A Multifaceted Approach to Job Title Analysis

Jimit Dholakia
Department of Computer Science
Stony Brook University
Stony Brook, NY, United States
jdholakia@cs.stonybrook.edu

Abstract

Corporate world is a rat-race where everyone tries to grab the highest Job Title and the Salary. However, not everyone can fetch the highest possible salary given their education levels and skills. Knowing one's worth before applying for jobs can help identify the right opportunities and land appropriate job. In addition to salary prediction, having the information about Job Titles based on similar skills can broaden opportunities for job-seekers.

However, analyzing Job Titles is not dependent only on a single factor. There are multiple factors affecting the Job Titles, Salaries and Job Satisfaction. Our work focuses on building a predictive model for salaries, clustering of Job titles based on skills, and analyzing the Job Satisfaction of employees based on a publicly available dataset.

1 Introduction

One of the major reasons people change their jobs is due to the salary difference. When employees churn out from their current organization, it becomes an hassle for both, the current employer as well as the employee. The current company struggles to find another employee to fit in the shoes of the current employee. The churned employee will need some time to adjust to the new organization's way of working and their culture.

A job title denotes a person's responsibility as well as the position he or she holds within an organization. Knowing the expected salary for a given Job Title given his/her education level will help both the employees and employers. Employees will be paid as per their qualifications, so they will be content with their current job and won't leave their company and thus the employers will have a lower employee attrition rate.

A problem often faced by job seekers is to not get the information about the jobs that he/she can apply to. An organization may have multiple job openings but the job seekers might not have information about all the available job titles given his/her skills. For e.g., a job seeker applying for the post of Software Development Engineer having skills such as Python may also apply for the job of Python Developer. Clustering these job titles based on their skills provides the job seekers with a wide range of available job titles and thus more opportunities.

Once a person is employed, his/her job satisfaction depends on a variety of factors such as age, department, travel distance, etc. Job satisfaction does not depend on a single factor but on an ensemble of factors, including professional and personal factors.

Our work focuses to provide insights on the above problem. The work consists of the following three parts:

1. Salary Prediction - Given the job title, skills required and the education level, what is the expected salary?

2. Job Clustering - What are the Job Title similar to a given Job Title based on the skills required?

3. Job Satisfaction Analysis - What are the major factors contributing to the Job Satisfaction of employees?

In addition to developing the above models and analysis, we have also developed an User Interface where the user can directly input the required fields and will get the output. More information, along with the screenshots of the UI, is provided in section 5.

2 Literature Survey

In this section, we summarize the findings by various authors.

In the paper by Royer (2010), the primary focus is on how job description is related to job titles.

The paper discusses how different tasks, skills, necessary technical knowledge and abilities and positional hierarchy affect people's perception of the job title. This helps us in figuring out a variety of parameters related to a job title and how we can leverage those details into our project.

Zhou et al. (2016) outlines various implementation methods to relate the skills required to different jobs and tries to come up with a general trend and model to match the jobs with different skill sets. They work upon various metrics to come up with an analytical solution. Reading this article throws light on the tasks and extent of the Skills Recommendation task that we are trying to implement.

Platis et al. (2015) explains how job satisfaction and job performance are related to each other and if one suffers then the other follows. Also, how job performance is directly related to over working implying that overtime leads to a low job satisfaction. Relationship status as well as health also matter and decide the attitude of an employee too. According to the research done in this paper, the most important features to job satisfaction are things like Performance Rating, Relationship Satisfaction, and how it is working under the Current Manager.

Job Satisfaction is the essential component for employee motivation and encouragement towards better performance. The research done by the Raziq and Maulabakhsh (2015) led to a conclusion that such working environments where employees are made a part of the overall decision-making process, being given flexible working hours, less work load, a team work approach and a supportive top management have positive impact on the performance of employees.

The above research papers throw light on various aspects of analyzing job title for multiple purposes like Salary Estimation, the effect of job positional hierarchy, skill sets particular to a job type, etc. It also opens a bunch of interesting questions related some of which are mentioned in the problem statement.

