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 Canadian 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, 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 do not 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.

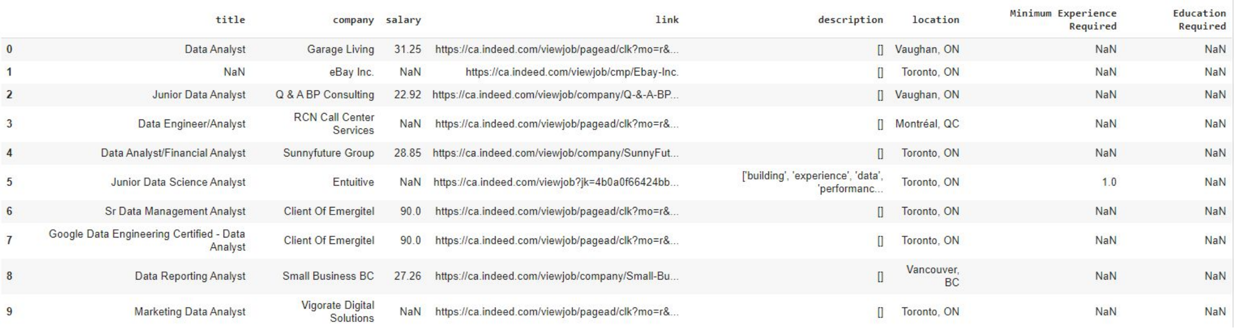
Data should consist of one unified standard and unify all benchmarks in the country where the company is located. For instance, salaries offered should be all listed in Canadian dollars.

Data analysis job posting could be found in the popular job post website, like indeed.com. Indeed is the first job site in the world[1] with over 250 million unique visitors[2] every month. They updated various jobs everyday. By taking advantage of indeed.com, the company can give priority to job search information from these websites when posting job postings. The realistic data collected from the website reflects the existing career market, helping the company and job seekers to know and understand the latest data analyst market.

Data Preprocessing

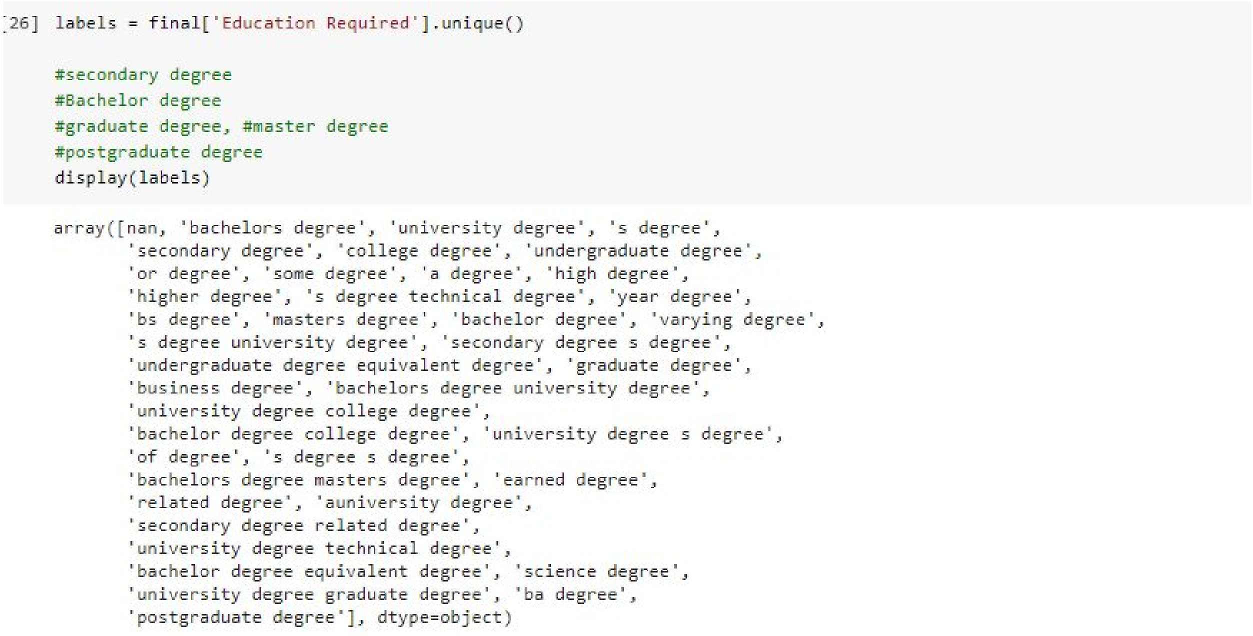
The picture below shows the first ten data we collected from page 0 to page 160 on November 27 2020.

