Analysis And Classification of Job Postings as Fake / Legitimate

(CSC 522)

Chetana Chunduru
cchetan2@ncsu.edu

Introduction

- Recession in IT sector
- Due to this, there is significant increase in the number of Fraudulent job postings
- People scammed out of $68 million in the first quarter of 2022 *

Goal:

Using a classification model, we want to determine in which category - fake or real - the job posting belongs.

* https://www.fastcompany.com/90803825/that-new-job-offer-may-be-a-scam-heres-what-to-look-out-for

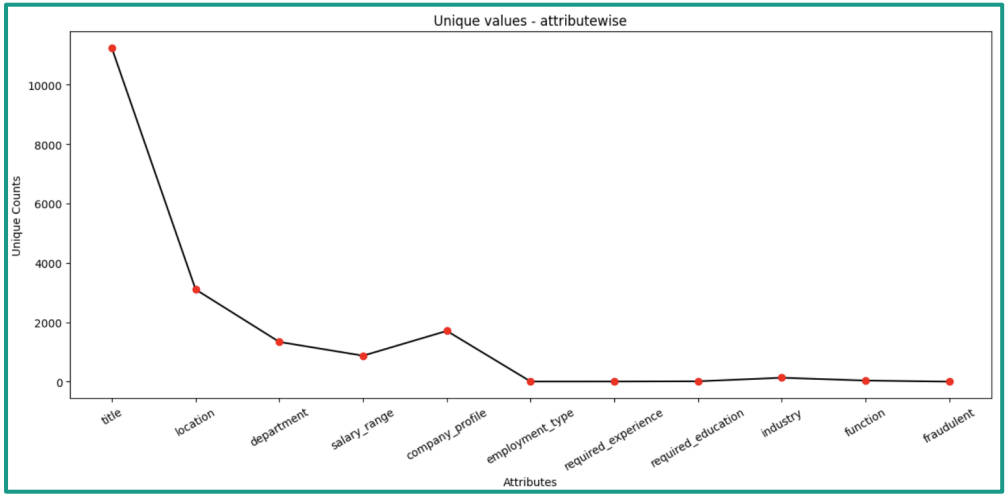
Data Selection

Independent variables

- Title
- Location
- Department
- Salary range
- Company profile
- Employment type
- Required experience
- Required education
- Industry
- Function

Dependent variable

- Fraudulent

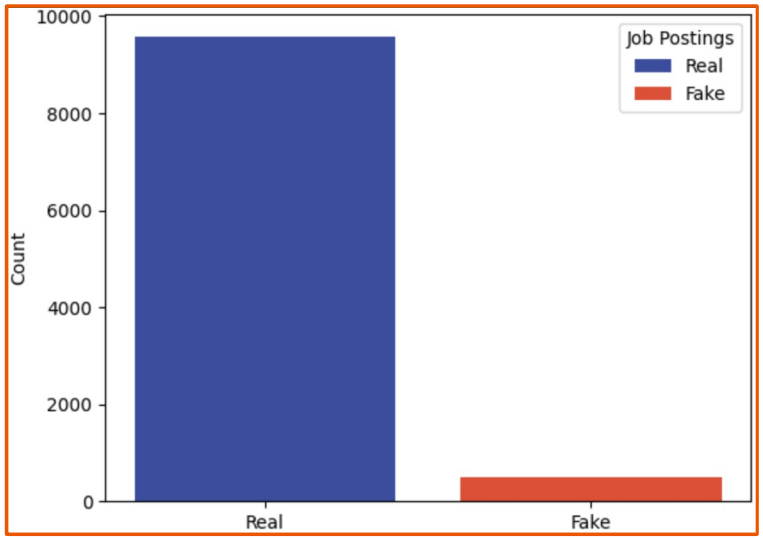


* https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction

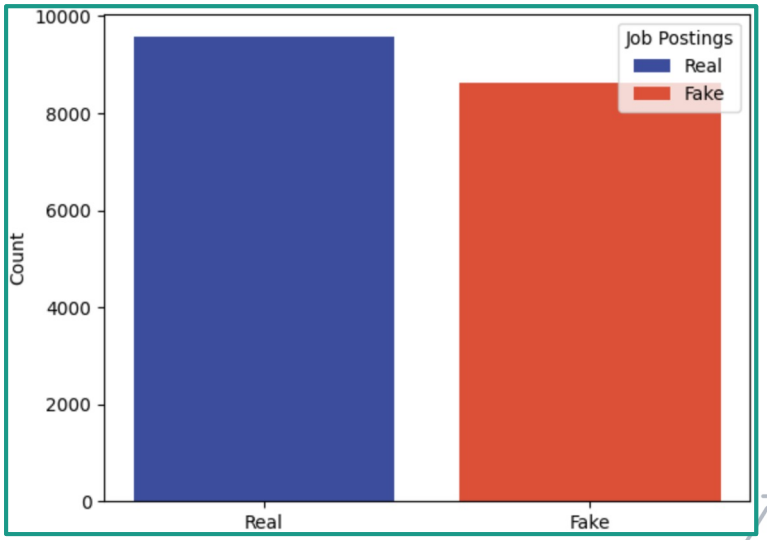
Data Preprocessing

SMOTE - Synthetic Minority Oversampling TEchnique

- To overcome the problem of imbalanced class distribution that existed in our dataset



→



* https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Models

K Nearest Neighbors

To estimate likelihood of a datapoint becoming member of a group or another based on which group that datapoint is nearest to

Decision Tree

To estimate which class datapoint belongs to based on a set of if-else conditions

Naive Bayes

To classify a datapoint assuming conditional independence between every pair of features given the value of the class variable