3 Data Collection and Cleaning

Due to the lack of any publicly available dataset which caters to all the three parts of the work, as mentioned in section 1, we divided the data collection process and used two different datasets for the implementation and the analysis which are relevant for each use-case.

The first dataset was obtained by web-scraping an employment website which is used for Salary Prediction and Job Clustering. The second dataset is obtained from Kaggle which is used to analyze the Job Satisfaction of employees. The details are as follows:

3.1 Web Scraping Employment Website

There is no publicly available dataset which provides the details of various Job Titles, their categories and their salaries. We were able to find an employment website (CareerBuilder, 2021) which contains the required information. However, the data is present is not available on a single page and the information about all job titles spans multiple pages.

We implemented a web scraper using Python and BeautifulSoup (Richardson, 2017) to get the required information and were able to construct the data for 1700+ job titles. The initial page contains the list of Main Categories of the Jobs (such as Accounting and Financial, Engineering and Technology etc.) and Sub-Categories of Jobs (such as Banking and Insurance, Finance, Engineering, Science, etc.). Then we perform a Depth-First Traversal for sub-categories and fetch the Job Titles under that sub-category. We then scrape the list of Recommended Skills and get the URL for Salary Data which provides the salary data for various Education Levels, as shown in Fig. 1.

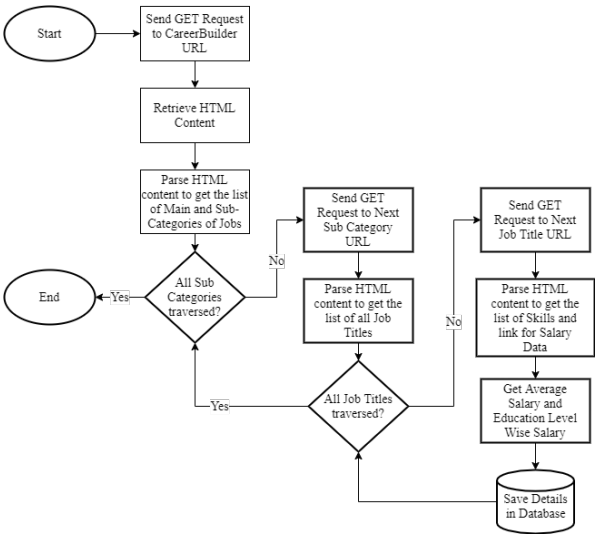

Figure 1: Web Scraping Process

The scraped data follows a hierarchical structure as shown in Fig. 2

A sample of the scraped data is as follows:

The Average Salary data scraped from the website is of the datatype 'string' which contains sym-

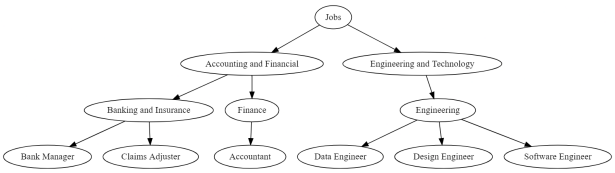

Figure 2: Job Hierarchy

Main Category	Sub Category	Job Title	Average Salary	Education Levels
Accounting and Financial Jobs	Banking and Insurance	Claims Adjuster	$69,500	{'VOCATIONAL': '$52,500', 'HIGH SCHOOL': '$53,500', 'ASSOCIATE': '$54,000', 'BACHELOR': '$63,500', 'MASTER': '$71,000', 'DOCTORATE': '$87,000'}
Accounting and Financial Jobs	Banking and Insurance	Bank Manager	$76,000	{'VOCATIONAL': '$54,500', 'HIGH SCHOOL': '$51,500', 'ASSOCIATE': '$56,500', 'BACHELOR': '$76,500', 'MASTER': '$93,000', 'DOCTORATE': '$103,500'}
Accounting and Financial Jobs	Finance	Accountant	$61,416	{'VOCATIONAL': '$53,000', 'HIGH SCHOOL': '$54,500', 'ASSOCIATE': '$57,500', 'BACHELOR': '$72,500', 'MASTER': '$80,000', 'DOCTORATE': '$79,000'}
Engineering and Technology Jobs	Engineering	Data Engineer	$98,912	{'VOCATIONAL': '$75,000', 'HIGH SCHOOL': '$74,500', 'ASSOCIATE': '$76,000', 'BACHELOR': '$91,000', 'MASTER': '$102,000', 'DOCTORATE': '$108,500'}
Engineering and Technology Jobs	Engineering	Design Engineer	$90,663	{'VOCATIONAL': '$80,500', 'HIGH SCHOOL': '$81,500', 'ASSOCIATE': '$81,000', 'BACHELOR': '$92,500', 'MASTER': '$103,500', 'DOCTORATE': '$110,500'}
Engineering and Technology Jobs	Engineering	Software Engineer	$113,885	{'VOCATIONAL': '$79,000', 'HIGH SCHOOL': '$79,000', 'ASSOCIATE': '$77,500', 'BACHELOR': '$92,000', 'MASTER': '$103,500', 'DOCTORATE': '$110,000'}

