NLP-based Classification of Fraudulent Job Postings on LinkedIn

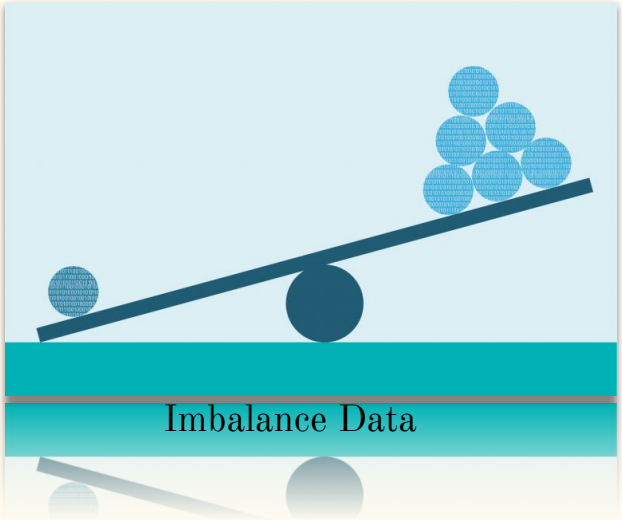
—

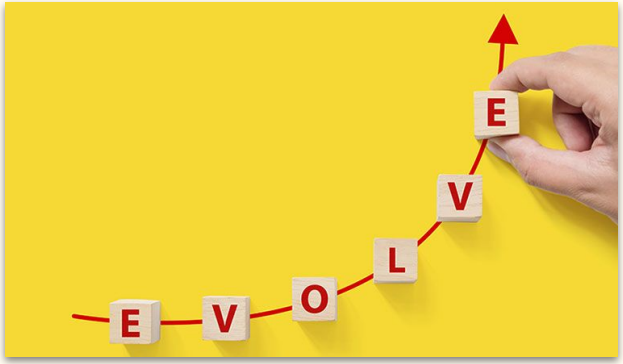
Presented by: Minh Duong
Springboard Capstone Project 2
Oct 2023

Problem Statement

- Linkedin: 60% of job seekers have come across counterfeit job postings
- BBB: 14 million people are exposed to scam job listings annually, more than $2 billion in direct losses

