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 worldwide. As the negative impacts due to job posting scammers grow every day, 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. This project aims 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 job\_id column, there are 16 independent attributes and one target class variable, fraudulent (Bansal, 2020).

**Data Preparation and Processing**

The downloaded data set was imported to Python through Jupyter Notebook under the 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 to fulfill the null values with valid values. First, the attribute location was processed to exclude non-US locations since we are only interested in the job posts from the United States. After excluding 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 values that are less meaningful and less significant. For employment type, blank values were filled with ‘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 ‘Not Applicable’ based on its analysis.

Chart, bar chart

Description automatically generated

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

Chart, bar chart

Description automatically generated

There are no blank values for the has\_company\_logo attribute, so no processing was needed. Moreover, when analyzing this attribute, an intriguing feature that not having a company logo significantly increases the probability of being a fake job post. A hypothesis that this attribute contributes to whether the job post is fake or not can be generated.

Chart, bar chart

Description automatically generated

Eventually, the dataset we worked on had 6181 rows with six columns with five 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 a 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.

The data set has a class imbalance problem since fraudulent job posts account for only a small portion of the dataset. There are only 402 fraudulent posts in the processed dataset among the 6181 datasets, 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, only 0.3% different from, 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. The processing and preparation tasks were performed in Python processed in the Jupyter Notebook under the Anaconda platform.

**Model Construction and Evaluation**

A model was generated using the training data set for each type of model. 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 a contingency table. Eight evaluation measurements were computed from the contingency table to validate the model with these measurements. The evaluation measurements in consideration are accuracy, error rate, sensitivity, specificity, precision, F1, F2, and F0.5scores.

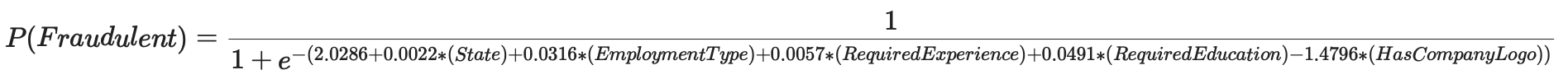
Accuracy is the overall measure of the proportion of correct classifications made by the model across all cells in the contingency table. In contrast, error rate is measures the proportion of incorrect classifications, and it is computed by subtracting accuracy value from 1 (Larose, 2019).

Sensitivity, or recall, is the measure of model’s ability to classify a record positively, in our case, non-fraudulent job posts, and specificity is the measure of model’s ability to classify a record negatively (Larose, 2019). The accuracy, error rate, sensitivity, and specificity of the five models were then compared, and the model that would work the best for prediction was determined. In addition to the three measurements, precision and F1, F2, and F0.5scores were also computed to assist in comparison and determination of the best predictive model.

Precision is the metric addressing what proportion of the records classified by the model as positive are true positives (Larose, 2019). Precision is often combined with sensitivity, or recall, measurement to compute F1, F2, and F0.5 scores. F1 score is the harmonic mean of precision and recall, where they are weighted equally. F2 score weights recall twice as high as precision, while F0.5 weights recall half as much as precision (Larose, 2019).

**Results and Discussion**

Five data mining algorithms were selected to construct models using training set: Naïve Bayes, decision tree, k-nearest neighbors, random forest, and logistic regression. Particularly, logistic regression model generation the equation as the following:



The has\_company\_logo attribute has the coefficient with highest absolute value (-1.4796) and the *p*-value less than 0.001. This indicates that the attribute has a statistically significant relationship with being a fraudulent job post, and this is not surprising because a significant relationship between this attribute and the target class variable was hypothesized, as seen in Figure 4. In addition, all other attributes have *p*-values greater than 0.1, which indicates that the attributes are statistically insignificant to the target variable so they should not be retained in the model (Larose, 2019).

After construction of model using the training data set and predicting the results from the attributes of the test set, the predicted result was compared with the values of the test set using a contingency table and computing the evaluation measurements. Table 1 summarizes the computation results.

**Table 1**

***Evaluation of the Models for Fraudulent Job Postings***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Evaluation Measurements** | **Naïve Bayes** | **Decision Tree** | **K-Nearest Neighbors** | **Random Forest** | **Logistic Regression** |
| **Accuracy** | 0.9444 | 0.9347 | 0.9185 | 0.9360 | 0.9444 |
| **Error Rate** | 0.0556 | 0.0653 | 0.0815 | 0.0640 | 0.0556 |
| **Sensitivity (Recall)** | 1.0000 | 0.9870 | 0.9623 | 0.9897 | 1.0000 |
| **Specificity** | 0.0000 | 0.0349 | 0.1744 | 0.0232 | 0.0000 |
| **Precision** | 0.9444 | 0.9456 | 0.9519 | 0.9451 | 0.9444 |
| **F1** | 0.9714 | 0.9659 | 0.9571 | 0.9669 | 0.9714 |
| **F2** | 0.9884 | 0.9784 | 0.9602 | 0.9804 | 0.9884 |
| **F0.5** | 0.9550 | 0.9536 | 0.9540 | 0.9537 | 0.9550 |

The result shows that Naïve Bayes and logistic regression models have the highest accuracy and K-nearest neighbors model has the lowest accuracy. However, high accuracy does not always indicate that the model is optimal because high accuracy or low error rate may lead to model overfitting, where the model was built upon the relationships in the training data set too much that the true nature of the statistical relationship is not reflected in the model. One of the results of model overfitting is that the target class prediction is mostly incorrect when test data set is applied. Model overfitting is mainly caused by limited training size and high model complexity (Tan et al., 2019).

**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.

Tan, P., Kumar, V., Karpatne, A., &amp; Steinbach, M. (2019). Introduction to data mining (Second). Pearson Education.