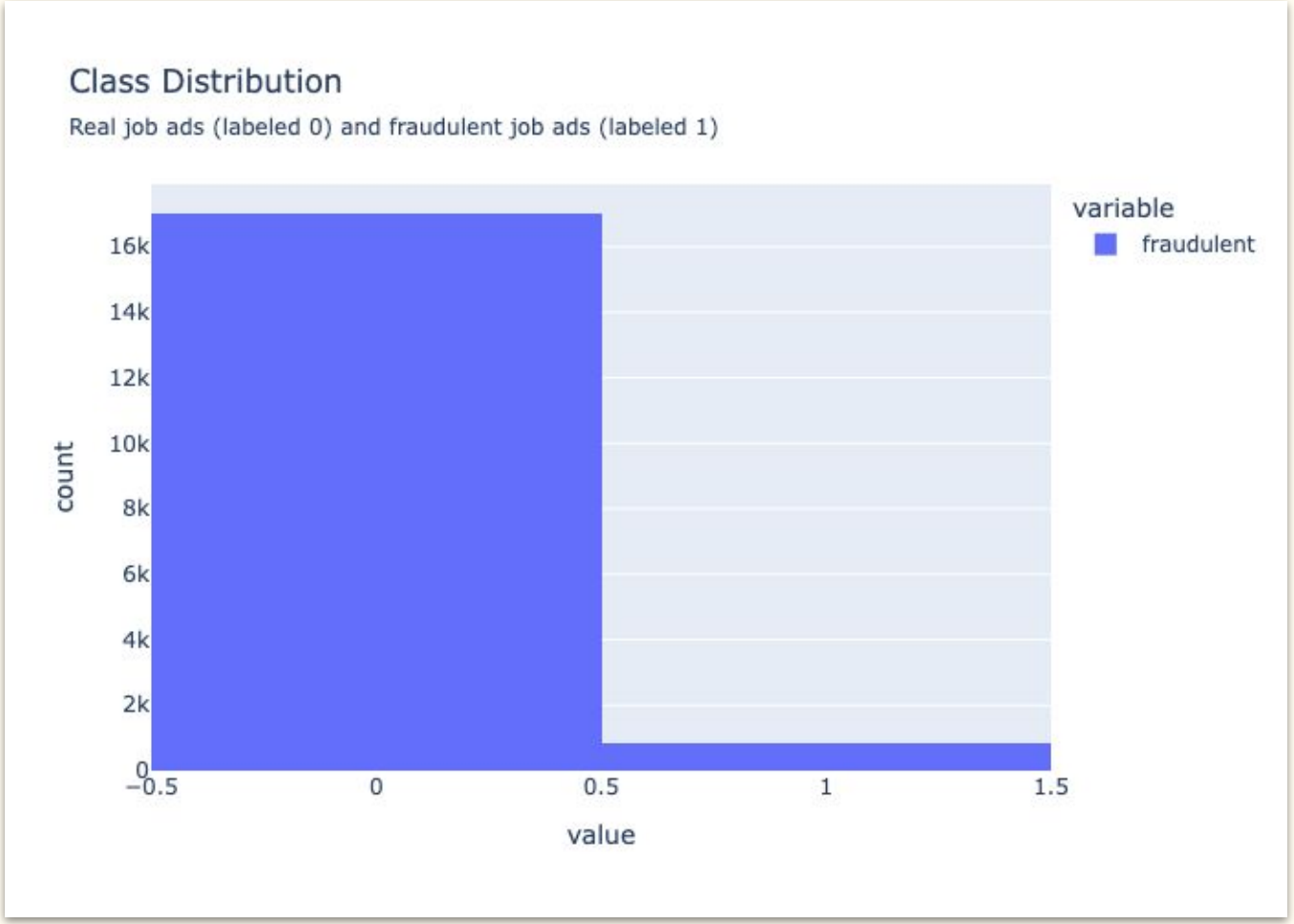
# Identifying Fake Job Ads

Erin De Pree, Ph.D.

