# The Biweekly Projects

Thursday 27/10/2021 Newsletter Edition: 001

## “Fake Jobs Posting Predection”

# INTRODUCTION

# The Dataset



## Fake Job posting dataset:

- 18 features.
- 17880 observations.

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

# The Biweekly Projects

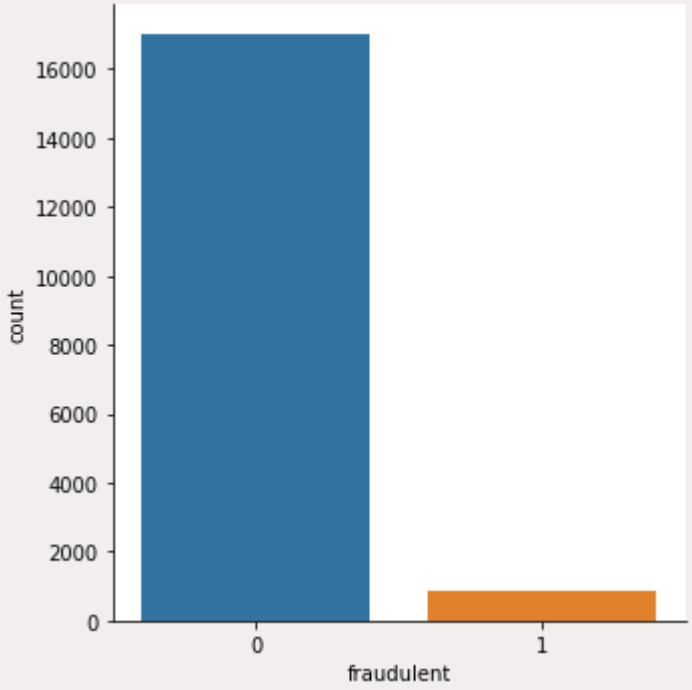
Thursday 27/10/2021 Newsletter Edition: 001

## 02

## EDA AND DATA CLEANING

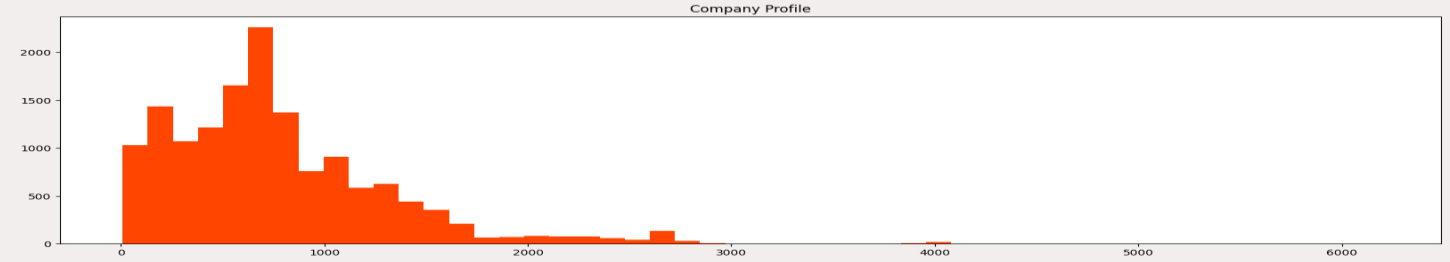
# EDA

# EDA

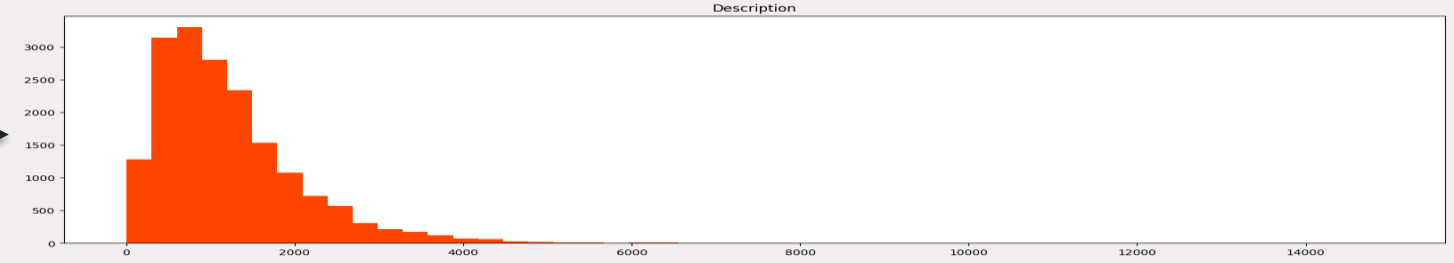



Likely to be fraudulent...

