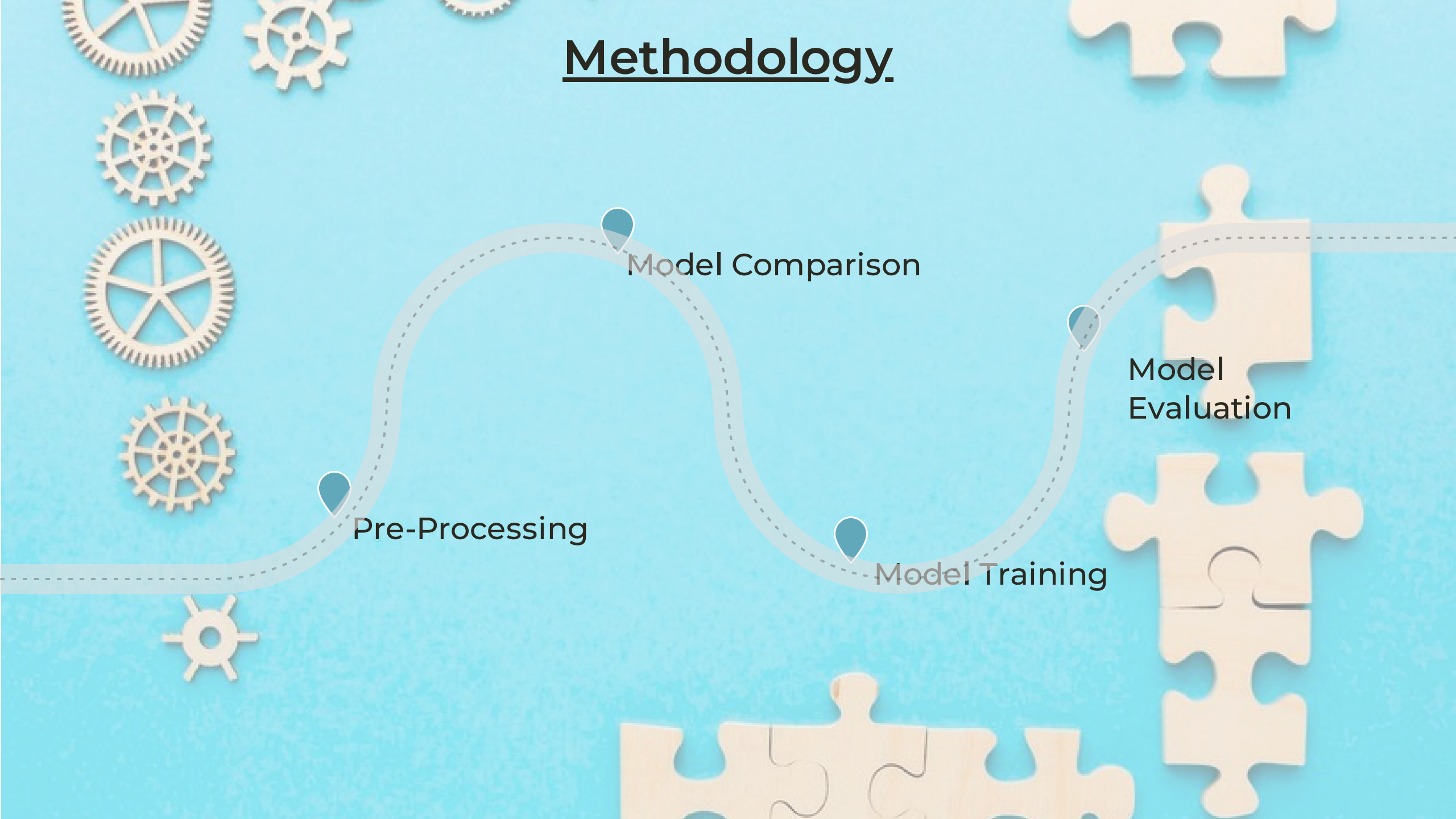
Machine Learning for Fraud Detection in Job Postings

Developing an ML model to identify fraudulent job postings, enhancing hiring integrity while protecting job seekers.

Methodology



Data Preprocessing

Converted all text to lowercase for uniformity

Used NLTK's RegexpTokenizer to break text into words, systematically excluding punctuation

Applied WordNetLemmatizer to reduce words to their lemma form, improving model's ability to understand standard word forms

Employed TfidfVectorizer to transform text into numeric format, calculating TF-IDF scores for words, phrases, and trigrams

Used SelectKBest with ANOVA F-test (f_classif) to select top features based on statistical significance to outcome variable, reducing feature space to most predictive features

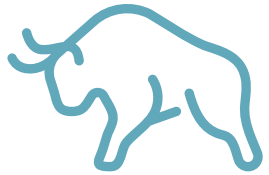
Model Comparison

Model	F1 Score	Training Time	Hyperparameters
SGDClassifier	0.701	5.55	ngram range: (1,3) max_features: 10000 max_num_iteration: 200 population_size: 50 mutation_probability: 0.2
RandomForest Classifier	0.62	8.43	n_estimators: 100, max_features: 'auto'
DecisionTree Classifier	0.64	6.27	max depth= 10 min_samples_split= 2

Model Training

SGDClassifier (Stochastic Gradient Descent Classifier) used

Efficient for handling large text datasets and binary classification tasks.



