Capstone Two: Final Report

Project Title:
Detecting Fraud in Job Postings

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

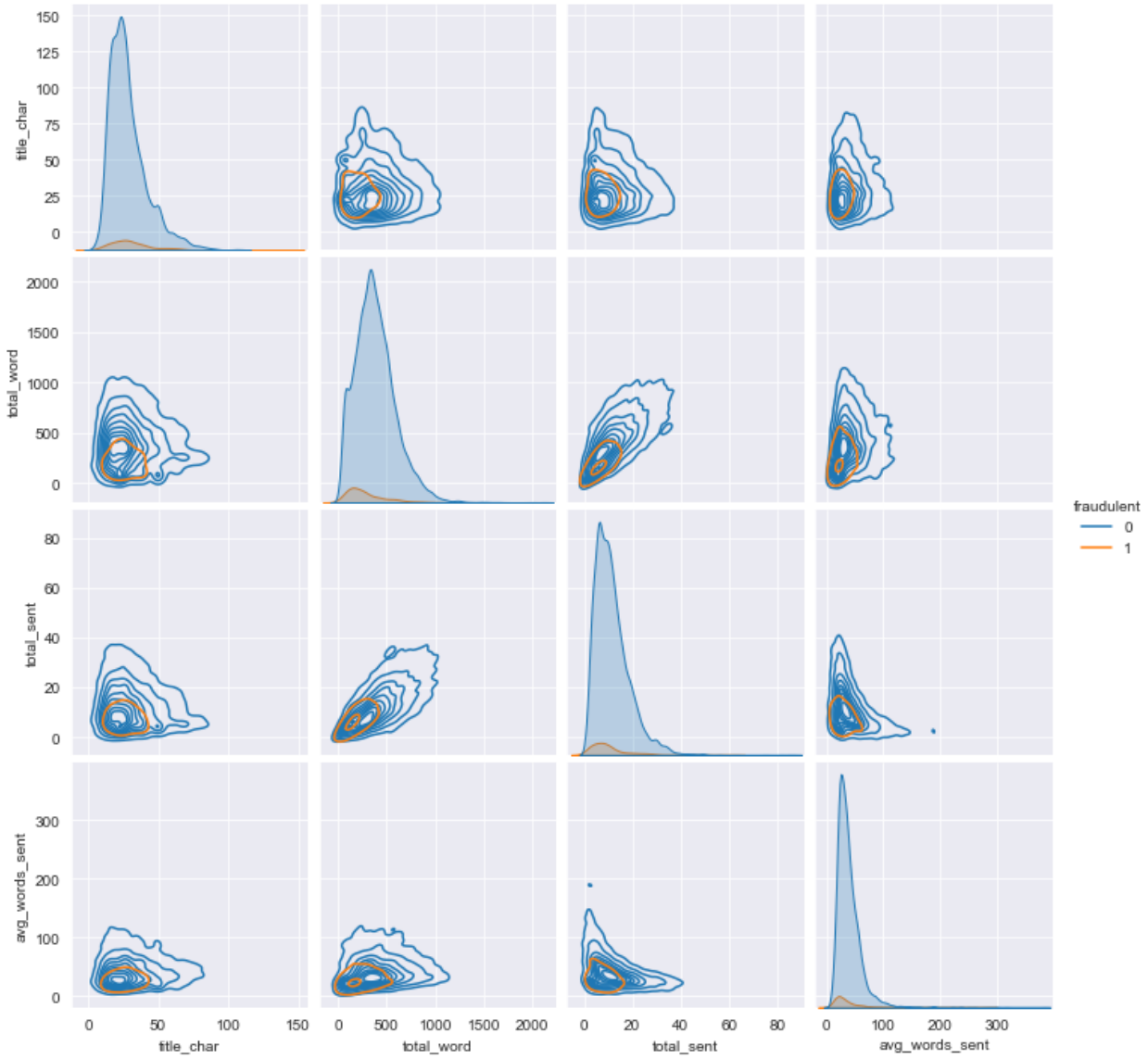
Problem Statement:
A company has requested a predictive model that can detect whether job postings are legitimate or fraudulent. Concerns have been raised from recruiters, advertisers, and subscription-based members about fraudulent job postings that undermine the reputation of the company's services.

Key Metric:
Recall of the minority class (minimizing false negatives)

Data Wrangling/Exploratory Data Analysis:
There were many null values for several explanatory variables, though no null values for the target variable 'fraudulent.' However, the classes for the target variable were not only very imbalanced (fraudulent job postings <5% of the data), but there were very few records for fraudulent job postings in total (856/17880). Thus, 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 columns with text using NLTK: total characters from the title column, total words (from all text columns combined), 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: title character lenghth, word count, sentence count, and average words per sentence are lower than non-fraudulent postings

Preprocessing:

After the train/test split of the data (in order to avoid data leakage) the combined text column was transformed using the TFIDF Vectorizer , chosen because frequency is likely an important factor for fraud detection. Numeric columns were transformed using the MinMaxScaler. Lastly, categorical variables were transformed using OneHotEncoder. All preprocessing steps were added to ColumnTransformer.

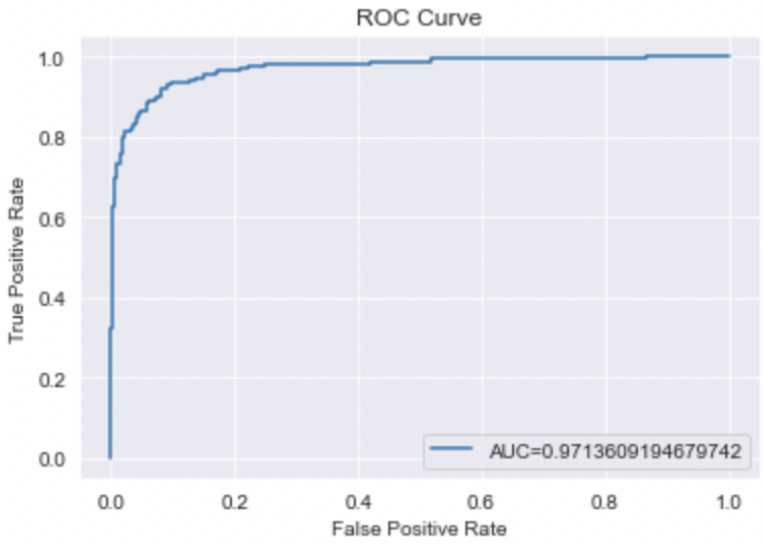
Modeling:

An imbalanced-learn pipeline was created in order to oversample the minority class via SMOTE before modeling. Complement Naive Bayes was chosen as the initial model for its relative strength in both text classification and imbalanced datasets.

There was strong out-of-the-box performance by ComplementNB for the key metric: recall of the minority class (0.87). Precision of the minority class was lower than recall but was also not as relevant to the problem statement. Additional models were not be considered, and instead the hyperparameter tuning was done via GridSearchCV for the ComplementNB model in order to find the optimal alpha smoothing parameter (0.9).

Metrics overall only improved slightly through hyperparameter tuning, indicating that testing additional values in the smoothing parameter would not likely to increase model performance.

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



Conclusions:
Model performance was strong on the test set. 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.