**Objective**

To calculate the accuracy of the ML model to predict candidate result.

**Process**

Create a test dataset with the appropriate columns required for the ML model and use the model to predict the results.

Compare these results with the client results for the candidates and calculate the accuracy score.

**Steps for creating test data set**

What do we have?

*Data set consisting of candidate names, and their selection result*

What do we require?

*Test data set containing the following columns:*

* *Candidate name*
* *Resume content in text format and JSON format*
* *Selection result*
* *Respective labelled data*
* *Appropriate job description*
* *Job order and title*

How to acquire?

*Test data set*

1. *Gather the resumes of the candidates for whom results are mentioned in the client results data set.*
2. *Extract the content of these resumes and create a text version and a JSON/ structured version.*
3. *Create additional columns as per requirements of the model.*
4. *Vectorize the data set if necessary.*

*Accuracy score*

1. *Train the model using training data.*
2. *Test the model using test data to obtain the test results.*
3. *Compare the predicted results with the client results for the candidates to obtain the accuracy score.*

**Instruction to run the package**

1. Gather the resumes and store it in local storage.
2. Create the folder/ file structure as directed and mention the specific paths in the main module.
3. Follow the steps mentioned in the doc strings.

**Folder structure**

Root/

│ check\_inconsistencies.py

│ main.py

│ pre\_process\_pipeline.py

│ README.md

│ requirements.txt

│ resume\_selector.py

│ score\_calc.py

├───datasets

│ ├───input

│ │ results\_from\_client.xlsx

│ └───output

│ │ missing\_resumes.py

│ └───predicted

├───files

│ ├───input

│ │ ├───structured\_rejected

│ │ └───structured\_selected

│ └───output

│ ├───rejected

│ └───selected

**Things to keep in mind**

* Candidate amount and order between data set and client results should be equal.
* Empty/ duplicate Job descriptions and structured resumes values for all candidates shouldn’t be present.
* The vectorized version of both training dataset and the test dataset must have the same number of columns.

**Future features**

* Azure storage takes time to download all the files. Using regex to search for the files online directly instead of downloading all 50000 resumes and searching amongst them would reduce time, and space.
* A python version of pdf to JSON converter.