Genetic algorithm for hyperparameter tuning

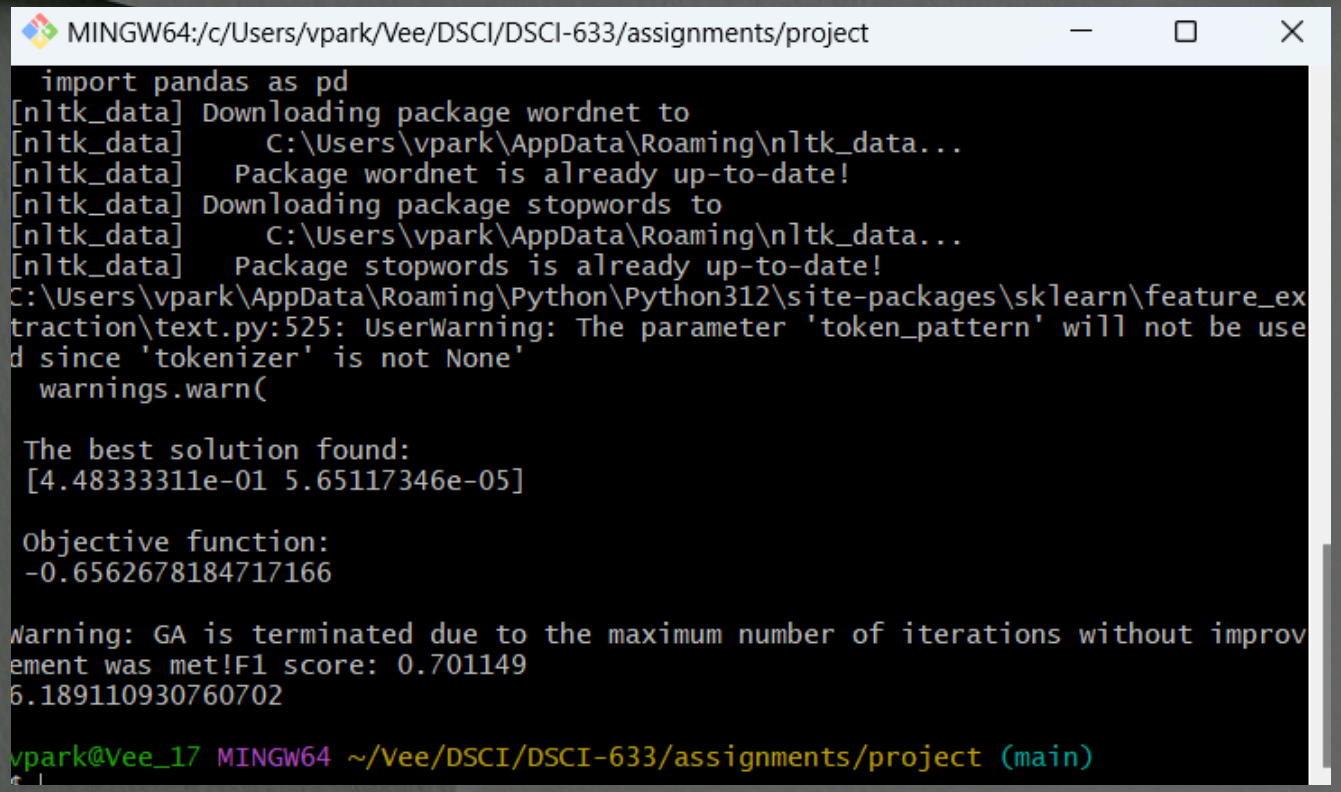
Systematically explored parameter space to maximize F1 score, balancing precision and recall.

Tuned regularization parameters

Optimized l2 regularization strength (alpha) to prevent overfitting and underfitting.

The SGD Classifier with optimized regularization achieved high performance in detecting fraudulent job postings from large text data.

Performance metrics



Using SGD Classifier for fraud detection in job postings attained an F1 score of 0.701149, indicating a reasonable balance between precision and recall. We conducted a granular analysis of the model's performance for both fake and legitimate job postings, along with a breakdown of the computational runtime.

The entire process, from data preprocessing to model training and prediction, took approximately 05.55 minutes, showcasing the efficiency of the selected techniques and algorithms.

CONCLUSION

1. Impact of Machine Learning - Protects job seekers and maintains platform integrity.

2. Scalability and Efficiency- Scalable approach with efficient hyperparameter tuning using genetic algorithms.

3. Future Directions - Explore advanced techniques like deep learning for enhanced fraud detection.