




IDENTIFY FAKE OR REAL JOB POSTINGS



Raha Soleymanzadeh
Supervisor: Dr. Ceni Babaoglu
Ryerson university, Summer 2020

Contents

Introduction:.....	2
Related works:	2
Approach.....	4
Explore the dataset.....	5
Data Analysis:.....	8
Resampling Techniques	10
Algorithms:	10
Model and discussion:.....	11
Resampling:	11
Feature Selection:.....	11
Category Encoders	12
Prediction and result	13
Conclusion	15
References.....	16

Introduction:

Online job postings are nowadays more and more frequent. There are plenty of job ads on the internet, including reputed work advertising websites, which never seem to be false. Often these job posts will have a website as well, and they will have a recruitment process that is like other companies in the industry. (1) A lot of candidates slip into their trap after selection. They often lose a lot of money because the so-called recruiters ask for money and the bank details. It would be helpful to recognize whether a work advertisement posted on the internet is real or false, but it is very hard to do it manually. So, applying machine learning to train a model for fake job classification would help to identify fake jobs accurately and it can be trained on the previews real and fake job postings.

In this study, we present our work to tackle the problem of fake job advertisements using Machine learning models performed on the [Employment Scam Aegean Dataset \(EMSCAD\)](#) which is provided publicly by the University of the Aegean Laboratory of Information & Communication Systems Security.

In machine learning, an individual classifier is not always in a position to provide the highest possible accuracy. Multiple classifiers are thus used to achieve as much accuracy as possible.

Section II covers the related work. Section III talks about the experimental approach of study, which contains Data description, Data Preparation, Data Analysis, Resampling Techniques, and Algorithms. Section IV describes the Modelling and testing. The last section discusses important findings in our research (Section V) and concludes the paper with the avenue of future work.

Related works:

In this section, we briefly review the related works and our focus is the detection of fake job postings. There are many similar studies related to fake job postings such as fake news detection, spam detection, rumor detection, and fake transaction detection.

[This literature](#) (2) is a study on the imbalanced dataset to predict the chance of fraudulent transactions and used supervised machine learning models. At this study, sensitivity, precision, and time used as a deciding parameter, but the accuracy were not used because it isn't sensitive to imbalanced data and doesn't provide a conclusive answer. The models used in this study was: kNN, Naive Bayes, Decision Tree, Logistic Regression, and Random Forest. In conclusion, the

sensitivity of KNN was higher but because the taken time was very large for KNN, the Decision tree was selected as a best-suited model.

The aim [of this study](#) (3) is to increase the efficiency and accuracy of the process of detecting the fraud transactions. In this article, the classification, regression, and feature selection were used to suggest the model for improving the detection of fraud transactions. The study proposes the use of artificial intelligence, geolocation, and data mining in the detection of fraud methods to reduce the weaknesses of used methods.

[This study](#) (4) is about analyzing a model to confirm real news collected from Twitter. The model which is based on deep learning is getting the idea from supervised models. In this study, all the information did not use for analyzing the dataset. The classification models which were used for this study were Bayesian Model, Logistic Regression & also Support Vector Machine, two most famous deep learning methods RNN Recurrent Neural Network, and Long Short-Term Memory were also used. What stands out from this study is that even the basic models, may find the proper result. The SVM Shows the highest accuracy with the TFIDF feature among all the classifiers were used.

[This study](#) (5) also used three methods of classification: Naive Bayes, Neural network, and Support Vector Machine (SVM). The result of the three methods was used to calculate and compare precision, recall, F-Measure, and accuracy. What stands out is that to recognize the fake news could be correctly identified by Machine learning methods.

[This paper](#) (6) is about measuring the efficiency and training speed of eight known machine learning algorithms (namely regression, support vector classification, multi-layer perceptron, Gaussian and multinomial naive Bayes, random forests, decision trees, and convolutional neural networks) against three public datasets namely” Liar, liar pants on fire: A new benchmark dataset for fake news detection”, “The signal media one-million news articles dataset” and ”Getting real about fake news”. As a result, the hundred-dimensional space is enough feature to get the required text feature and get high detection accuracy.

[The study](#) (7) provides an additional feature to solve the limitation of a deep learning model based on NLP. These features are source domains of the article, author names, etc. And the study was compared the result of deep learning models based on FNN and LSTM were built in combination

with different word vector representations with and without using these features. As a result, the performance of all models was better when combined with data mining sections.

[In this study](#) (8) , the news of Facebook was analyzed for recognition of fake or real information. The LSTM model was applied, and the ANN used for classification. The dataset of this study is publicly available on Kaggle and the used method had 93% accuracy.

