Authenticity in Job Posting

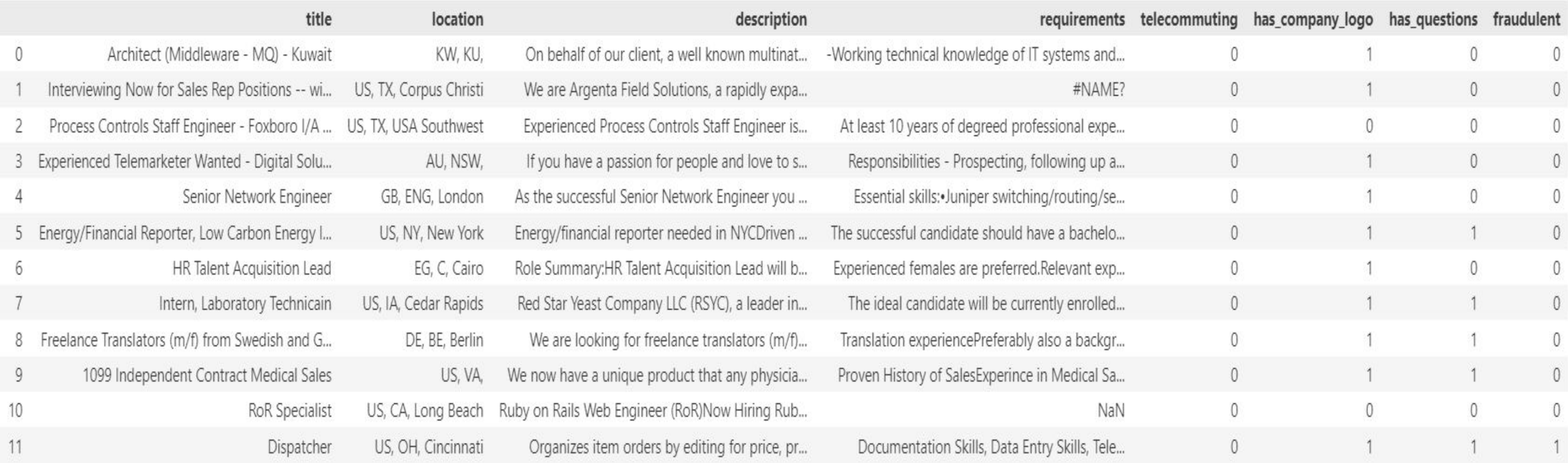
Data-Driven Techniques for Real vs Fake Classification

Presented By

Onkar Shelar

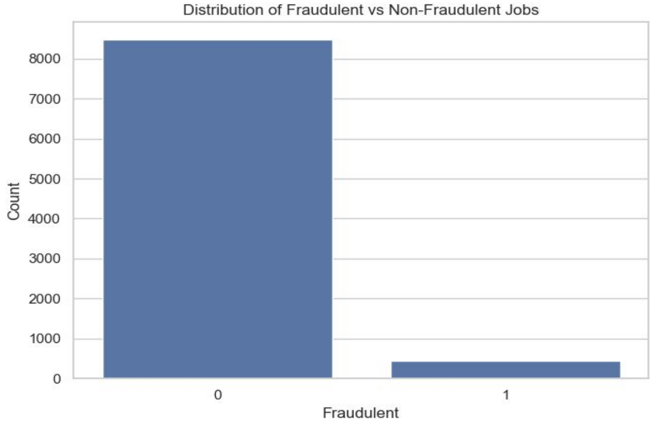
Golisano College of Computing and Information Sciences

