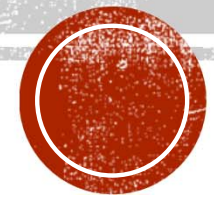
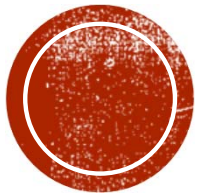
PREDICTING A JOB POST AS REAL OR FAKE

Jayme Kirchner – CSC535 – Deep Learning

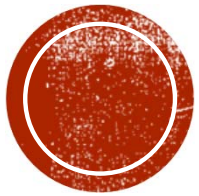




PROBLEM DEFINITION AND PROJECT GOALS

PROBLEM TO ADDRESS

- Predict whether a job posting is real or fake based on the posting's text, categorical variables, and the salary range
- Data downloaded from a csv file on Kaggle
  - https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction
  - 17,880 total job postings
  - 18 variables in total
  - Imbalanced dataset with roughly 5% of fake postings (866 of 17,880)



DATA EXPLORATION AND PRE-PROCESSING

TEXT VARIABLES (UNIQUE VALUES)

- **Title**: title of the job ad entry (11,231)
- **Location**: geographical location of the job ad (3105)
- **Department**: corporate department (1337)
- **Company Profile**: brief company description (1709)
- **Description**: details description of the job ad (14,801)
- **Requirements**: enlisted requirements for the job opening (11,968)
- **Benefits**: enlisted benefits offered by the employer (6205)
- **Salary Range***: indicative salary range (874)

*Note: this value is a string in the original dataset but will be split into two numeric values during pre-processing

CATEGORICAL VARIABLES (UNIQUE VALUES)

- **Employment Type**: full-time, part-time, etc. (5)
- **Required Experience**: entry-level, etc. (7)
- **Required Education**: Bachelor, etc. (13)
- **Industry**: automotive, IT, etc. (131)
- **Function**: consulting, research, etc. (37)

All of these variables will be one-hot-encoded for the initial baseline RNN model

NUMERIC VARIABLES

- **Job Id**: identification number for the job posting (17,880)
- **Telecommuting**: indicates whether posting mentions telecommuting (2)
- **Has Company Logo**: indicates whether company logo is present (2)
- **Has Questions**: indicates whether screening questions are present (2)
- **Fraudulent**: indicates whether posting is fake (2)

EXPLORING THE TEXT DATA

- Maximum length of characters in each text column:
  - 'title': 142
  - 'location': 161
  - 'department': 255
  - 'company_profile': 6178
  - 'description': 14,907
  - 'requirements': 10,864
  - 'benefits': 4429
- There is only 1 row where description is blank
- There are 19 rows where all text variables except title and description are blank
  - 12 of these are fake

CLEANING THE DATA

- Replace missing values in categorical columns with a new 'No Data' label
  - Distinguish between the absence of data and existing 'Not Applicable' or 'Other' labels
- Combine all text columns into one called 'full_text' and drop the relevant columns
  - Minimum length of full_text is 33 characters
  - Maximum length of full_text is 14,964 characters
- Split the string variable salary_range into two columns: salary_low and salary_high
  - If string value was converted to a month, update cell to be the month's corresponding numeric value (i.e., 'Oct' becomes '10')
- Replace NA values with integer -1 to add columns into main data frame, then replace -1 by NaN and cast into Int64 datatype
- Update instances of 'US' to 'USA' so it is distinct country and not word 'us'
- Separate combined words with a space (i.e., PinterestLoves → Pinterest Loves)
- Move the fraudulent column to the end of the dataframe

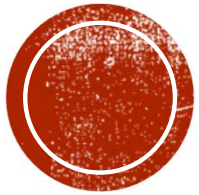
REPLACE NA/MISSING VALUES IN SALARY COLUMNS

Strategy

Using only the training data, the mean low and high salary for each employment type will replace missing values in the respective salary column