- Title - job title
- Company – the job company offered
- Salary – hourly Wage in Canadian dollar
- Link - Indeed link
- Description – the summarized keyword from job post content
- Minimum experience required – the minimum experience the company asked
- Education required - the education the job post mentioned, (should) taking the lowest one



	title	company	salary	link	description	location	Minimum Experience Required	Education Required
0	Data Analyst	Garage Living	31.25	https://ca.indeed.com/viewjob/pagead/clk?mo=r&...	[]	Vaughan, ON	NaN	NaN
1	NaN	eBay Inc.	NaN	https://ca.indeed.com/viewjob/cmp/Ebay-Inc.	[]	Toronto, ON	NaN	NaN
2	Junior Data Analyst	Q & A BP Consulting	22.92	https://ca.indeed.com/viewjob/company/Q-&-A-BP...	[]	Vaughan, ON	NaN	NaN
3	Data Engineer/Analyst	RCN Call Center Services	NaN	https://ca.indeed.com/viewjob/pagead/clk?mo=r&...	[]	Montréal, QC	NaN	NaN
4	Data Analyst/Financial Analyst	Sunnyfuture Group	28.85	https://ca.indeed.com/viewjob/company/SunnyFut...	[]	Toronto, ON	NaN	NaN
5	Junior Data Science Analyst	Entuitive	NaN	https://ca.indeed.com/viewjob?jk=4b0a0f66424bb...	['building', 'experience', 'data', 'performanc...	Toronto, ON	1.0	NaN
6	Sr Data Management Analyst	Client Of Emergitel	90.0	https://ca.indeed.com/viewjob/pagead/clk?mo=r&...	[]	Toronto, ON	NaN	NaN
7	Google Data Engineering Certified - Data Analyst	Client Of Emergitel	90.0	https://ca.indeed.com/viewjob/pagead/clk?mo=r&...	[]	Toronto, ON	NaN	NaN
8	Data Reporting Analyst	Small Business BC	27.26	https://ca.indeed.com/viewjob/company/Small-Bu...	[]	Vancouver, BC	NaN	NaN
9	Marketing Data Analyst	Vigorate Digital Solutions	NaN	https://ca.indeed.com/viewjob/pagead/clk?mo=r&...	[]	Toronto, ON	NaN	NaN