# Have You Seen this Ad?

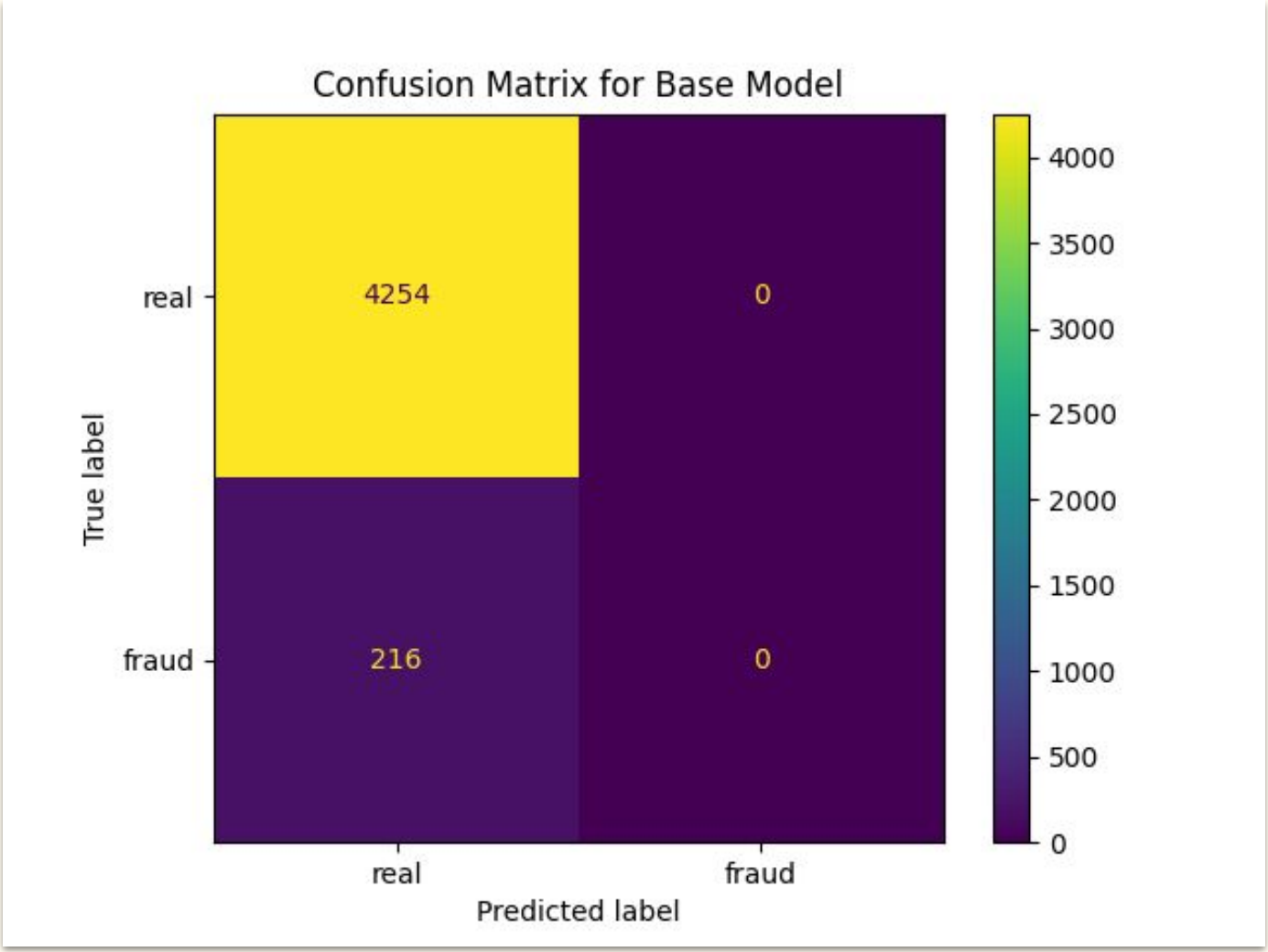
SALES ASSOCIATES - CHANGE YOUR LIFE Unemployed, underemployed or unhappy because of todays changed economy?  Learn to take control of your life. We offer a Multi Award winning, accredited Success Education Program and we require talented sales professionals to keep up with the demands of those products in the market place. While no experience is necessary, as we offer full training to qualified candidates, it will be important that you conduct yourself in a professional and positive manner.  Influence what you achieve.  Start now. Change your Life.  We will give you the tools to succeed.  I look forward to working with you on your next journey.  Reach out to me on and I will come back to you as soon as you send your details.

