Job title Classification by industry

1- Cleaning the data

- Checked the nun values in the data
- Checked the imbalance in the data, and we will deal with it later
- Use TF-IDF vectorizer to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction

2- Choosing the classifier

- Used Multinomial Naive Bayes classifier
  We do have other alternatives when coping with NLP problems, such as Support Vector Machine (SVM) and neural networks. However, the simple design of Naive Bayes classifiers make them very attractive for such classifiers. Moreover, they have been demonstrated to be fast, reliable and accurate in a number of applications of NLP.

3- Dealing with the imbalance data

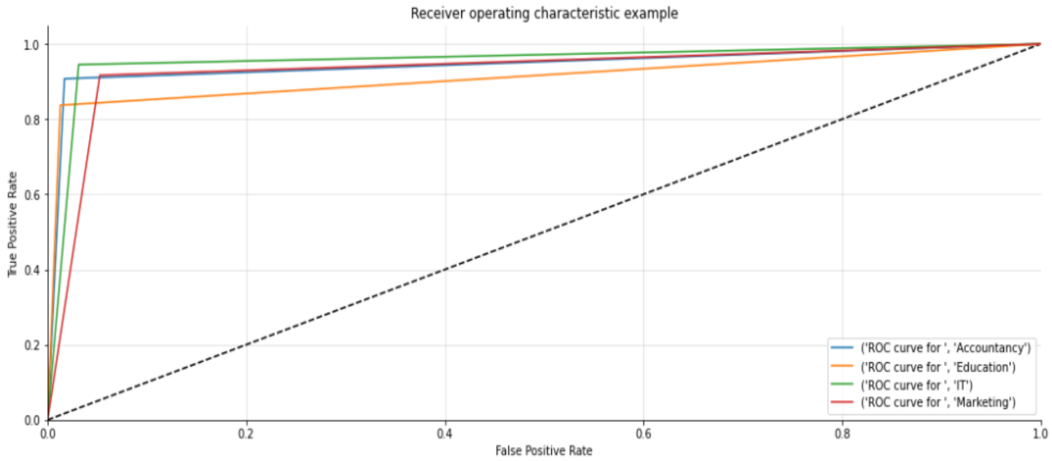
- Over-sampling: SMOTE
  - SMOTE (Synthetic Minority Oversampling Technique) consists of synthesizing elements for the minority class, based on those that already exist. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.
  - Mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances
  - No loss of useful information

4- How can we extend the model to have better performance?

- Collect more data
  - Having more data is always a good idea. It allows the “data to tell for itself,” instead of relying on assumptions and weak correlations. Presence of more data results in better and accurate models.
  - Most of time we don’t have an option to add more data but if possible. This will reduce your pain of working on limited data sets.
- Algorithm Tuning
- Ensemble methods
  - This technique simply combines the result of multiple weak models and produce better results.
- Dimensionality reduction
  - Using PCA

5- Evaluate the model

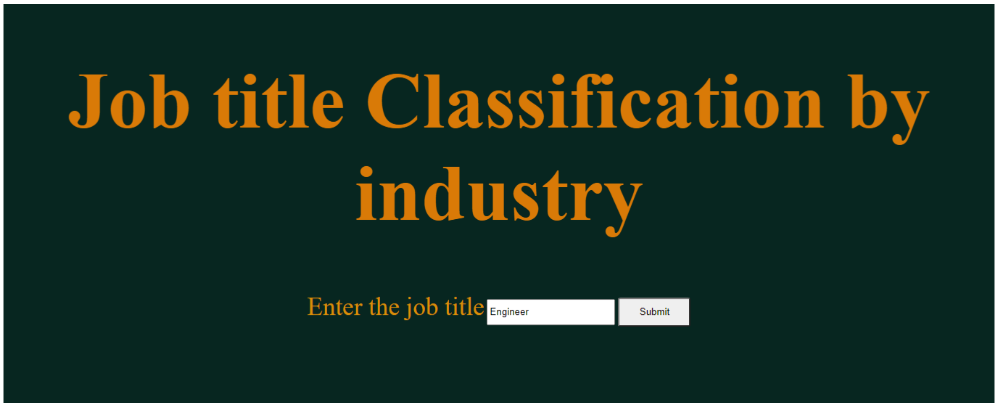
- Checked the performance of the model by the confusion matrix
- Calculated precision, recall, f1 score
- Plot the ROC curve



6- What are the limitations of this methodology or where does this approach fails?

- Syntax error
- Foreign languages
- Ambiguity
  - Ambiguity in NLP refers to sentences and phrases that potentially have two or more possible interpretations.
- Synonyms
  - There are many different words to express the same idea

7- FLASK



Job title Classification by industry

Enter the job title Engineer