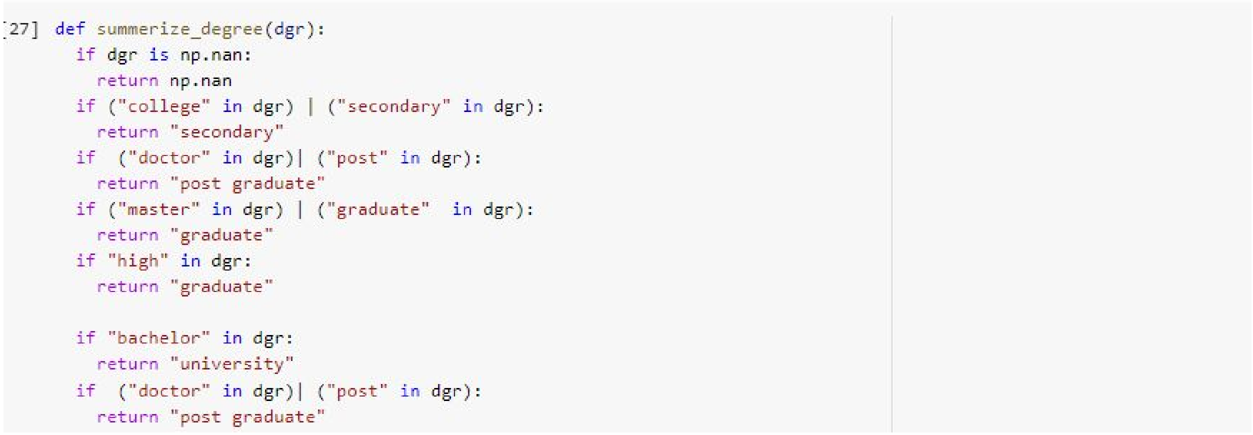
There are multiple text about degree.



```
[26] labels = final['Education Required'].unique()

     #secondary degree
     #Bachelor degree
     #graduate degree, #master degree
     #postgraduate degree
     display(labels)
```

```
     array([nan, 'bachelors degree', 'university degree', 's degree',
            'secondary degree', 'college degree', 'undergraduate degree',
            'or degree', 'some degree', 'a degree', 'high degree',
            'higher degree', 's degree technical degree', 'year degree',
            'bs degree', 'masters degree', 'bachelor degree', 'varying degree',
            's degree university degree', 'secondary degree s degree',
            'undergraduate degree equivalent degree', 'graduate degree',
            'business degree', 'bachelors degree university degree',
            'university degree college degree',
            'bachelor degree college degree', 'university degree s degree',
            'of degree', 's degree s degree',
            'bachelors degree masters degree', 'earned degree',
            'related degree', 'auniversity degree',
            'secondary degree related degree',
            'university degree technical degree',
            'bachelor degree equivalent degree', 'science degree',
            'university degree graduate degree', 'ba degree',
            'postgraduate degree'], dtype=object)
```



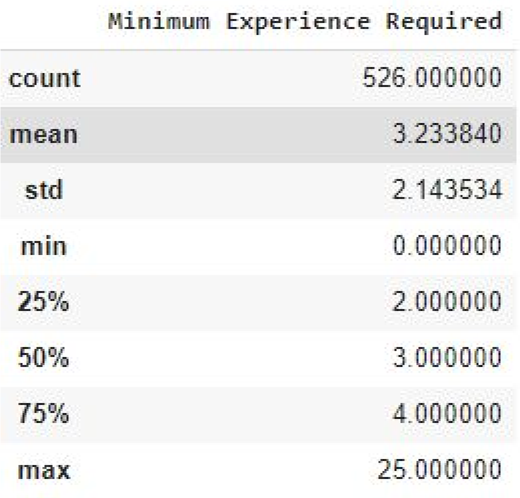
```
[27] def summerize_degree(dgr):
       if dgr is np.nan:
         return np.nan
       if ("college" in dgr) | ("secondary" in dgr):
         return "secondary"
       if  ("doctor" in dgr)| ("post" in dgr):
         return "post graduate"
       if ("master" in dgr) | ("graduate"  in dgr):
         return "graduate"
       if "high" in dgr:
         return "graduate"

       if "bachelor" in dgr:
         return "university"
       if  ("doctor" in dgr)| ("post" in dgr):
         return "post graduate"
```

By using unique() function, all education requirements can be summarized. summerize_degree() function classifies data into five main categories, including NaN, secondary, graduate, university and postgraduate degrees.