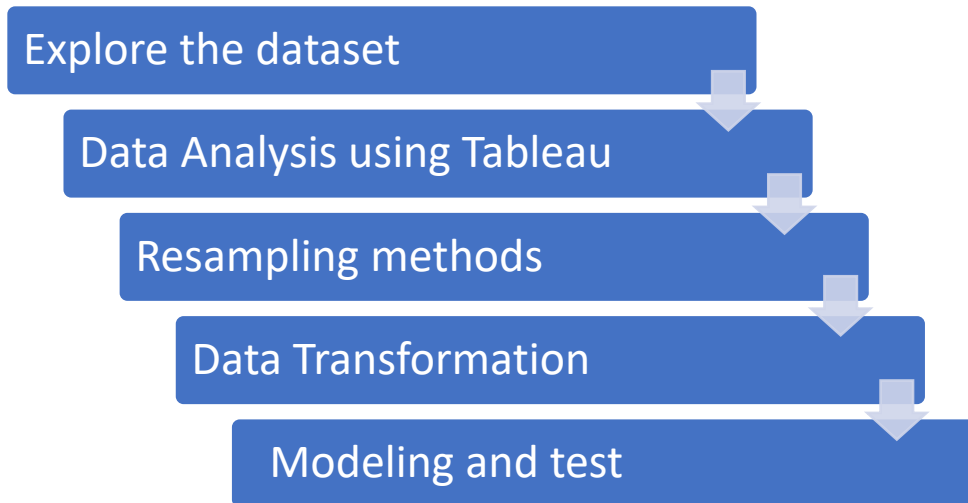
[This study](#) (9) was done on recognition of real or fake news collected from various users via Twitter and media sources such as PolitiFact. The study compares the s convolutional neural networks (CNNs), long short-term memories (LSTMs), ensemble methods, and attention mechanisms and as a result, the ensemble network of CNN and bidirectional LSTM performed the highest accuracy compared to other models such as logistic regression and SVM.

Based on [this study](#) (10), the under-sampling method was used for a highly imbalanced dataset and 70% of the data was used for training purposes. To compare the models they used accuracy, F1-Score, Recall, Precision, and specificity. The stacking classifier provides a better result compare to other models.

[This article](#) (11) suggests a method to find fake news based on different structure types of the news article. Deep learning (GRU) was used for word part analysis and make the dependency tree providing unique features for actual and fake news.

Approach

We use binary classification to distinguish fake and real job postings. The flowchart applied methodology is as bellow:



Explore the dataset

In this section, the dataset is described, and its features are introduced. Since there is not much work available in this dataset the cleaning procedures are applied to have a workable dataset.

Data description

The dataset used is the Employment Scam Aegean Dataset (EMSCAD) which is publicly supported by the Information & Communication Systems Security Laboratory of the University of Aegean. This dataset includes 17,880 real-life job postings in which 17,014 are real and 866 are fake. This dataset is further processed and uploaded on the Kaggle and is available publicly. The variables of the dataset include binary, String, HTML fragment, and Nominal. (12)

Binary

Telecommuting	True (=1) for telecommuting positions.
Company logo	True (=1) if the company logo is present.
Questions	True (=1) if screening questions are present.
Fraudulent	Classification attribute.(0= Real , 1= Fake Job postings)

String

Title	The title of the job ad entry.
Location	Geographical location of the job ad.
Department	Corporate department (e.g. sales).
Salary range	Indicative salary range (e.g. \$50,000-\$60,000)

HTML fragment

Company profile	A brief company description.
Description	The details description of the job ad.
Requirements	Enlisted requirements for the job opening.
Benefits	Enlisted offered benefits by the employer.

Nominal

Employment type	Full-type, Part-time, Contract, etc.
Required experience	Executive, Entry level, Intern, etc.
Required education	Doctorate, Master's Degree, Bachelor, etc.
Industry	Automotive, IT, Health care, Real estate, etc.
Function	Consulting, Engineering, Research, Sales, etc.

Data Cleaning

Filling missing values is an important task in the data cleaning process. There are many ways to overcome this issue, such as removing, filling, or ignoring them.

What we used for this dataset is removing features with a high percentage of missing (Figure1). For example, the 'salary_range', 'department', and 'benefits' features were removed from the dataset.

```
df.isna().sum()/ len(df)
```

job_id	0.000000
title	0.000000
location	0.019351
department	0.645805
salary_range	0.839597
company_profile	0.185011
description	0.000056
requirements	0.150727
benefits	0.403244
telecommuting	0.000000
has_company_logo	0.000000
has_questions	0.000000
employment_type	0.194128
required_experience	0.394295
required_education	0.453300
industry	0.274217
function	0.361018
fraudulent	0.000000

Figure 1(Missing values of the dataset)

Also, for this dataset, the backward method is used to fill the missing values for some features such as 'employment_type', 'required_experience', 'required_education', 'industry', and 'function'. But before filling them all dataset was sorted based on title feature.

We also add all the text columns in one to study them in detail. We looked at the top 20 most common words being used in both fake and real job postings (Figures 2 and 3). It is shown that most frequent words are almost the same, and it is tough to differentiate between fake and real ones.

Finally, we separated the Location column into two cities and country names and keep only country names because the number of categories for city names was high and most of them were missing.

