**Job Salary Analysis Based on Text Classification and Location Feature**

ISGB/BYGB 7977-002: Text Analytics

Instructor: Professor Yilu Zhou

May 5, 2022

Section 2 - Group 11

Maoxi Li

Qi Wang

Huiye Huang

Nanzanxuan Chen

Chengqi Xie

# **Executive Summary**

The goal of this project is to help companies find the appropriate salary for the employees to increase their market competitiveness and help job seekers find the key factors that affect the salaries.

By doing exploratory data analysis for raw data crawled from Glassdoor, completing the data cleaning process and then extracting the top skill requirements from text content, our project did a statistical analysis to gain the distribution of different features and explored the relationship between each independent variable and salary level. After vectorizing the whole job description content, the result as input was used for a traditional machine learning model to conduct a text classification and predict the salary level. Then the project used the top 10 skills for each job title, job location and other features as the combination input for the same traditional machine learning model to classify the salary level and predict it. Finally, evaluate the model and gain the insights from the result.

After evaluating these models, the model combining TF-IDF and Random Forest performs the best prediction. Skills are the most essential factors to affect job salary. Other factors, such as location, founding year and rating, also have influence on the prediction result.

# **Business Goal Analysis**

## **Background**

*Determining the “right” compensation can be tricky. Not only is money a touchy subject, but so many factors(e.g., experience, location and skills) play into determining compensation rates that are both fair and* [*competitive*](https://resources.careerbuilder.com/featured-stories/what-s-a-competitive-wage)*. (“5 essential factors for determining compensation”,2022)*

There are a variety of articles talking about job salaries on the Internet. In these articles, a lot of factors can greatly affect job salaries.

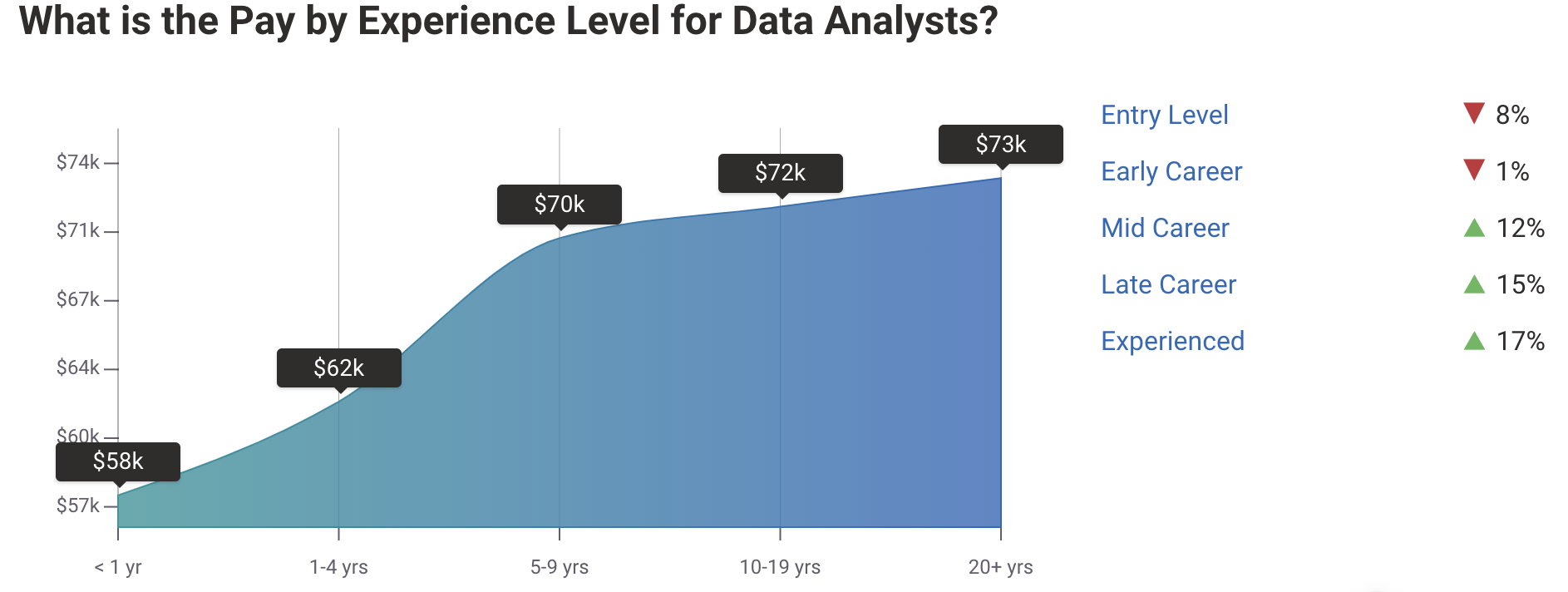


Figure1. Salary by Experience Level for Data Analysts

Figure 1 shows that job salary is related to the experience level, but doesn’t change a lot when the working year is more than 9 years.