Key Challenges

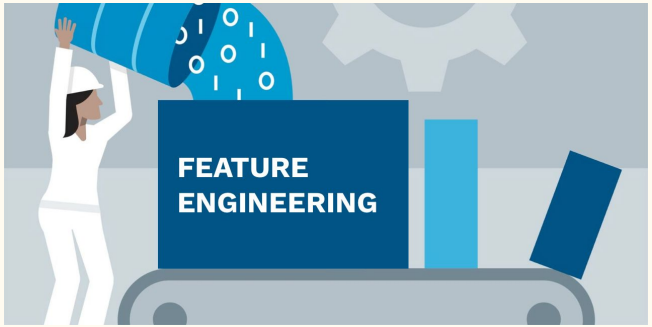


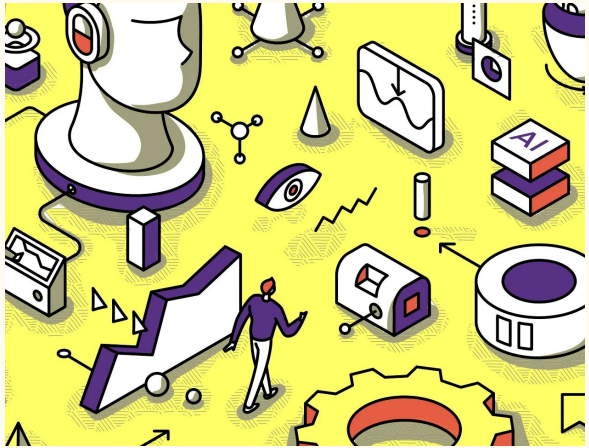


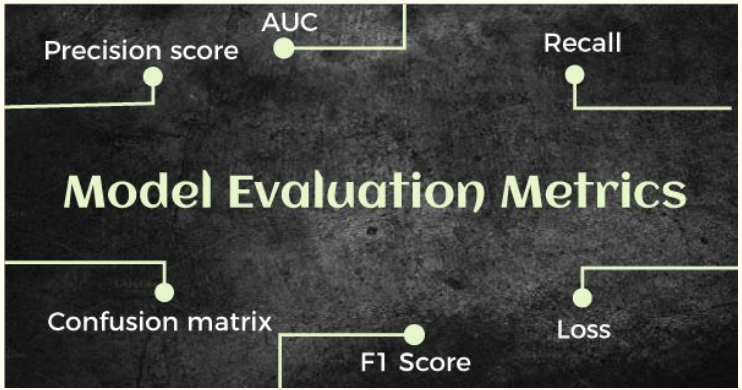
Evolving Tactics

Multimodal

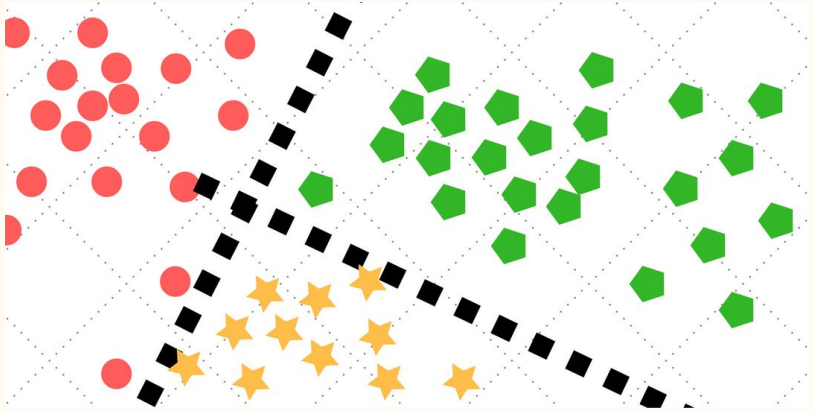
Scope of the Project