# Dataset

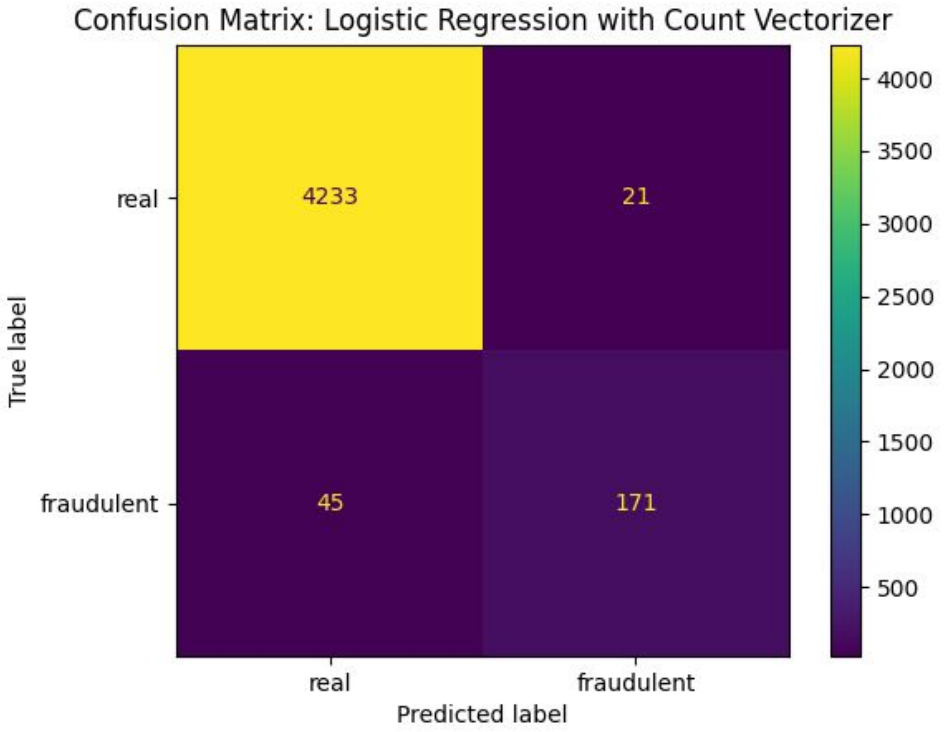

Class Distribution
Real job ads (labeled 0) and fraudulent job ads (labeled 1)
variable
fraudulent
16k
14k
12k
10k
8k
6k
4k
2k
0
count
−0.5
0
0.5
1
1.5
value