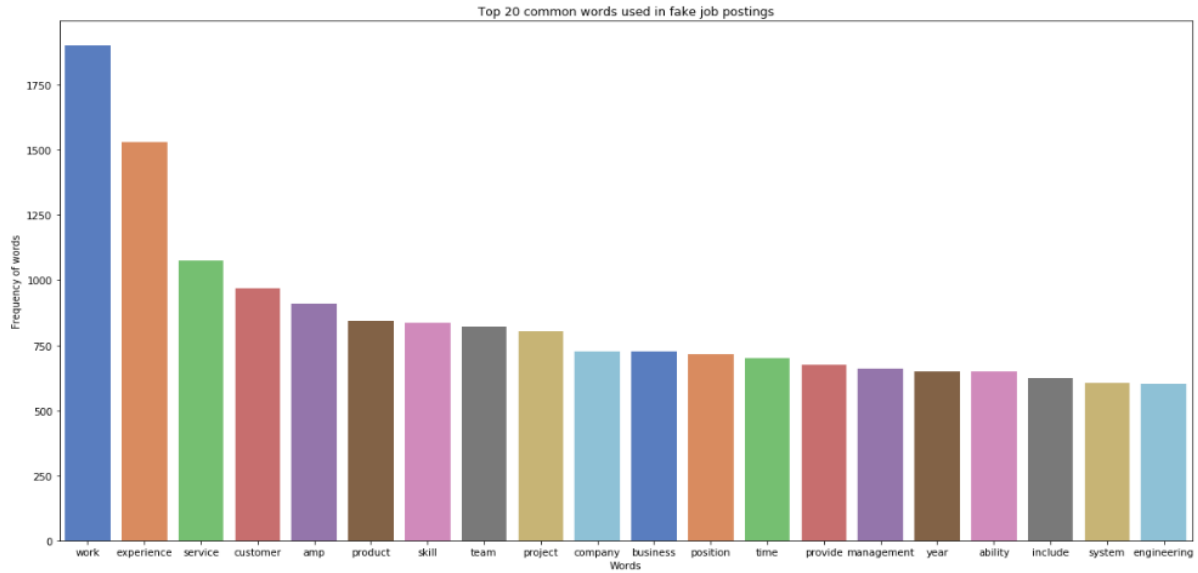


Figure 2 (Top 20 common words used in fake job postings)

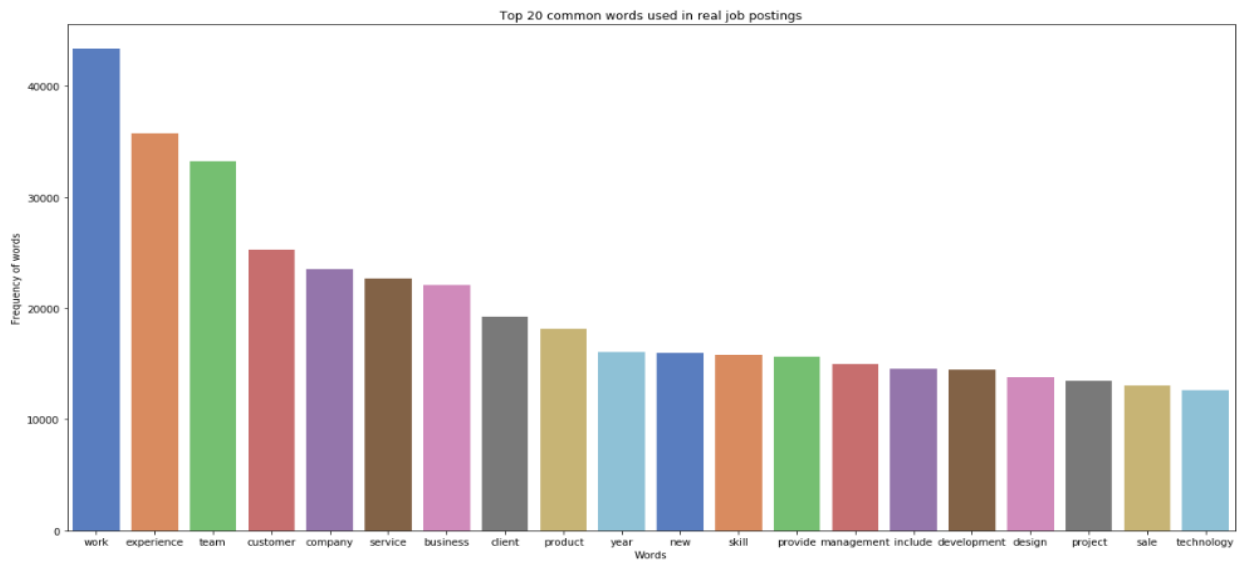


Figure 3 (Top 20 common words used in fake job postings)

Data Analysis:

The distribution of the target variable shows that the dataset is highly imbalanced with 16391 real job posting and 834 fake job advertisements (Figure 4).