Analysis of Dataset

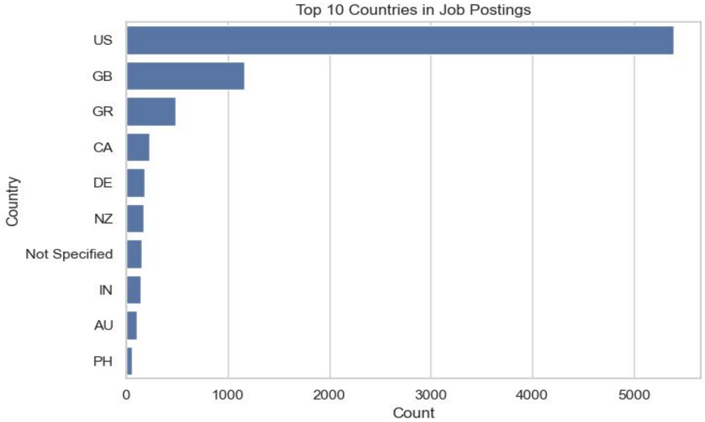


	title	location	description	requirements	telecommuting	has_company_logo	has_questions	fraudulent
0	Architect (Middleware - MQ) - Kuwait	KW, KU,	On behalf of our client, a well known multinat...	-Working technical knowledge of IT systems and...	0	1	0	0
1	Interviewing Now for Sales Rep Positions -- wi...	US, TX, Corpus Christi	We are Argenta Field Solutions, a rapidly expa...	#NAME?	0	1	0	0
2	Process Controls Staff Engineer - Foxboro I/A ...	US, TX, USA Southwest	Experienced Process Controls Staff Engineer is...	At least 10 years of degreed professional expe...	0	0	0	0
3	Experienced Telemarketer Wanted - Digital Solu...	AU, NSW,	If you have a passion for people and love to s...	Responsibilities - Prospecting, following up a...	0	1	0	0
4	Senior Network Engineer	GB, ENG, London	As the successful Senior Network Engineer you ...	Essential skills:•Juniper switching/routing/se...	0	1	0	0
5	Energy/Financial Reporter, Low Carbon Energy I...	US, NY, New York	Energy/financial reporter needed in NYCDriven ...	The successful candidate should have a bachelo...	0	1	1	0
6	HR Talent Acquisition Lead	EG, C, Cairo	Role Summary:HR Talent Acquisition Lead will b...	Experienced females are preferred.Relevant exp...	0	1	0	0
7	Intern, Laboratory Technicain	US, IA, Cedar Rapids	Red Star Yeast Company LLC (RSYC), a leader in...	The ideal candidate will be currently enrolled...	0	1	1	0
8	Freelance Translators (m/f) from Swedish and G...	DE, BE, Berlin	We are looking for freelance translators (m/f)...	Translation experiencePreferably also a backgr...	0	1	1	0
9	1099 Independent Contract Medical Sales	US, VA,	We now have a unique product that any physicia...	Proven History of SalesExperince in Medical Sa...	0	1	1	0
10	RoR Specialist	US, CA, Long Beach	Ruby on Rails Web Engineer (RoR)Now Hiring Rub...	NaN	0	0	0	0
11	Dispatcher	US, OH, Cincinnati	Organizes item orders by editing for price, pr...	Documentation Skills, Data Entry Skills, Tele...	0	1	1	1

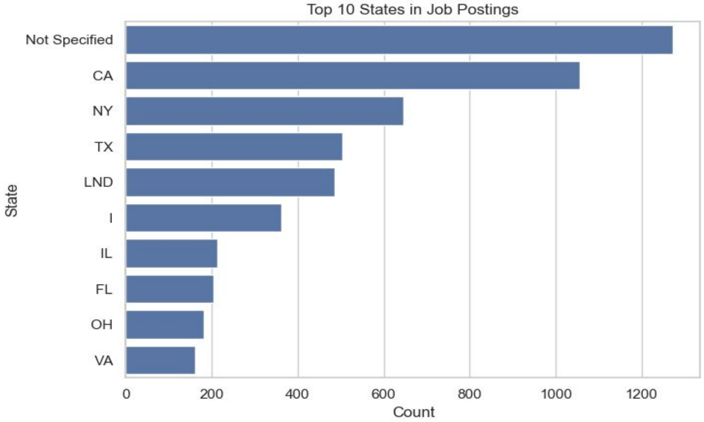
- Location, Description and Requirements columns have missing values
- The dataset contains job posting records for more than 6100 unique positions, 2100 unique locations, 75 unique countries, 268 unique states and 1682 unique cities



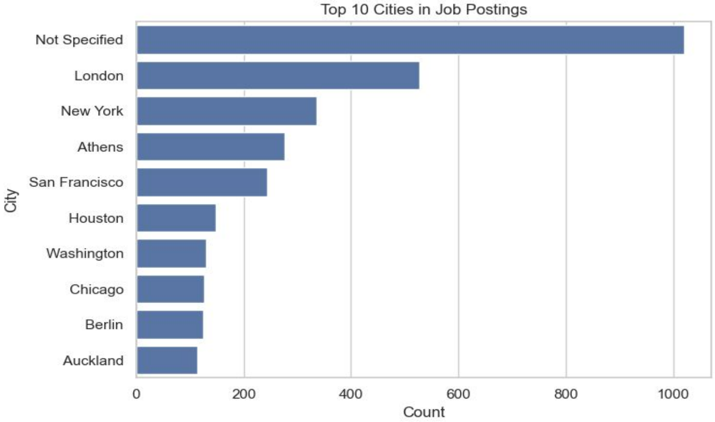
Distribution of Fraudulent vs Non-Fraudulent Jobs



