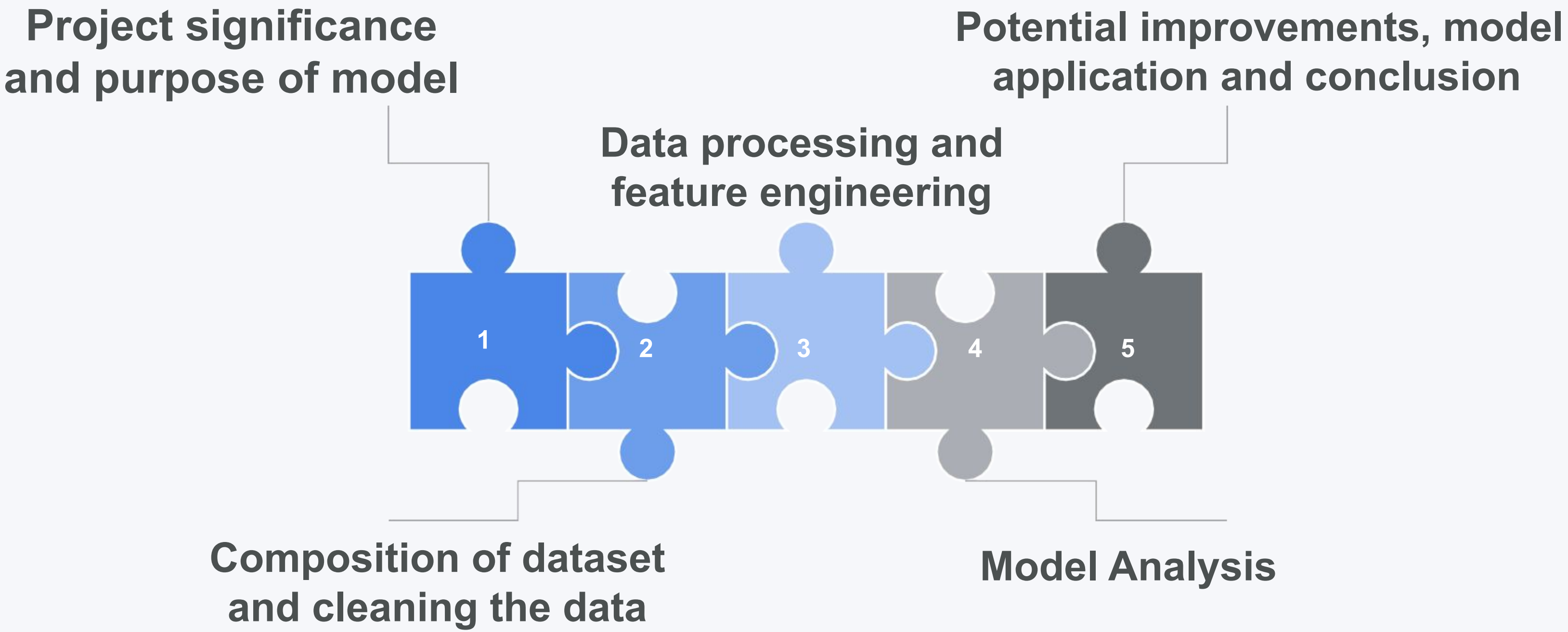
BA476 Final Presentation

Predicting Fraudulent Job Postings

John Antony, Rawin Bunajinda, Neeraja Mehta,
Arnav Misra, Theodore Phua

PRESENTATION AGENDA



PROJECT SIGNIFICANCE

ML model will help identify Fake Job Posting based on a jobs meta information

ML MODEL

- Waste of time and energy
- Scam applicants of personal identity
- Scam applicants for money
- Selling Applicant info involuntarily

April 21, 2021

FBI Warns Cyber Criminals Are Using Fake Job Listings to Target Applicants' Personally Identifiable Information

WORK

Americans lost $68 million to job scams this year—here's what to look out for

Published Fri, Jun 10 2022 11:00 AM EDT

PURPOSE OF MODELS

Determine whether a job is fake or real through specific parameters

Find common trends among fake job postings

Identify keywords and traits

Finding the best baseline model for future evaluation

COMPOSITION OF ORIGINAL DATASET

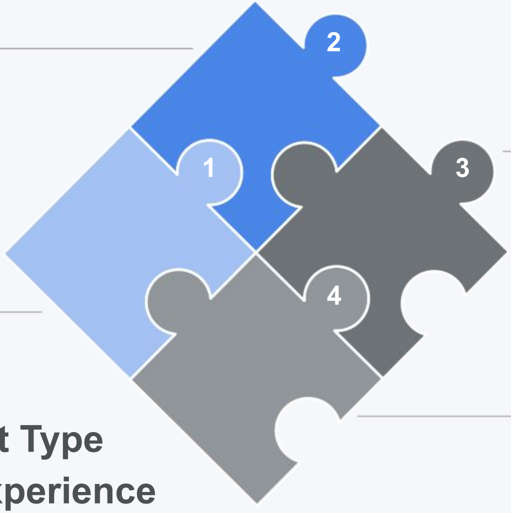
18,000 total job posting prediction dataset from Kaggle

BOOLEAN PREDICTORS

- Telecommuting
- Company Logo
- Questions

TEXT PREDICTORS

- Title
- Location
- Department
- Salary Range
- Company Profile
- Description
- Benefits
- Employment Type
- Required Experience
- Required Education
- Industry
- Function



NUMERICAL PREDICTORS

- Job ID

TESTING OUTCOME

- Fraudulent

DATA CLEANING

Processing of acquired dataset of real and fake job posting predictors

Check for insignificant data points

- Parameters that are unique but do not affect ML model were dropped i.e. “Job ID”