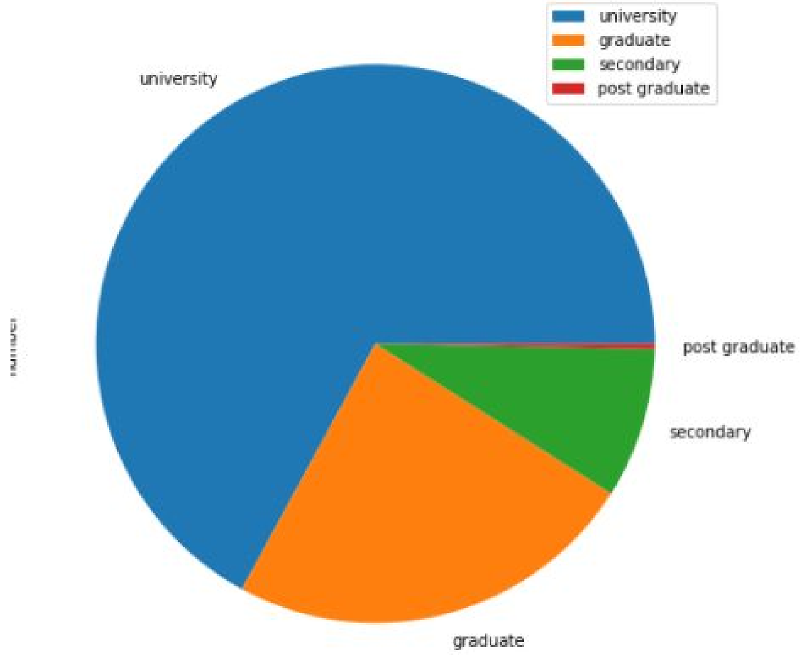
EDA

Without removing all null data, there are 526 posts indicating their minimum experience requirements. Most of them require 2 to 4 years experience. However, the detection function may not correctly find out all experience requirements. For example, the max one, 25 years, is not the minimum experience required. It’s about the company's history.

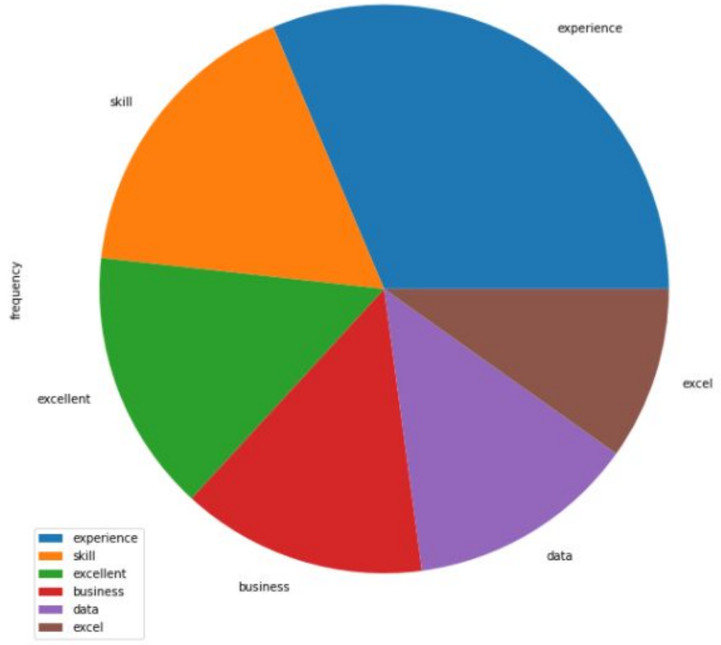


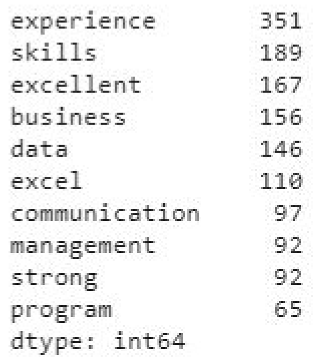
	Minimum Experience Required
count	526.000000
mean	3.233840
std	2.143534
min	0.000000
25%	2.000000
50%	3.000000
75%	4.000000
max	25.000000

