Analysis and Predictions on Fraudulent Job Postings

University of San Diego

Applied Data Science – ADS502

Team 4

Mahalakshmi Jayapal

Yi Wang

Jonathan Yoon

**Introduction**

Federal Bureau of Investigation (FBI) reports that 16,012 people reported being victims of employment scams in 2020, with losses totaling more than $59 million (2021). This number significantly increased in 2021, as the pandemic from COVID-19 continued world-wide. As the negative impacts due to job posting scammers becomes more devastating, investigations on these postings and finding their patterns are considered crucial and important to prevent additional negative consequences caused by the scams. More studies and investigations would be needed to create strategies to prevent further damages. Our aim in this project is to find models that could define the statistical relationships between some attributes of normal and fraudulent job postings and to determine which of the found models shows the best performance in predicting fraudulent job posts.

**Methodology**

The main goal of this project is to find statistical relationships of several attributes of job posts and to obtain predictability of fraudulent job posts based on these relationships.

**Dataset**

For the dataset, fake\_job\_posting.csv was obtained from Kaggle (<https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction>) and it has 17880 samples with 18 columns. Excluding the first column because it represents the job\_id, there are 16 independent attributes and 1 target class variable, fraudulent (Bansal, 2020).

**Data Preparation and Processing**

The downloaded data set was imported to Python through Jupyter Notebook under Anaconda platform. After the import, the dataset was investigated, and significant attributes were selected for building the models. In our case, location, employment type, required experience, required education, and whether there is a company logo in the post or not were the selected attributes to build models to predict fraudulent job posts.

The selected attributes were processed so the null values are filled with valid values. First, the attribute location was processed to exclude non-US location since we are only interested in the job posts from the United States. After exclusion of non-US posts, only the states were extracted to simplify the dataset.

For the rest of the categorical attributes, the blank values are filled with the values that are less meaningful and less significant. For employment type, the blank values were filled with the value ‘Other’ based on the overall analysis of the attribute in the data.

Chart, bar chart

Description automatically generated with medium confidence

For required experience blank values were filled with the value ‘Not Applicable’ based on its analysis because ‘.

Chart, bar chart

Description automatically generated

Finally, the blank values of required education attribute were filled with the value ‘Unspecified’.

Chart, bar chart

Description automatically generated

There are no blank values for has\_company\_logo attribute so no processing was needed. Eventually, the dataset we worked on has 6181 rows with 6 columns with 5 independent attributes and the target class variable, fraudulent.

The selected attributes were then encoded into numerical values to ease the model construction process.

Table

Description automatically generated

Table

Description automatically generated

**Data Set Split into Training and Test Sets**

There is only single data set provided so the data needs to be split into two sets with a training set to establish the model and a test set to validate the model. 75% of the processed data set was randomly assigned as the training dataset, and the rest 25% was assigned as the test set.

Class imbalance problem was noted in the dataset since fraudulent job posts account for only a small portion of the dataset. In fact, there are only 402 fraudulent posts in the processed dataset among the 6181 dataset, and this only constitutes 6.5% of the entire dataset. Comparably, 6.8% of the training data set is fraudulent so rebalancing is omitted because the proportion of fraudulent data in the training data set is similar to that in the original data set.

After data processing and preparation of training and test sets, we were ready to construct models using five different data mining techniques: Naïve Bayes, Decision Tree, K-nearest Neighbors, Random Forest, and Logistic Regression. All the processing and preparation tasks were performed in Python processed in Jupyter Notebook under Anaconda platform.

**Model Construction and Evaluation**

For each type of model, a model was generated using the training data set. Then, the independent variables from the test set were applied to the constructed model for prediction. The predicted result from the model was then compared with the fraudulent class variable from the original test set using contingency table. To validate the model, three different measures from the contingency table were computed: Accuracy, Sensitivity, Specificity.

Accuracy is the overall measure of the proportion of correct classifications made by the model across all cells in the contingency table. Sensitivity, or recall, is the measure of model’s ability to classify a record positively, in our case non-fraudulent job post, and specificity is the measure of model’s ability to classify a record negatively (Larose, 2019). Accuracy, sensitivity, and specificity of the five models were then compared with one another, and the model that would work the best for prediction was determined.

**Results and Discussion**

****

**Conclusion**

**References**

Bansal, S. (2020, February 29). Real / fake job posting prediction. Real / Fake Job Posting Prediction. Retrieved from <https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction>.

FBI. (2021, April 21). *FBI warns Cyber Criminals are using fake job listings to target applicants' personally identifiable information*. FBI. <https://www.fbi.gov/contact-us/field-offices/elpaso/news/press-releases/fbi-warns-cyber-criminals-are-using-fake-job-listings-to-target-applicants-personally-identifiable-information>.

Larose, C. D., & Larose, D. T. (2019). Data Science Using Python and R. Wiley.