job offers which will then help to estimate the employment satisfaction range. The company related data needed for the analysis will be gathered from various websites such as indeed.com, careerbuilder.com, glassdoor.com, monster.com, and other job search engines. The data for housing expenses will be gathered from apartmentlist.com.

Methodology

The method best suited for this project due to its flexibility and ease is the web scraping technique using Python. For the pilot phase, we only scraped for data from three websites namely indeed.com, apartmentlist.com, and kununu.com for companies with entry level data science jobs currently available in the United States.

Python scripting language was used to carry out web extraction because the methodology is pivotal in converting unstructured or noisy data mainly in HTML (Hypertext Markup Language) format on the web to structured data (spreadsheet). The PyPI (Python Package Index) is a source for the python scripting language and there are over 10,000 packages in this repository [1].

The packages used in the analysis of the data retrieved were Beautifulsoup, Requests, Pandas, CSV, Nltk, Textblob, Numpy, and Matplotlib.

For the data extraction and storage, a few guidelines were followed to limit the extraction to specific contents such as company reviews, job position, annual salary, and annual living expenses. After the extraction was done, the next line of action was to remove the chaff from the wheat by performing data preprocessing steps and then dumping the useful dataset in a CSV (comma-separated value) and text format. There were multiple datasets generated and set for analysis.

In order to analyze the dataset, we used the classification data mining technique which is a supervised machine learning method [2].

The supervised machine learning algorithm used to implement this project is the Naïve Bayes algorithm. Before implementing the Naïve Bayes classifier rule we performed the tokenizing of words, casefolding and created a list of stop words in the corpus using the normalization process where you either leave the words how it is or take out the

suffix in a word. This is also known as stemming or using pos tags. Also, there was the Textblob package in the PyPI that did the sentiment analysis automatically by categorizing the company reviews into positive, negative, and neutral reviews. The Textblob package is known to perform sentiment analysis on the words and then determine the polarity of the word in the text document [3].

Result

In order to better understand the extracted data, we decided to visualize it. The data retrieved from the websites were stored in both comma separated value and text format. The comma separated value document was used to import the data into the tableau visualization software. For this phase, the initial comparison was between various job offers based on job title, location and average annual salary.

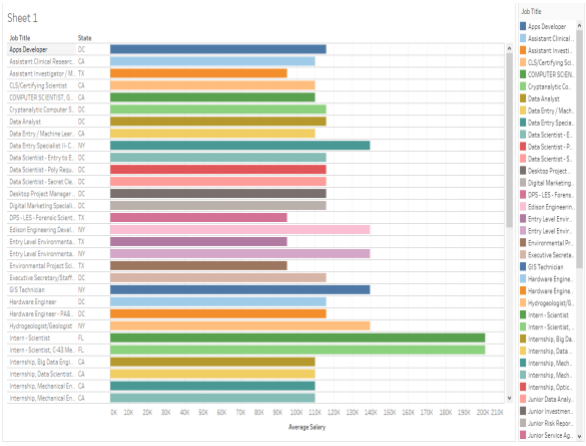


Fig. 1.Tableau visualization comparison for job offers in five states using job title and average annual salary

The company review data which was collected was then used to perform the sentiment analysis to determine the percentages of positive to negative reviews. For a better understanding of these reviews and to highlight the most important and frequent words mentioned about a company, python was used to create a word cloud image.

To illustrate this, the company reviews text data for PwC was used to create the word cloud image shown below.



Fig. 2. Word cloud image showing reviews for PwC

The word cloud image above shows the most frequent words from the review of PwC organization. A user with a job offer from PwC can understand that PwC has great culture and work opportunities.

Conclusion

The visualized results shown above are small pieces of the entire interactive analytical system. As mentioned above, only the pilot phase of analysis has been completed. The pilot phase was carried out to determine the viability and feasibility of building the system for recommending the acceptance or rejection of an employment offer based on the criteria provided by the user. The next part will be to gather more data such as data on various company work cultures, diversity ratio, and retention rate amongst others. These data will then be integrated together and used to develop a predictive model that can forecast a job satisfaction range when a user provides their job offer information. This result given by the interactive engine can surely help the user to gain a better understanding and possibly influence the decision to choose an employment offer.

References

[1] M. Driscoll, *Python 101*. 1st ed. Lean Publishing, 2014, p.192.

[2] J. Han, M. Kamber and J. Pei, *Data mining*, 3rd ed. Burlington: Elsevier Science, 2011.

[3] P. Baid, A. Gupta and N. Chaplot, "Sentiment Analysis of Movie Reviews using Machine Learning Techniques", *International Journal of Computer Applications*, vol. 179, no. 7, pp. 45-49, 2017.