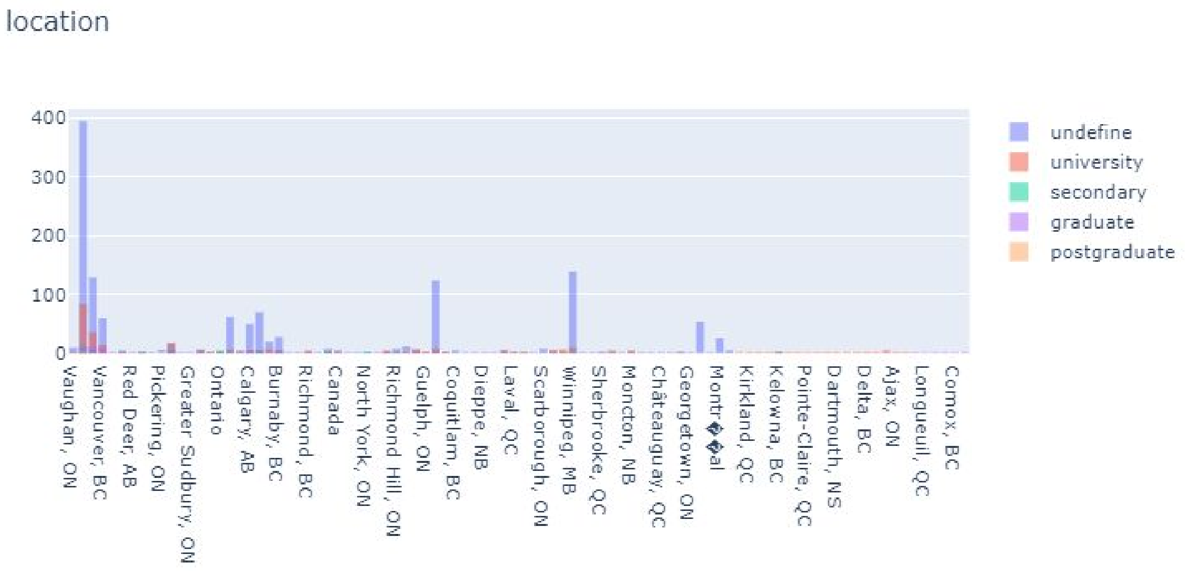
A pie chart represents the degree requirements for data analysts. Most of the jobs ask for a university degree or higher. Also, there are a significant number of companies ask for master degrees.



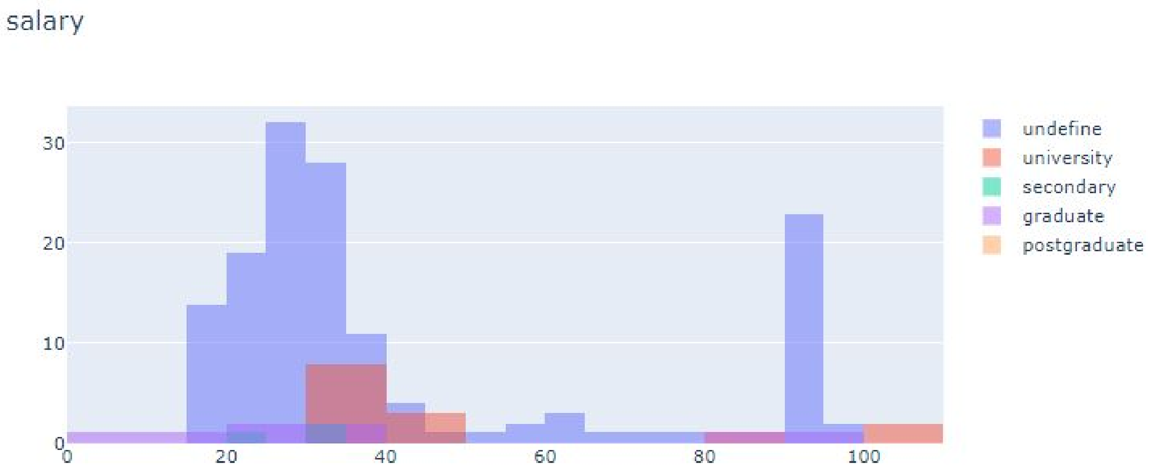
Keywords in the job posts show the key qualifications of labour needed for the company.




```
experience       351
skills           189
excellent        167
business         156
data             146
excel            110
communication     97
management        92
strong            92
program           65
dtype: int64
```