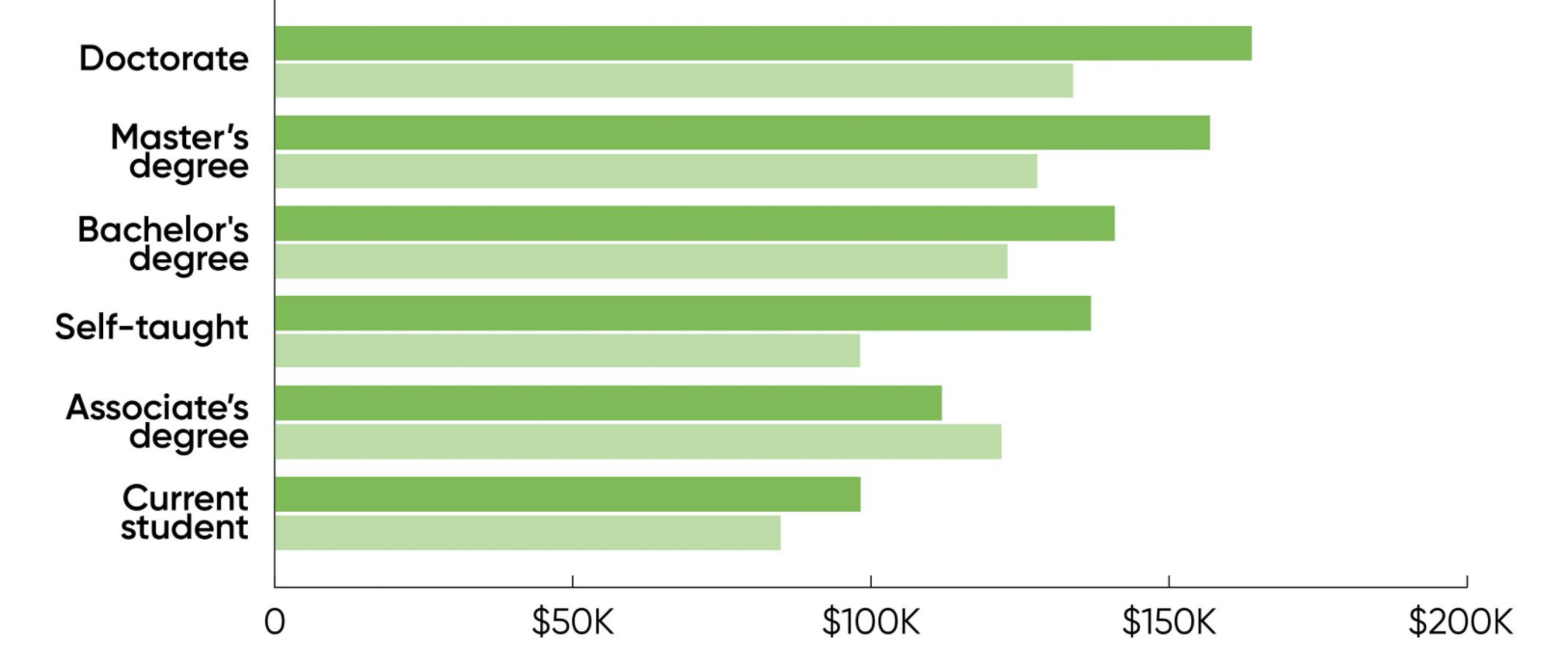


Figure 2. The Difference between Different Degree and Gender to Salary

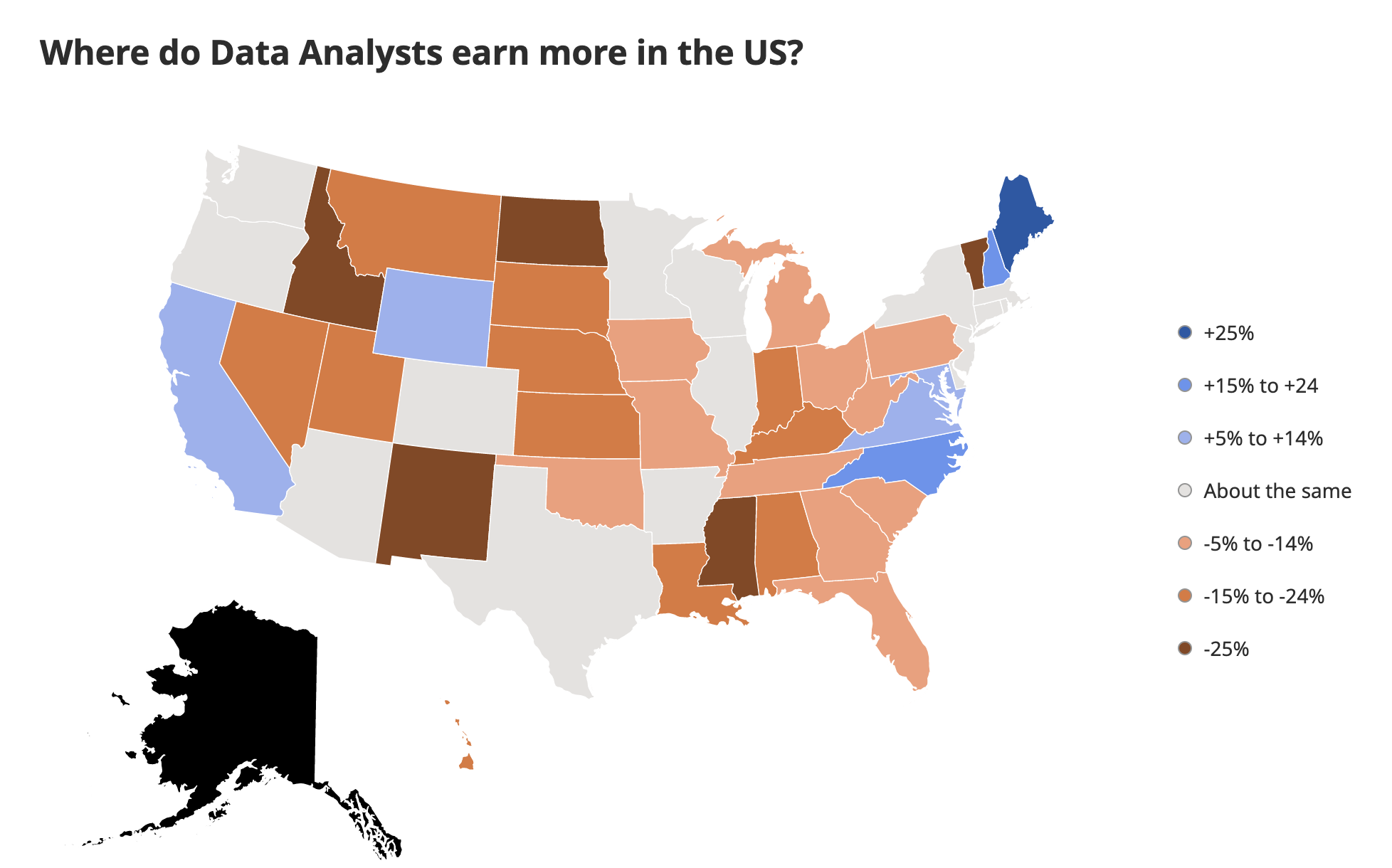
Figure 2 shows the job salary is related to the degree. 

Figure 3. Different Salary into Different States

Figure 3 shows the relationship between location state and job salary, there is a large difference in salary between different states.

Through reading these articles, it’s really interesting to see the features that can affect job salaries a lot.

## **Business Goal**

The result of this project will be beneficial to the companies and job seekers who want to have an insight for salaries.

On one hand, this prediction model will help companies to find the appropriate salary to the employer and increase their market competitiveness.

On the other hand, this prediction model will be helpful for job seekers to find the key factors which affect the salaries. Thus, job seekers can find ways to improve their skills to get well-paid jobs.

# **Dataset Description**

This dataset contains job postings from companies in 49 states for data-related positions, such as Data Scientist, Data Analyst, Data Engineer and Business Analyst.

The dataset is from Github, and the data in the dataset was crawled from Glassdoor by [Selenium](https://selenium-python.readthedocs.io/) and [Beautiful Soup](https://www.crummy.com/software/BeautifulSoup/doc). This dataset is 49.2MB in size which includes 15761 job posting information. There are fifteen features to describe the job detail which includes *Job Title, Salary Estimate, Job Description, Rating, Company Name, Location, Headquarters, Size, Founded, Type of ownership, Industry, Sector, Revenue, Competitors* and *Whether Easy Apply*.

The data types in this dataset are numerical , categorical and text(unstructured).

More and more companies are posting jobs with increasingly large salary ranges to meet the different conditions of the candidates. Using the last two years of job information as the dataset will reduce the final prediction result of the model. So in this study, Glassdoor 2019 full-year data-related job within the United States was chosen as the dataset.



Figure 4. Dataset Example

# **System Design**

First, define the business problem of this research and determine the business goal and expected output. Then did an exploratory data analysis to find out missing value and descriptive statistics for numeric features. After that, data preprocessing step (dealing with missing value, cleaning outliers and discretizing continuous data) was used to clean the raw data.

After the cleaned data was ready, statistical analysis was used to uncover the deep relationships contained in the data, such as the distribution of different features, the relationship between different independent variables and dependent variables and correlation analysis between independent variables. Text extraction processing for *Job Description* should be done at the same time to find out the top skills for each job title.

Then, vectorize the text content in the *job description* in three ways and use a traditional machine learning model(including Naive Bayes, Decision Tree, Random Forest and Neural Network) to predict the salary level. And then use top 10 skills, location and other features as the input to predict the salary level.

Finally, compare the result between these two models and evaluate the model result to get a project conclusion.

