**Statement of Work**

Yingshu Wang 100614552

**Introduction**

This statement of work outlines the general scope of the project named "Data Analysts’ salaries". The main objective of this project is to help both employers and employees to understand the existing US data analyst career market. Employers can take advantage of the salaries’ solutions to avoid **overpaying** employees, attract high levels of data analysts and reduce the employment turnover rate. It also guides the job seekers to pick a fair paid offer.

**Problem Statement**

An overpaying is an amount of money paid to an employee to which they are not entitled.

It affects the employee motivation,the development of the company and the Employee-company relationship. At the first, overpaying may limit the development of employees. They will consider the extra wages paid as part of their due. They will be satisfied with what they have now and refuse to continue working hard. Then, Excessive wages restrict the company’s own resources and slow down the company’s development. In addition, Once the company asks the employees to recover more than their income, the employees are very likely to become dissatisfied and reduce their loyalty to the company.

There is a possible solution that giving a reasonable salary mechanism and a data engineer with a good algorithm to reduce the possibility of error

Employee turnover, or employee turnover rate, is the measurement of the number of employees who leave an organization during a specified time period, typically one year. It affects company revenue and profitability, employee morale and Product or Service Quality. The first thing is the company's revenue and profitability. Especially if a severance package is paid, this will be an expense with no return on investment. The cost of making a job-placement service will have no return. The second impact is that it will result in low workplace morale. As employees leave, this will bring increased workloads and responsibilities to remaining employees. They may suffer from low morale from this working environment. Finally, as employees leave, the low productivity and quality of work can result from a disruption by daily high workload employees.

The company should hire the people right to it and in the meanwhile fire the people who don’t fit. In addition, keeping a compensation and benefits current can also help with low employee turnover.

**Data Requirements and limitation**

All the contents within the data analyst hiring post are required. The expected essential data should include the company's salaries offered, the sector it belongs to, company name, company size, date of the company found, working location, employee’s certificate requirement and skills requirement. Dataset is supposed to have various floats and sentences.

In consideration of company security, all sensitive information about the company should be masked during job recruitment. For example, the company’s working environment, job bonus and other benefits. Only after they have confirmed their identity in the company then they can be posted to the employees.

Data should consist of one unified standard and unify all benchmarks in the country where the company is located. For instance, salaries offered should be all listed in US dollars.

Data analysis job data could be found in the popular job post website, like indeed.com. Also, there are many professional big raw data websites. Therefore, the company can give priority to job search information from these websites when posting job postings.

**Validation process**

The solution will be tested by several validation methods. Firstly, unrepresentative data will be removed. Outliers increase the variability of data, thereby reducing statistical power. Therefore, excluding outliers may make your results statistically meaningful. Then, it’s expected to do data normalization. The goal of normalization is to change the value of a numeric column in the data set to a common scale without distorting the difference in the value range. By having the same scale, data is easier to be compared. It can also increase the model's accuracy. Also

The project is using multiple methods to validate the model, including k-nearest neighbors algorithm(KNN), decision tree and random forest. Using “score(X, y, sample\_weight=None)”, simply checking the mean accuracy on the given test data and labels. Plotting Validation Curves shows training scores and validation scores in a line graph.

**Reference:**

<https://www.businessanalystlearnings.com/ba-techniques/2017/6/27/what-is-a-problem-statement#:~:text=A%20problem%20statement%20defines%20the,up%20with%20a%20product%20vision.&text=Solution%3A%20Include%20your%20recommendation%20for%20solving%20the%20problem>.

<https://smallbusiness.chron.com/negative-impacts-high-turnover-rate-20269.html>