Identify parameters with null values

- Created list of parameters and calculated number of null values; dropped “Department” and “Salary Range”

US-based model

- Dropped non-US jobs; foreign language increases complexity of model

Dropped null values in existing parameters

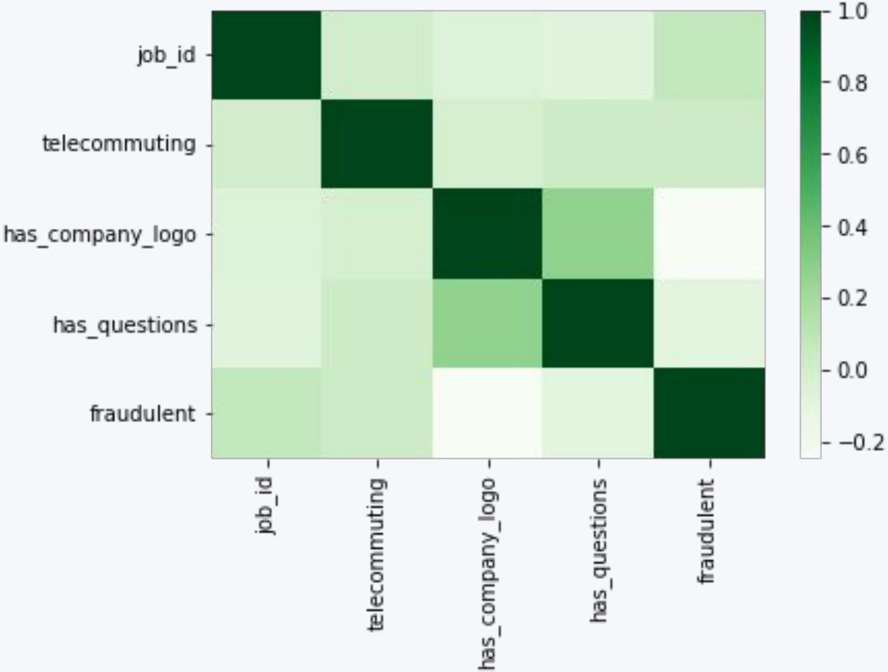
- Drop any null values existing in remaining dataset (for any parameter)

Cleaned dataset

- 9868 data points
- 722 fraudulent job postings

CORRELATION MATRIX FOR NUMERICAL FEATURES

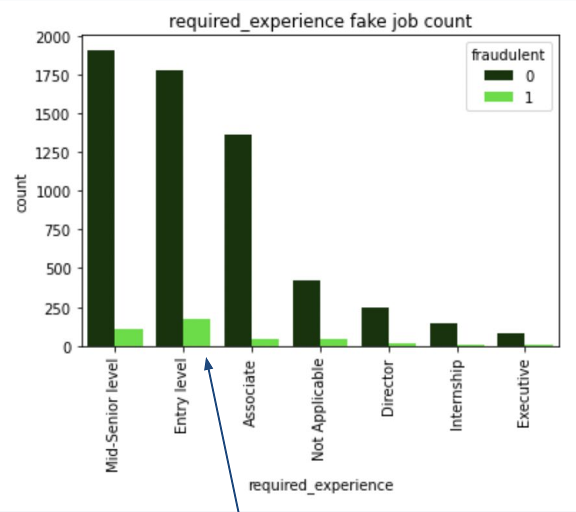
Numerical features consist of all non-textual features including boolean features

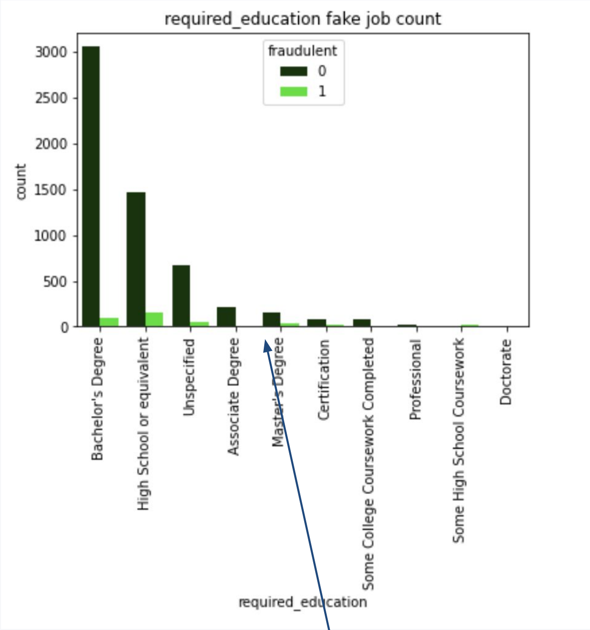


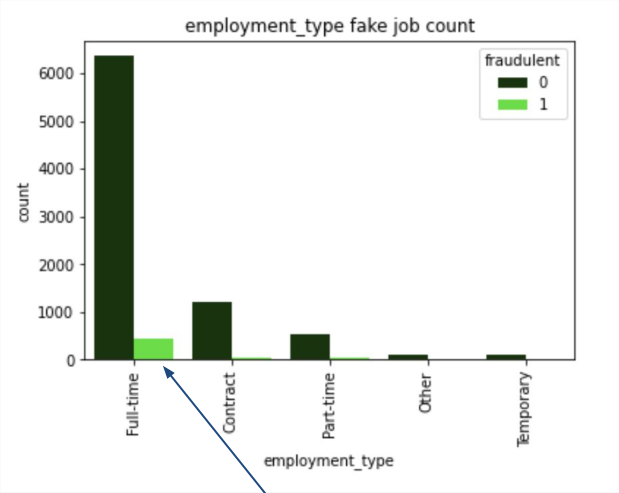
1. No dropped values as correlation values are not highly positive or negative
2. Fraudulent and has_company_logo has greatest negative correlation
3. Greatest correlating features with fraudulent are has_questions and telecommuting