Top 10 Countries in Job Postings

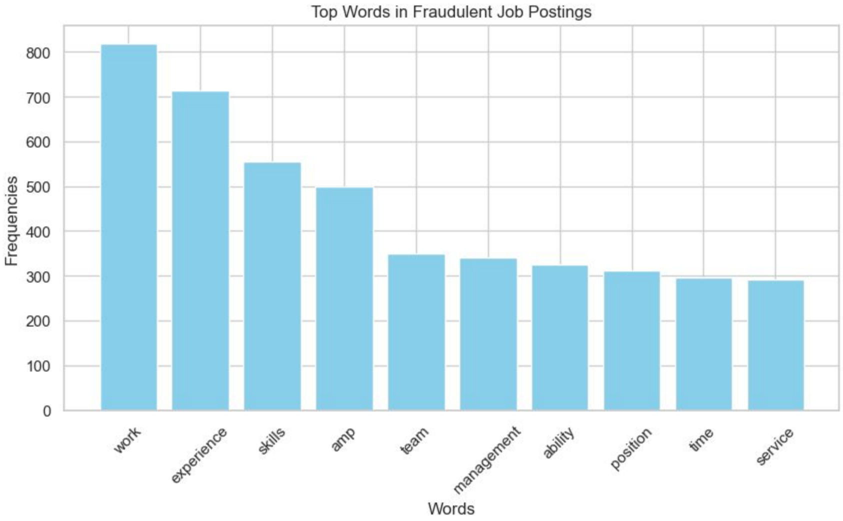


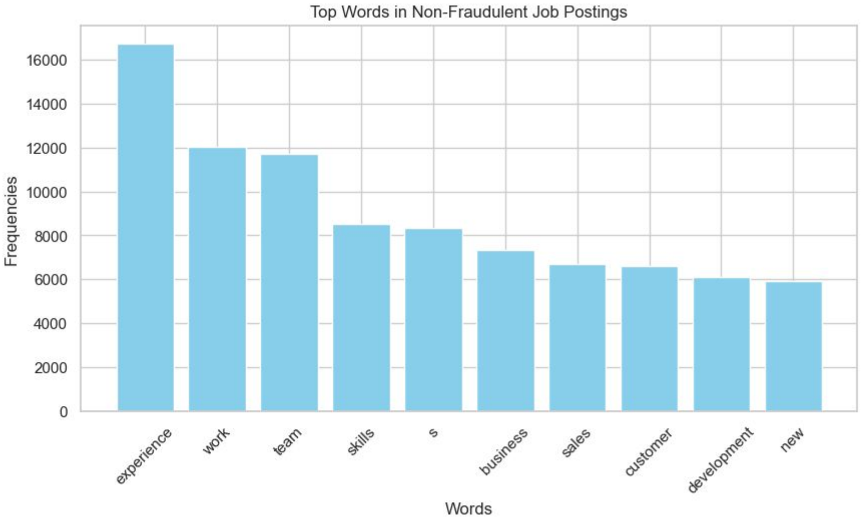
Top 10 States in Job Postings



Top 10 Cities in Job Postings

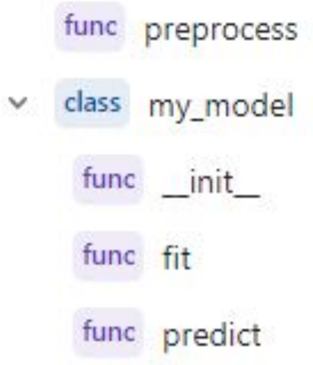
- Postings without company logo and skimming questions are likely to be fraudulent ones
- Fraudulent job postings have shorter description and requirements compared to non-fraudulent ones
- Fraudulent job postings have more generic terms comparatively