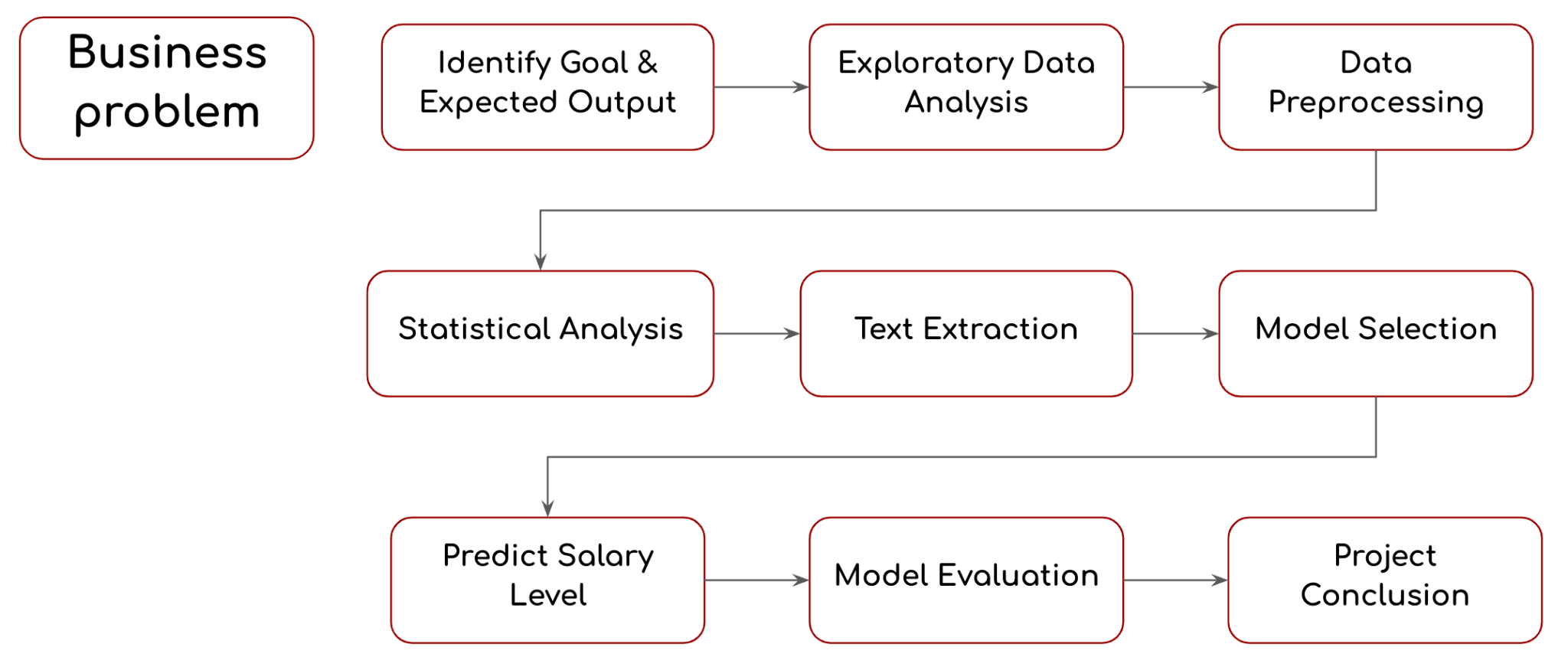


Figure 5. System Design Workflow

# **System Implementation**

## **Data Preprocessing**

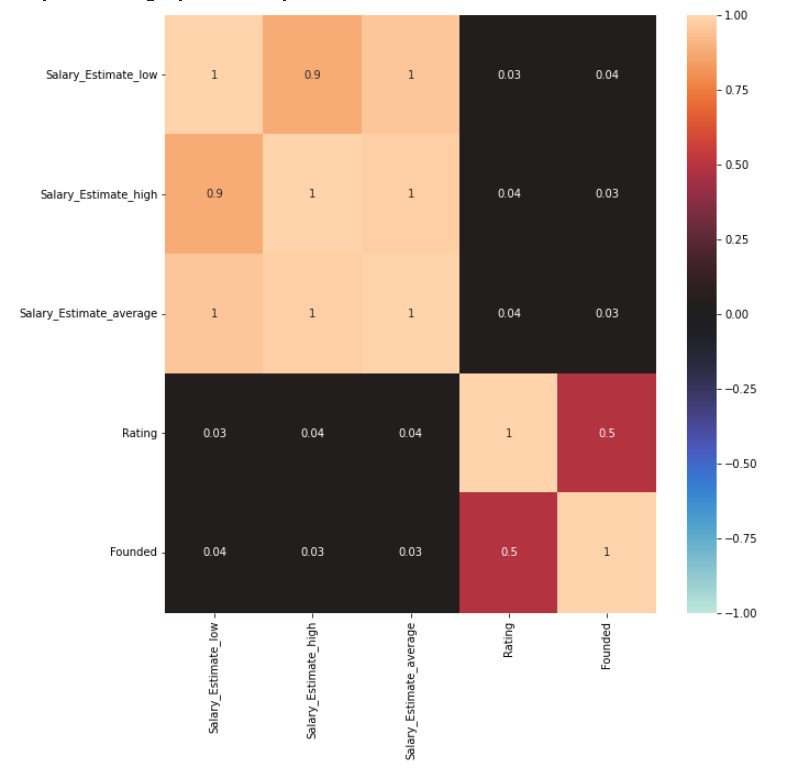


Figure 6. The Correlation Matrix for Numerical Features

For data pre-processing, this project used correlation matrix to drop irrelative features. The following figure is a numerical feature correlation matrix. For nominal features, this project assigns different numbers to draw nominal features correlation matrix and find out the irrelative features, e.g., Headquarters and Industry.

