Individual Assignment

Professor Taha Havakhor

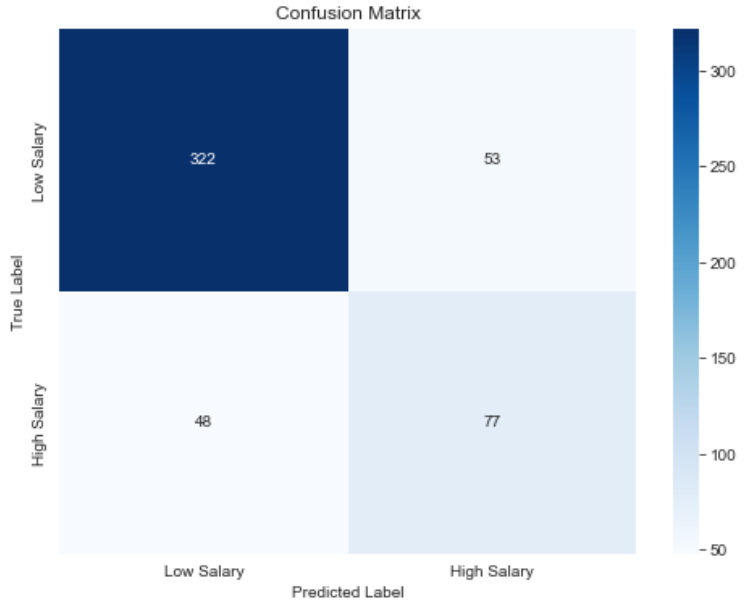
By
Julien Palummo (260946408)

INSY 669 – Text Analytics - Section 076

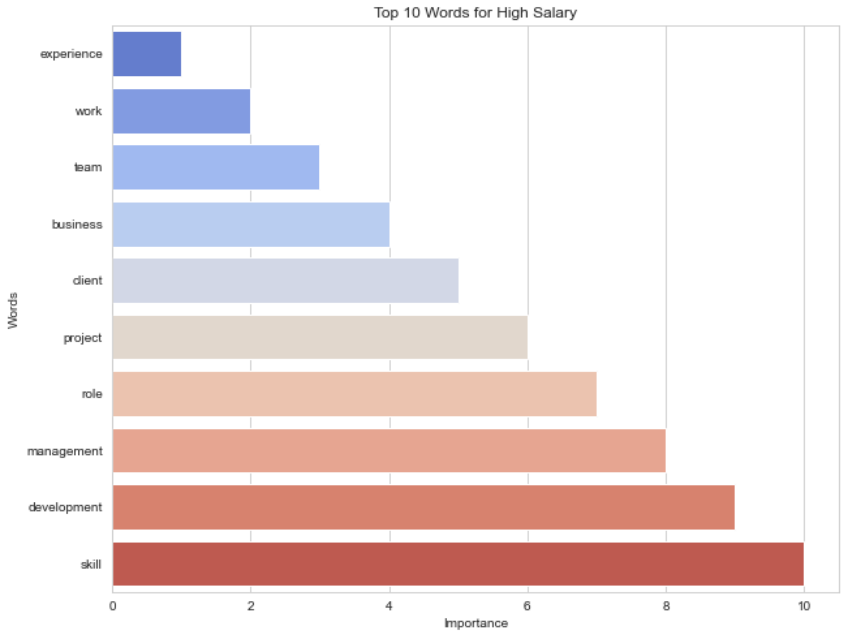
McGill University - Desautels Faculty of Management
February 20th 2024

1. Build a classification model with text (full job description) as the predictor. What is the accuracy of your model? Show the confusion matrix. Also show the top 10 words (excluding stopwords) that are most indicative of (i) high salary, and (ii) low salary.