Model Architecture

- preprocess()
  - Handles cleaning, feature extraction, final dataset preparation
- My_model
  - Encapsulates machine learning pipeline
- __init__
  - Setting up TF-IDF vectorizer and RandomForest classifier
- fit()
  - Training the model on preprocessed data
- predict()
  - Making the predictions on new data



func preprocess

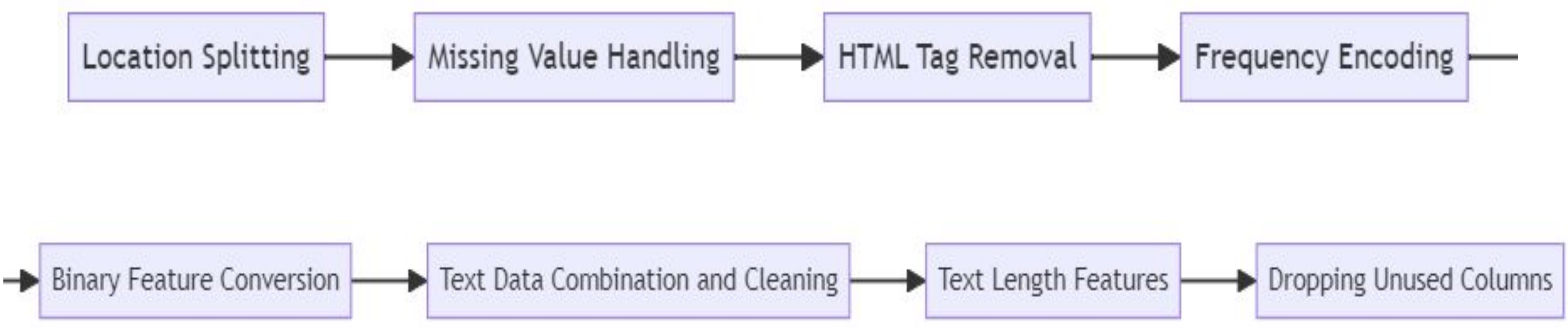
class my_model

func __init__

func fit

func predict

preprocessing()