Figure 3: Sample of Scraped Data

bols like $ and comma. Our first data cleaning step was to remove such special characters and convert it into numeric datatype.

The Salary data for various Education Levels is in the form of JSON data. We expanded these records into proper format and performed the cleaning step for Salary data, as mentioned above.

There were around 3% of the records for which the salary data is missing. For the scope of the Project Progress Report, we discarded such values but will be considered and imputed with appropriate values in the Final Report.

Also, we noticed that some Job Titles were present in multiple sub-categories. So in order to remove redundancy we dropped those duplicate records.

3.2 Kaggle Dataset

The dataset on Kaggle by pavansubhash (2017) is created by the Data Scientists at IBM which contains data about the employee attrition and contains features such as Education, Performance Rating, Gender, Job Involvement, Relationship Satisfaction, Work Life Balance, Job Satisfaction, etc. We use this dataset to analyze the job satisfaction of employees and provide insights on them.

4 Exploratory Data Analysis

We perform Exploratory Data Analysis (EDA) on the data collected (as mentioned in section 3). The following insights are revealed from the EDA.

Fig. 4 shows the mean salary distribution among different job categories. We can see that jobs related to Information Technology and Software have

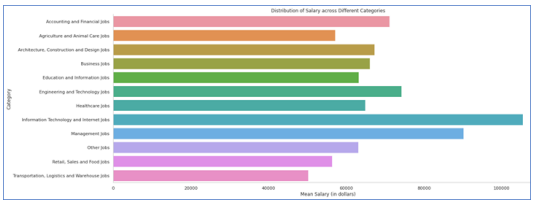

Figure 4: Distribution of Salaries across Different Categories

the highest mean salary of $100,000+. It's followed by Management jobs at around a mean salary of $90,000. Jobs related to Logistics, Agriculture, and Retail seem to have the least salaries.

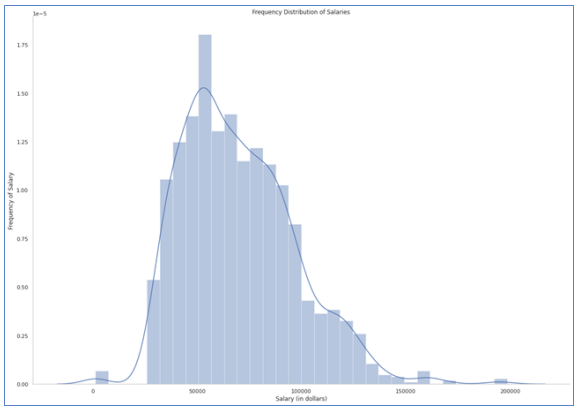

Figure 5: Frequency Distribution of Salaries

Fig. 5 denotes the frequency of different Salaries. We can notice that salary in the range of $50,000-$60,000 is the most common salary for the categories mentioned above. The most common range for salaries seems to be between $35,000 and $90,000.

5 Implementation and Results

As mentioned in the section 1, the work consists of 3 parts. Each part is explained in the separate subsection below.

5.1 Salary Prediction

The task is to predict the salary based on the Job Title, Skills and the Education Level. The dataset consists of mixed data types - textual (string) data, categorical data and numeric data (target variable). We have used Natural Language Processing (NLP) to handle the textual data (explained in detail in the following sub-section) and concatenated it with the categorical data to train the model.