Regarding the feature Location, because this feature contains both city and state information, this project needs to split Location into Location\_city and Location\_state.

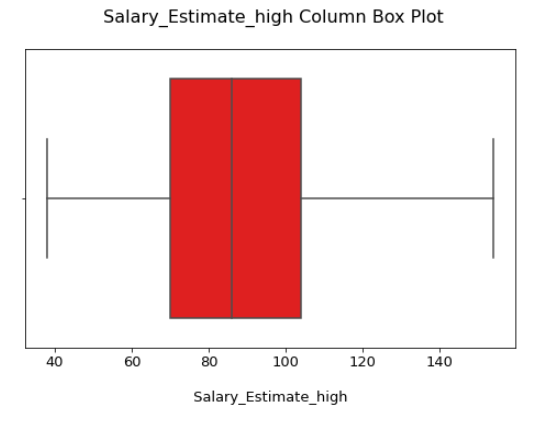
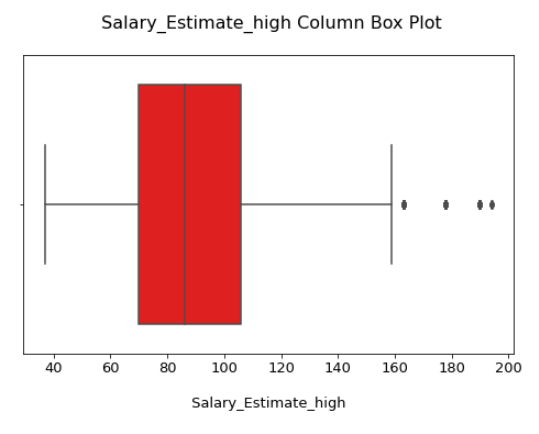


Figure 7(Left). The Box Plot of Salary With High Range

Figure 7(Right). The Box Plot of Salary With High Range After Remove Outliers

Because the salary ranges of the 4 occupations after integration are different, if the project uses the integrated salary as the base to remove outliers directly, the result will not be relevant. Hence, this project needs to divide the occupations first, and then set 3 intervals for each occupation's salary range, such as high predicted salary, low predicted salary, and average predicted salary. The next step is to draw 12 box plots corresponding to three salary range intervals of 4 different occupations and use the outlier formula to identify outliers and eliminate them.

## **Statistical Analysis**

## 

Figure 8. The Average Salary for Different States

## 

Figure 9. The Min and Max Salary for Different States

In Figure 8, the first is the top 10 average salaries of states, of which California, Oregon, and New York these three states have the highest average salary. After analysis, we find that 7 of 10 states are coastal states.

Figure 9 shows the top 10 MIN & MAX salary states. This chart is done with the top 10 MIN salary as the base, of which MAX salary matches 8 states, in other words, these 8 states are on the MIN & MAX salary list at the same time, where the black line represents their MIN & MAX salary interval. The colored bar chart represents their average. From this graph, we can analyze that CA's min and max salary are the highest, and NY's black line length is smaller than other states, which means that NY's salary distribution will be more concentrated relative to other states.

## 

Figure 10. The Violin Plot for Different Job Titles in NY

In this project the salary distribution of different job titles is analyzed at a deeper level, using NY as an example, the salary distribution of 4 job titles in NY is plotted out using the violin plot.

The orange color is Data Scientist, by its distribution can be seen that its salary range is large, and its distribution is very uniform, it does not have a significant mode.

Data Engineer's average salary ranking is second. After the analysis, by its shape, most of its salary distributions have a high degree of overlap with Data Scientist.

DataAnalyst and BusinessAnalyst have similar shapes and similar average salaries, and their average salaries are also the two lowest among the 4 jobs.

To sum up, the salary distribution of the four job titles is very different and has significant differences.

## **Text Extraction**

In the text extraction part, we mainly focused on the “Job Description” column of the data, gaining insights on the requirements of skills, year of experience and degree.

### 1. Skill

To extract skills, the text of all job descriptions is tokenized. Based on the observation of tokens frequency and domain knowledge, we identified and extracted top 30+ skills, which are shown below:

'python', 'sql', 'r', 'excel', 'tableau', 'bi', 'c', 'c++' , 'java', 'hadoop', 'scala', 'aws', 'flask', 'pandas', 'spark', 'scikit', 'numpy', 'php', 'mysql', 'css', 'mongdb', 'mango', 'nltk', 'flink', 'fastai', 'keras', 'pytorch', 'tensor', 'tensorflow', 'linux', 'ruby', 'javascript', 'django', 'react', 'reactjs', 'ai', 'ui'

As a result, the frequency of top 10 skills for 4 job titles are generated.

|  | **sql** | **python** | **r** | **aws** | **spark** | **tableau** | **excel** | **hadoop** | **java** | **bi** | **c** | **Java script** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data Scientist** | 2983 | 2629 | 1502 | 1163 | 1133 | 856 | 792 | 782 | 770 | 716 | 295 | 182 |
| **Data Engineer** | 2144 | 1925 | 316 | 1501 | 1184 | 363 | 224 | 903 | 1082 | 442 | 326 | 292 |
| **Business Analyst** | 1630 | 341 | 243 | 135 | 32 | 639 | 1396 | 53 | 132 | 563 | 152 | 84 |
| **Data Analyst** | 5338 | 1743 | 1267 | 436 | 211 | 1928 | 3095 | 339 | 335 | 1543 | 247 | 431 |

Figure 11. The Heart Map of Most Frequent Skills

The heat map (see Figure 11) indicates that SQL and Python are the top skills in demand. In addition, Cloud Computing skills such as AWS and visualization skills could add value for candidates.

