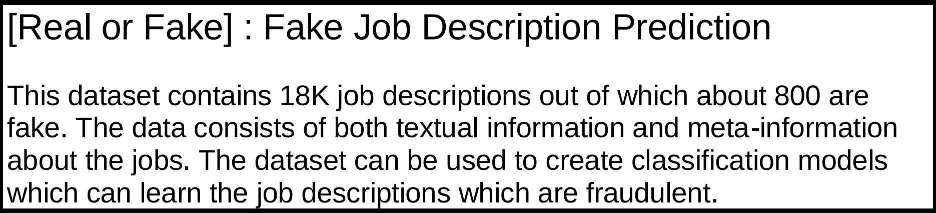
JOB DESCRIPTOR DATASET:

============================

===================

1. CONTENT



[Real or Fake] : Fake Job Description Prediction

This dataset contains 18K job descriptions out of which about 800 are fake. The data consists of both textual information and meta-information about the jobs. The dataset can be used to create classification models which can learn the job descriptions which are fraudulent.

2. ABOUT THIS FILE

- job_id- Unique Job ID
- title-The title of the job ad entry.
- Location-Geographical location of the job ad.
- Department-Corporate department (e.g. sales).
- salary_range-Indicative salary range (e.g. $50,000-$60,000)
- company_profile- A brief company description.
- Description-The details description of the job ad.
- Requirements-Enlisted requirements for the job opening.
- Benefits-Enlisted offered benefits by the employer.
- Telecommuting-True for telecommuting positions.
- has_company_logo-True if company logo is present.

- has_questions-True if screening questions are present.
- employment_type- Full-type, Part-time, Contract, etc.
- required_experience- Executive, Entry level, Intern, etc.
- required_education- Doctorate, Master's Degree, Bachelor, etc.
- industry-Automotive, IT, Health care, Real estate, etc.
- function- Consulting, Engineering, Research, Sales etc.
- fraudulent- target & Classification attribute.

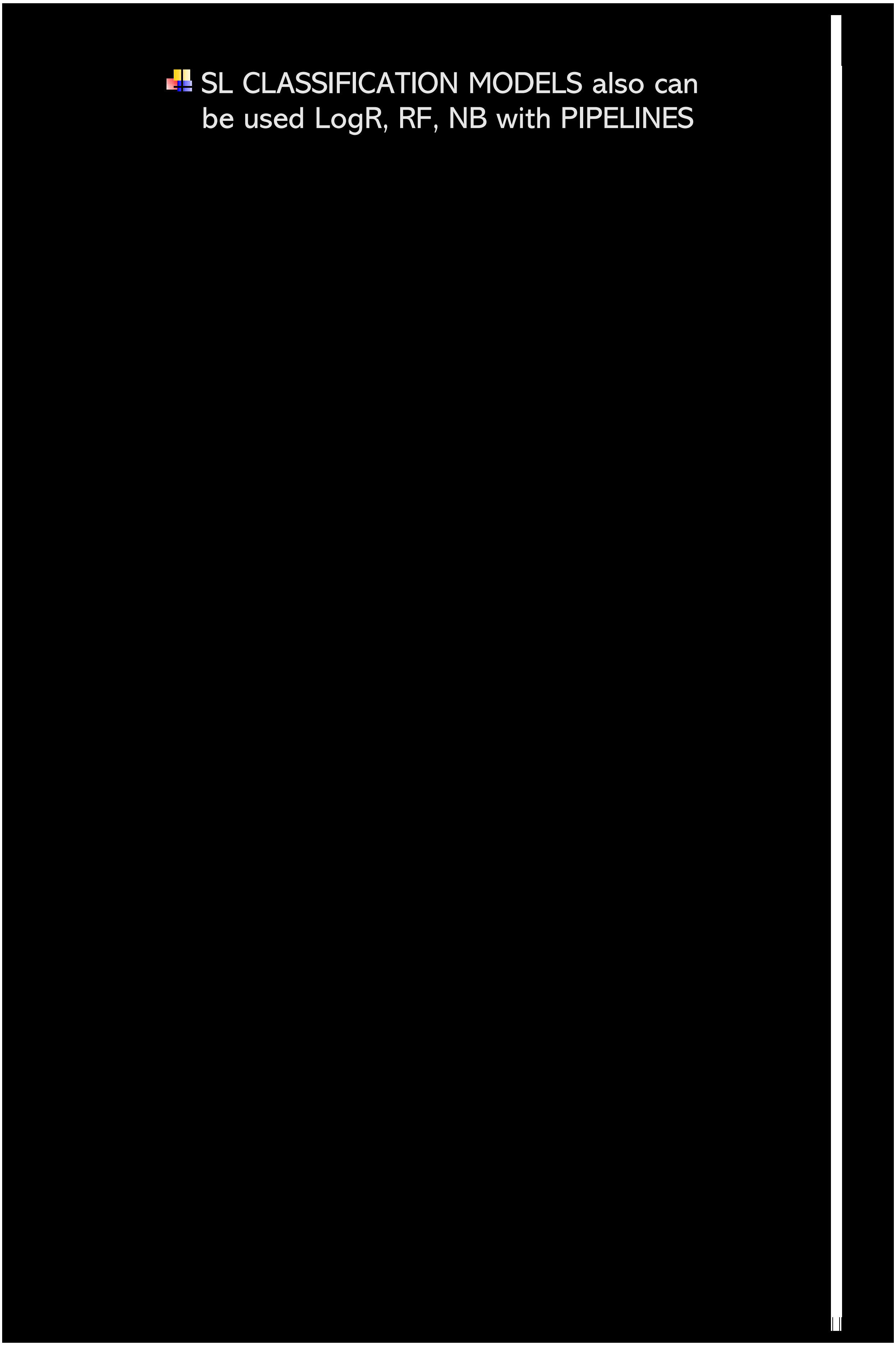
3. PROBLEM STATEMENT

The dataset is very valuable as it can be used to answer the following questions:

1. Create a classification model that uses text data features and meta-features and predict which job description are fraudulent or real.
2. Identify key traits/features (words, entities, phrases) of job descriptions which are fraudulent in nature.
3. Run a contextual embedding model to identify the most similar job descriptions.
4. Perform Exploratory Data Analysis on the dataset to identify interesting insights from this dataset.

4. ROUGH APPROACH

- NLP used for feature extraction, text mining



SL CLASSIFICATION MODELS also can be used LogR, RF, NB with PIPELINES