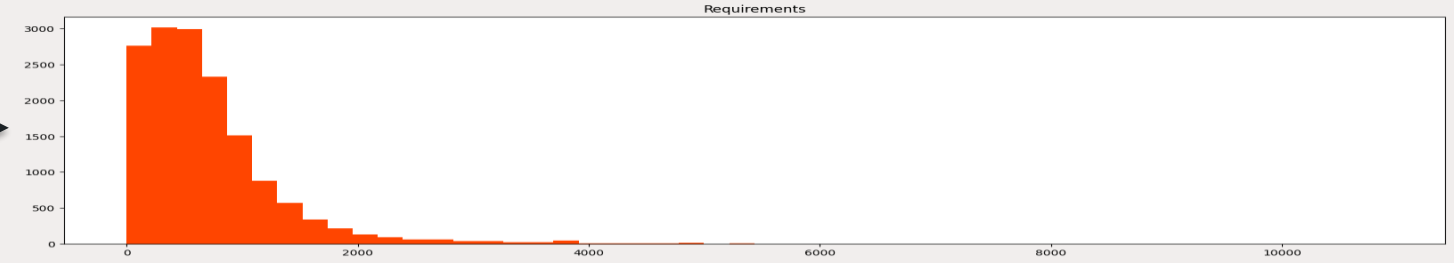
Company Profile



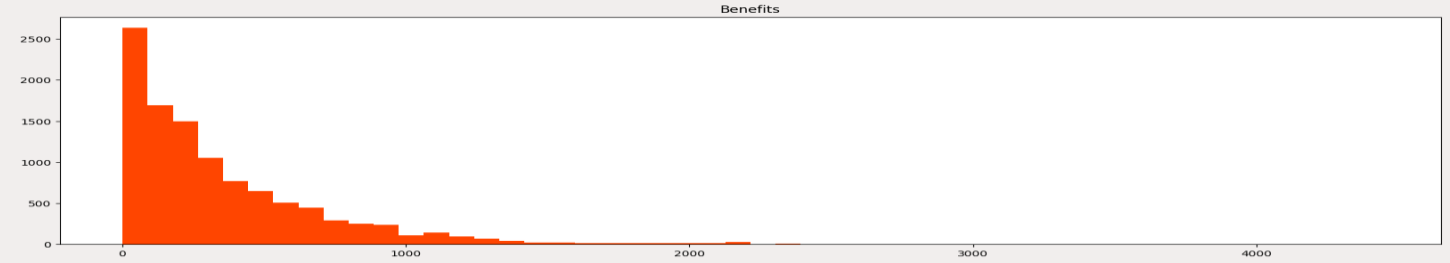

Description



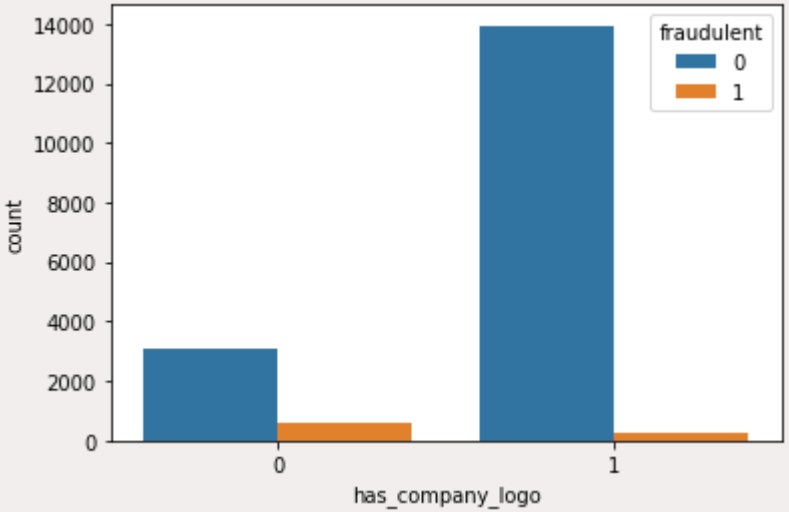

Requirements

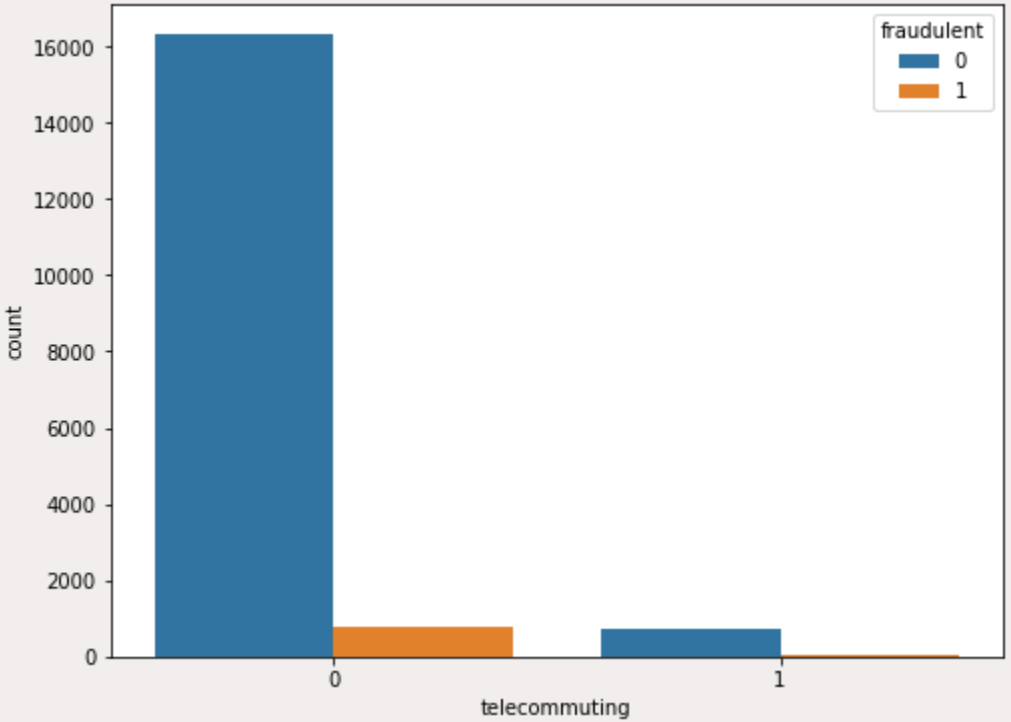



Benefits



# EDA


fraudulent
0
1
count
14000
12000
10000
8000
6000
4000
2000
0
0
1
has_company_logo


fraudulent
0
1
count
16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
telecommuting

# Data Cleaning

Thursday 27/10/2021 Newsletter Edition: 001

```
Original:           (17880, 18) (866, 18) (17014, 18)
Description:        (17879, 18) (865, 18) (17014, 18)
Company Profile:    (14572, 18) (279, 18) (14293, 18)
Requirements:       (15185, 18) (712, 18) (14473, 18)
Benefits:           (10670, 18) (502, 18) (10168, 18)
Salary Range:       (2868, 18) (223, 18) (2645, 18)
Telecommuting:      (17880, 18) (866, 18) (17014, 18)
Has Company Logo:   (17880, 18) (866, 18) (17014, 18)
Has Questions:      (17880, 18) (866, 18) (17014, 18)
Employment Type:    (14409, 18) (625, 18) (13784, 18)
Required Education: (9775, 18) (415, 18) (9360, 18)
Required Experience: (10830, 18) (431, 18) (10399, 18)
All:                (17880, 18) (866, 18) (17014, 18)
```

- Deal with null values.
- Combine all textual features into one.
- Drop unnecessary features.
- Ordinal Encode some features.