Some insights can be gained further according to different skill requirements among job titles. For instance, for beginners who are not good at python, it's considerable to start with an Analyst position. Data Engines who plan to transition to the Data Scientist positions are much more likely to require to equip with the R language, which they may not have learned.

### 2. Years of Experience

## Another aspect we dug into is the years of experience, for which we searched and located text strings about years of experience required within the “Job Description” column.

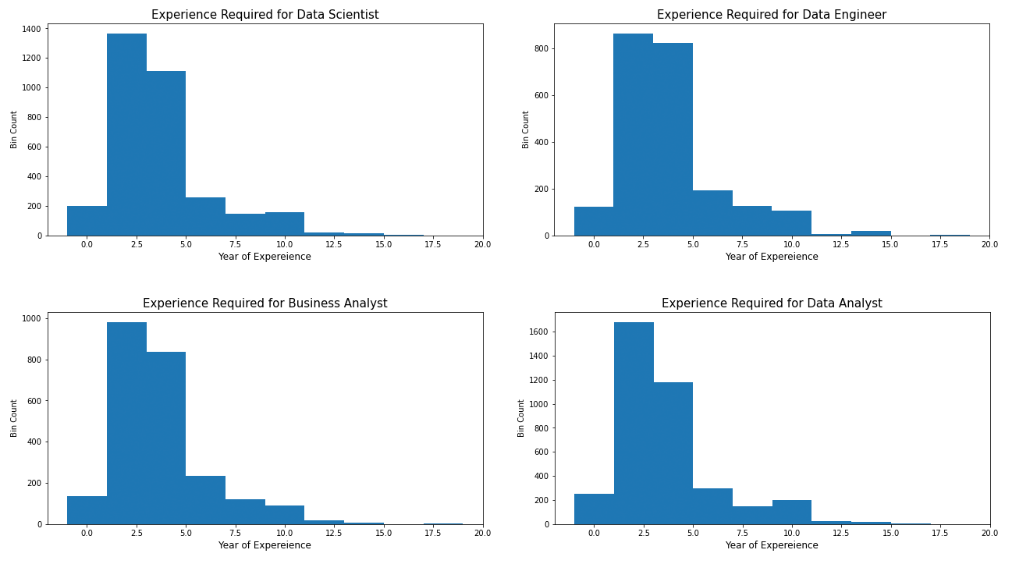


Figure 12. The Bar Charts of Distribution of Years of Experience Required

The years of experience required show similar distribution on different job titles (see Figure 12). It is easy to observe that a large proportion of experience required is less than 5 years. In addition, the average numbers are very close, 4.25, 4.42, 4.29, and 4.17 respectively, revealing that the length of experience may not distinguish these four job titles.

### 3. Degree

Likewise, the same methods are used to extract the degree requirement.

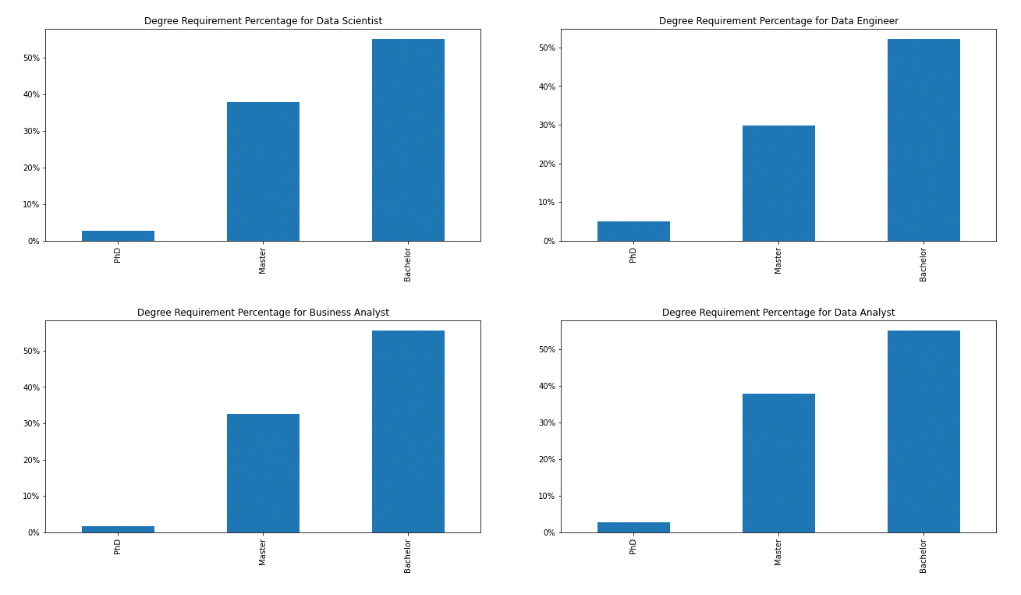


Figure 13. The Bar Charts of Degree Requirement Percentage

The distribution of the bar charts (see Figure 13) reveals that most jobs expect candidates holding a Master’s or Bachelor’s degree. Data Scientist could be considered to be most technical among all titles with the highest percentage of Ph.D degree required.