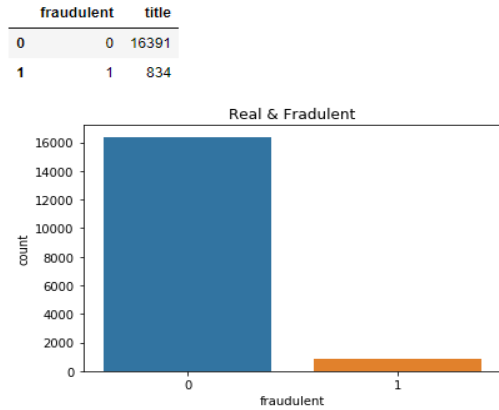
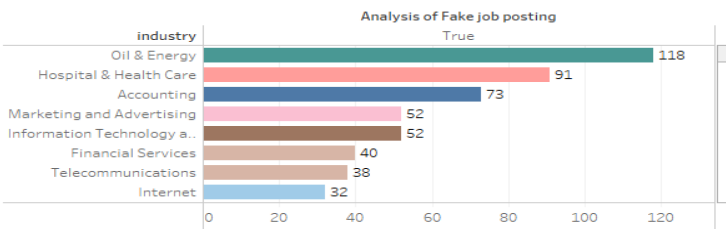


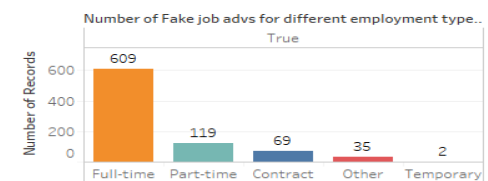
Figure 4 (Distribution of target variable)

Based on Figure 5 scammers are targeting mostly full-time jobs. And the United States has the highest number of fake job postings. The Oil & Energy industry is targeted mostly by so-called recruiters, additionally, the Entry-level job seekers are targeted mostly. It is also obvious the number of fake job postings without an interview is more than the ones with the interview process.

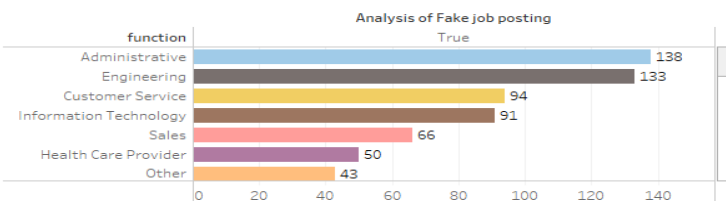
Number fake job posting for different industries



Number of Fake job advs for different employment types



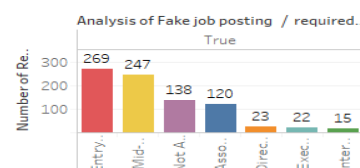
Number fake job posting for different functions



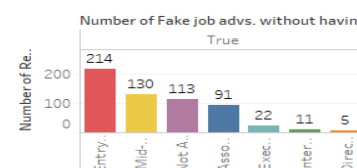
Number of Fake job Posting for different countries



Number of Fake job posting for different Levels



Number of Fake job posting without having Question



Number of fake job posting for having question or not.

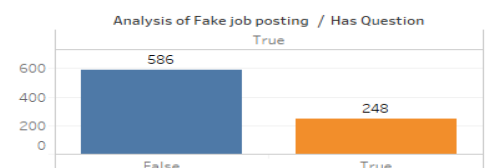


Figure 5 (Fake job posting analysis)

Resampling Techniques

The dataset is highly imbalanced because it carries more real job postings as compared to fake job postings (17) with 16391 real job posting and 834 fake job advertisements (Figure 3). That means prediction will get very high accuracy without detecting fake job postings. To handle this, we conducted under-sampling and over-sampling by reducing the majority occurrences and by raising the minority occurrences respectively. (18)

Algorithms:

In our model, we used 4 different types of machine learning algorithms and for the implementation work, we used Python 3.8.3 as our programmable language.

The classification models that we implemented using the above- mentioned dataset are Random Forest, KNN, logistic regression, and decision tree.

Random forest, the algorithm constructs hundreds or thousands of deep decision trees wherein every single tree act as a weak learner, but all together they make a robust learner. is an ensemble method that can be used to build predictive models for both classification and regression problems. (19)

k- Nearest Neighbor model is one the simplest but most effective models. KNN algorithm can be used for both classification and regression problems. The KNN algorithm uses ‘feature similarity’ to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set. (20)

A decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. As per the name of the model, this is built in the form of a tree-like structure. (20)

Logistic Regression is a classification, not a regression algorithm. It is used to predict a binary outcome given a set of independent variables. (20)

Model and discussion:

80% of the dataset used for training and 20% used for testing. Data was balanced by two under-sampling and over-sampling methods. So, the Accuracy, F1 Score, and AUC - ROC Curve are used to compare the models. The Figure 6 shows the following steps after splitting the dataset.