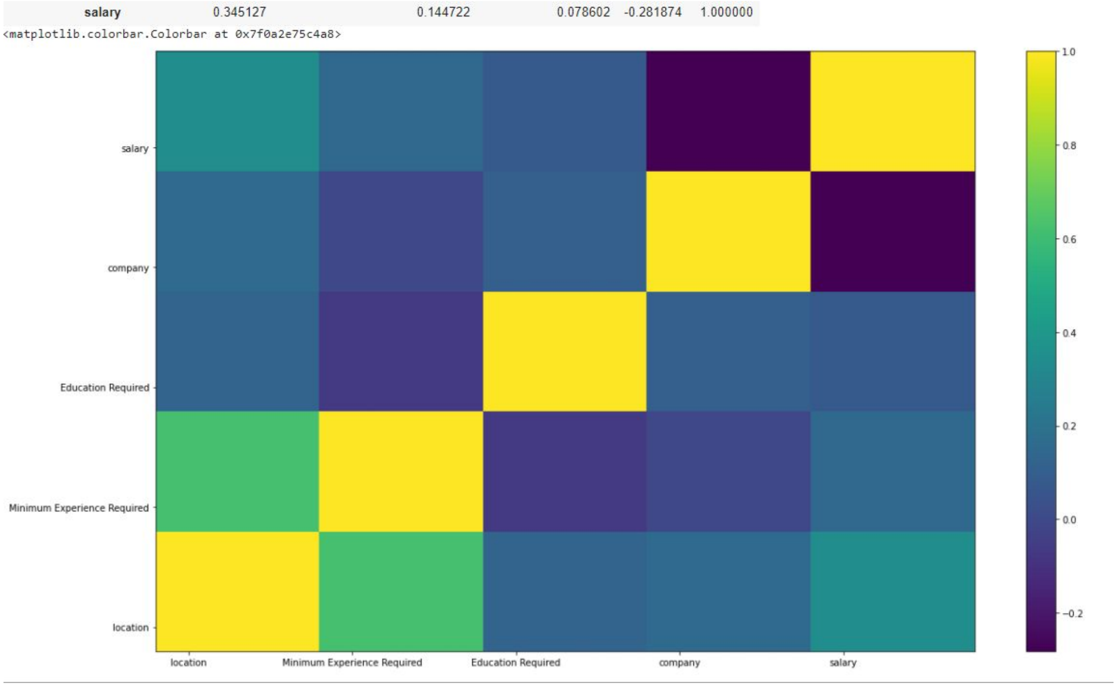


Most of the job positions available in Toronto, Vancouver, Edmonton and Montreal.



Most job posts which show the salary, do not clear out the degree they required. In this graph, degree requirements and salary do not have an obvious relationship.

Feature Engineering is creating features for raw data. Feature engineering allows programs to reformulate nonlinear problems as linear problems, enhancing the predicted result’s accuracy. Because features have different ranges, it is necessary to normalize the data. Therefore, there is no distorting difference in values. After normalizing data, an correlationship graph is generated.



```
salary    0.345127    0.144722    0.078602   -0.281874   1.000000
<matplotlib.colorbar.Colorbar at 0x7f0a2e75c4a8>
```

Salary offered has medium relationship with location and company, but not education requ iment and experience requirement. However, it’s interesting to find that the job location and minimum experience required has a strong relationship.

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 is 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