OBSERVATIONS FROM TEXTUAL FEATURES

Plots to visualize and analyze trends within textual features of the dataset







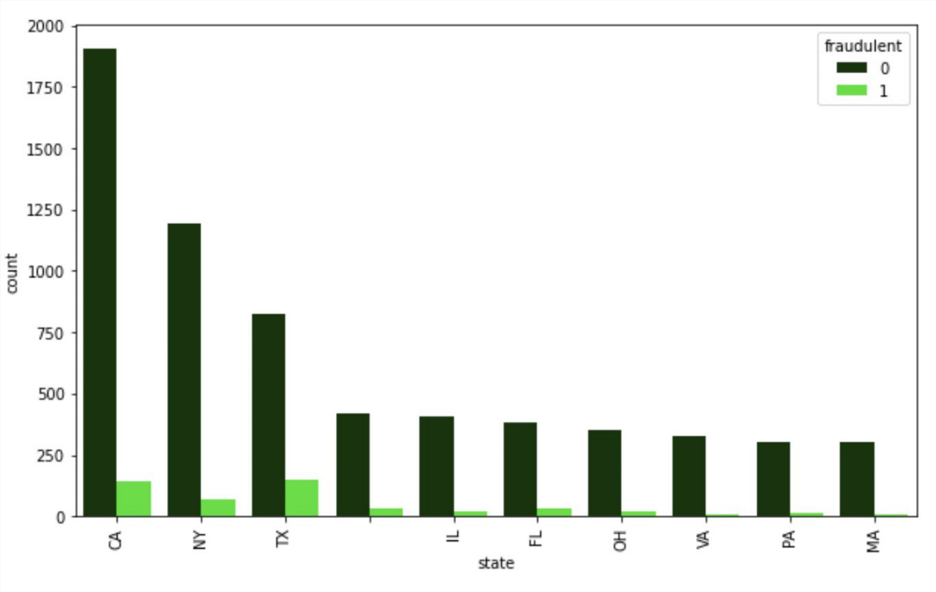
Most fraudulent jobs present in entry level

No fraudulent jobs present in education levels higher than associate degrees

Fraudulent jobs are targeted at full time positions

FEATURE ENGINEERING FOR LOCATION PARAMETER

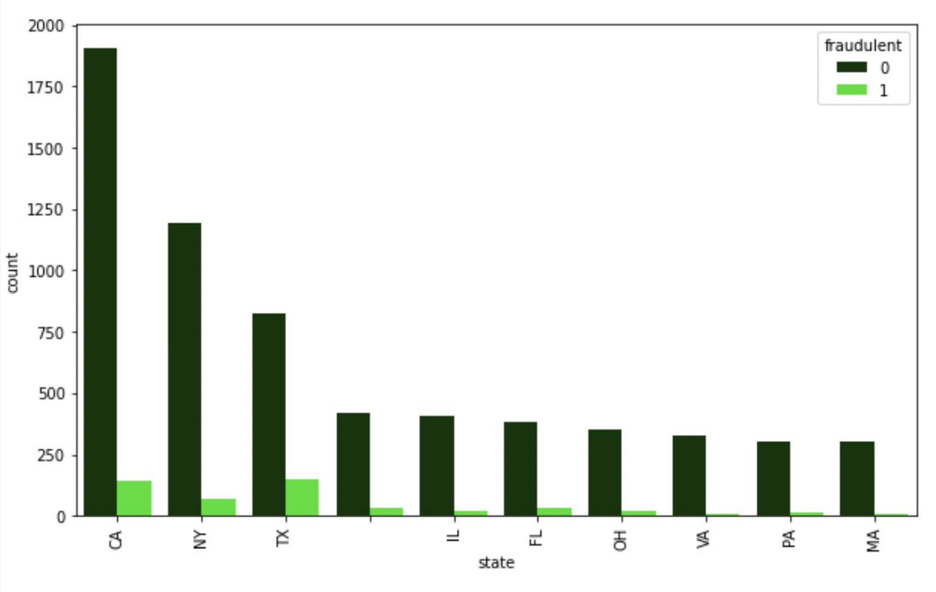
Using Location parameter to visualize distribution of fake jobs

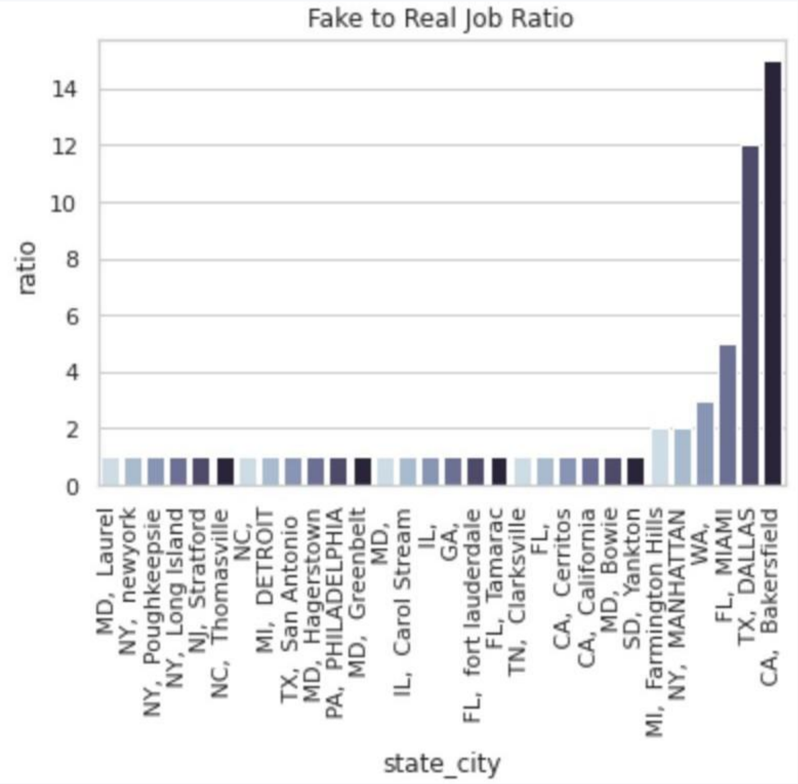


1. States with lowest amount of fake postings are Maryland, New York
2. States with highest amount of fake job postings are California and Texas
3. Utilized to create "Ratio" - parameter applying different probabilistic weights to features