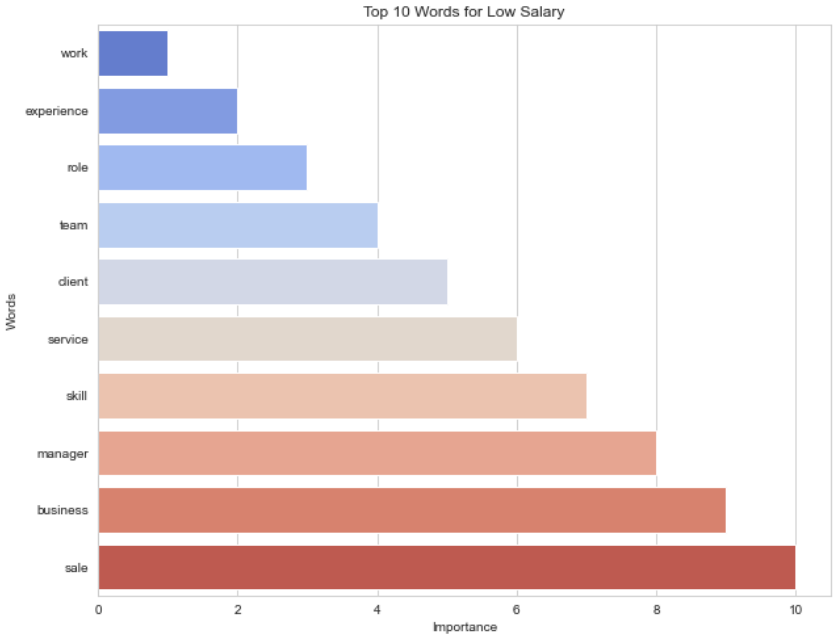
After training the model and running it on the test set, the accuracy is **0.798**. The confusion matrix shows the amount of true label and predicted label. We can see that the model performed particularly well on predicting low salaries (48 "mistakes" on a total of 370 values). However, the performance to predict high salaries was significantly lower (53 "mistakes" on 130 values). Overall, we notice that our sample dataset is unbalanced and skewed with low salaries values which ultimately hindered the results.



Top 10 words that are most indicative of high salary:



Top 10 words that are most indicative of low salary:



Above are the graphs indicating the words most indicative of high and low salaries.

The high salary words in order from most to less indicative:

- Skill
- Development
- Management
- Role
- Project
- Client
- Business
- Team
- Work
- Experience

The low salary words in order from most to less indicative:

- Sale
- Business
- Manager
- Skill
- Service
- Client
- Learn
- Role
- Experience
- Work

Since the words for both high salaries and low salaries are pretty similar, we added a function to remove frequent words from the analysis.
vectorizer = CountVectorizer(stop_words='english', max_df=0.2).
The max_df parameter in the code above is set to 0.2. It means that terms that appear in more than 20% of the documents will be ignored. This is useful for eliminating terms that are so common across documents that they may not carry much meaningful information for analysis or modeling. These might include not only explicit stop words but also terms that are common in a given dataset but not included in the typical stop words list.

Our new accuracy increased to **0.8** and we got the new words below.
High salary: marketing, financial, software, developer, senior, engineer, solution, product, technical, design.
Low salary: hour, maintain, standard, people, engineer, product, staff, account, design, care.

2. If you wanted to increase the accuracy of the model above, how can you accomplish this using the dataset you have?

To enhance the accuracy of the model, we could incorporate bigrams or trigrams alongside unigrams as features could significantly boost its ability to capture contextual nuances in text data, providing a richer understanding of the content. I tried this in the analysis but for some reason the overall accuracy decreased so I decided not to include it but it could still be explored further.
Additionally, we can expand the feature set to include relevant non-textual information, such as job titles, locations, and company sizes, which may offer valuable insights that further refine the model's salary predictions.
Finally, as the sample dataset seems to exhibit class imbalance, we can employ techniques like the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples of the underrepresented class, or adjust the class weight parameters in the model,which could improve its performance by ensuring it learns equally from both high and low salary instances, leading to a more accurate and robust prediction model.