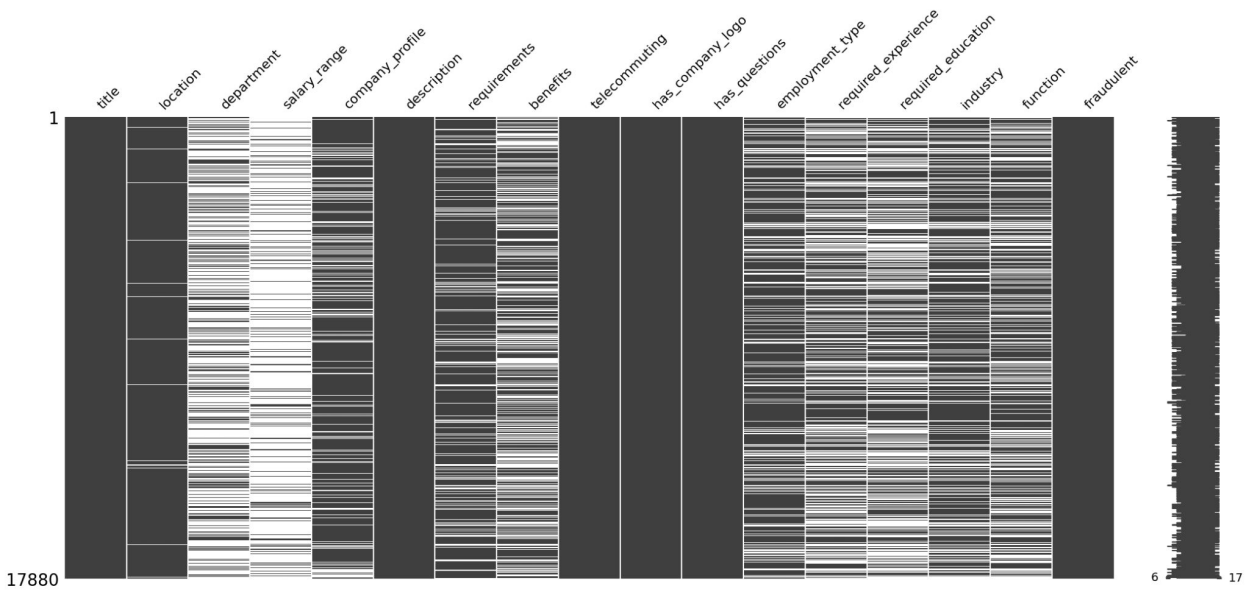
Expected Outcomes



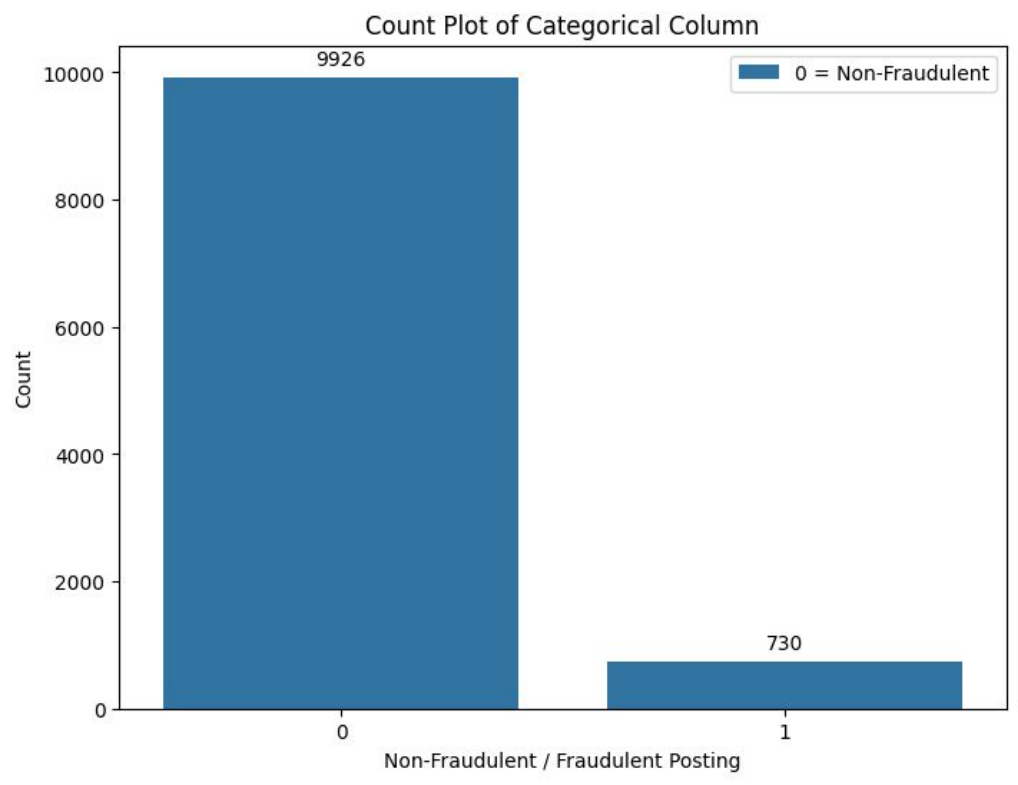
Data Collection and Data Source

Real / Fake Job Posting Prediction:
https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction/data