An overview of the Training Process is explained in Fig. 6.

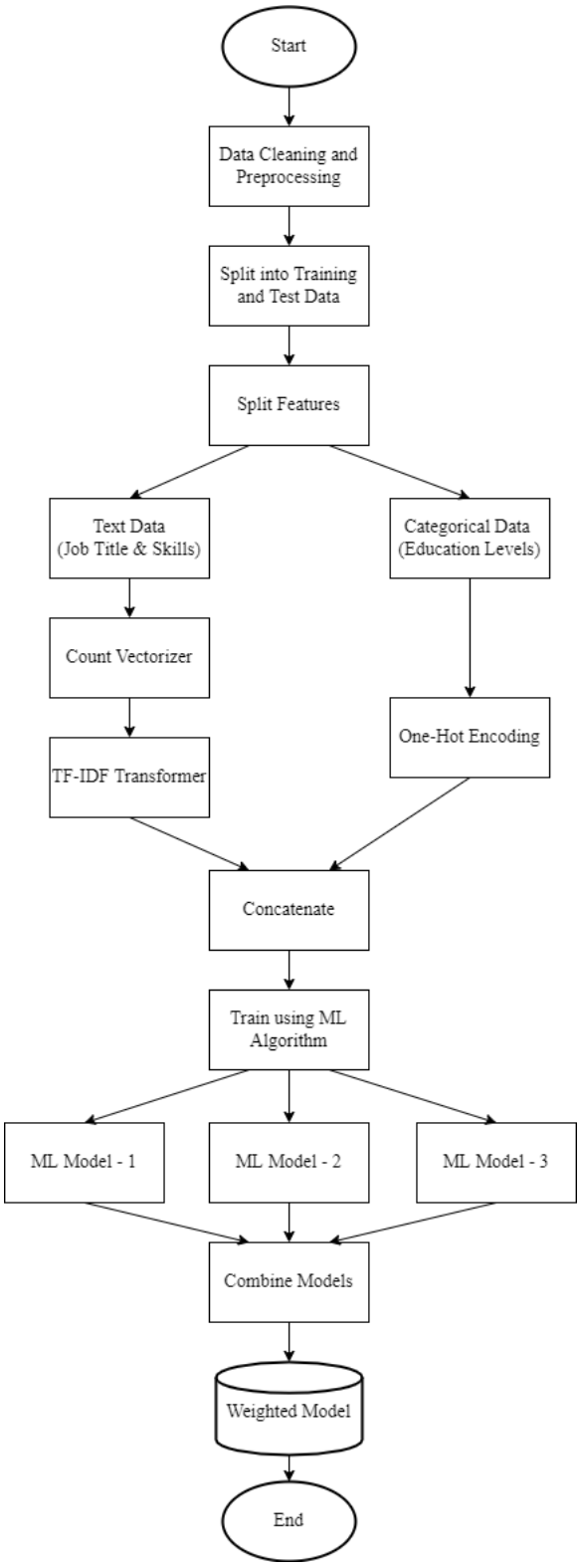

Figure 6: Salary Prediction - Model Training Process

5.1.1 Handling Textual Data

Since a Machine Learning Algorithm cannot handle textual data, so we need to convert it into a numeric representation. To do so, we have used TF-IDF (Term Frequency - Inverse Document Frequency) which is a statistical measure that evaluates how relevant a word is to a document in a collection of documents.

However, TF-IDF Transformer transforms a count matrix to a normalized tf or tf-idf representation. To get the count matrix, we used Count Vectorizer to convert a collection of text documents to a matrix of token counts. Count Vectorizer returns a sparse representation of the counts which is passed to the TF-IDF Transformer to get the normalized representation of the words i.e. Job Titles.

The weight of a term that occurs in a document is simply proportional to the term frequency. The Term Frequency is $tf(f, d)$ is defined in the equation 1.

$$tf(t, d) = \frac{f_{t,d}}{\Sigma_{t' \in d} f_{t',d}} \tag{1}$$

where $f_{t,d}$ is the raw count of a term in a document i.e., the number of times that term t occurs in document d.