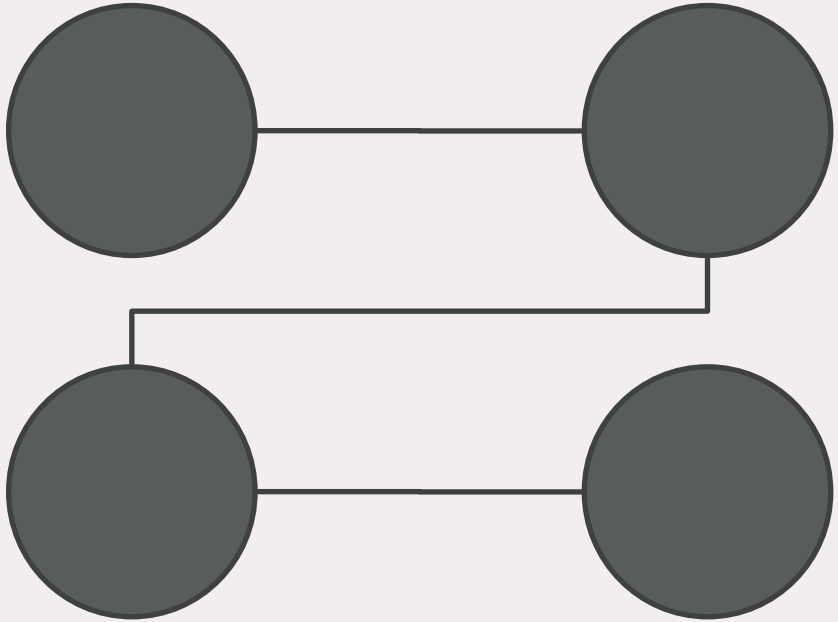
# Dealing With Texts

Thursday 27/10/2021 Newsletter Edition: 001

## 1-LOWERCASE

## 2-TOKENIZATION

breaking up a piece of text into smaller parts



## 3- STOP WORDS

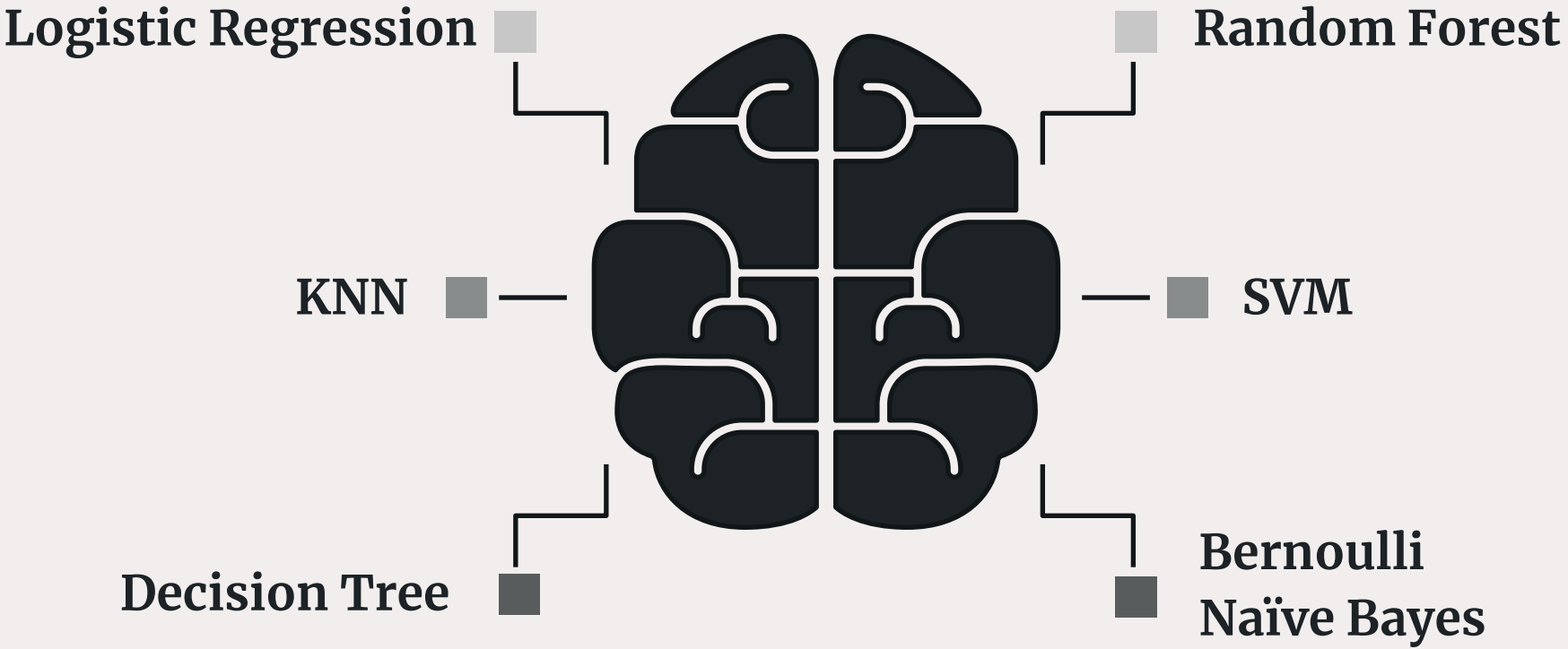
drop the words that are present in text but do not contribute to the meaning of a sentence

## 4- STEMMING

extract the base form of the words by removing affixes from them

# The Biweekly Projects

Thursday 27/10/2021 Newsletter Edition: 001


Logistic Regression
Random Forest
KNN
SVM
Decision Tree
Bernoulli
Naïve Bayes

# The Biweekly Projects

Thursday 27/10/2021 Newsletter Edition: 001

# Experiemnts

Thursday 27/10/2021 Newsletter Edition: 001

**TFIDF with max featuers = 2000**

**CountVectorizer**

**TFIDF With unigrams and bigrams**

**One_Hot**

# TFIDF  with max features = 2000



| Models | Precision | Recall | F1 score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Logistic regression | 0.755 | 0.782 | 0.768 | 0.884 | 0.977 |
| Decision trees | 0.719 | 0.694 | 0.706 | 0.840 | 0.972 |
| Random Forest | 0.989 | 0.541 | 0.699 | 0.770 | 0.977 |
| SVM | 0.392 | 0.876 | 0.541 | 0.904 | 0.929 |
| Bernoulli Naïve Bayes | 0.204 | 0.823 | 0.328 | 0.832 | 0.839 |