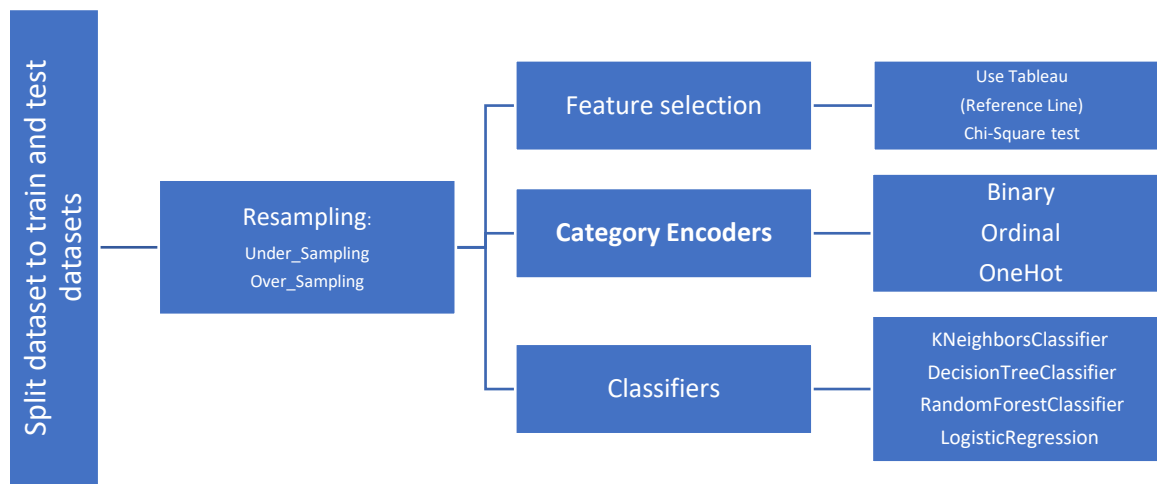


Figure 6 (Steps after splitting dataset)

Resampling:

To make a balanced dataset the under_sampling and over_sampling method is used. Table 1 shows the dimension of the train dataset for both methods.

Under_Sampling	Over_sampling
<code>1 667</code> <code>0 667</code> <code>Name: fraudulent</code>	<code>1 13113</code> <code>0 13113</code> <code>Name: fraudulent</code>

Table 1

Feature Selection:

Because the dataset contains categorical data, the Chi-square test is used to find out the relationship between the features. (21)

Null Hypothesis (H0): Two variables are independent.

Alternate Hypothesis (H1): Two variables are not independent.

With 95% confidence that is $\alpha = 0.05$, we will check the calculated Chi-Square value falls in the acceptance or rejection region, and based on Figure 7, we reject the Null Hypothesis for all categorical data.

	Column	Hypothesis
0	employment_type	Reject Null Hypothesis
1	required_experience	Reject Null Hypothesis
2	required_education	Reject Null Hypothesis
3	industry	Reject Null Hypothesis
4	function	Reject Null Hypothesis
5	country_name	Reject Null Hypothesis

Figure 7(Result of Chi_Square test)

In this section by using the visualization tool Tableau and applying the A/B test, it is possible to compare the influence of the dataset variables with the target variable in the prediction. Figure 8 shows the reference line for each feature and what stands out is all features have a significant role because the attributes increase the reference line for most of their values.

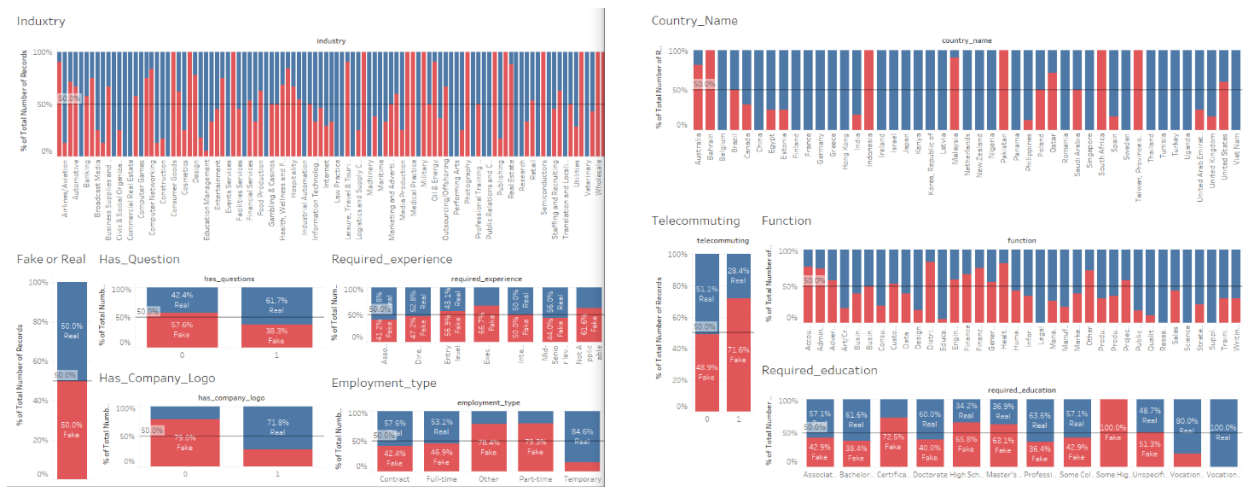


Figure 8 (Reference line for features)