Location Splitting

Missing Value Handling

HTML Tag Removal

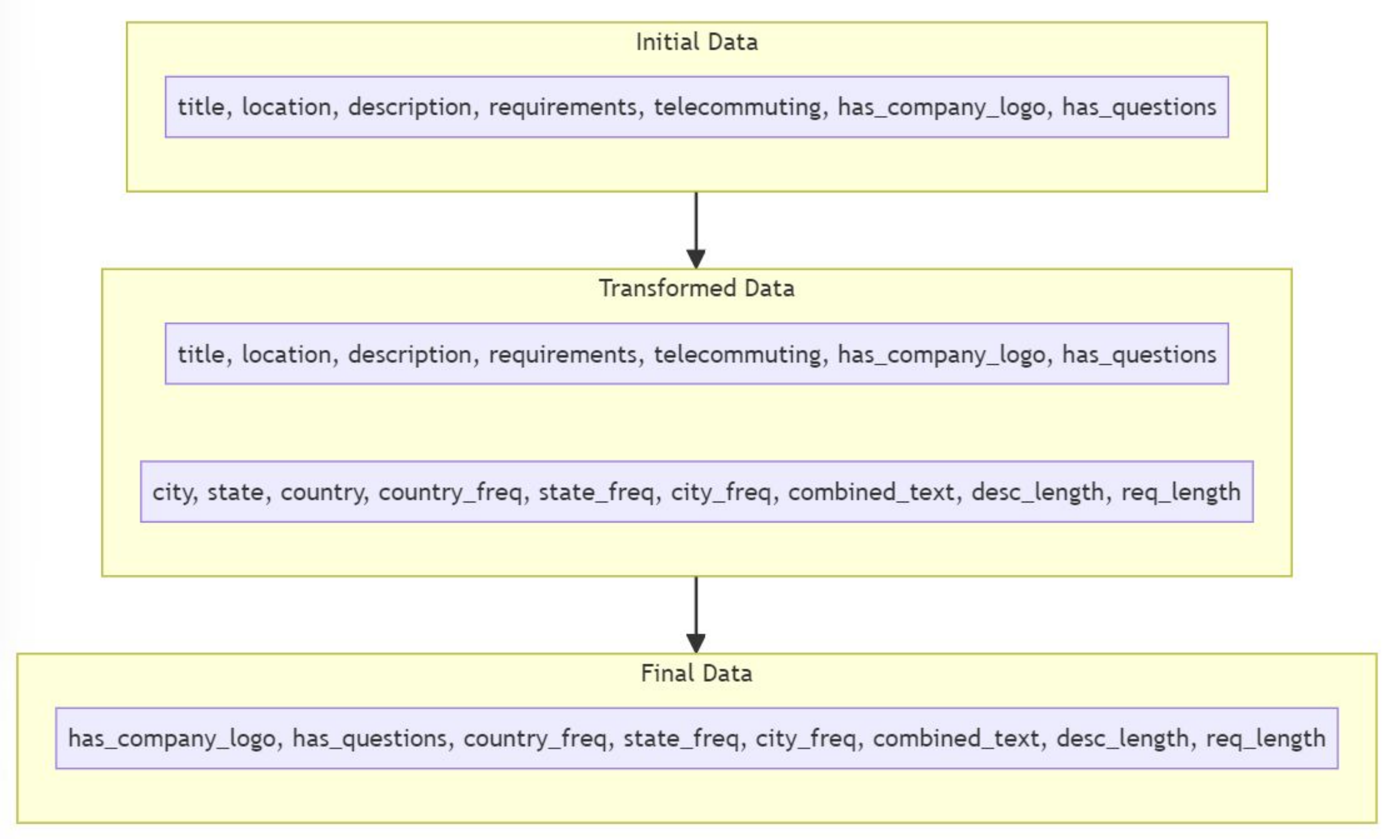
Frequency Encoding

Binary Feature Conversion

Text Data Combination and Cleaning

Text Length Features

Dropping Unused Columns



Initial Data

title, location, description, requirements, telecommuting, has_company_logo, has_questions

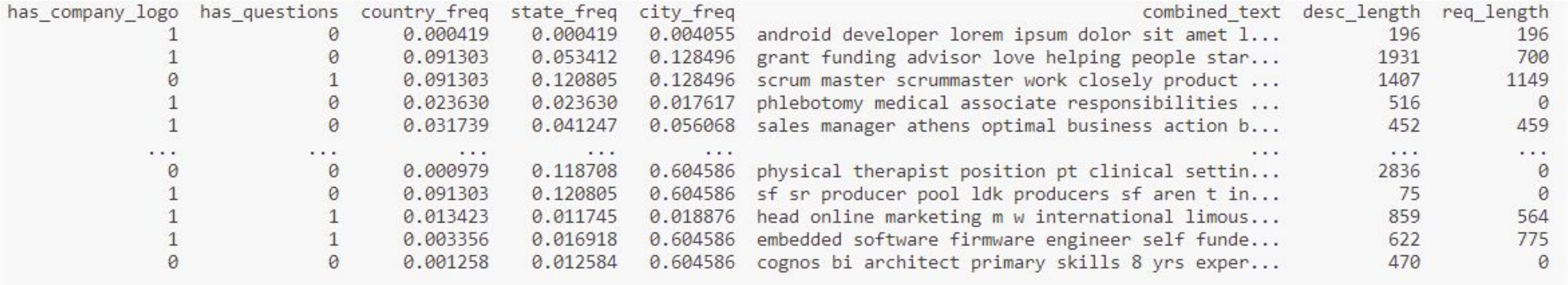
Transformed Data

title, location, description, requirements, telecommuting, has_company_logo, has_questions

city, state, country, country_freq, state_freq, city_freq, combined_text, desc_length, req_length

