CS657 October 23, 2024

HW2:Predicting Fake Job Posting

Antara Tewary (G01413546),Ankit Kumar (G01436204)

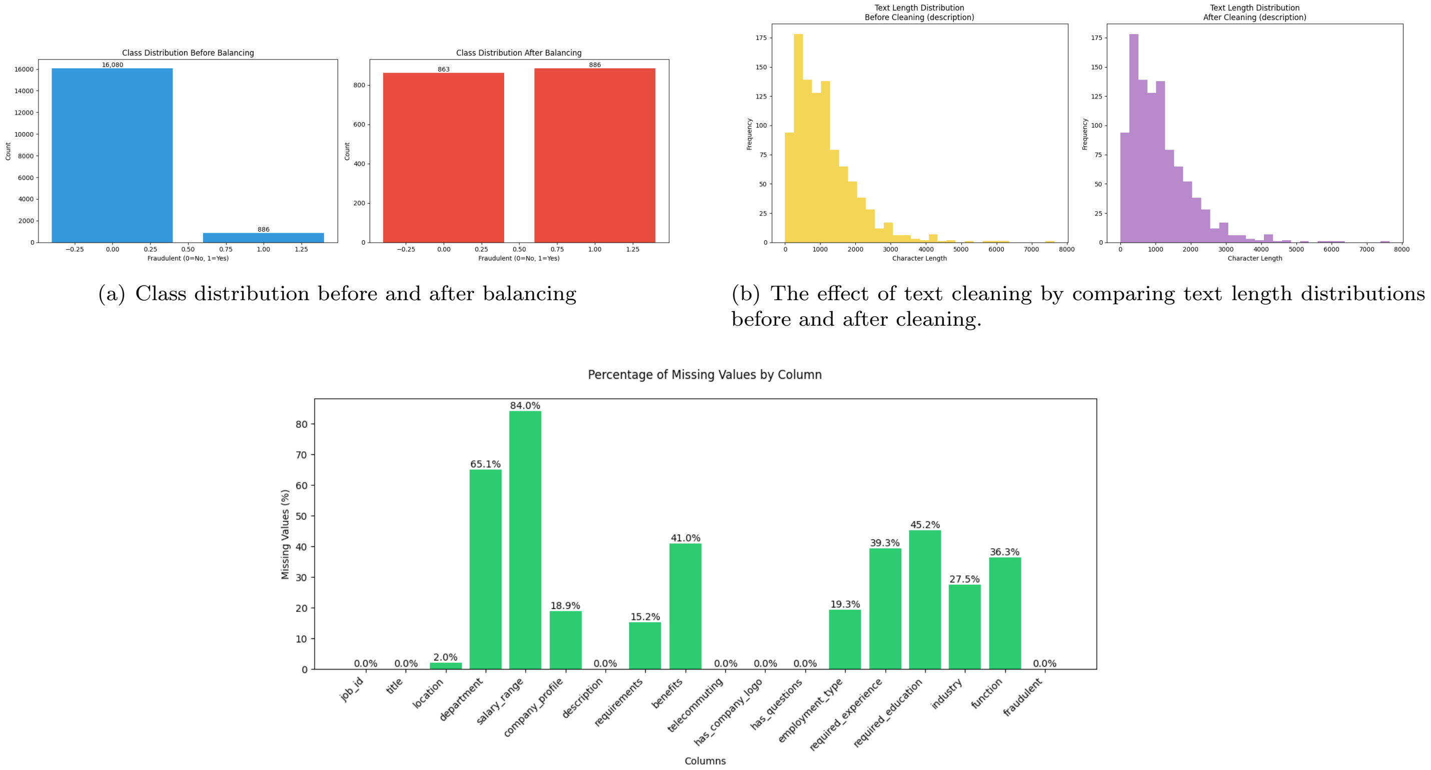
1. Introduction

This assignment aims to create a system that flags suspicious job postings on Indeed.com, addressing the challenge of fraudulent postings that mislead the efforts of job seekers. We are using Fake Job Prediction Dataset, which contains 18000 job descriptions out of which 800 are fraudulent samples. The goal is to create an automated model which prioritizes which posting to review based on the likelihood of it being fraudulent.

2. Data Preparation

- *Label Encoding and Filtering Invalid Data*: The target column fraudulent was cast to integer type to represent legitimate (0) and fake (1) job postings. Any records with invalid values were filtered out.
- *Handling Missing Values*:We identified missing values by calculating percentage of null or NaN values across columns and dropped those with more than 1% null values to reduce noise.
- *Text Cleaning*: The job descriptions were cleaned by removing special characters, converting text to lowercase, and removing stop words to ensure uniformity.
- *Balancing the Dataset*:To address class imbalance, we undersampled the majority class (legitimate postings) to match the minority class (fake postings), ensuring balanced data for model training. We used sampleBy function in PySpark to match the number of legitimate postings to the number of fraudulent postings. Balancing the dataset removes the risk of bias towards majority class during model training.

Here are some statistics of the dataset after cleaning:



(a) Class distribution before and after balancing

(b) The effect of text cleaning by comparing text length distributions before and after cleaning.

(c) % of missing values for each column

Figure 1: Data Preparation Statistics

3. Feature Engineering