Category Encoders

Because the dataset contains categorical data, it is used an encoder. The category encoder library has a set of transformers. The transformers provide a wide variety of methods to transform

categorical data. (13) In this study, three categorical encoders were used for the purpose of comparison namely: BinaryEncoder, OneHotEncoder, and OrdinalEncoder.

Binary encoding for categorical variables, like one hot, but store categories as binary bitstrings. (14)

Onehot (or dummy) coding for categorical features, produces one feature per category, each binary. (15)

Ordinal Encodes categorical features as ordinal, in one ordered feature. (16)

Prediction and result

When the training and test data were ready, we train the machine learning model to classify the fake and real job postings We used 4 different types of machine learning algorithms in our model namely: RandomForestClassifier, DecisionTreeClassifier, KNeighborsClassifier, and Logistic Regression.

First, the accuracy of models when we used Onehot encoder to transform our categorical features performed better for both under-sampling and over-sampling methods. (Figure8,9)

Besides, the accuracy of models for the over_sampling method was higher in comparison to the under-sampling method. (Figure9,10)

Also, if we transform the categorical data by one-hot encoding and use the over_sampling method to train the data the Random forest classifier gives the highest accuracy in comparison to other classifiers.

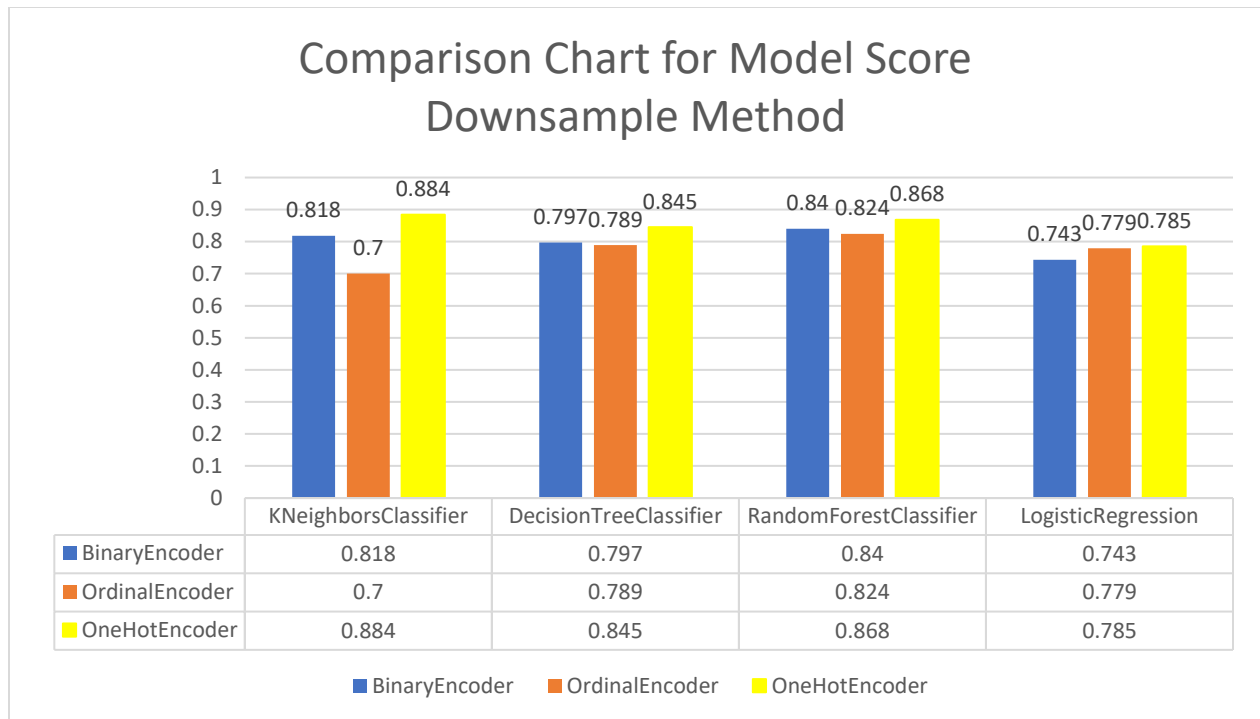


Figure 9 (Comparison chart for model scores based on downsampling)

