Text Analytics Individual Assignment

Seunghyun Park

260686853

1. The accuracy of the model is 79.2% using a list of the 2000 most frequent words in the overall description.

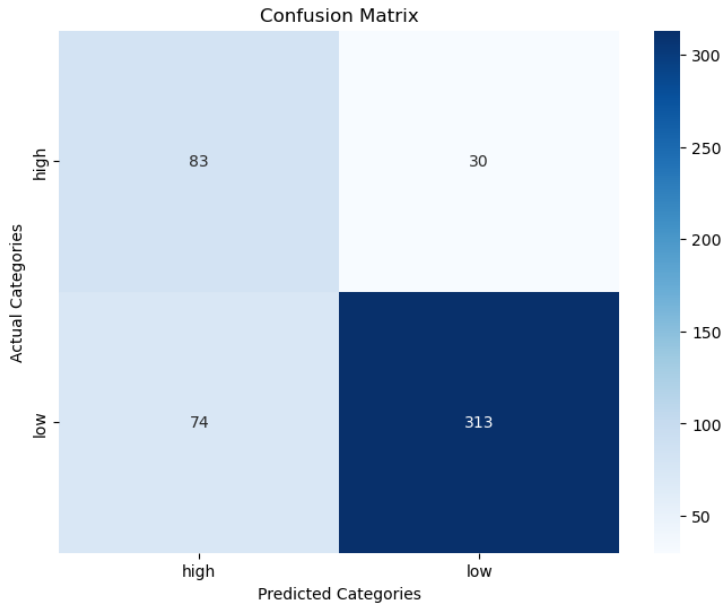

Figure 1: Confusion Matrix

For the 'high' salary category, the precision is 52.87% and the recall is 73.45%.

For the 'low' salary category, the precision is 80.88% and the recall is 91.25%.

Top 10 words that are most indicative of high salary.

Architect: 20.3 times more likely to be associated with high salary.

Allegis: 15.4 times more likely to be associated with high salary.

Collaborative: 8.8 times more likely to be associated with high salary.

Scrum: 8.5 times more likely to be associated with high salary.

Instrumentation: 8.0 times more likely to be associated with high salary.

Transformation: 7.9 times more likely to be associated with high salary.

ACA: 7.4 times more likely to be associated with high salary.

Equity: 6.9 times more likely to be associated with high salary.

Cloud: 6.8 times more likely to be associated with high salary.

Analogue: 5.8 times more likely to be associated with high salary.

Top 10 words that are most indicative of low salary.

Restaurant: 13.5 times more likely to be associated with low salary.

Repair: 11.2 times more likely to be associated with low salary.

We: 10.5 times more likely to be associated with low salary.

Teacher: 9.2 times more likely to be associated with low salary.

NMC: 9.0 times more likely to be associated with low salary.

Assistant: 8.8 times more likely to be associated with low salary.

Hospitality: 8.6 times more likely to be associated with low salary.

Invoice: 8.3 times more likely to be associated with low salary.

School: 8.3 times more likely to be associated with low salary.

Kitchen: 7.9 times more likely to be associated with low salary.

2.

To improve the accuracy of the Model, several effective strategies can be adopted.

Firstly, optimizing the number of features can enhance the accuracy of the model. By selecting the right number of features (list of the top words based on frequency), the model has access to a broader spectrum of information, potentially capturing more nuances related to salary distinctions.

Also, the dataset has an imbalance between the classes, which hampers the model's performance. By addressing this imbalance, the model's performance is likely to improve.

Applying a TF-IDF transformation to the job description can also be a potential solution to increase the model accuracy. TF-IDF emphasizes words that are particularly relevant to a document, thus

helping to highlight the features that are most informative for the classification task.

