Fake Job Posting Prediction

Anshupriya Srivastava

6/16/2020

Domain Background - Employment scams are on the rise. According to CNBC, the number of employment scams doubled in 2018 as compared to 2017. The current market situation has led to high unemployment. Economic stress and the impact of coronavirus have significantly reduced job availability and the loss of jobs for many individuals. A case like this presents an appropriate opportunity for scammers. Many people are falling prey to these scammers using the desperation that is caused by an unprecedented incident. I am a university student, and I have received several such scam emails. The scammers provide users with a very lucrative job opportunity and later ask for money in return. Or they require investment from the job seeker with the promise of a job. This is a dangerous problem that can be addressed through Machine Learning techniques.

Problem Statement - Job postings that are not authentic pose the danger of scamming individuals. Most of these postings demand a certain sum of money. Machine Learning techniques can be used to identify such deceptive listings and protect individuals from falling for such scams.

Datasets and Inputs - The dataset that will be used for this analysis is publicly available on Kaggle (Link - https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction). This data has been collected by the Laboratory of Information & Communication Systems Security, University of Aegean (Link - http://emscad.samos.aegean.gr/). It contains about 18,000 real-life job advertisements. The data is in both textual and meta-information about jobs. The dataset is extremely imbalanced, with only about 4% of the cases being fraudulent.

Solution Statement - Classification algorithms such as Logistic Regression, Decision Trees, K-Nearest Neighbors, etc. can be used to identify fraudulent classes.

Benchmark Model - Logistic Regression will be the benchmark model for this analysis. All other Models will be compared to the output quality of Logistic Regression.

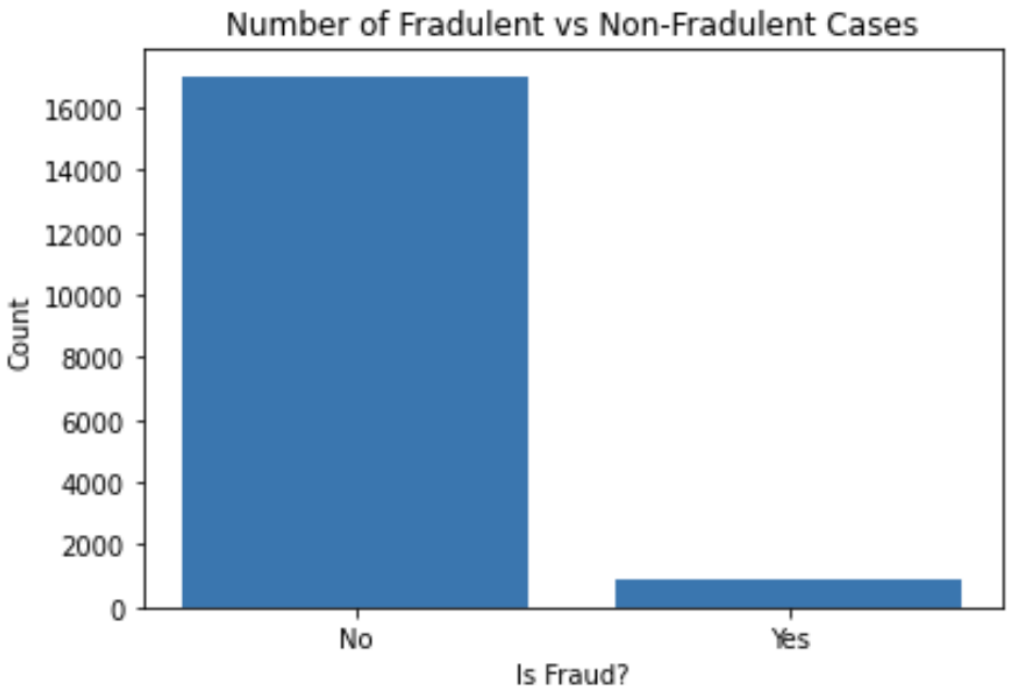
Evaluation Metrics - Since this dataset is meant to identify fraudulent classes, **Recall** is the appropriate metric. This is because the dataset will be heavily imbalanced. The formula for Recall or Sensitivity or True Positive Rate is -

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

This analysis requires the minimization of false negatives. Recall is the model's ability to identify all points of interest in the dataset. This project can be later extended to find an appropriate balance between recall and precision by using F1-score. The F1-score will help identify the best model to minimize both false positives and false negatives. However, that is beyond the scope of the current analysis.

Project Design - Step1: Exploring the dataset and identifying the relevant columns. An appropriate method to deal with missing data also needs to be identified. Also, the relationship between the target variable and the other variables in the dataset will be explored.

Step2: The second step is to balance the two classes by using methods such as over-sampling and under-sampling. This is imperative because Machine learning algorithms tend to favor the class with the largest proportion of observations (known as majority class), which may lead to misleading accuracies. This is particularly problematic when we are interested in the correct classification of a "rare" category (also known as minority class). However, we find high accuracies, which are the product of the correct classification of the majority class (i.e., are the reflection of the underlying class distribution).



Step3: Comparing various models and selecting the one that performs the best based on the selected Metric.

Step4: Creating a web app for this analysis. The web app will be able to produce results on the authenticity of a job posting.