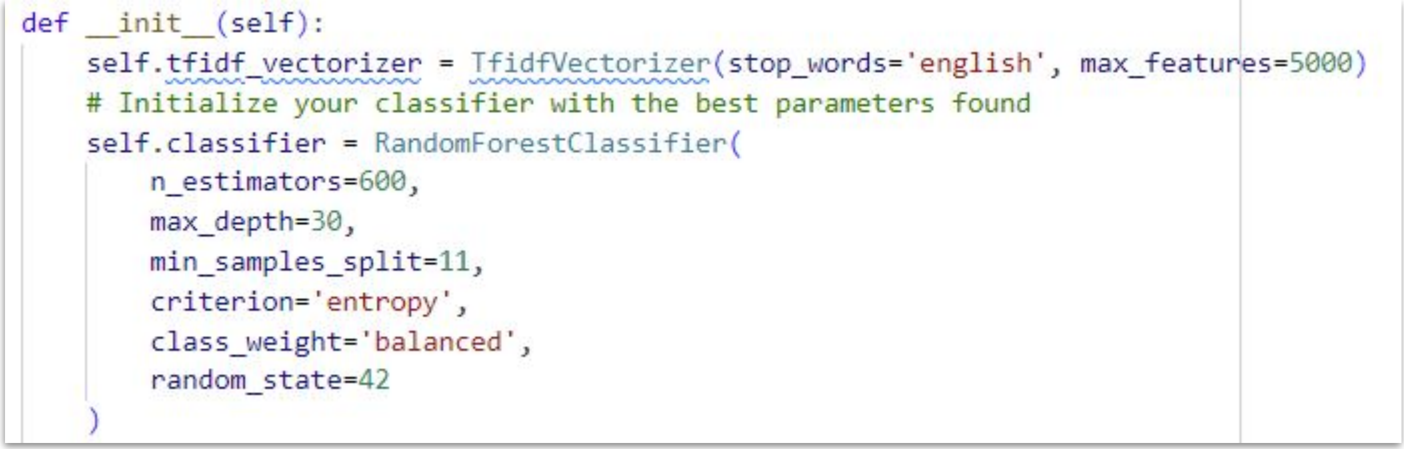
Final Data

has_company_logo, has_questions, country_freq, state_freq, city_freq, combined_text, desc_length, req_length



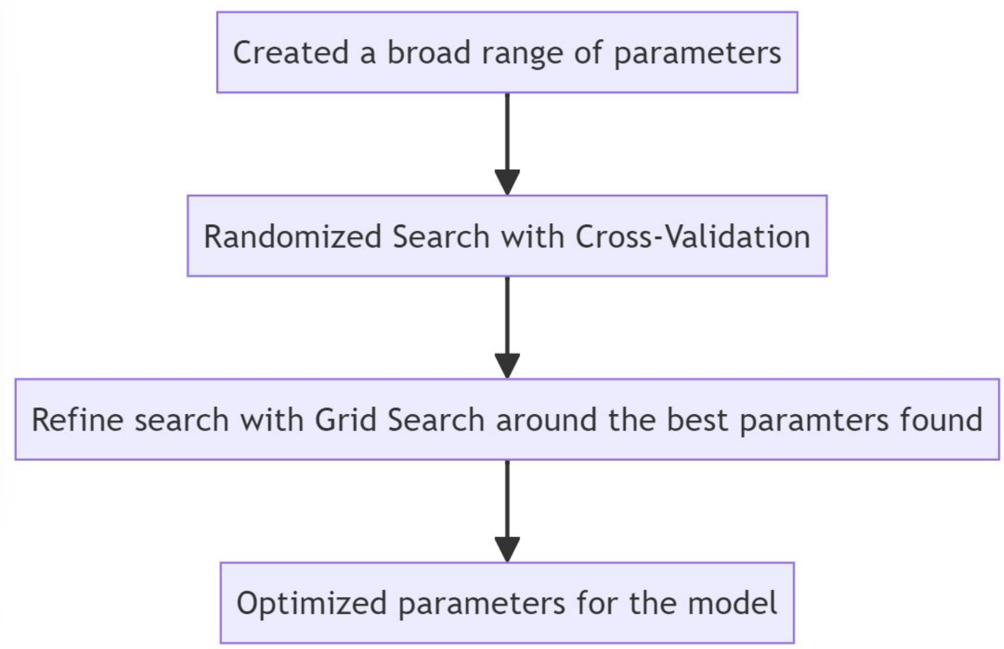
```
has_company_logo  has_questions  country_freq  state_freq  city_freq                                     combined_text  desc_length  req_length
               1              0      0.000419    0.000419   0.004055  android developer lorem ipsum dolor sit amet l...          196         196
               1              0      0.091303    0.053412   0.128496  grant funding advisor love helping people star...         1931         700
               0              1      0.091303    0.120805   0.128496  scrum master scrummaster work closely product ...         1407        1149
               1              0      0.023630    0.023630   0.017617  phlebotomy medical associate responsibilities ...          516           0
               1              0      0.031739    0.041247   0.056068  sales manager athens optimal business action b...          452         459
             ...            ...           ...         ...        ...                                               ...          ...         ...
               0              0      0.000979    0.118708   0.604586  physical therapist position pt clinical settin...         2836           0
               1              0      0.091303    0.120805   0.604586  sf sr producer pool ldk producers sf aren t in...           75           0
               1              1      0.013423    0.011745   0.018876  head online marketing m w international limous...          859         564
               1              1      0.003356    0.016918   0.604586  embedded software firmware engineer self funde...          622         775
               0              0      0.001258    0.012584   0.604586  cognos bi architect primary skills 8 yrs exper...          470           0
```