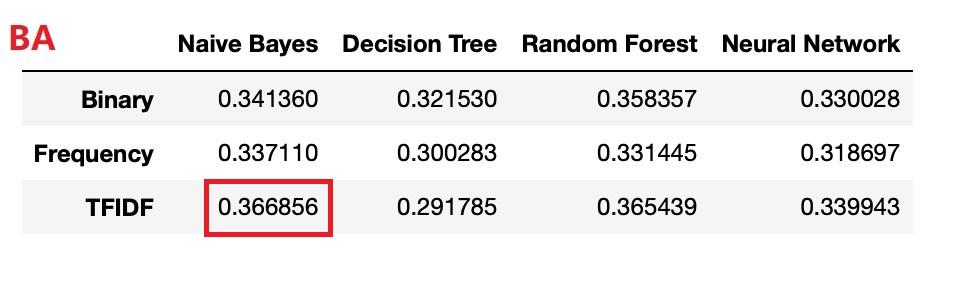
# **Model and Evaluation**

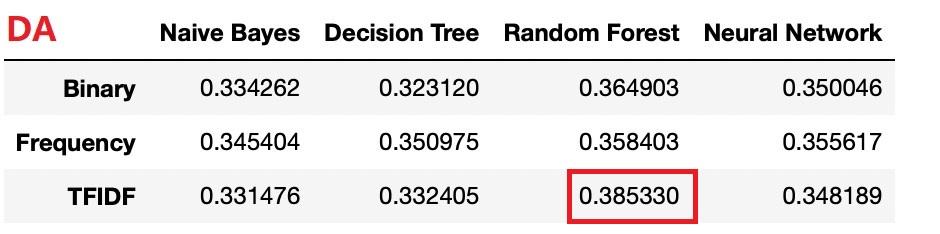
## **Model 1**

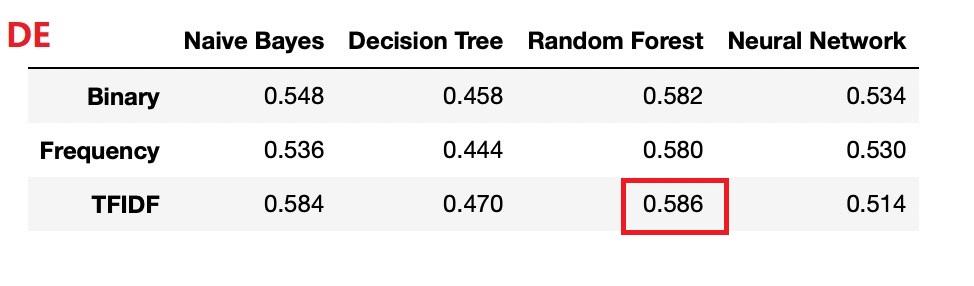
In our initial model, the entire text from the ‘Job Description’ column is vectorized using three different methods: binary, frequency, and TF-IDF.

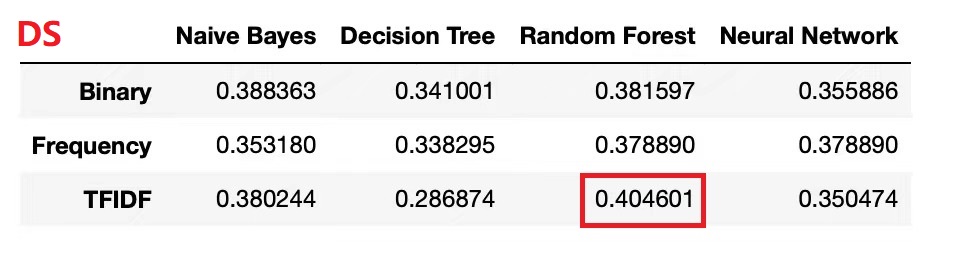
Then the text is used as variables to predict the salaries for four different job types using four algorithms: Naive Bayes, Decision Tree, Random Forest, and Neural Network.

Performance matrix for each job type (data analyst as DA, data scientist as DS, data engineer as DE, business analyst as BA) are shown below. The best performing method and algorithm combination is marked in red.









The average accuracy is between 0.35 to 0.4. Model 2 is the improved model based on model 1.