# TFIDF With unigrams and bigrams



| Models | Precision | Recall | F1 score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Logistic regression | 0.677 | 0.807 | 0.737 | 0.893 | 0.97 |
| Decision trees | 0.68 | 0.723 | 0.701 | 0.852 | 0.969 |
| Random Forest | 1.0 | 0.542 | 0.703 | 0.771 | 0.977 |
| SVM | 0.371 | 0.870 | 0.520 | 0.896 | 0.920 |
| Bernoulli Naïve Bayes | 0.212 | 0.836 | 0.339 | 0.837 | 0.838 |

# CountVectorizer

Thursday 27/10/2021 Newsletter Edition: 001

| Models | Precision | Recall | F1 score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Logistic regression | 0.710 | 0.803 | 0.75 | 0.893 | 0.975 |
| Decision trees | 0.753 | 0.726 | 0.739 | 0.857 | 0.975 |
| Random Forest | 1.0 | 0.571 | 0.727 | 0.785 | 0.979 |
| SVM | 0.653 | 0.839 | 0.734 | 0.909 | 0.971 |
| Bernoulli Naïve Bayes | 0.298 | 0.893 | 0.447 | 0.895 | 0.896 |

# One_Hot

Thursday 27/10/2021 Newsletter Edition: 001

| Models | Precision | Recall | F1 score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Logistic regression | 0.710 | 0.803 | 0.754 | 0.894 | 0.975 |
| Decision trees | 0.735 | 0.744 | 0.739 | 0.865 | 0.975 |
| Random Forest | 0.989 | 0.577 | 0.729 | 0.788 | 0.979 |
| SVM | 0.653 | 0.839 | 0.734 | 0.908 | 0.971 |
| Bernoulli Naïve Bayes | 0.298 | 0.892 | 0.447 | 0.895 | 0.896 |

# EVALUATION



| | Model and Texts Dealing Method |
|---|---|
| Precision | Random Forest – Countvectorizer, TFIDF unigrams and bigrams |
| Recall | Bernoulli Naïve Bayes – Countvectorizer |
| F1 Score | Logistic Regression – TFIDF |
| ROC AUC | SVM - Countvectorizer |

# EXTRA EXPERIMENTS

Thursday 27/10/2021 Newsletter Edition: 001

## SMOTE

## ADASYN

# SMOTE

Thursday 27/10/2021 Newsletter Edition: 001

| Models | Precision | Recall | F1 score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Logistic regression | 0.359 | 1.0 | 0.526 | 0.5 | 0.357 |
| KNN | 0.774 | 0.696 | 0.733 | 0.791 | 0.819 |
| Decision trees | 0.771 | 0.823 | 0.796 | 0.844 | 0.85 |
| Random Forest | 0.772 | 0.823 | 0.797 | 0.844 | 0.85 |
| SVM | 0.476 | 0.981 | 0.641 | 0.691 | 0.608 |
| Bernoulli Naïve Bayes | 0.691 | 0.640 | 0.665 | 0.741 | 0.769 |

# ADASYN

Thursday 27/10/2021 Newsletter Edition: 001

| Models | Precision | Recall | F1 score | ROC AUC | Accuracy |
|---|---|---|---|---|---|
| Logistic regression | 0.499 | 1.0 | 0.665 | 0.5 | 0.498 |
| KNN | 0.824 | 0.851 | 0.837 | 0.835 | 0.835 |
| Decision trees | 0.827 | 0.878 | 0.852 | 0.847 | 0.847 |
| Random Forest | 0.826 | 0.878 | 0.815 | 0.847 | 0.847 |
| SVM | 0.551 | 0.996 | 0.710 | 0.595 | 0.594 |
| Bernoulli Naïve Bayes | 0.771 | 0.620 | 0.687 | 0.718 | 0.719 |

# The Biweekly Projects

Thursday 27/10/2021 Newsletter Edition: 001

## 05

## PROBLEMS FACED

# PROBLEMS FACED



Couldn’t oversample textual data

TFIDF With unigrams and bigrams returns more than half a million column

Computationally expensive to run

# The Biweekly Projects

Thursday 27/10/2021 Newsletter Edition: 001

# THANK YOU FOR LISTENING