FEATURE ENGINEERING FOR LOCATION PARAMETER

Using Location parameter to visualize distribution of fake jobs





FEATURE ENGINEERING FOR TEXT PARAMETERS

Combine text parameters into new two parameters Text and Character Count

	telecommuting	fraudulent	ratio	text
0	0	0	0.03	Marketing Intern US, NY, New York We're Food52...
1	0	0	0.03	Visual Designer US, NY, New York Kettle is an ...
2	0	0	0.03	Payroll Tax Specialist US, NY, New York Namely...
3	0	0	0.03	Marketing Manager US, NY, New York Super Socce...
4	0	0	0.03	English Teacher Abroad US, NY, New York We hel...
...	...	...	...	...
10588	1	1	0.00	Military Benefits Counselor US, , chicago Anth...
10589	0	0	0.00	Sr.Business Intelligence Technical Architect U...
10590	0	0	0.00	Licensed Practical Nurse (LPN)- Private Duty U...
10591	0	0	0.00	SAS Grid Developer US, NJ, Berkeley Heights ...
10592	0	0	0.00	Sr. Scm Web Development Technical Lead US, CA,...

10593 rows × 4 columns

1. Amalgamated all text datatype predictors into Text predictor which will be used for further text analysis
2. Tokenization, removal of stopwords and lemmatization - created a word-cloud
3. Transform text into vector matrix based on frequency of words in the text through a countvectorizer

WORD CLOUD FROM TEXT PARAMETER

Utilizing word_net lemmatizer to create word clouds to identify frequently used words for real and fake jobs

10 high frequency words for fake jobs

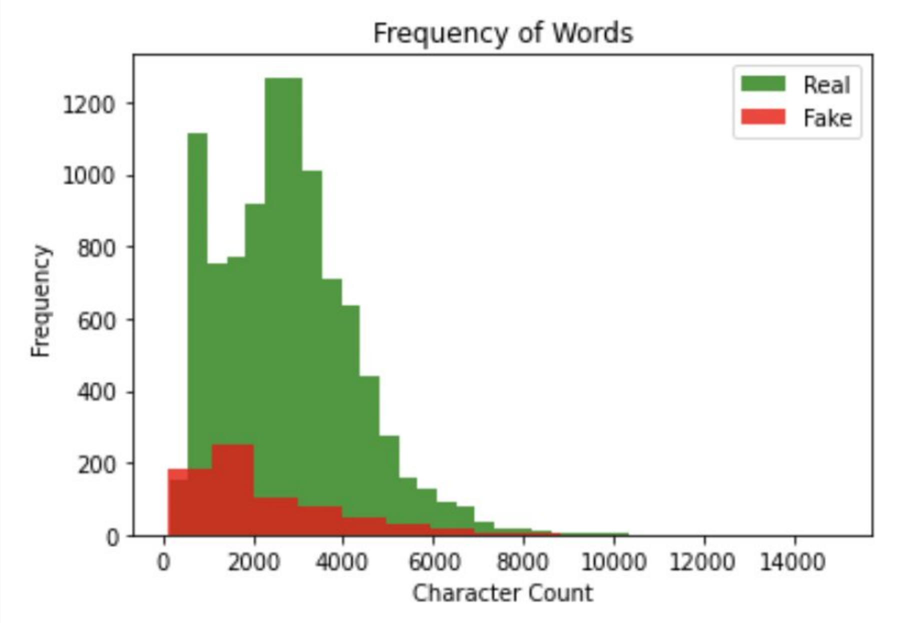
10 high frequency words for genuine jobs