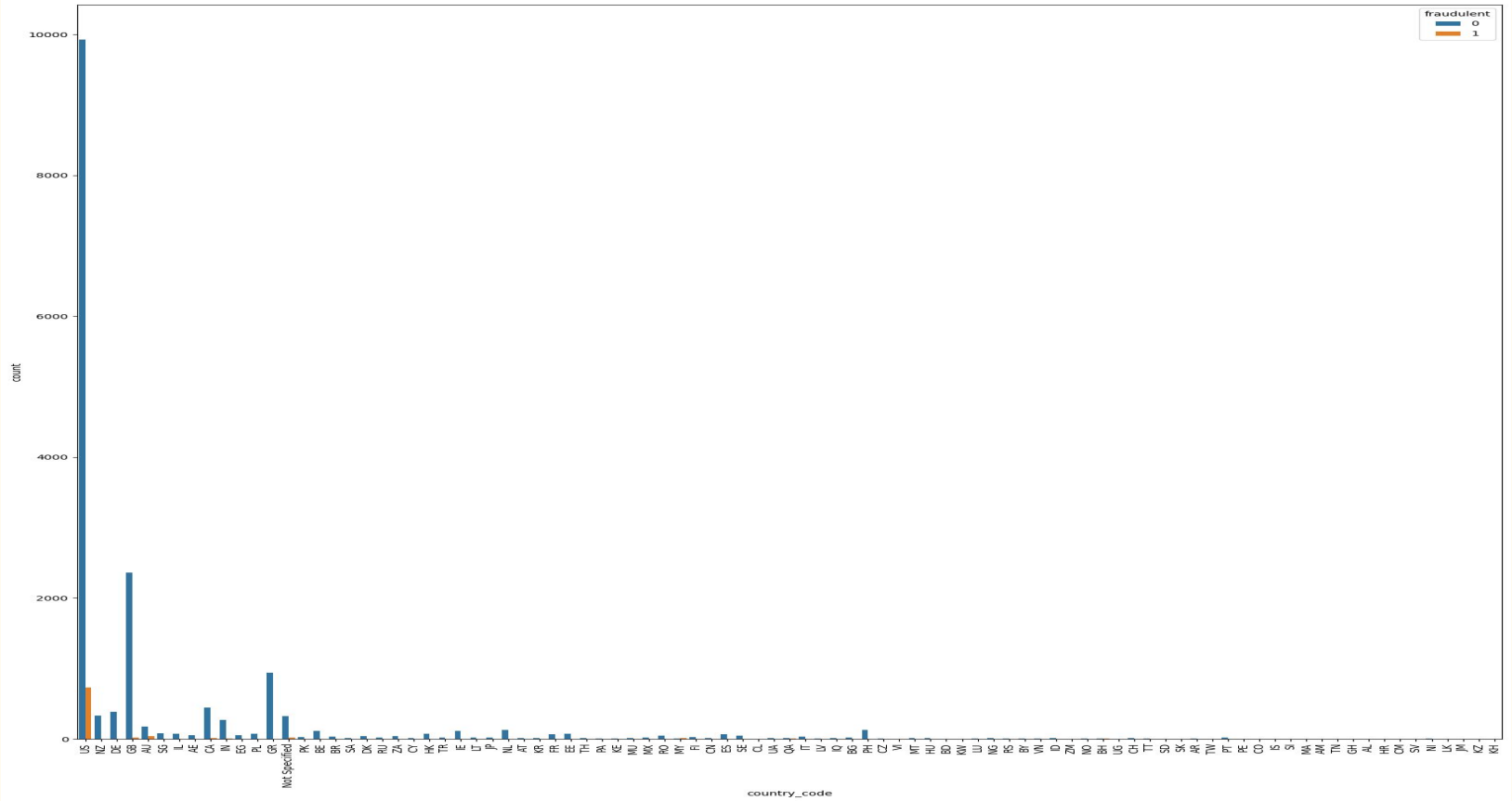
Dealing with missing values



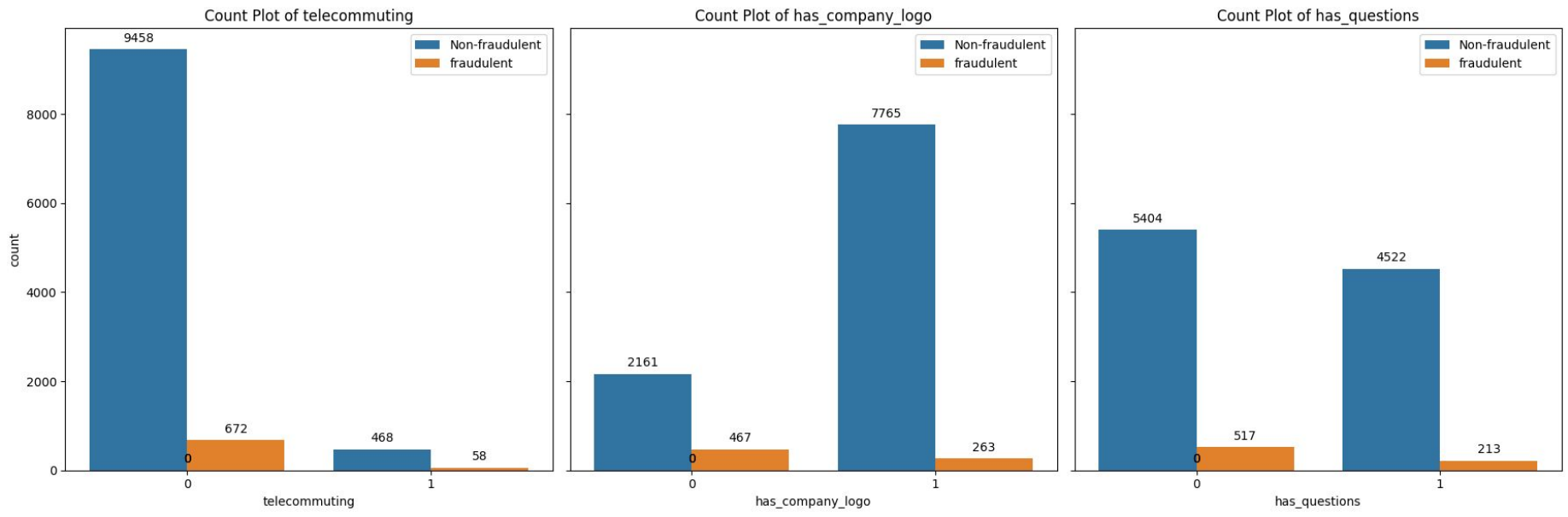
Imbalance Data



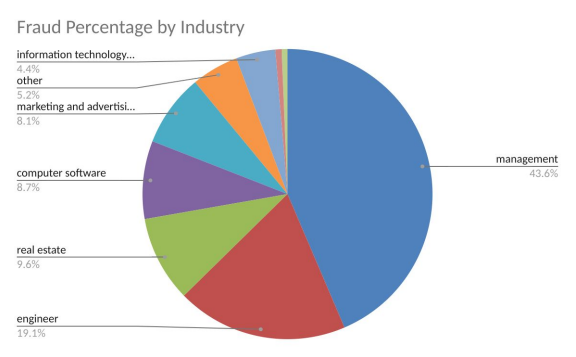
EDA - Job Postings by Country

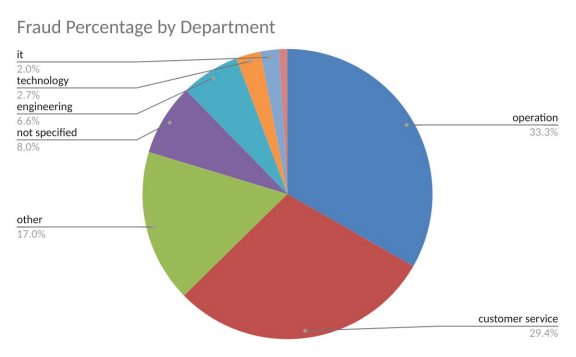