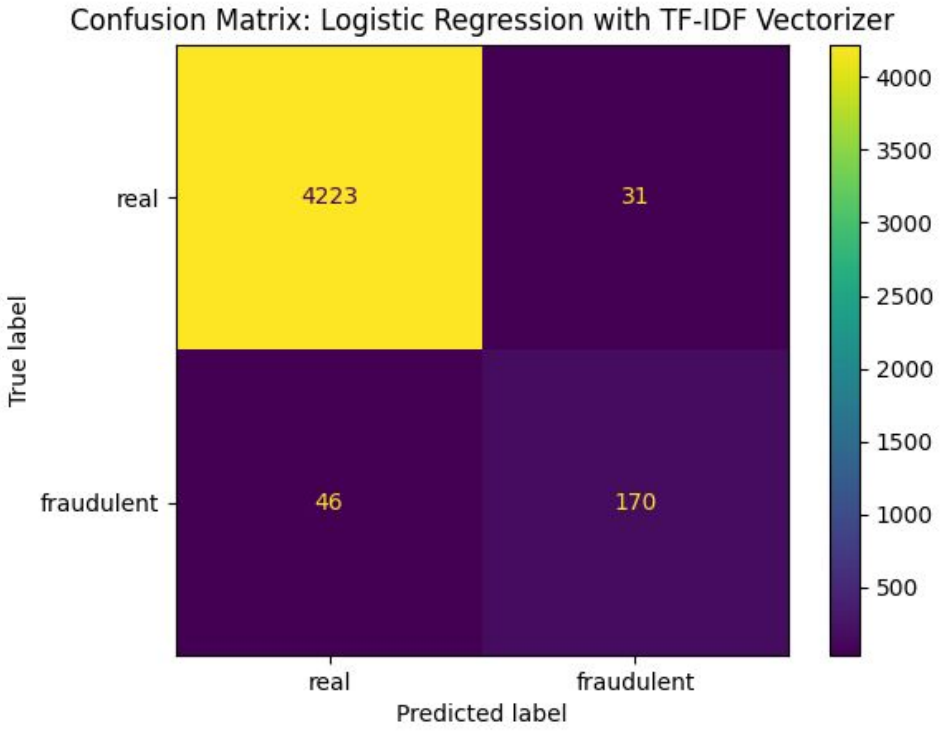
# Baseline

They're all real!



# Count Vectorizer vs. Tf-Idf Vectorizer



# Stop Words

| Stop Words | Accuracy | Precision | Recall | F1 | ROC-AUC |
|---|---|---|---|---|---|
| **Base** | 0.9517 | 0.0 | 0.0 | 0.0 | 0.5 |
| **None** | 0.9852 | 0.8906 | **0.7917** | 0.8382 | **0.8934** |
| **SciKit-Learn** | 0.9861 | 0.9140 | 0.7870 | 0.8458 | 0.8916 |
| **NLTK** | **0.9864** | **0.9330** | 0.7731 | 0.8456 | 0.8852 |
| **spaCy** | 0.9859 | 0.9135 | 0.7824 | 0.8429 | 0.8893 |
| **Gensim** | **0.9864** | 0.9282 | 0.7778 | **0.8463** | 0.8874 |

# Regularization

| Stop Words | Penalty | Accuracy | Precision | Recall | F1 | ROC-AUC |
|---|---|---|---|---|---|---|
| **Base** | n/a | 0.9517 | 0.0 | 0.0 | 0.0 | 0.5 |
| **None** | none | 0.9852 | 0.8906 | **0.7917** | 0.8382 | **0.8934** |
| **Gensim** | none | **0.9864** | 0.9282 | 0.7778 | 0.8463 | 0.8874 |
| **None** | **Ridge** | 0.9846 | **0.9623** | 0.7083 | 0.8160 | 0.8535 |
| **Gensim** | **Ridge** | 0.9864 | 0.9429 | 0.7639 | **0.8499** | 0.8808 |
| **None** | LASSO | 0.9767 | 0.9000 | 0.5833 | 0.7079 | 0.7900 |
| **Gensim** | LASSO | 0.9843 | 0.8925 | 0.7685 | 0.8259 | 0.8819 |

# Advanced Models

- Random Forest
- Multinomial Bayes
- Complement Bayes
- Support Vector Classifier (SVC)
- Stochastic Gradient Descent (SGD) Classifier
- K–Neighbors

# DistilBERT

- Thinks all ads are real
- Needs some reworking

# Train-Test-Split

| Random | Accuracy | Precision | Recall | F1 | ROC-AUC |
|---|---|---|---|---|---|
| 1613 | 0.9852 | 0.9573 | 0.72685 | 0.8263 | 0.8626 |
| 25966 | 0.9852 | 0.9689 | 0.7189 | 0.8254 | 0.8589 |
| 156 | 0.9846 | 0.9401 | 0.7269 | 0.8198 | 0.8623 |
| 698 | 0.9852 | 0.9747 | 0.7130 | 0.8235 | 0.8560 |
| 5 | 0.9875 | 0.9653 | **0.7696** | 0.8564 | **0.8841** |
| 4 | **0.9879** | 0.9765 | 0.7686 | **0.8601** | 0.8838 |
| 6 | 0.9832 | 0.9490 | 0.6898 | 0.7989 | 0.8440 |
| 123 | 0.9848 | **0.9933** | 0.6898 | 0.8142 | 0.8448 |

# Path Forward

- Additional tokenizers (Word2Vec, nltk)
- Preprocessing, text prep / cleaning
- BERT models
- Streamlit app
- Multiple Train-Test splits

Sometimes,
simpler is better!