A REPORT

ON

Fake Job Posting Prediction

BY
AMAN VOHRA
777001503
av8920@g.rit.edu
Prepared in Fulfillment of
Project Course: DSCI -633

AT

Rochester Institute of Technology,
Rochester, NY,
United States of America.

(August 2021 -December 2021)

Introduction

The first step of people finding jobs is looking for job postings, either online or offline. With the emergence of the Internet and several platforms (e.g. Scouted, Jora, Indeed), job seekers have more access to free, available job postings. However, this convenience meanwhile increases the risk of encountering fraudulent postings, harming privacy and security.

Everyday thousands of postings ,are made on online forums , for which prospective candidates submit their applications.Unfortunately, some of these job postings are fake and used in order to illegally misuse candidate personal information without their knowledge.In this project an effort has been made to deal with this problem by training and evaluating the performance of machine learning models to identify the fake job postings.

ACKNOWLEDGEMENTS.

I would sincerely like to thank Prof..Nidhi Rastogi for assigning me this design project to apply my knowledge of programming for Machine Learning, and also for her constant supervision and mentorship. This allowed me to gain a significant amount of knowledge in the topic.

Lastly, I would like to thank the TA, Rigved, for his input and constructive advice, which allowed me to improve and fine-tune my final project.

CONTENTS

I. Development of Question

.

Employment scams have been on the rise in recent years. According to a study, the number of employment scams doubled in 2018 as compared to 2017. Due to high unemployment and the impact of COVID-19, job availability has significantly reduced.Economic stress and the coronavirus's impact have significantly reduced job availability and job loss 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. Most scammers do this to get personal information from the person they are scamming. Personal information can contain addresses, bank account details, social security numbers, etc. I myself, being a university student, have received several such scam emails. The scammers provide users with a very lucrative job opportunity and in turn get hold of the users' important personal data. This is a dangerous problem that can be addressed through Machine Learning techniques.

Machine Learning has many practical applications that can benefit us in solving practical problems. One of these is to efficiently help reduce fraud.Using predictive modeling we can try to tackle this problem by defining algorithms to try and accurately identify these fake postings.
Therefore our motivation is to create a classifier to identify if a job posting is fraudulent.

II. Data Research

The problem at hand is that of identifying a fake job posting. Identifying a fake job posting from a group of legitimate job postings is a good problem space for utilizing Machine Learning to perform binary classification.

The dataset that has been used for this analysis is publically available on kaggle. This data has been collected by the Laboratory of Information and Communication Systems Security, University of Aegaean. It contains about 17,800 real life job advertisements. The data has both textual and meta- information about jobs. This data set consists of 18 features including the target class called 'Fraudulent'. The target class consists of binary values:

0 ⇒Not a fake Job Posting

1 ⇒A fake job posting

Just like an actual real-world scenario, the job postings dataset is extremely unbalanced. In a total of 17,800 postings, only about 800 are classified as fraudulent.

#	Variable	Datatype	Description
1	job_id	int	Identification number given to each job posting
2	title	text	A name that describes the position or job
3	location	text	Information about where the job is located
4	department	text	Information about the department this job is offered by
5	salary_range	text	Expected salary range
6	company_profile	text	Information about the company
7	description	text	A brief description about the position offered
8	requirements	text	Pre-requisites to qualify for the job
9	benefits	text	Benefits provided by the job
10	telecommuting	boolean	Is work from home or remote work allowed
11	has_company_logo	boolean	Does the job posting have a company logo
12	has_questions	boolean	Does the job posting have any questions
13	employment_type	text	5 categories – Full-time, part-time, contract, temporary and other
14	required_experience	text	Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable
15	required_education	text	Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational
16	Industry	text	The industry the job posting is relevant to
17	Function	text	The umbrella term to determining a job's functionality
18	Fraudulent	boolean	The target variable → 0: Real, 1: Fake

III. Literature Review

Models

The following models were used for the purpose of this project-:

1. Naive Bayes Classification

Naive Bayes Classifier Algorithm is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of a feature. Bayes theorem calculates probability P(c|x) where c is the class of the possible outcomes and x is the given instance which has to be classified, representing some certain features.

$P(c|x) = P(x|c) * P(c) / P(x)$

Popular uses of naive Bayes classifiers include spam filters, text analysis and medical diagnosis.

2. K - Nearest Neighbor

The k-Nearest Neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. It is the most accepted classification method due to its ease and practical efficiency. In a more complicated approach, k-NN classification, finds a group of k objects in the training set that are nearest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood.

3. Passive Aggressive Classifier

Passive-Aggressive algorithms are generally used for large-scale learning.This is very useful in situations where there is a huge amount of data and it is computationally infeasible to train the entire dataset because of the sheer size of the data.

4. Random Forest Classification

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

5. XGBoost Classification

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

Problem Choice

In the current scenario with the lack of jobs and high rate of unemployment, the issue with the fraudulent job postings is one that needs to be addressed. Therefore, implying to the importance of our problem- Creating a classifier to identify if a job posting is fraudulent.

IV. Analysis Strategy

Since our model is supposed to classify false negatives, recall was the most important metric.We also looked at Accuracy,precision, F-1 Score

The following steps were followed after collecting the data-:

1. **Data Cleaning.**
2. **Data Exploration.**
3. **Model Building**
4. **Evaluation by means of appropriate metrics.**
5. **Model Comparison.**

The data was found to be highly imbalanced and the target class label was the minority class. A combination of Over-Sampling of minority class and Under-Sampling of majority class , **SmoteTomek** was used to resample the data and the model was trained again.

V. Analysis Code

Data Cleaning

An observational study of the data revealed that there was a significant amount of missing data points and noise in the dataset.

For all features with missing values, to add substantial textual data, empty strings were added to the features.For textual data, Term Frequency-Inverse Document Frequency(TF-IDF) processing was used.

We ended up dropping the department and industry columns.

Data Exploration

We treated the null values , dropped duplicate data points, dropped features that were converted. The numerical and categorical variables were treated individually while treating the null values.

Model Building

The following models were built with clean data.

1. Naive Bayes Classifier
2. KNN Classifier
3. Passive Aggressive Classifier
4. Random Forest Classifier

Model Comparison

	Naive Bayes	PAC	KNN	Random Forest
Accuracy	0.98	0.99	0.98	1.00
Precision	0.89	0.88	0.82	1.00
Recall	0.58	0.76	0.80	0.99

VI. Conclusion

Due to the data being highly imbalanced, the models tend to overfit the data. This is a major issue when dealing with class imbalance. That being said, models built with Passive Aggressive and KNN classifier gave exceptional results on the test dataset. Moreover, using SmoteTomek and retraining the Random Forest Classifier, gave us the recall of 0.99 and precision of 1.0.

The best model therefore, is the **Random Forest Classifier.**