FEATURE ENGINEERING FOR TEXT PARAMETERS

Summarizing shift of parameters from previously Text to now Character Count

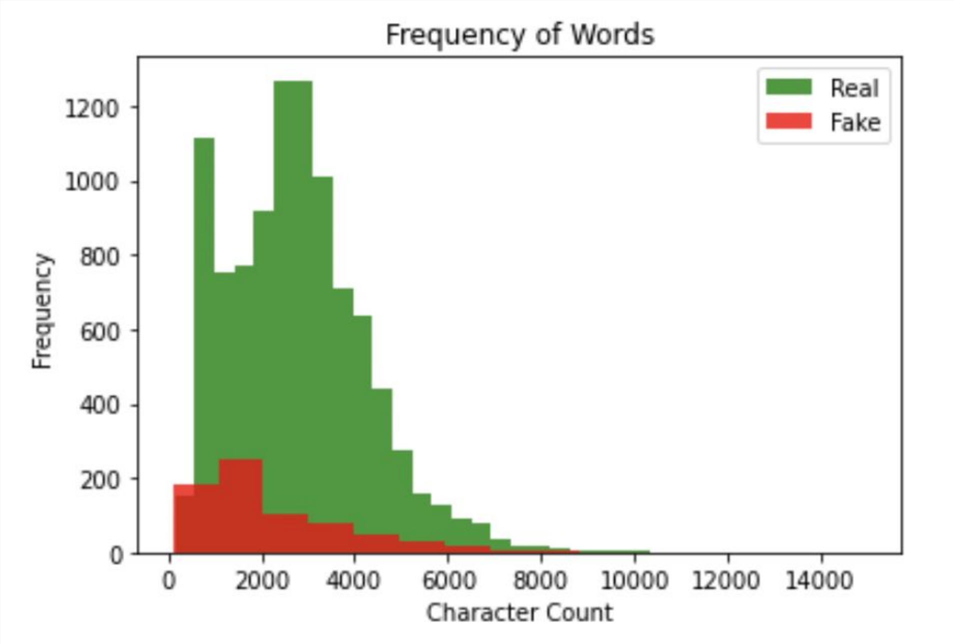
	telecommuting	fraudulent	ratio	text
0	0	0	0.03	Marketing Intern US, NY, New York We're Food52...
1	0	0	0.03	Visual Designer US, NY, New York Kettle is an ...
2	0	0	0.03	Payroll Tax Specialist US, NY, New York Namely...
3	0	0	0.03	Marketing Manager US, NY, New York Super Socce...
4	0	0	0.03	English Teacher Abroad US, NY, New York We hel...
...	...	...	...	...
10588	1	1	0.00	Military Benefits Counselor US, , chicago Anth...
10589	0	0	0.00	Sr.Business Intelligence Technical Architect U...
10590	0	0	0.00	Licensed Practical Nurse (LPN)- Private Duty U...
10591	0	0	0.00	SAS Grid Developer US, NJ, Berkeley Heights ...
10592	0	0	0.00	Sr. Scm Web Development Technical Lead US, CA,...

10593 rows × 4 columns



FEATURE ENGINEERING FOR TEXT PARAMETERS

Combine text parameters into new two parameters Text and Character Count



1. Character count is relatively similar distributed in both fake and real jobs
2. Word frequency is higher among real jobs as compared to fake jobs

FEATURE PROCESSING

Utilized Tokenization, Stopword Removal and Lemmatization to create 4 main concluding parameters:
Telecommuting, Fraudulent, Ratio, Text

job_id	title	location	department	salary_ran ge	company_p rofile	description	requirements	benefits	telecommuting	...	has_q uestions	employment_type	required_experience	required_education	industry	function	fraudulent	state	city	state_city

telecommuting	fraudulent	ratio	text	character_count

P VALUE SIGNIFICANCE TEST

Determining which features are important

```
Optimization terminated successfully.
         Current function value: 0.258066
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:            fraudulent   No. Observations:                 7097
Model:                         Logit   Df Residuals:                     7094
Method:                          MLE   Df Model:                            2
Date:               Thu, 01 Dec 2022   Pseudo R-squ.:                -0.02155
Time:                       19:59:47   Log-Likelihood:                -1831.5
converged:                      True   LL-Null:                       -1792.9
Covariance Type:           nonrobust   LLR p-value:                     1.000
===================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------
telecommuting       0.1272      0.188      0.676      0.499      -0.241       0.496
ratio               2.4433      0.225     10.840      0.000       2.002       2.885
character_count    -0.0012   2.77e-05    -42.908      0.000      -0.001      -0.001
===================================================================================
```

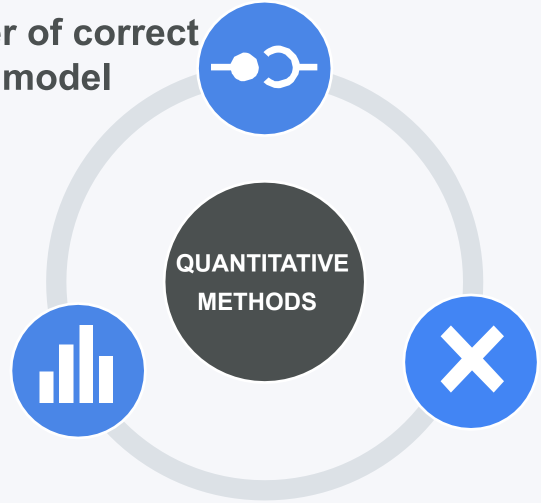
Telecommuting has a p value > 0.05 and is statistically insignificant; this feature is dropped from the model

MEASURING ACCURACY QUANTITATIVELY

Below are the following quantitative methods used to determine the best model for classifying fraudulent jobs

ACCURACY SCORES:

Measures the number of correct predictions from the model



K-FOLD CROSS VALIDATION:

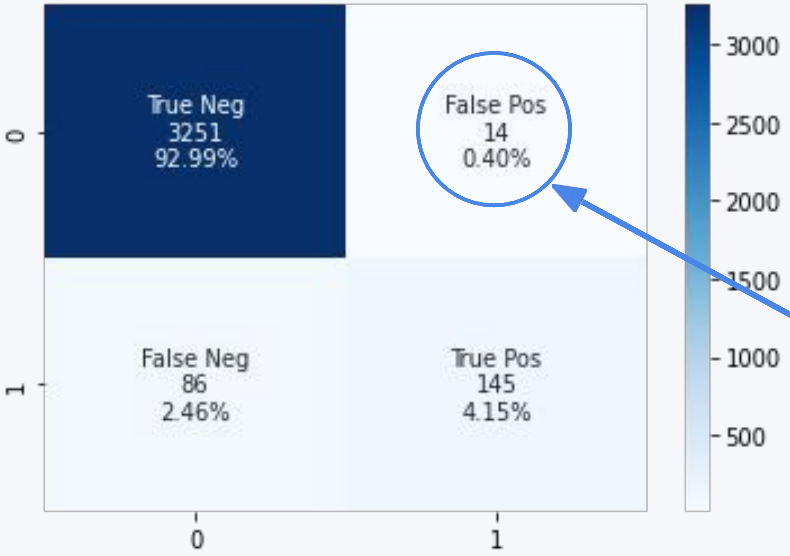
Separate dataset into training (k-1) and testing (k) and run both datasets k-times