__init__()



```python
def __init__(self):
    self.tfidf_vectorizer = TfidfVectorizer(stop_words='english', max_features=5000)
    # Initialize your classifier with the best parameters found
    self.classifier = RandomForestClassifier(
        n_estimators=600,
        max_depth=30,
        min_samples_split=11,
        criterion='entropy',
        class_weight='balanced',
        random_state=42
    )
```

Model Optimization

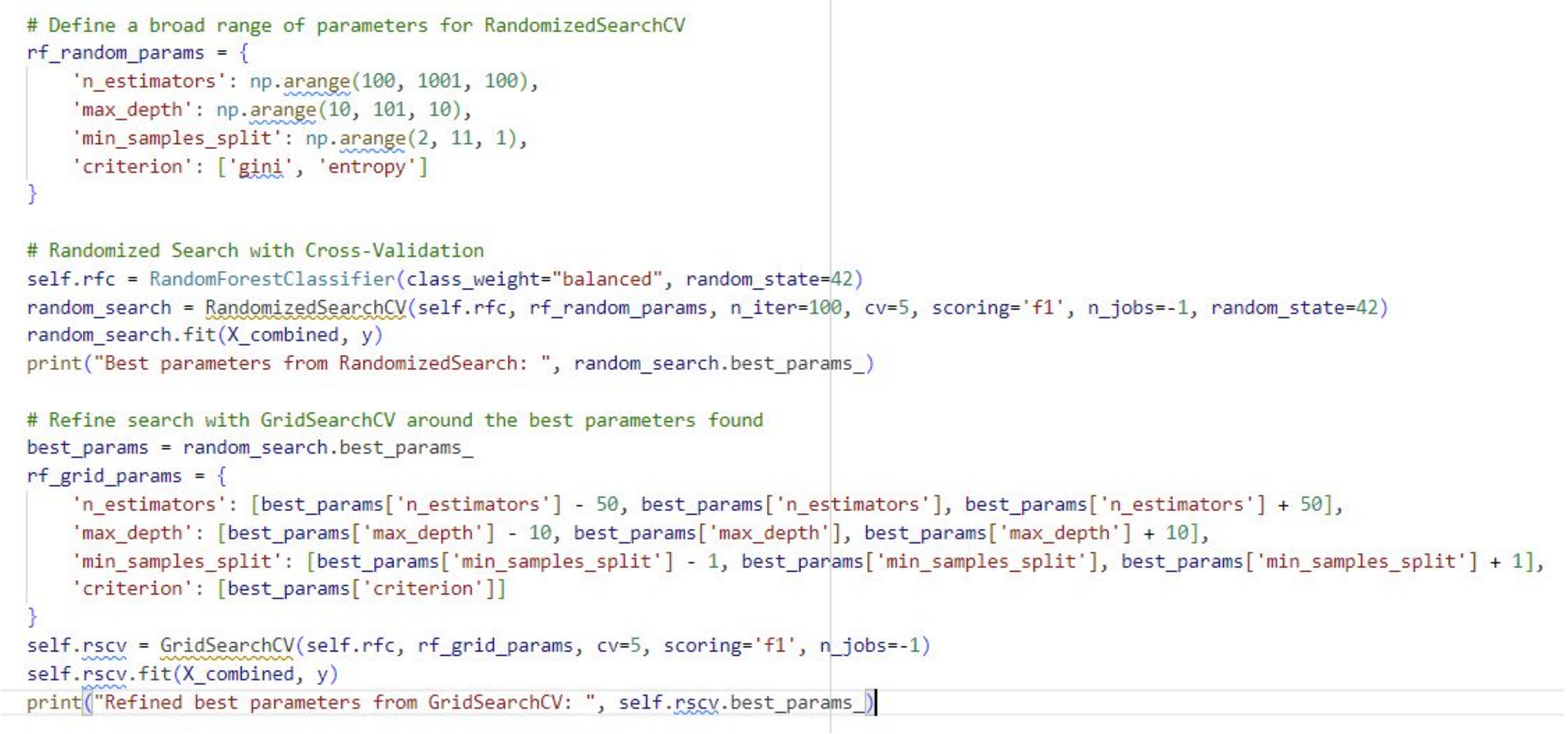


Created a broad range of parameters

Randomized Search with Cross-Validation

Refine search with Grid Search around the best paramters found

Optimized parameters for the model

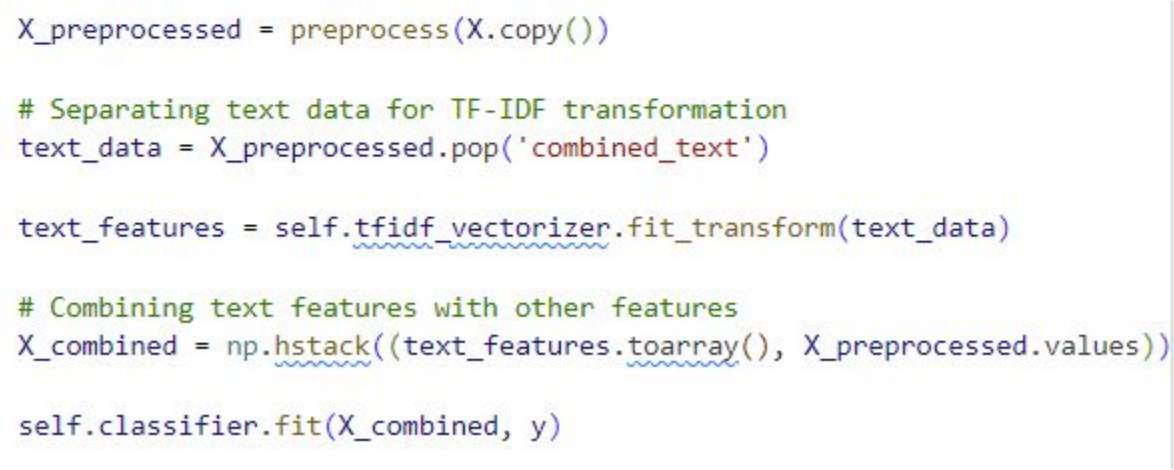


```python
# Define a broad range of parameters for RandomizedSearchCV
rf_random_params = {
    'n_estimators': np.arange(100, 1001, 100),
    'max_depth': np.arange(10, 101, 10),
    'min_samples_split': np.arange(2, 11, 1),
    'criterion': ['gini', 'entropy']
}

# Randomized Search with Cross-Validation
self.rfc = RandomForestClassifier(class_weight="balanced", random_state=42)
random_search = RandomizedSearchCV(self.rfc, rf_random_params, n_iter=100, cv=5, scoring='f1', n_jobs=-1, random_state=42)
random_search.fit(X_combined, y)
print("Best parameters from RandomizedSearch: ", random_search.best_params_)

# Refine search with GridSearchCV around the best parameters found
best_params = random_search.best_params_