The specificity of a term can be quantified as an inverse function of the number of documents in which it occurs.

The Inverse Document Frequency is defined in the equation 2.

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

where N is the total number of documents in the corpus $N = |D|$ and $|\{d \in D : t \in d\}|$ is number of documents where the term t appears

Then tf-idf is calculated as

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D) \tag{3}$$

A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms.

5.1.2 Handling Categorical Data

The subset of data containing the Categorical Values is shown in Table 1.

We have used One-Hot Encoding to encode the Education Levels as a one-hot numeric array.

Table 1: Sample of Categorical Data

Job Title	Education Level	Salary
Bank Manager	HIGH SCHOOL	$51,500
Bank Manager	VOCATIONAL	$54,500
Bank Manager	ASSOCIATE	$56,500
Bank Manager	BACHELOR	$76,500
Bank Manager	MASTER	$93,000
Bank Manager	DOCTORATE	$103,500

5.1.3 Model Training and Evaluation

Now all the features are converted to a numeric representation. After transforming the textual (Job Titles and Skills) features into numeric representation, we concatenate them with the one-hot encoded categorical features to train the models. Since salary is a continuous target variable, so we have used Regression Algorithms to train the model. Multiple Regression Algorithms such as Linear Regressor, Stochastic Gradient Descent Regressor, Random Forest Regressor, KNN Regressor, XGBoost Regressor and Extra Trees Regressor were trained.

The most popular and widely used metrics for Regression - R^2 Score and Root Mean Squared Error (RMSE) are used as evaluation metrics. The results of the individual models are shown in table 2.

Table 2: Salary Prediction - Results

Model (Regressor)	R^2 Score	RMSE
Linear Regression	0.890	6862.82
SGD Regressor	0.890	6877.58
XG Boost	0.880	7169.18
Random Forest	0.765	10050.43
Extra Trees	0.754	10268.27
Gradient Boosting	0.706	11237.91
K Neighbors	0.638	12472.71
Elastic Net	0.147	19134.62

To further improve the results, we create an ensemble of the top 3 models and use Voting Regressor for predictions. Since the R^2 Score and RMSE for the top 3 models are quite similar, we use uniform weights for all the 3 models in Voting Regressor.

After using Voting Regressor, the predictions were boosted to provide an R^2 Score of 0.901 and RMSE of 6515.37, which means the average predicted salaries have a deviation of $6515.37 from the actual salaries which is very quite less considering that the salaries are predicted on a yearly basis. On an hourly basis, the deviation will be around $3.4.

5.1.4 User Interface

The user interface along with a test case is shown in Fig. 7

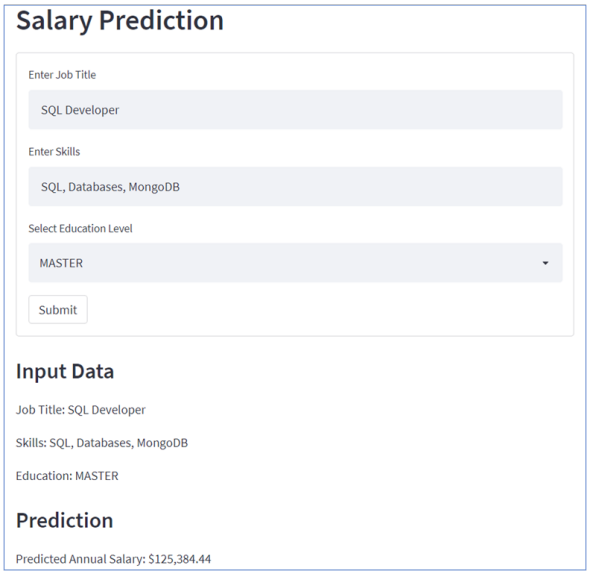

Figure 7: Salary Prediction - User Interface

5.2 Job Clustering

In this task, we cluster the Job Titles based on the required Skills. We use the same dataset of the previous task i.e. the dataset scraped from the employment website (as explained in section 3.1). The flowchart of the Clustering Process is explained in Fig. 8.