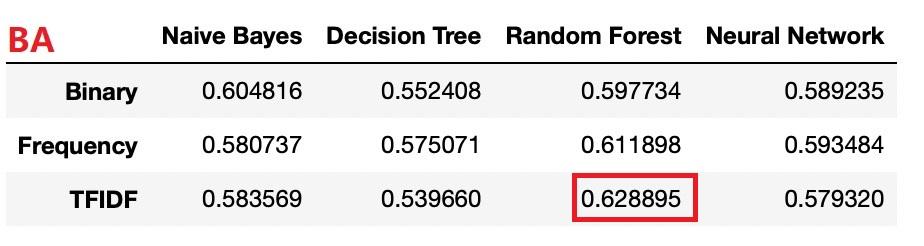
## **Model 2**

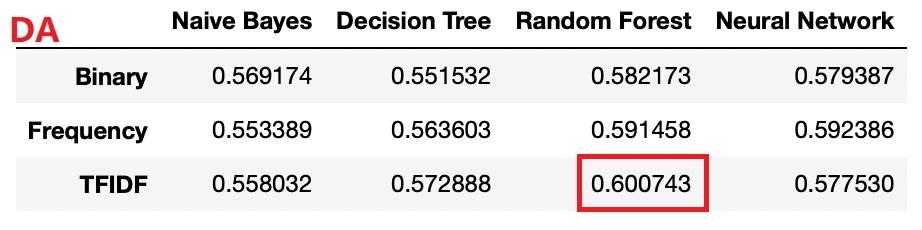
In model 2, top 10 skills and their frequencies extracted from the ‘job description’ column of each job title in the text extraction step are added as new variables.

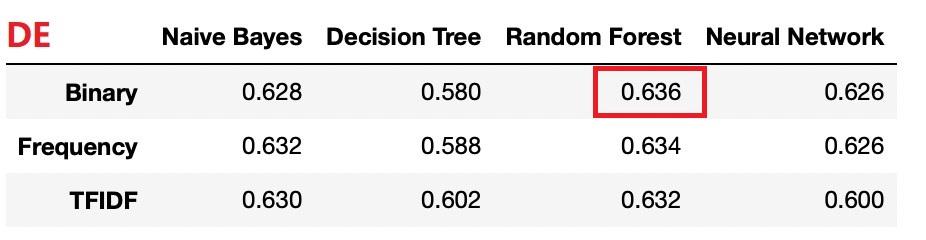
Also ‘location\_state’, ‘rating’ and ‘founded’ columns are added into model 2. The variables in model 2 are shown as below:

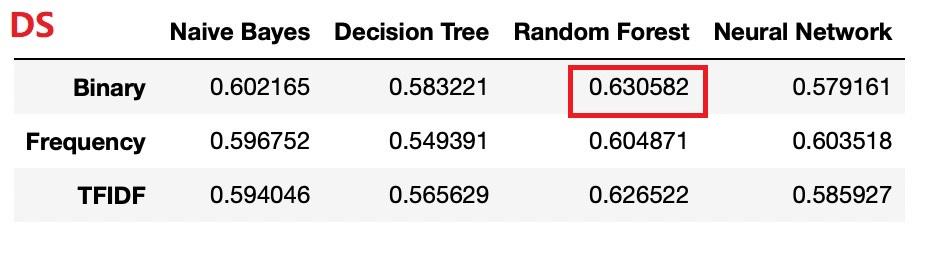
[‘sql’, ‘python’, ‘excel’, ‘tableau’, ‘r’, ‘bi’, ‘aws’, ‘spark’, ‘java’, ‘hadoop’, ‘Location\_state’, ‘Rating’, ‘Founded’]

These variables are used to predict the salaries for four different job types with three vectorization methods (binary, frequency, TF-IDF) and four algorithms (Naive Bayes, Decision Tree, Random Forest, Neural Network). Performance matrix for each job type are shown below:









The model 2 performance has greatly improved and the average accuracy is between 0.55 to 0.6.

The best performing combination is TF-IDF + Random Forest considering all four job types.

In the first model, only the text extracted from the column ‘Job Description’ is used for model prediction. Even if data preprocessing steps are carried out, the model performance is still under expectation.

In the second model, top keywords were added instead of the whole text and domain knowledge was also included, leading to a much better model performance.

# **Conclusion & Future Direction**

## **Conclusion**

After comparing two models, the model combining TF-IDF and Random Forest performs the best prediction. In our model, skills are most essential factors to affect job salary. Other factors, such as location, founding year and rating, have influence on our prediction result. Job seekers who are looking for jobs need to consider the location factor. Given that location is correlated to the cost of living, balancing the cost of living and salary is significant to consider.

## **Future Direction**

On the one hand, employers could pay more attention to location and rating instead of job title. However, the skillset as a factor is more important than the job title. On the other hand, job seekers could achieve a better salary by improving their skills or getting a higher degree. As a result, they could be more competitive and more likely to land high-paying jobs.

# References

https://resources.careerbuilder.com/recruiting-solutions/how-to-build-employee-compensation-programs

https://www.kaggle.com/code/ryanvu/data-science-job-market-eda-and-text-analysis/notebook#Data-Science-Job-Market-EDA-and-Text-Analysis

https://www.kaggle.com/code/nikhilbhathi/100-insights-data-science-jobs-eda/notebook#8.-Skills-Required-by-Companies-for-Each-Job-Title