Steps

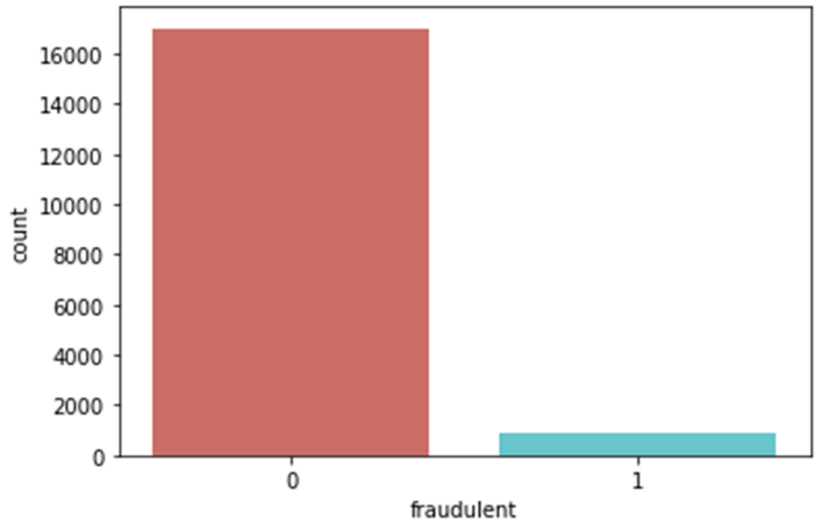
- Split data frame into train/test and then train/val – 80/20 split for both
- Create dictionary with employment type as key, and respective mean values for salary_low and salary_high
- For each employment type,
  - Create a temp data frame
  - Fill missing values with relevant mean
  - Update original data frame
  - Cast into Int64 datatype