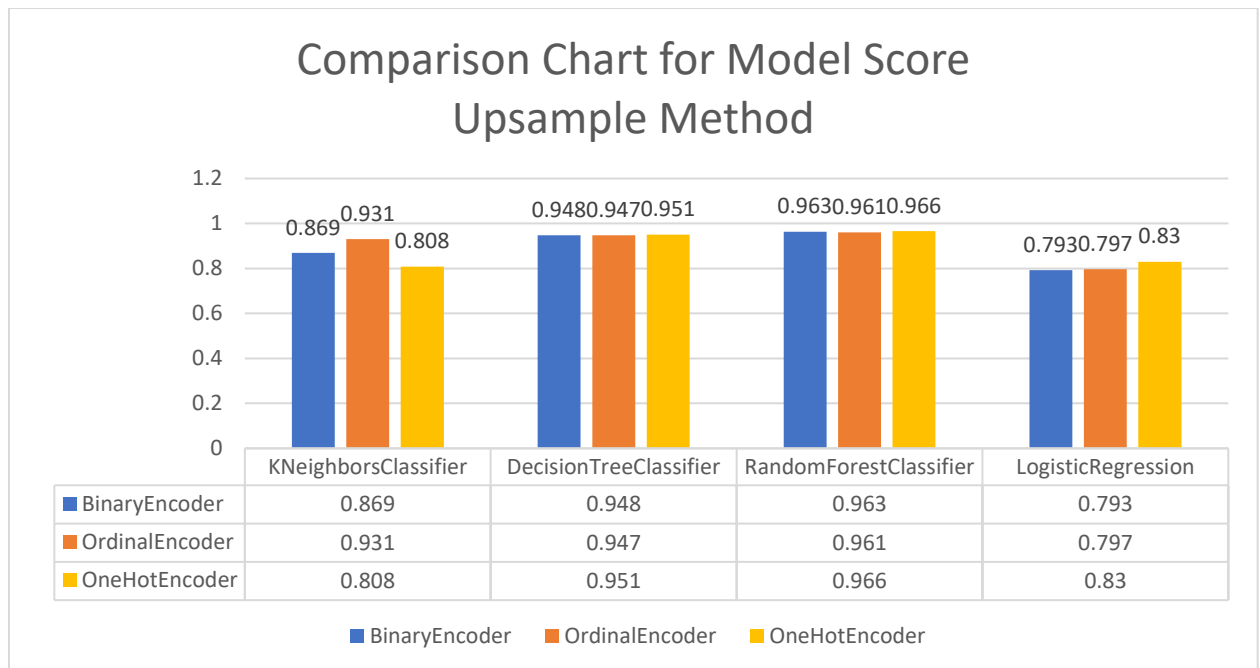


Figure 10 (Comparison chart for model scores based on upsampling)

Finally, other evaluation metrics such as ROC-AUC and F1 score also provides that the RandomForestClassifier did better than other classifier algorithms for this dataset Figure 2.

