Project :

Identify fake job posting using machine learning

BY

PRANJAL SHUKLA

Year : 2020

- Introduction:

The project “Identifying fake job postings” uses machine learning algorithms which will help determine the fraud job postings advertised through newspapers, posters.

This project will help the people searching for jobs through various forms of medium like newspapers, advertisements,

Posters to differentiate between a real job advertisement and a fraud job advertisement .

This project model can be used by different people seeking jobs under various fields of medicine, engineering, government, business and others.

- **Understand and define the problem :**

I. Problem Statement :-

Using machine learning algorithms develop a model to predict whether a job posting is fraudulent or not.

II. Objectives :-

a. To identify a job posting as fraudulent or valid
b. To find machine learning algorithm having the best accuracy of predictions

III. Scope :-

a. This model can be used by graduates looking for jobs
b. This model can be used in the field of research to find out valid job postings in country

IV. Data Sources:

The various data sources used are:

a. CSV data
b. Excel data
c. MySql data

V. Tools and Techniques :

The various tools and techniques used are:

a. Pandas library
b. Numpy library
c. Matplotlib and seaborn library for data visualization
d. Sklearn library for machine learning algorithms implementation

- ## Dataset Preparation and Preprocessing

In this stage of project implementation, focus is put on data collection, data selection, data preprocessing, and data transformation.

I. Data collection :-

The data used in this project is in the form of csv file consisting of rows and columns in the form of table.

II. Data Visualization :-

In this project we are using seaborn and matplotlib library of python for data visualization.

The various graphs plotted in project are boxplot,heatmap,barplot

III. Labeling

As supervised learning is implemented in the model we have a target column in our dataset named “Fraudulent” which consist of values “0” meaning the job posting is not fraudulent and “1” for the job is fraudulent

IV. Data Selection:-

The dataset used in the project contains insignificant columns which need to be dropped like : job_id , telecommuting , has_company_logo , has_questions.

V. Data Preprocessing

The purpose of preprocessing is to convert raw data into a form that is useful in training and

testing the ML model. The structured and clean data produces more precise results. In short,

good quality data when fed to the ML model, it produces better results.

The Preprocessing technique includes data formatting, cleaning, and sampling techniques.

a. Data Formatting : -

   We have used the Standard Scaler function to normalize our dataset

b. Data cleaning:

   As our dataset did not contain null values for numerical columns and it also did not contain null values after converting categorical columns to numerical, our dataset is clean

c. Data anonymization: -

   The dataset used in project do not contain any sensitive information.

d. Data Transformation:-

We have converted categorical columns to numerical columns using the LabelEncoder function .

Scaling: The dataset columns are normalized using StandardScaler function

Feature Extraction: Some of the existing features are combined to create new features which are useful for ML modeling.

VI. Dataset Splitting:

The given dataset is split into three parts: training, testing, and validation sets. The ration of training and testing sets is typically 80 to 20 percent. The 20 percent of the training set is further split as a validation set.

- ## Model Training:-

In this stage, the training data is fed to the ML algorithm to build and train.

The different algorithms used to train the models are:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. SVM
5. KNN

- ## Model Testing and Evaluation

The goal of this step is to develop the simplest, reliable and efficient model.

We have used the above mentioned algorithms to train the model and also calculated the best model with highest accuracy.

From the results, We can say that the Decision Tree algorithm gives the best accuracy for our project of 0.97

- ## Improving Predictions with Ensemble Methods

As the Decision Tree algorithm provides the best accuracy for our model we are using the following technique to improve predictions:

1. Bagging: In this case, the models of the same type are combined in sequential manner. The training dataset is split into subsets. Then the models are trained on each of these subsets.

- **Model Deployment**

We are using the Web Service based Deployment in which the prediction is done continuously. Mostly the private or public cloud is used for deployment.

- **Conclusion and Further Development:**

This project explains the stages of machine learning used to make the model which mainly are:

1. Data Collection
2. Data Preprocessing
3. Data visualization
4. Splitting the dataset
5. Training the model
6. Testing the model
7. Deployment of model

I would like to make changes in model which will help to automatically delete the fake job postings and only store the valid job postings in future.