**Approach**

1. Read the data using pandas
2. Did some basic count checks of the data to understand the cardinality of the categorical variables.
3. Objective is to predict the sub functional area hence I used the word2vec pre trained google model for the embedding
4. NLTK wordnet lemmatizer was used to create lemma for all the strings in the data.
5. Use the word2vec embedding to create each word embedding and for multiple word string assuming the words are similar, we create an average 300 dimension word embedding for the strings.
6. Create the embedding for industry, skills and job title
7. As the cardinality of the categorical variables is very high, we can’t use one hot encoding vector. Hence I used the frequency encoding here. Bayesian encoding can be tried as well but I couldn’t try due to less time.
8. Create one more variable for number of skills for each candidate.
9. For missing data, impute the data with -99 as we are using trees so should be fine.
10. Create indicators for variable that have missing values.
11. Also use k-means clustering for the word2vec variables and encode them as cluster numbers
12. Split the data into train and test (All the variables created earlier didn’t use any information from rows hence no issue)
13. Use RF for variable selection (total 908 variables were created and 327 were selected, most were w2v variables)
14. Use grid search for selecting best hyper-parameters for GBM
15. Build a GBM model on the data and test on the test data.

**Could have been tried:**

1. Glove embedding/NN embedding could have been tried (but didn’t have GPU access)
2. Spell check was not done due to lack of time, could have given some more accuracy
3. Bayesian encoding of categorical variables
4. Didn’t tune the parameters/feature selection much due to lack of time.

**Prediction Accuracy:**

Test Accuracy Table

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Actual Values** | | | |  |
| **Predicted** | Network / System Administration | Telecom Network Design / Management | Hardware / Telecom Equipment Design | Embedded, VLSI | % Accuracy |
| Network / System Administration | 847 | 47 | 125 | 35 | 80% |
| Telecom Network Design / Management | 69 | 886 | 232 | 22 | 73% |
| Hardware / Telecom Equipment Design | 152 | 157 | 691 | 84 | 64% |
| Embedded, VLSI | 30 | 15 | 59 | 689 | 87% |