F1 VALUES:

Weighted average of precision and recall

BASELINE MODEL: NAIVE BAYES MODEL

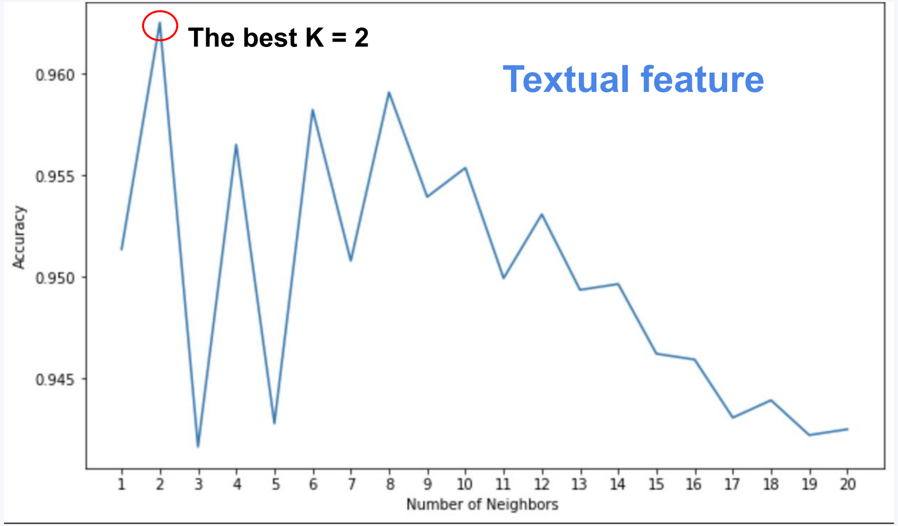
Model has an accuracy rate of 97.1% and F1 value of 0.7435