5.2.1 Handling Textual Data

Since the data for this task consists completely of Textual Data i.e. the Job Titles and the Skills, so we followed the same methodology mentioned in section 5.1.1 to handle the textual data and convert it into numeric representation.

5.2.2 Clustering

Since the number of clusters cannot be decided beforehand for this task, so cannot use K-means clustering. Instead, we use DBSCAN Clustering Algorithm to cluster the Job Titles.

DBSCAN (Ester et al., 1996) is a clustering algorithm that defines clusters as continuous regions of high density and works well if all the clusters are dense enough and well separated by low-density regions

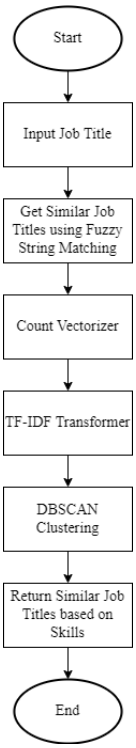

Figure 8: Job Title Clustering - Process Flow

In the case of DBSCAN, instead of guessing the number of clusters, will define two hyperparameters: epsilon and minPoints to arrive at clusters.

1. *Epsilon* (ϵ): A distance measure that will be used to locate the points/to check the density in the neighbourhood of any point.

2. $minPoints$ (n): The minimum number of points (a threshold) clustered together for a region to be considered dense.

5.2.3 Evaluation

We used Silhouette Coefficient (Rousseeuw, 1987) to evaluate the quality of the clusters. The best value is 1 and the worst value is -1.

The model received a Silhouette Coefficient of 0.134, indicating decent clusters with non-overlapping clusters or mislabeled data points.

5.2.4 User Interface

The user interface along with a test case is shown in Fig. 9

5.3 Job Satisfaction Analysis

Another aspect to Job Title Analysis is to analyze the Job Satisfaction of employees. In this task, we aim to analyze the IBM HR Analytics Employee

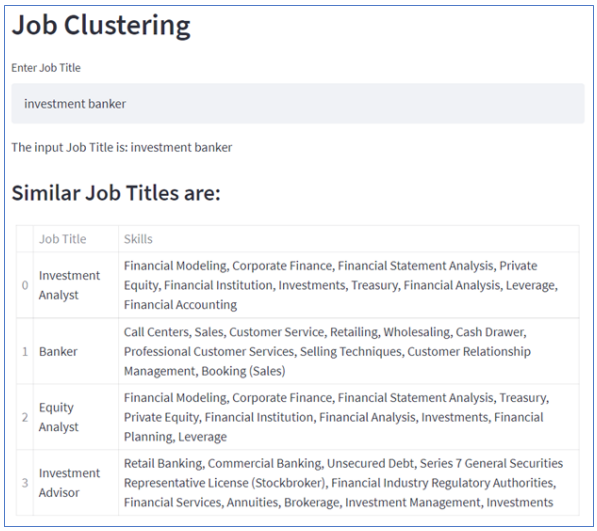

Figure 9: Job Clustering - User Interface

Attrition & Performance dataset available on Kaggle (as explained in section 3.2) and derive some insights from it.

We train Random Forest Classifier twice - one with 'Attrition' as the target label and second time with 'Job Satisfaction' as the target label.

The importance of the attributes with respect to 'Job Satisfaction' are shown in Fig. 10 and feature importances with respect to 'Attrition' are shown in Fig. 11.

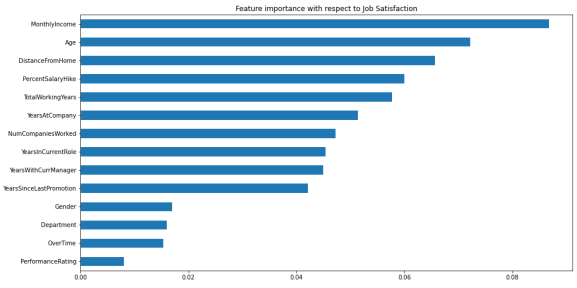

Figure 10: Job Satisfaction - Feature Importance

Fig. 10 and Fig. 11 denote the impact of various features on Job Satisfaction and Attrition. As we can see from the graph, the most important features contributing to Job Satisfaction and Attrition are Monthly Income and Age. Salary Hike has more importance on the Job Satisfaction of an employee but not on the Attrition. Over Time has a much higher importance for Attrition on an employee but does not contribute much to the Job Satisfaction. Performance Rating and Department contribute among the lowest importances to both