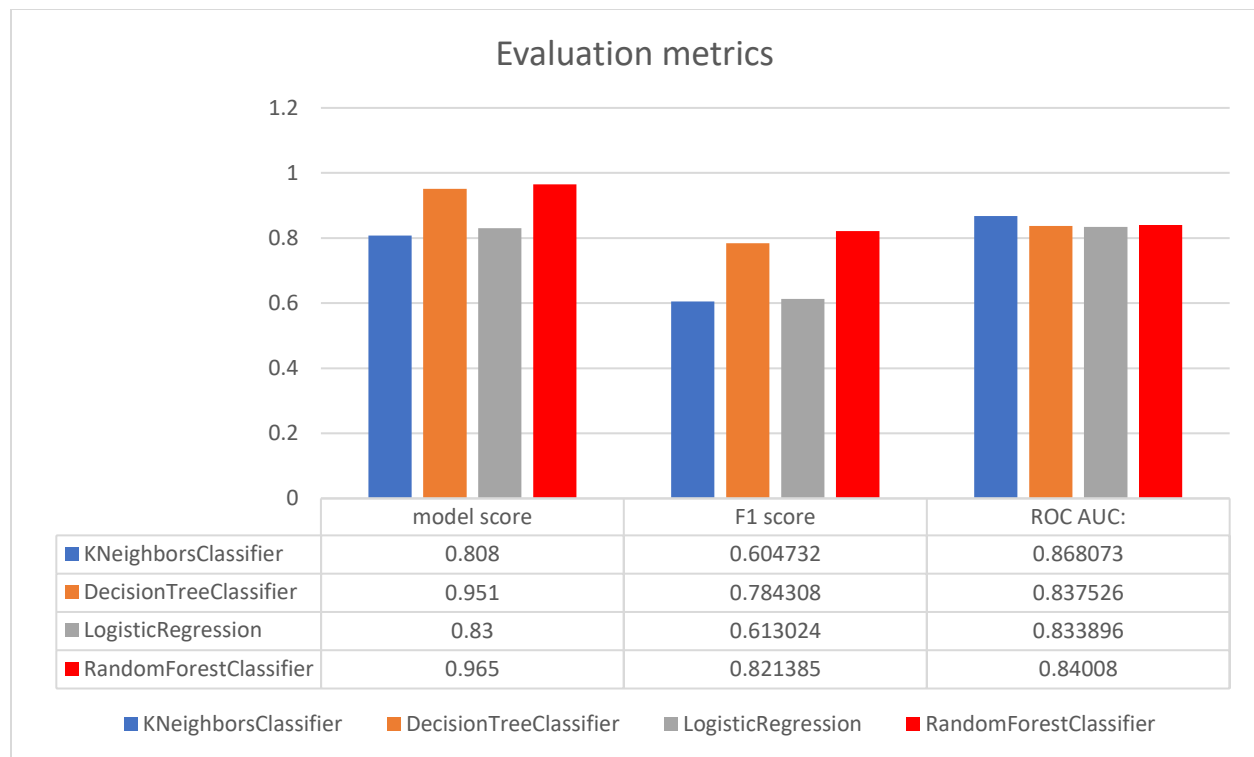


Figure 3

Conclusion

In this study, we used different methods for preparing the dataset. Because the dataset includes both categorical and numerical data the encoders used to transform the categorical features and OneHot encoder performed better in comparison to the binary and ordinal encoder. The dataset was highly imbalanced and between two downsampling and upsampling, the evaluation metrics were high in the upsampling method. Finally, the RandomForestClassifier has given an accuracy score of 96.6% and for 3445 test observations, it has correctly predicted the class labels for 3326 job postings.

Notice that only the categorical and numerical columns for the feature alongside the target column which is the fraudulent column are used columns and the text data type columns did not use for this study. The future study could be done on text columns either itself or with others as well.

References

1. **FISHER, ANNE.** Fortune. [Online] March 2020. <https://fortune.com/2020/03/02/fake-job-postings-hiring-scams/>.
2. **Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal.** Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison. April 09, 2020.
3. **Busisizwe Kelvin Nkomo, Thayne Breetzke.** A conceptual model for the use of artificial intelligence for credit card fraud detection in banks. April 30, 2020.
4. **Abdullah-All-Tanvir, Ehesas Mia Mahir, Saima Akhter, Mohammad Rezwanul Huq.** Detecting Fake News using Machine Learning and Deep Learning Algorithms. September 19, 2019.
5. **Supanya Aphiwongsophon, Prabhas Chongstitvatana.** Detecting Fake News with Machine Learning Method. January 21, 2019.
6. **Dimitrios Katsaros, George Stavropoulos, Dimitrios Papakostas.** Which machine learning paradigm for fake news detection? November 25, 2019.
7. **Deepak S, Bhadrachalam Chitturi.** Deep neural approach to Fake-News identification. April 16, 2020.
8. **Majumder, Tanisha, et al.** ANALYSIS OF FAKE DATA ON SOCIAL MEDIA. May 2020.
9. **Sachin Kumar, Rohan Asthana, Shashwat Upadhyay, Nidhi Upreti, Mohammad Akbar.** Fake news detection using deep learning models: A novel approach. November 05, 2019.
10. **Sahil Dhankhad, Emad Mohammed, Behrouz Far.** Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. August 06, 2018.
11. **Anmol Uppal, Vipul Sachdeva , Seema Sharma.** Fake news detection using discourse segment structure analysis. April 09, 2020.
12. **Aegean, University of the.** Employment Scam Aegean Dataset. [Online] <http://emscad.samos.aegean.gr/>.

13. *Detecting Credit Card Fraud Using Selected Machine Learning Algorithms*. **Puh, Maja, and Brkić, Ljiljana**. 2019.
14. **Dilip Singh Sisodia, Nerella Keerthana Reddy, Shivangi Bhandari**. Performance evaluation of class balancing techniques for credit card fraud detection. June 21, 2018.
15. *Credit Card Fraud Detection Techniques* – A. **Nikita Shirodkar, Pratikesh Mandrekar, Rohit Shet Mandrekar, Rahul Sakhalkar, K.M. Chaman Kumar, Shailendra Aswale**. s.l. : International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020.
16. *Supervised Machine Learning Algorithms for Credit Card Fraud Detection: A Comparison*. **Samidha Khatri, Aishwarya Arora, Arun Prakash Agrawal**. s.l. : 10th International Conference on Cloud Computing, Data Science & Engineering, 2020.
17. **Wikipedia**. [Online] https://en.wikipedia.org/wiki/Chi-squared_test.
18. *Survey on categorical data for neural networks*. **Hancock, John T, and Khoshgoftaar, Taghi M**. 2020, *Journal of Big Data*, Vol. Vol. 7 (1).
19. **Binary**. [Online] http://contrib.scikit-learn.org/category_encoders/binary.html.
20. **One Hot**. [Online] http://contrib.scikit-learn.org/category_encoders/onehot.html.
21. **Ordinal**. [Online] http://contrib.scikit-learn.org/category_encoders/ordinal.html.