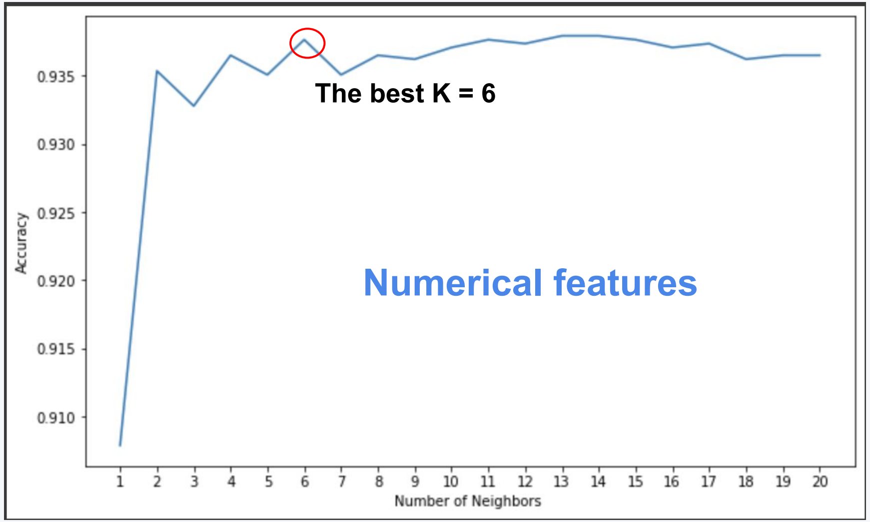


Model with lowest number of false positives

MODEL: K-NEAREST NEIGHBORS

Determining the best K for the KNN model

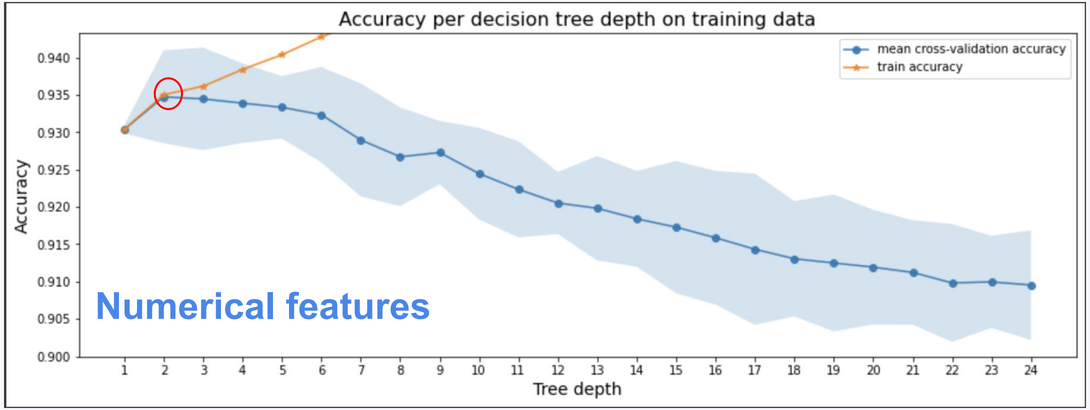




Accuracy: 96.25%
F1: 0.7171

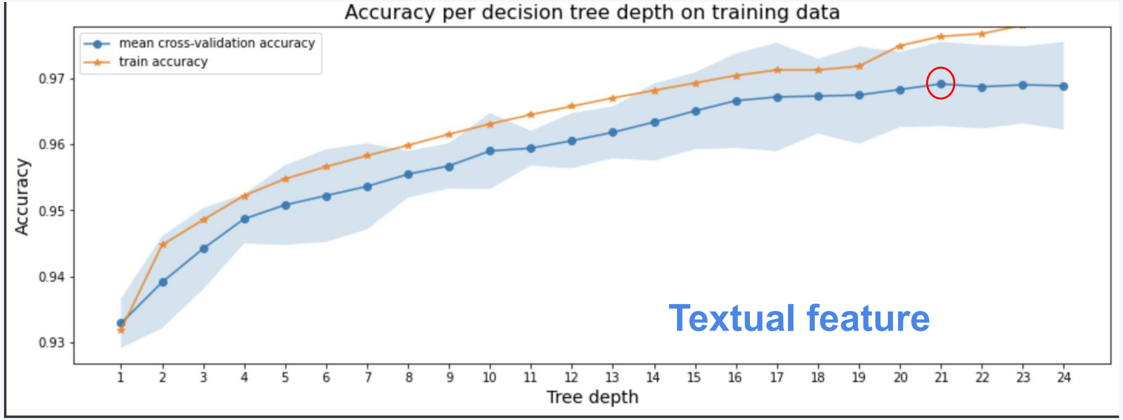
Accuracy: 93.54%
F1: 0.1630

MODEL: DECISION TREES (Depth)



Tree with depth of 2 has highest accuracies = 93.48%

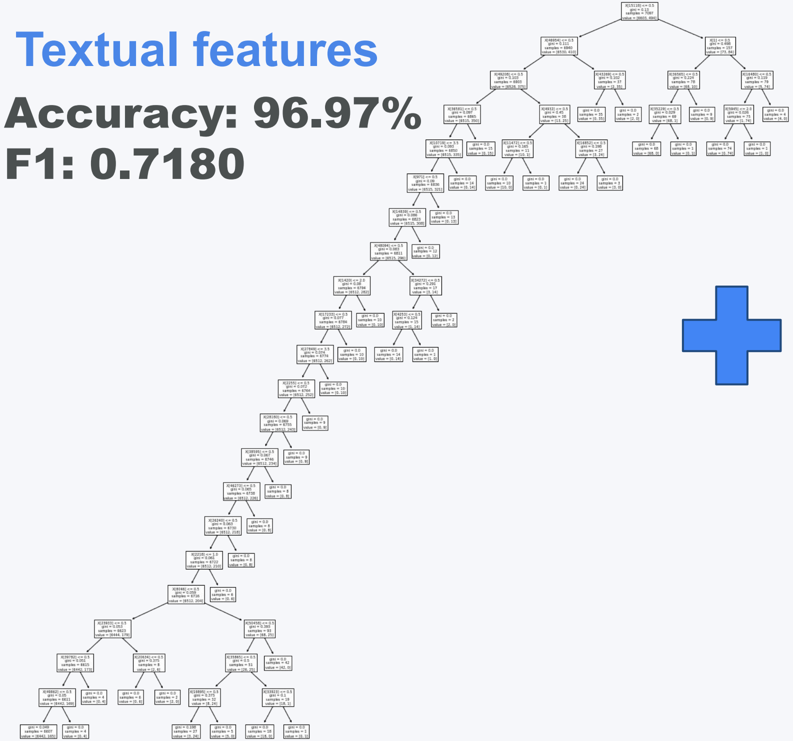
Tree with depth of 21 has highest accuracies = 96.91%

