Detecting Fraud in Job Postings

2022 April 26

Problem Statement:

- A company has requested a predictive model that can detect whether job postings are legitimate or fraudulent.
- Concerns have been raised from recruiters, advertisers, and subscription-based members about fraudulent job postings that undermine the reputation of the company's services.

Data Source:

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

Key Metric:

- Recall of the minority class (minimizing false negatives)
- Precision is less of a concern given that the company has a well-developed appeals process for flagged legitimate job postings

Data Wrangling/Exploratory Data Analysis:

- Very imbalanced classes (less than 5% of the positive class)
- Very few positive class records in total (less than 1000)
- Many columns with null values

↓

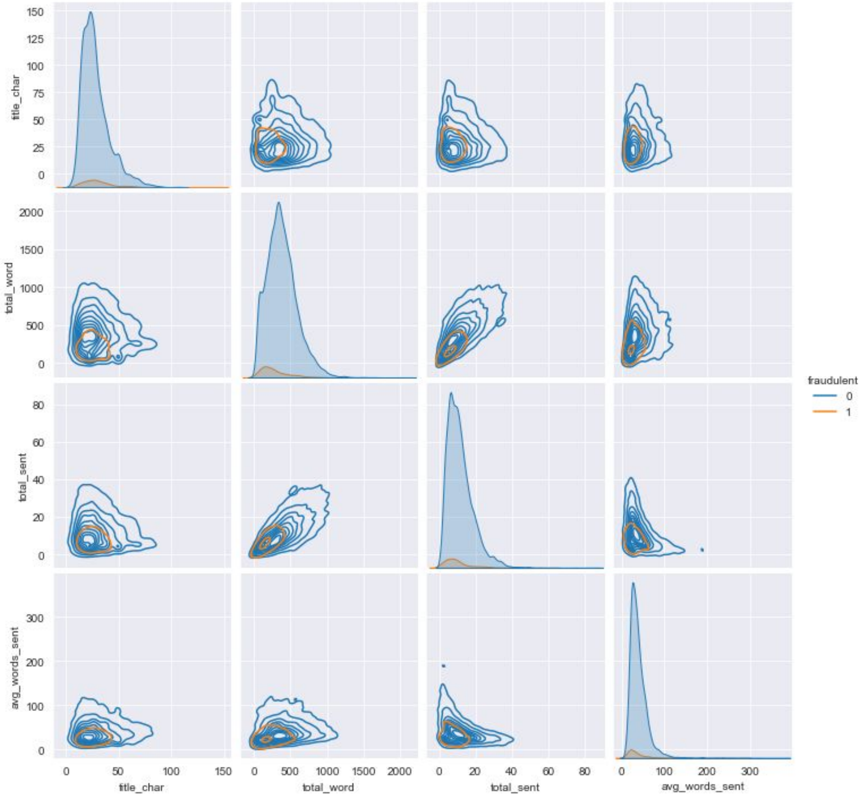
- Simple imputation strategies were avoided in order to retain as much information as possible for the extremely small minority class.

Natural Language Processing:

- Numeric features were engineered from the five text columns using NLTK
  - Total characters from the title column, total words, total sentences, and average words per sentence

- Additional natural language processing was considered but rejected
- Further tokenization/lemmatization/stemming to standardize the text may have inadvertently compromised features from the text that actually predict fraud
  - Ex. Fraudulent job postings may include more errors than legitimate job postings

Pair Plot of Numeric Features From Text Columns:



- In general, it looks like fraudulent job postings tend to be shorter overall

Preprocessing

- ColumnTransformer
  - Text variables: TfidfVectorizer
  - Numeric variables: MinMaxScaler
  - Categorical variables: OneHotEncoder

Modeling

- Imbalanced-learn Pipeline:
  - SMOTE (oversampling the minority class)

- Naive Bayes Model:
  - ComplementNB
  - Chosen as the initial model for its relative strength in both text classification and imbalanced datasets

Initial Model Performance

- Key metric (recall of the minority class): 0.87
- Precision was lower overall than recall (but also not as relevant to the problem statement)

↓

- Additional models were not be considered
- Hyperparameter tuning: GridSearchCV

Hyperparameter Tuning

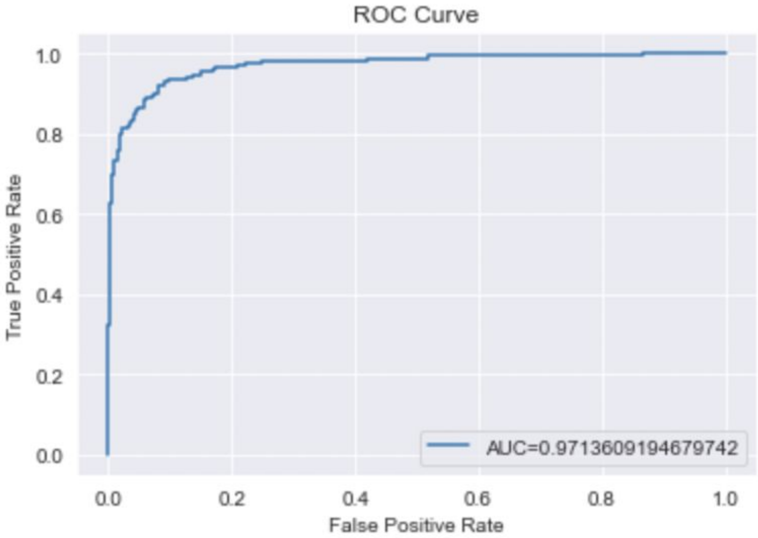
- Complement NB: alpha parameter (smoothing for potential zero probability values in the test set)
- Custom Scorer: Recall of the minority class
- Optimal value: 0.9
- Metrics overall only improved slightly

↓

- Additional tuning not likely to increase model performance

Final Model Performance

- Key metric (recall of the minority class): 87%
- Overall accuracy: 94%



Conclusion

- Only 23 fraudulent job postings (out of the 171) were missed (misclassified as legitimate) in the test set
- Model performance was likely affected but unseen text values, though a higher smoothing parameter value (0.9) has helped.
- Given that the company is primarily concerned with identifying as many of the fraudulent job postings as possible (even at the cost flagging some legitimate job postings), the final model was selected despite lower precision scores