ANALYSIS AND CLASSIFICATION OF JOB POSTINGS AS FAKE / LEGITIMATE

P12 : Abhishek Firodiya, Vaidehi Koppolu, Amey Meher, Chetana Chunduru
Department of Computer Science, North Carolina State University
Raleigh, NC 27695
[asfirodi, vkoppol, avmeher, cchetan2] @ncsu.edu

1. Data Set

The dataset consists of 18000 data points, of which 800 are classified as fake job postings under the 'fraudulent' attribute. It also consists of data elements such as job title, location of work, department of the job role, salary range and other metadata related to job profile. There are 13 string attributes for which transformation needs to be done so as to collect valuable information out of it. The dataset also contains NULL elements in some of the attributes for which proper cleaning technique needs to be applied. The dataset can be accessed on this link.

2. Project Idea

Our project goal is to be able to classify the job postings as legitimate or not based on the data attributes on which the model is trained. We will implement and compare the performance of various classification algorithms for the above task, which includes greedy learning algorithms such as Decision Trees as well as lazy learning algorithms such as K-Nearest Neighbors. Also, we aim to discover valuable patterns in the data by employing different visualization techniques so as to get a better understanding of the data.

3. Software to Write

We will write modular files in Python for each process of the data mining pipeline, such as feature_selection.py, feature_transformation.py, train_decision_tree.py and so on. We will also write files such as evaluate_decision_tree.py to evaluate the performance of our model. All the files functionality would be integrated in a driver file such as job_classifier.py. The code will be written in Python using Jupyter Notebook and Spyder. For the functionality, we'll also need python libraries like Pandas, Numpy, sklearn, matplotlib, seaborn etc.

4. References

[1] Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries - Syed Mahbub,Eric Pardede,A.S.M Kayes.

[2] A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques - Sultana Umme Habiba, Md.Khairul Islam,Farzana Tasnim

[3] K Nearest Neighbor Algorithm for Learning and Classification - Kashvi Taunk,Srishti verma,Sanjukta De

5. Work Distribution

- Abhishek & Amey: Work related to analyzing by visualization of the data for selection of relevant features
- Vaidehi & Chetana: Preprocessing and Transformation of data
- Amey & Vaidehi: Implementation of classification based on Greedy learning algorithms
- Abhishek & Chetana: Implementation of classification based on Lazy learning algorithms

6. Midterm Milestone

By the Midterm checkpoint, based on the data mining pipeline, our milestone would be completing the activites of data selection, preprocessing, transformation and implementing as well as evaluating the accuracy for atleast one classification algorithm. The outcome for the midterm milestone would be the accuracy scores for one of the classification algorithm.