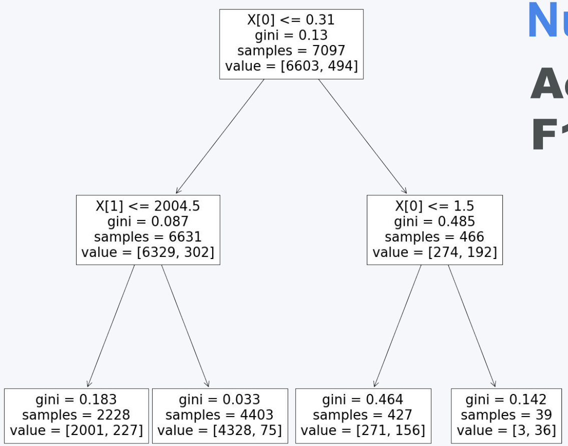


NEXT STEP: STACKING TREES

Textual features

Accuracy: 96.97%
F1: 0.7180





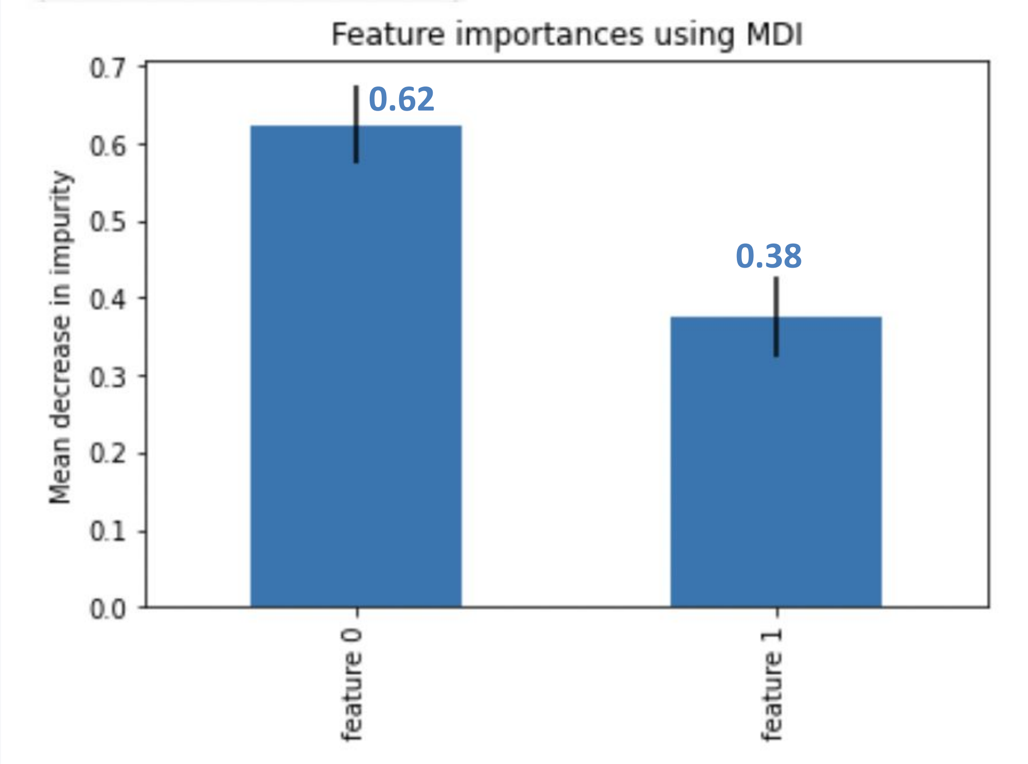
Numerical features

Accuracy: 93.79%
F1: 0.1423

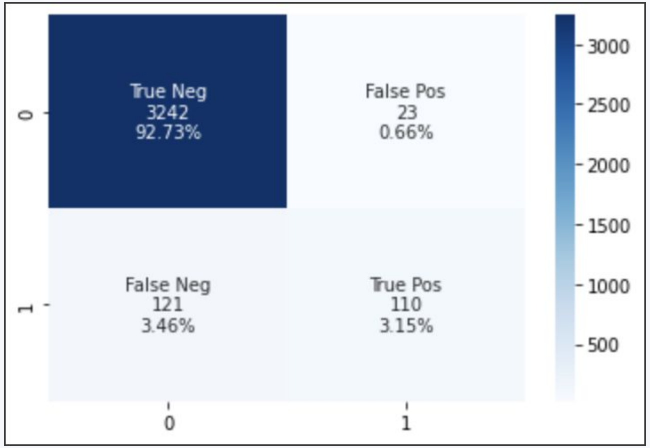
Combined features

Accuracy: 96.97%
F1: 0.7239

MODEL: RANDOM FORESTS



- 100 trees in Random Forest.



Legend
feature 0 = ratio
feature 1 = character_count

LOGISTIC REGRESSION MODEL COMPARISON

	Lasso	Ridge	Elastic Net
Accuracy Rate (%)	97.78	95.62	97.11
False Positives (%)	0.60	2.83	1.12
False Negatives (%)	1.63	1.86	1.77
F1 Value	0.8169	.6693	.7897

MODEL COMPARISON

	NAIVE BAYES	Lasso	KNN	Decision Tree	Random Forest
Accuracy Rate (%)	97.12	97.78	96.25	96.97	93.79
False Positives (%)	0.40	0.60	1.89	0.40	0.11
False Negatives (%)	2.46	1.63	1.86	3.98	6.99
F1 Value	0.7435	0.8169	0.7910	0.7180	0.1422

PROJECT CONCLUSION

ML MODEL

Tradeoff between F1 and accuracy

SGD with log loss and lasso penalty balances both

Used as baseline for future model evaluation

APPLICATION

A high accuracy ML model can be integrated at the back end of job posting websites

A 0.4% FN implies 996 out of 1000 fraudulent jobs get classified as fraudulent; significant prevention of fraud

DATA VISUALIZATION

Ratio and employment type have significant influence on predicting fraudulent jobs

Extra scrutiny in locations with high fake-to-real job ratios, or at entry level; help reduce number of fraudulent job posts

PROJECT CHALLENGES

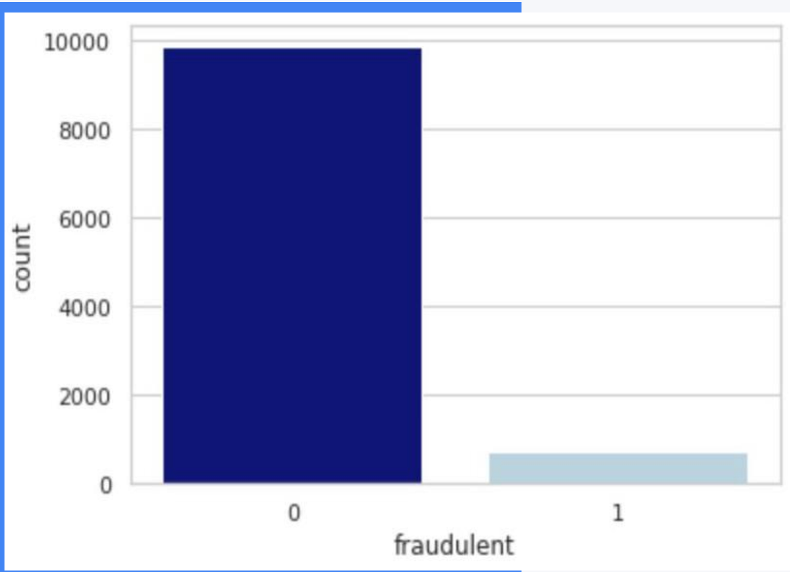
- Large number of null values
- Unbalanced dataset (5% fraudulent)
- Some text stored included miscellaneous characters (translated from different languages) potentially causes loss of information
- Interpreting models from text data
- Simultaneously using Colab is difficult to follow new changes.

POTENTIAL IMPROVEMENTS

How can we make the model predict better?

Distribution of Target Variables

The dataset is unbalanced with a significantly large number of real jobs as compared to the number of fraudulent job postings



Feature selection using PCA

Generate synthetic instances through synthetic minority oversampling technique

Find a more balanced dataset

Impute null values and predictors with median or mean

Q&A