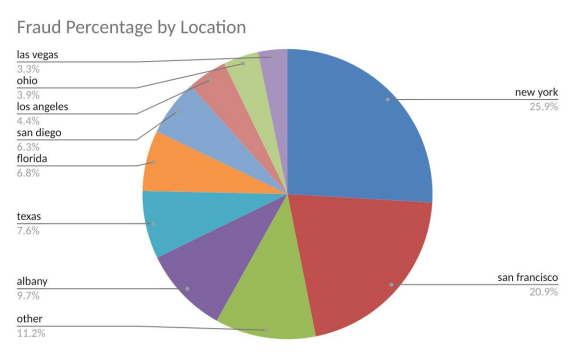
EDA - Numerical Features

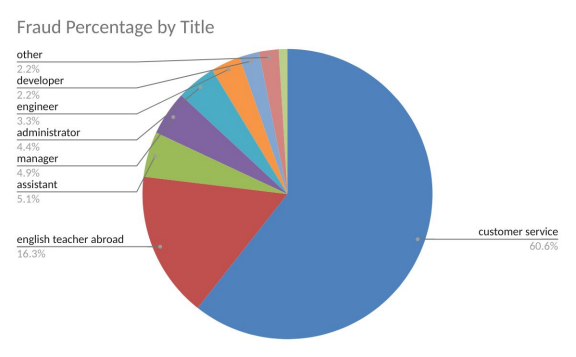


EDA - Categorical Features

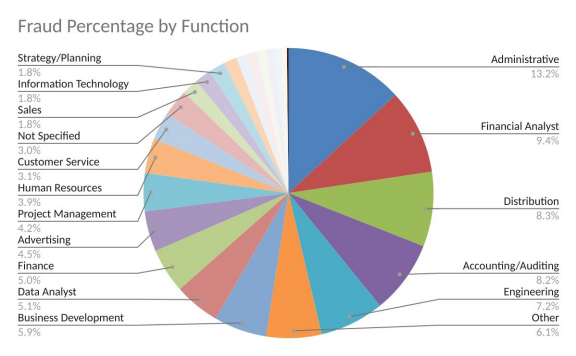


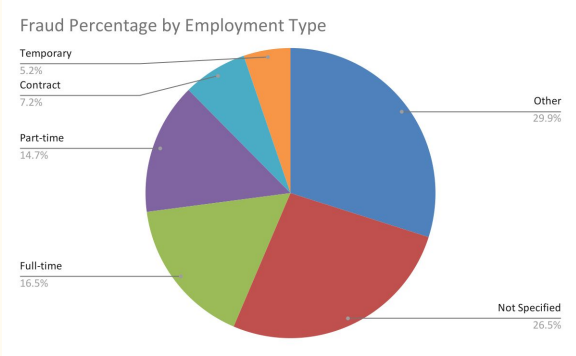


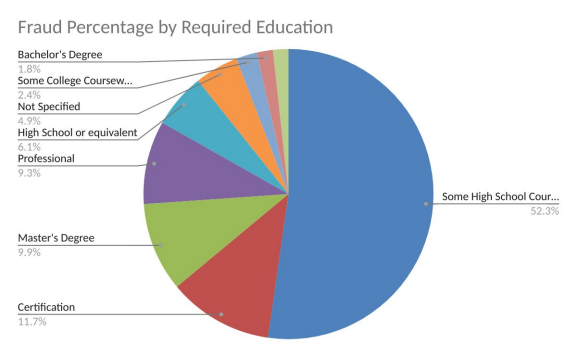


