An Interactive Data Analytics System for Employment Acceptance Recommendation

Opeyemi Fadairo, David Olatunbosun, Premsagar Ravula, and Manpreet Kaur
School of Computer Information Systems
Virginia International University

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.

Keywords – Data analytics, predictive analytics, employment offer decision support system, recommender system, big data application, job satisfaction forecasting system.

Introduction
As students in the data science field, we are looking to work in organizations where we can contribute and achieve the set goals and stated mission. Also, as humans with the desire for self-fulfillment we would like to work for organizations that offer great career growth, conducive and enjoyable work atmosphere with outstanding benefits. Typically, when individuals get 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 idea for developing this interactive system is to make data and predictive analytics instrumental in forecasting a job satisfaction degree or range. Also, to make it possible for users to visualize the comparison of multiple employment offers and make decisions based off numbers and figures. For this pilot phase of development, the company related data needed for the analysis was gathered from various websites such as indeed.com, careerbuilder.com, glassdoor.com, monster.com, and other job search engines. The data for housing expenses was extracted from apartmentlist.com. The employment offer dataset is generated by data extraction using the web scraping technique. The dataset contained the attributes city name, state name, company name, job title, job summary, company reviews, company rating, average salary, and housing expenses around the job location. This dataset will be analyzed and used to compare various jobs which will then help to build a predictive model that can estimate the employment satisfaction range for potential job offers.

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 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 housing 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.

In order to analyze the dataset, we used the classification data mining technique which is a supervised machine learning method [2]. The 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 they are 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 a percentage of positive, negative, and neutral reviews. The TextBlob package is known to perform sentiment analysis on the words in which classification is used to determine the polarity of the words in the text document [3].

**Result**
The interactive system will also provide visualization comparison for users by comparing the job offers using the attributes from the dataset. Therefore, to gain an insight on how job offers weigh differently depending on these factors, visualization was performed using Tableau. The CSV document was used to import the dataset into the software. Below is the result of comparing various jobs using their location and average annual salary.

![Fig. 1. Tableau visualization comparison for job offers in five states using job title and average annual salary.](image)

Also, another way we can determine the job satisfaction degree for a prospective job is by analyzing the company reviews of the company offering the job. Sentiment analysis function was performed on the data retrieved to determine the percentages of positive to negative reviews for each company. This percentage will be used as a variable in building the predictive model. Users will also see the difference in the sentiment result percentages for each company and use it to support their decision. To better understand these reviews and to highlight the most important and frequent words mentioned about a company, python was used to create a word cloud. Word cloud are graphical representations or visualizations that perform exploratory textual analysis. The larger the word the more common it is in the text document. To illustrate this, the company reviews text data for PwC was used to create the word cloud to communicate the most salient words. This word cloud image shows the most frequent words in the reviews of PwC organization.

![Fig. 2. Word cloud image showing reviews for PwC](image)

A user with a job offer from PwC will see that PwC can offer them a great work culture, value and opportunities.

**Conclusion**
The results discussed above are small pieces of the entire interactive analytics recommendation system. As mentioned previously, 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. This paper therefore focuses only on the data analytics part which was basically to identify our data sources, extract data from these sources, clean the data and then analyze and visualize it to gain intelligence. The next part will focus on developing the predictive analytics side of the system. We will gather more data such as data on various company work cultures, diversity ratio, and retention rate among others. The input variables that will be used to develop the predictive model are company rating, percentage of positive reviews, company recognition factor, retention rate, salary grade, diversity ratio, and location expenses. To develop the predictive model, the classification machine learning technique will be used. Various algorithms such as Naïve Bayes, Random forest, K -Nearest Neighbor and others will be used to develop a model with the highest classification accuracy. The target variable to be determined by this model is the job satisfaction degree which is a numerical value ranging from 1 to 5 where 1 represents a low satisfaction level and 5 as a high level of satisfaction. The result generated by the interactive engine along with the visualizations should help the user to gain a better understanding and possibly influence their decision.

**References**