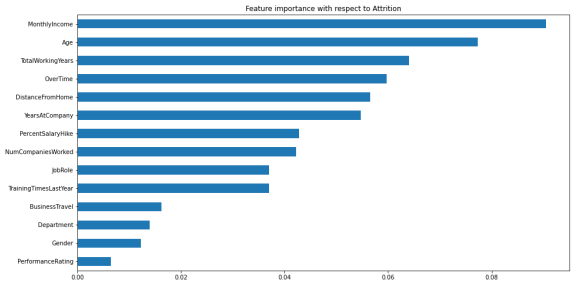

Figure 11: Attrition - Feature Importance

the Job Satisfaction and Attrition.

Fig. 12 denotes the distribution of Job Satisfaction and the work-life balance across different age buckets. As we can see from graph, the Job Satisfaction is directly proportional to the work-life balance. Better work-life balance indicates better Job Satisfaction.

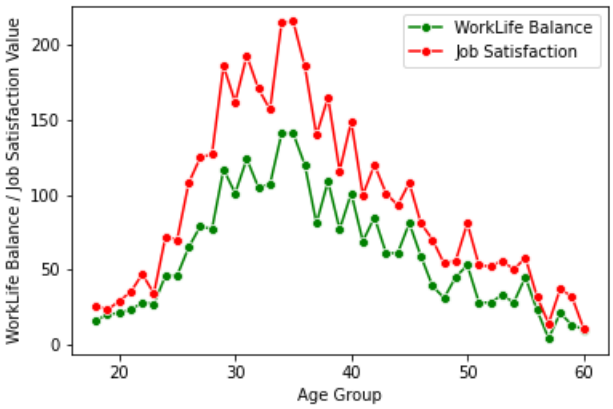

Figure 12: Distribution of Work-Life Balance and Job Satisfaction

6 Conclusion

Analyzing Job Titles cannot be viewed as one dimension problem since there are many contributing factors. In this paper, we presented a multi-dimensional approach for analyzing jobs such as Salary Prediction, Job Title Clustering and provided insights on the Job Satisfaction. We are able to successfully develop model for Salary Prediction with an R^2 Score of 0.901 and used DBSCAN clustering for grouping Job Titles with similar skills.

7 Future Scope

The scope of this work can be extended by including additional features such the number of candidates and the number of jobs available to improve the results of Salary Prediction model.

Job clustering can be improved by trying various other clustering algorithms and improving the numeric representations by considering advanced techniques such word2vec and gensim embeddings which can capture the semantic meanings to create better representations.

References

CareerBuilder. 2021. Browse all us jobs. Accessed: 2021-10-31.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231.

pavansubhash. 2017. Ibm hr analytics employee attrition & performance. Accessed: 2021-11-15.

Ch. Platis, P. Reklitis, and S. Zimeras. 2015. Relation between job satisfaction and job performance in healthcare services. *Procedia - Social and Behavioral Sciences*, 175:480–487. Proceedings of the 3rd International Conference on Strategic Innovative Marketing (IC-SIM 2014).

Abdul Raziq and Raheela Maulabakhsh. 2015. Impact of working environment on job satisfaction. *Procedia Economics and Finance*, 23:717–725. 2nd GLOBAL CONFERENCE on BUSINESS, ECONOMICS, MANAGEMENT and TOURISM.

Leonard Richardson. 2017. Beautiful soup documentation.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

K. P. Royer. 2010. *Job descriptions and job analyses in practice: How research and application differ*. Ph.D. thesis, College of Liberal Arts & Social Sciences, DePaul University.

Wenjun Zhou, Yun Zhu, Faizan Javed, Mahmudur Rahman, Janani Balaji, and Matt McNair. 2016. Quantifying skill relevance to job titles. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1532–1541.