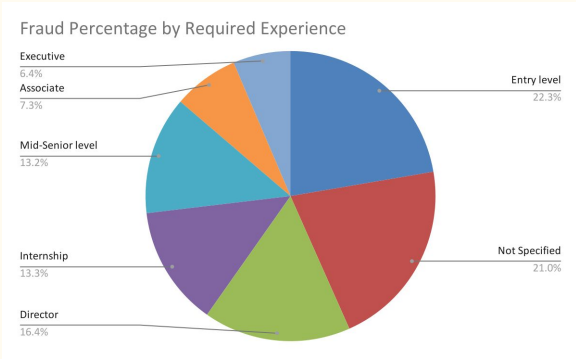


EDA - Categorical Features

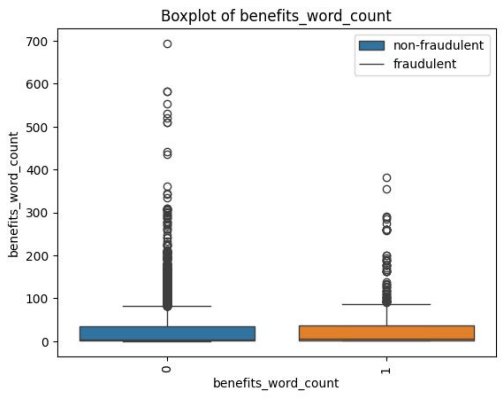


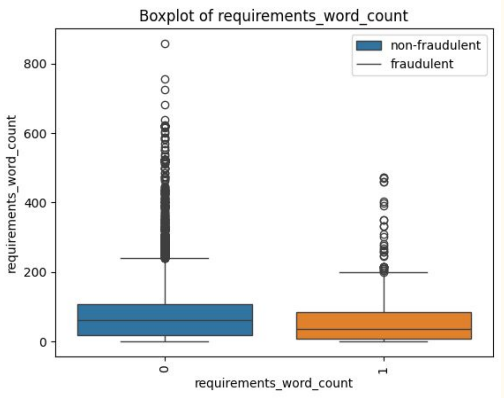


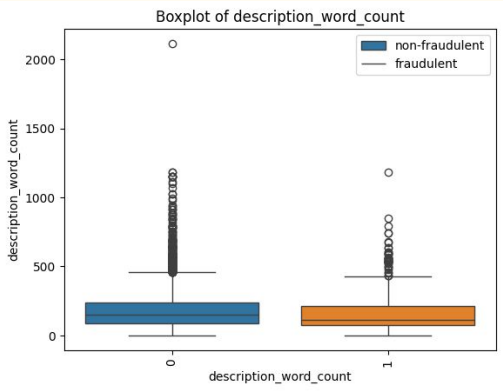


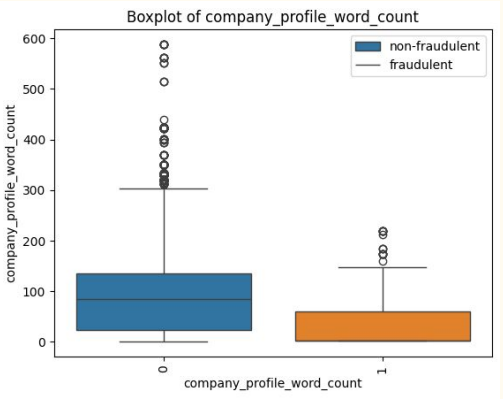