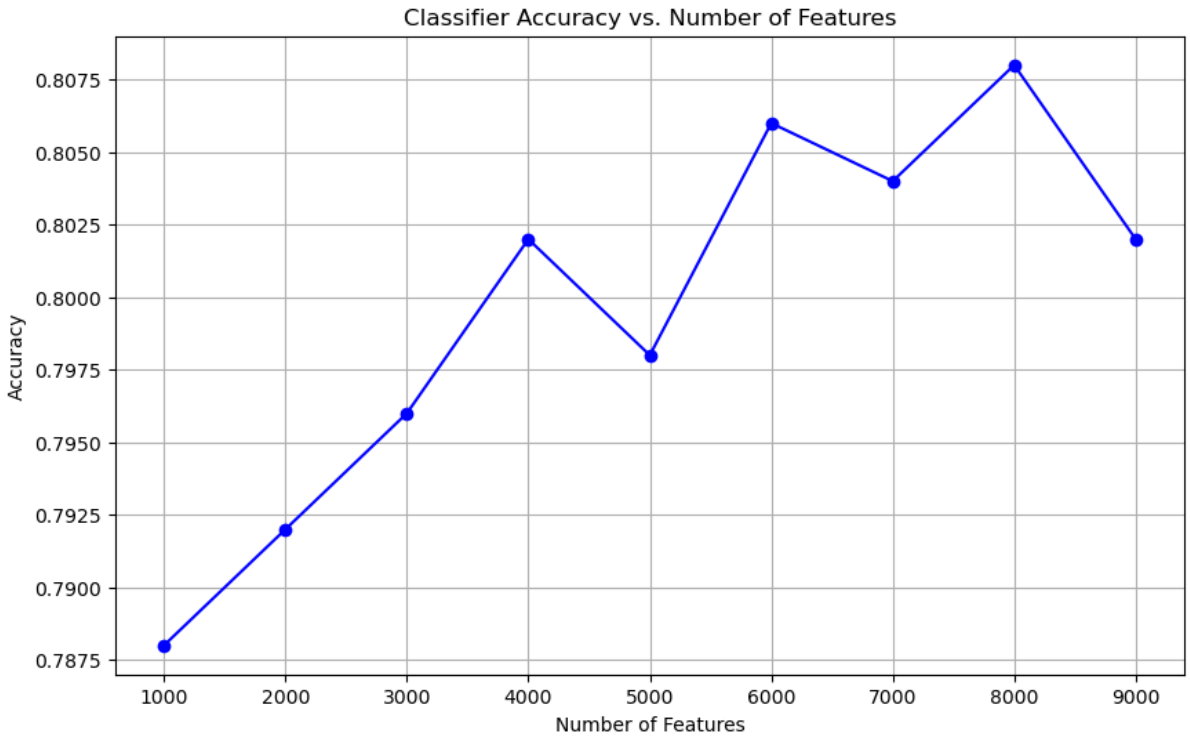

Figure 2: Model Accuracy vs Number of Words for the Model

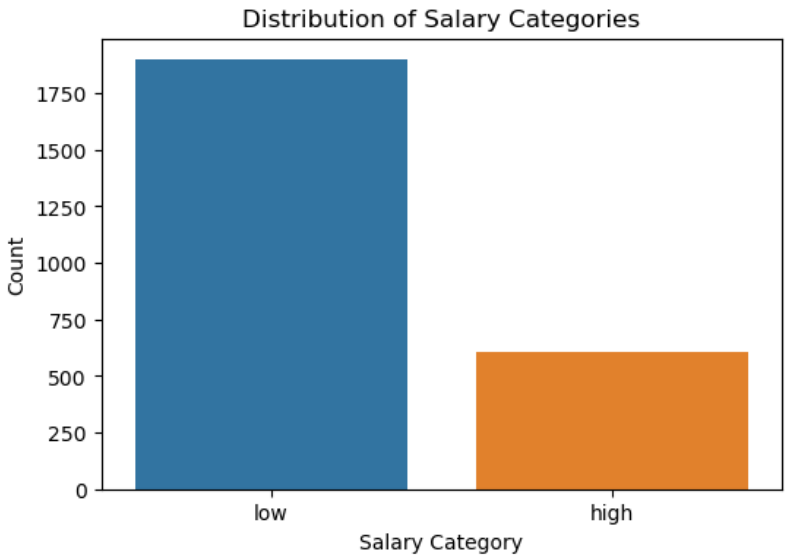

Figure 3: Dataset Class Imbalance