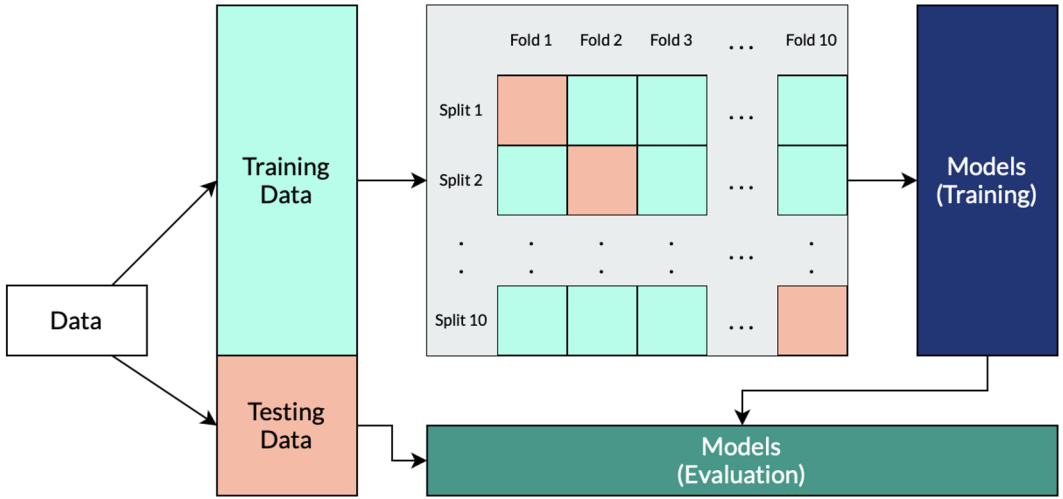
Logistic Regression

To predict the relationship between the binary-natured class (Y) and the independent variables i.e. features (X)

Model Evaluation

Cross Validation

- Implemented 10 fold cross validation technique on our dataset for different model parameters.



* https://machinelearningmastery.com/k-fold-cross-validation/

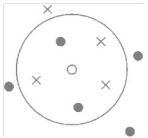
K Nearest Neighbors

Main idea:

Uses proximity to make classifications or predictions about the grouping of an individual data point.

Parameters to consider:

- Distance metric (Manhattan, Euclidean or other)
- Value of the number of means (k)
- Initial values or seed values for the means



K Nearest Neighbors

Advantages:

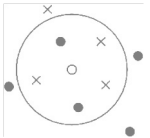
- Easy to implement
- Adapts easily
- Few hyperparameters

Disadvantages:

- Does not scale well, since lazy evaluation
- Curse of dimensionality
- Prone to overfitting

Results:

	Accuracy	Recall	Precision
1NN	0.910826	0.555556	0.283912
3NN	0.885774	0.586420	0.231144
5NN	0.877423	0.629630	0.225166
10NN	0.870564	0.672840	0.222449



Decision Trees

Main idea:

Split the dataset as a tree, based on a set of rules and conditions.

Parameters to consider:

- Need to consider the criteria for the split (Gini, Entropy)
- Need to handle missing data in training as well as testing data properly
- Can consider a max depth of tree parameter to limit overfitting the tree

Decision Trees

Advantages:

- Highly intuitive and easy to understand.
- White box model, easy to understand the conditions for the decisions of the model.
- Cost of using the tree is dependent upon the depth of the tree.

Disadvantages:

- Are unstable. Some changes in data can dramatically alter entire predictions.
- Can create over-complex trees that do not generalize the data and overfit to the dataset.

Results:

	Accuracy	Recall	Precision
DT Val	0.931106	0.679012	0.380623

Naive Bayes

Main idea:

Uses the Bayes theorem with an assumption that the presence of a particular feature in a class is unrelated to the presence of any other feature, which helps in reducing the computation for classification task

Parameters to consider:

- Need to stratify data set properly to avoid the zero probability phenomena
- Apply smoothing technique if zero probability phenomena is frequent or unavoidable

Naive Bayes

Advantages:

- Simple, Fast in processing, and effective in predicting the class of the dataset.
- It performs well in case of text analytics problems
- Easy to obtain the estimated probability for a prediction.

Disadvantages:

- Not ideal for data sets with a large number of numerical attributes.
- It relies on an often-faulty assumption of equally important and independent features which results in biased posterior probabilities
- Smoothing techniques are required to deal with unknown categorical data which are not present in training data.

Results:

	Accuracy	Recall	Precision
NB Val	0.6272	0.54321	0.06962

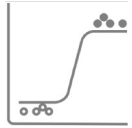
Logistic Regression

Main idea:

Logistic Regression is a classification technique which uses a logistic function to model the dependent variable. It is basically used to calculate or predict the probability of a binary (yes/no) event occurring.

Parameters to consider:

- Can set different values for solvers ('newton-cg', 'lbfgs', 'liblinear', 'sag', 'saga')
- Can set different values for C (penalty strength)



Logistic Regression

Advantages:

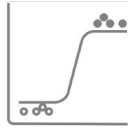
- Logistic Regression is easier to implement
- It is very fast at classifying unknown records.
- It performs well when datasets are linearly separable.

Disadvantages:

- The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
- It constructs linear boundaries.
- It can only be used to predict discrete functions

Results:

	Accuracy	Recall	Precision
LR Val	0.755443	0.462963	0.092822



Conclusion and Inference

We evaluated our models based on Recall, since cost for predicting a fake job as legitimate would carry more cost to it since it would adversely affect the user. Since, this is a case of false negative, we are preferring the Recall metric for our evaluation of the models.

Thus, based on "**RECALL**" values, we chose Decision Tree Classifier.

Results:

	Accuracy	Recall	Precision
Test (Final)	0.942729	0.728111	0.44507