EDA - Text Features



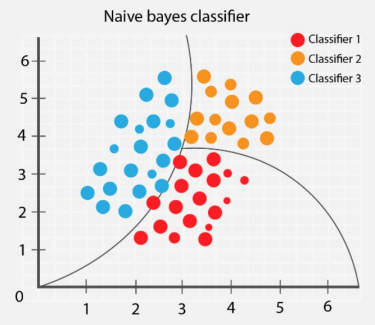


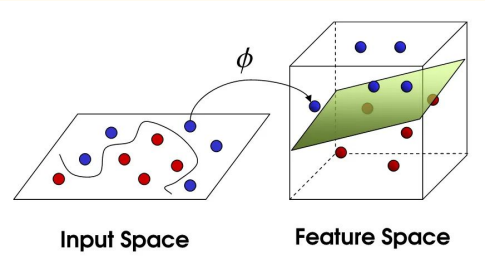


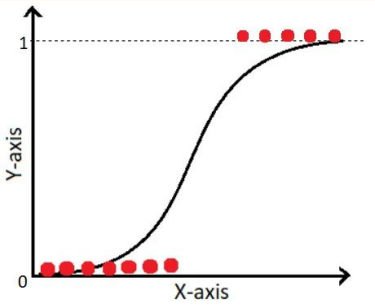


Modeling

- Naive Bayes Classifier for Multinomial
- Linear Support Vector Machine
- Logistic Regression





