Documentation for Resume Categorization Model

1. Data Preprocessing:

Here I use some methods include loading data

- Text cleaning through tokenization
- Lowercasing
- Stopword removal
- Non-alphanumeric removal.

These steps help me to prepare the text data by reducing noise and focusing on relevant words.

2. Feature Extraction:

For feature extraction I used is TF-IDF Vectorization, which measures the originality of a word by comparing the number of times a word appears in a document with the number of documents the word appears in. This helps in accentuating important words and downplaying common words across documents.

3. Model Selection and Training:

I chosen RandomForest Classifier Model. I think for this dataset RandomForest Classifier is more suitable than another model because its robustness to high dimensional data, general high performance and effective management of bias and variance and its ability to provide insights on feature importance. Here I mention some point why I choose it:

- **Robustness**: It can handles high dimensional data well and is less likely to overfit compared to other algorithms.
- **Performance:** It provides high accuracy and handles the balance between bias and variance effectively.
- **Feature Importance:** This classifer offers good insights into which features are most influencing the predictions, which is valuable for text data.

These are the main reason for choosing this classifier.

I use train validation and test ratio is 80% : 10% : 10%

For visualization of precision, recall, and F1-score metrics for each category,

Here is the classification report:

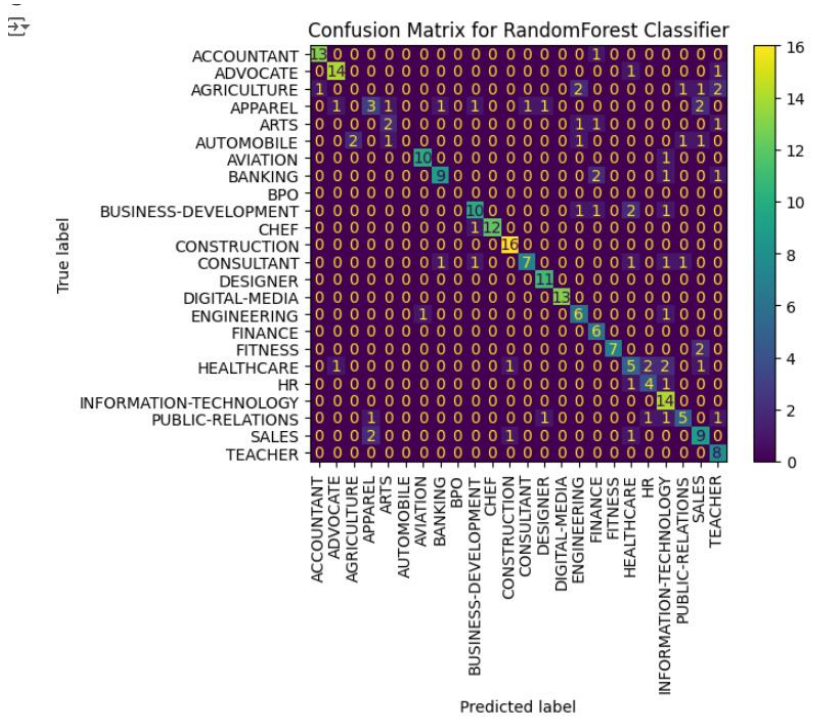
Classification Report:

	precision	recall	f1-score	support
ACCOUNTANT	0.93	0.93	0.93	14
ADVOCATE	0.88	0.88	0.88	16
AGRICULTURE	0.00	0.00	0.00	7
APPAREL	0.50	0.27	0.35	11
ARTS	0.50	0.40	0.44	5
AUTOMOBILE	0.00	0.00	0.00	6
AVIATION	0.91	0.91	0.91	11
BANKING	0.82	0.69	0.75	13
BUSINESS-DEVELOPMENT	0.77	0.67	0.71	15
CHEF	1.00	0.92	0.96	13
CONSTRUCTION	0.89	1.00	0.94	16
CONSULTANT	0.88	0.58	0.70	12
DESIGNER	0.85	1.00	0.92	11
DIGITAL-MEDIA	1.00	1.00	1.00	13
ENGINEERING	0.55	0.75	0.63	8
FINANCE	0.55	1.00	0.71	6
FITNESS	1.00	0.78	0.88	9
HEALTHCARE	0.45	0.42	0.43	12
HR	0.57	0.67	0.62	6
INFORMATION-TECHNOLOGY	0.61	1.00	0.76	14
PUBLIC-RELATIONS	0.62	0.50	0.56	10
SALES	0.56	0.69	0.62	13
TEACHER	0.57	1.00	0.73	8
accuracy			0.74	249
macro avg	0.67	0.70	0.67	249
weighted avg	0.72	0.74	0.72	249