rf_grid_params = {
    'n_estimators': [best_params['n_estimators'] - 50, best_params['n_estimators'], best_params['n_estimators'] + 50],
    'max_depth': [best_params['max_depth'] - 10, best_params['max_depth'], best_params['max_depth'] + 10],
    'min_samples_split': [best_params['min_samples_split'] - 1, best_params['min_samples_split'], best_params['min_samples_split'] + 1],
    'criterion': [best_params['criterion']]
}
self.rscv = GridSearchCV(self.rfc, rf_grid_params, cv=5, scoring='f1', n_jobs=-1)
self.rscv.fit(X_combined, y)
print("Refined best parameters from GridSearchCV: ", self.rscv.best_params_)
```

fit() & predict()



```
X_preprocessed = preprocess(X.copy())

# Separating text data for TF-IDF transformation
text_data = X_preprocessed.pop('combined_text')

text_features = self.tfidf_vectorizer.fit_transform(text_data)

# Combining text features with other features
X_combined = np.hstack((text_features.toarray(), X_preprocessed.values))

self.classifier.fit(X_combined, y)
```

- Average computation time is 1.5 minutes
- Average F1 score is 0.72

Next Steps for Enhancement

- Algorithmic Expansion
- Hyperparameter Optimization
- Model Interpretability

Project Files

Data Analysis - https://github.com/Onkar2102/DSCI-633/blob/main/assignments/project/DataUnderstanding.ipynb

ML Model - https://github.com/Onkar2102/DSCI-633/blob/main/assignments/project/project.py

Thank You!