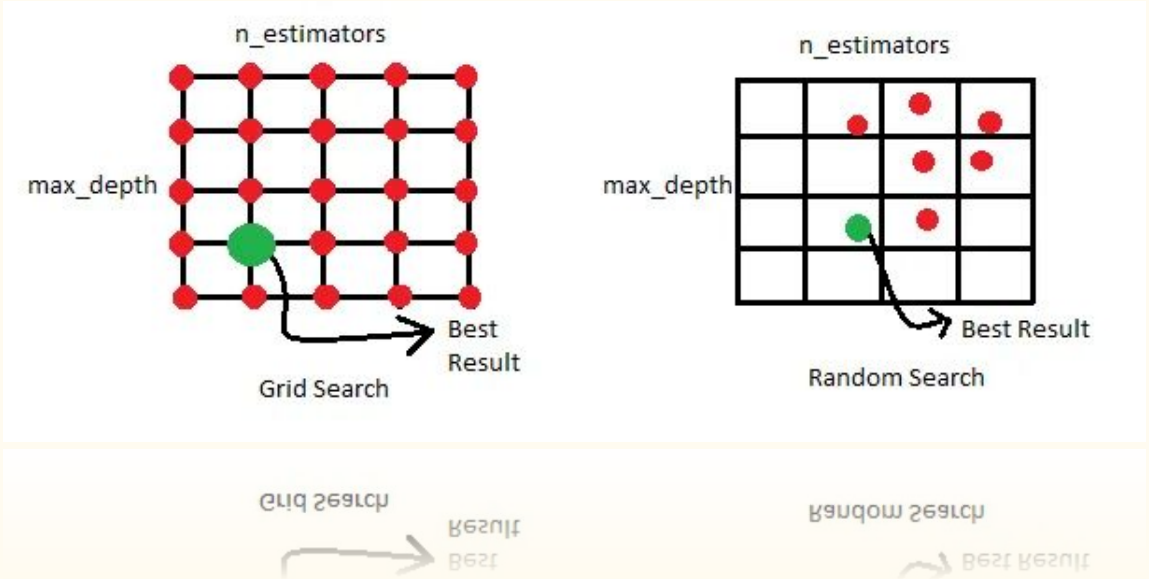


Modeling - Model Assessment

Performance Metrics	**NB**	**SVM**	**Logit**
Recall	0.73	0.6	0.84
F1	0.73	0.73	0.87
Precision	0.73	0.95	0.90
Accuracy	0.96	0.97	0.98

Hyper-parameters Tuning

- GridSearchCV
- RandomizedSearchCV



Hyper-parameters Tuning

GridSearchCV

Wait time	CPU times: user 27.2 s, sys: 5.72 s, total: 32.9 s Wall time: 17min 22s
Optimal Hyperparameters	warm_start: False solver: liblinear C: 10
Recall	0.8424657534246576
F1	0.87

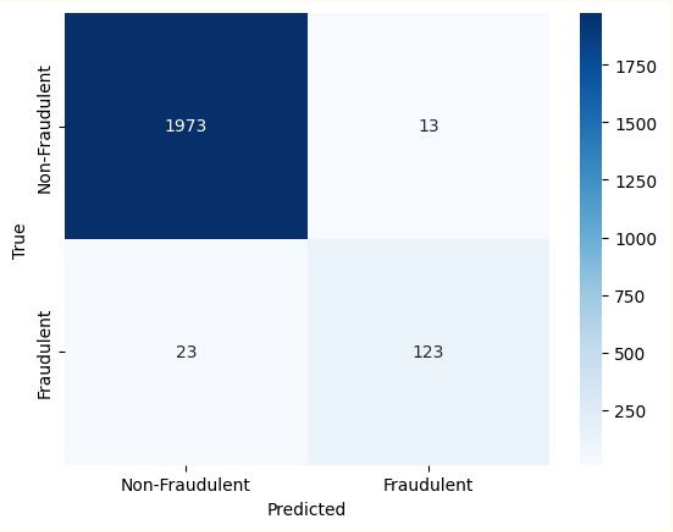
RandomizedSearchCV

Wait time	CPU times: user 15min 55s, sys: 818 ms, total: 15min 56s Wall time: 6min 29s
Optimal Hyperparameters	C: 100 solver: lbfgs warm_start: True
Recall	0.8424657534246576
F1	0.87

Summary

Logistics Regression Model

Parameters	Value
C	100
class_weight	None
dual	False
fit_intercept	True
intercept_scaling	1
l1_ratio	None
max_iter	100
multi_class	auto
n_jobs	None
penalty	l2
random_state	1
solver	lbfgs
tol	0.0001
verbose	0
warm_start	True



Future Works