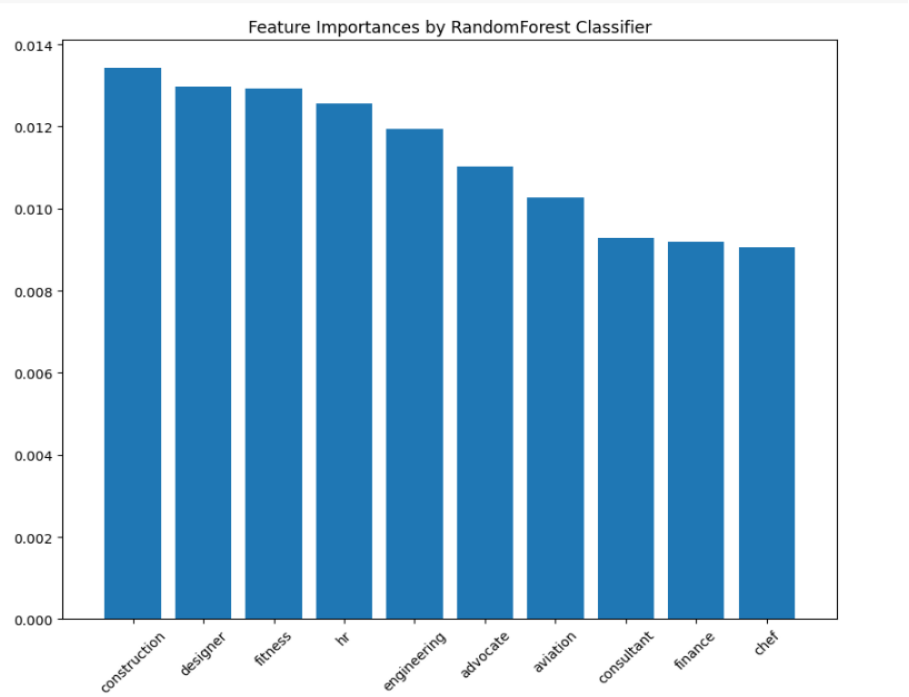
4. Result analysis and Evaluation:

For evaluation here I give confusion matrix, ROC curve and feature importance curve etc.

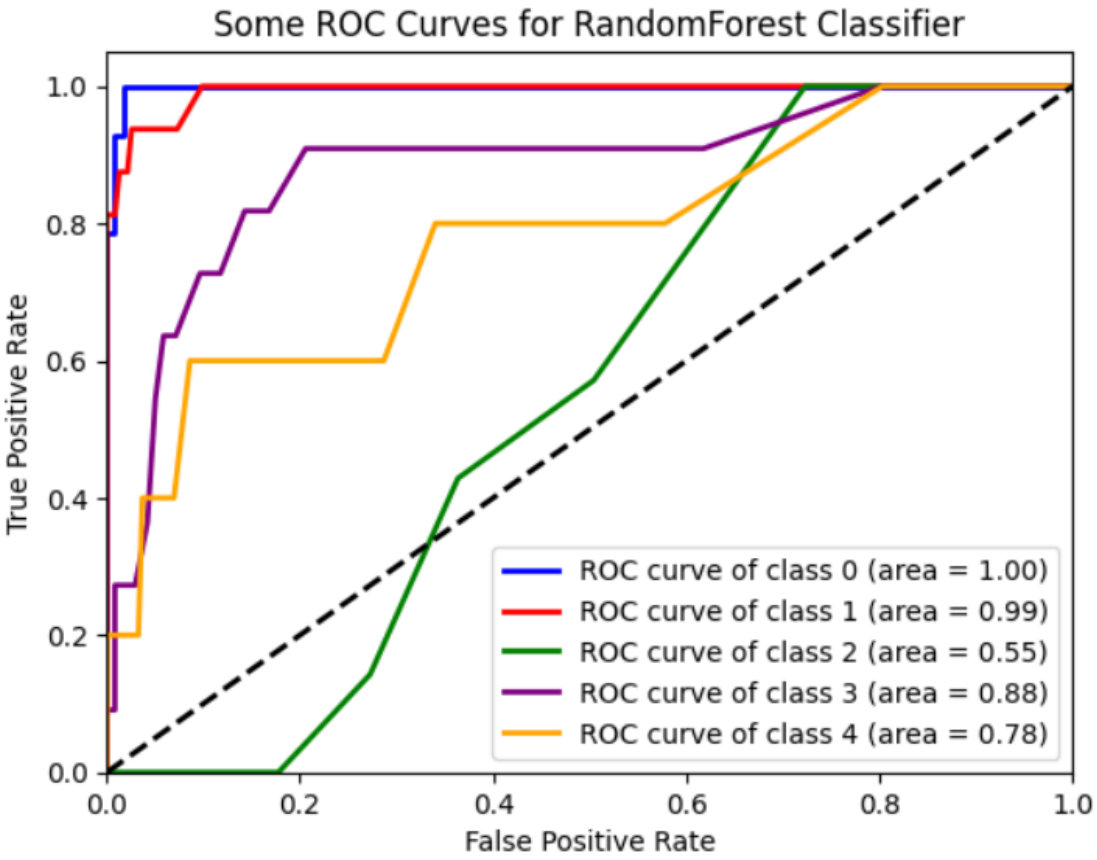
Confusion Matrix:



Feature Importance curve:



ROC curve:



5. Deployment

I uploaded the dataset into my drive (publicly accessable). I do all code in google colab . If you want to run the whole code please go to this link where I put my code:

https://colab.research.google.com/drive/1Bn0UI92yi-KXCJvrw-McXieZeVlBGClc?usp=sharing