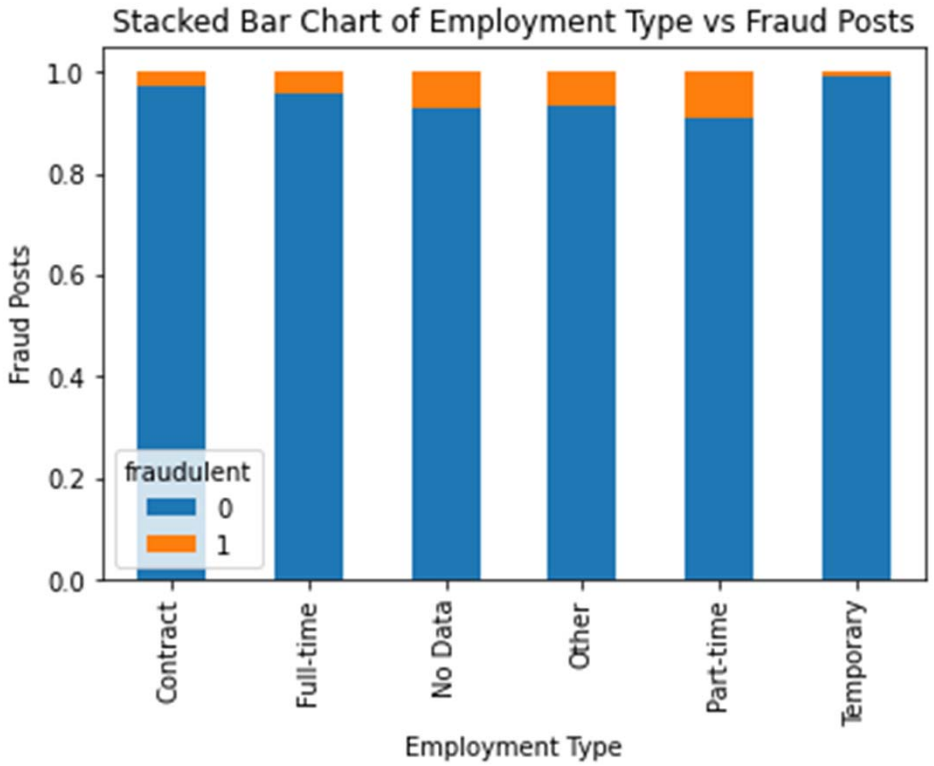
VISUALIZING THE DATA

DISTRIBUTION OF FRAUDULENT POSTINGS

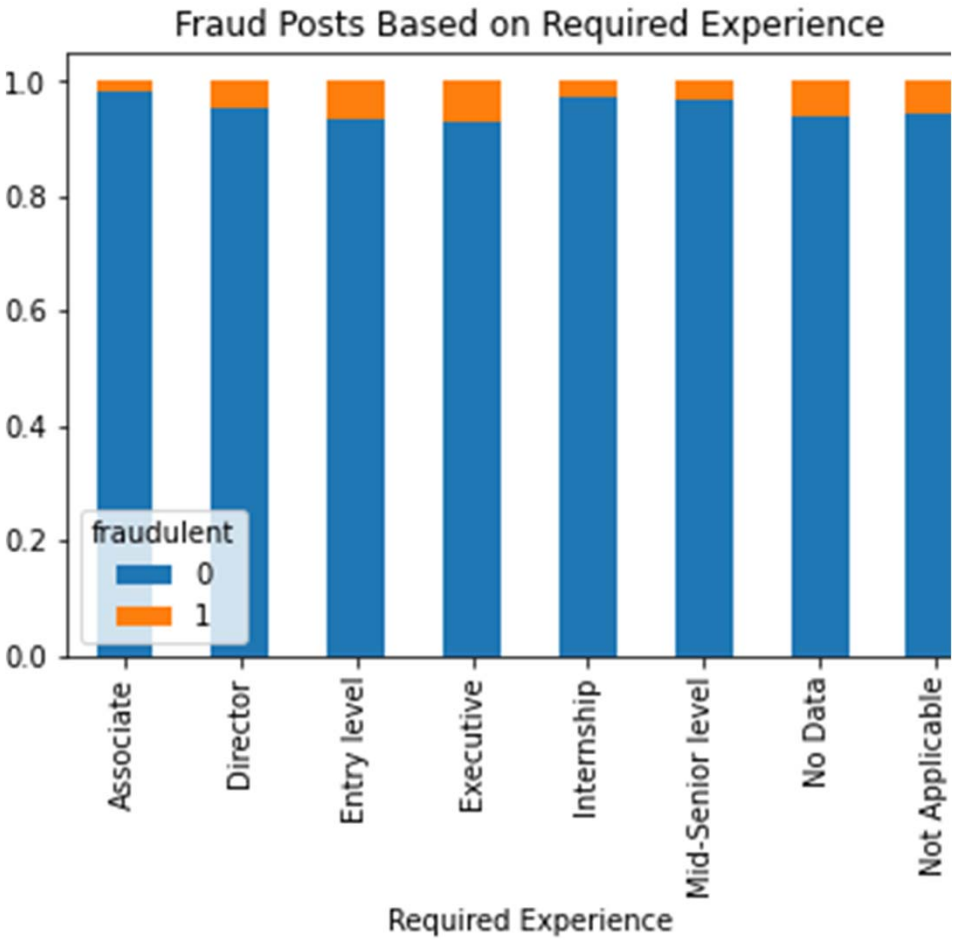
- There are 17,880 real postings and 866 fake postings
- Fake postings do not start until index 98 and the last is found at index 17,831
- The data is very imbalanced with 95% as real posts and only 5% as fake



EMPLOYMENT TYPE

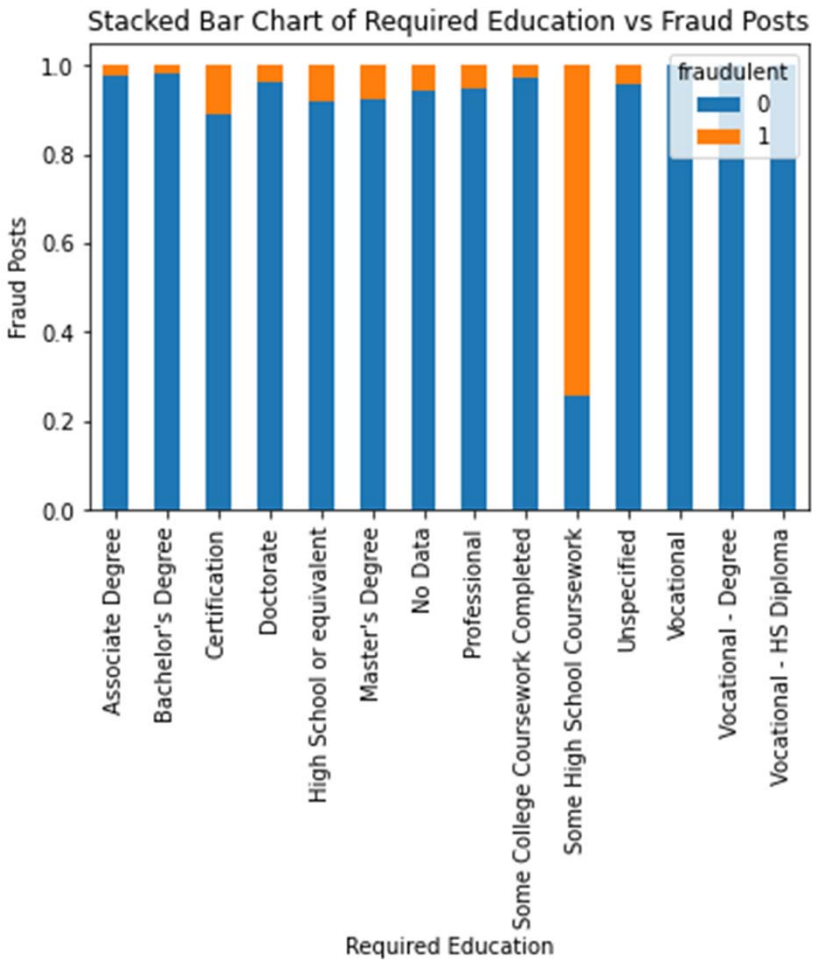


- Full-time: 11,620
- No Data: 3471
- Contract: 1524
- Part-time: 797
- Temporary: 241
- Other: 227



REQUIRED EXPERIENCE

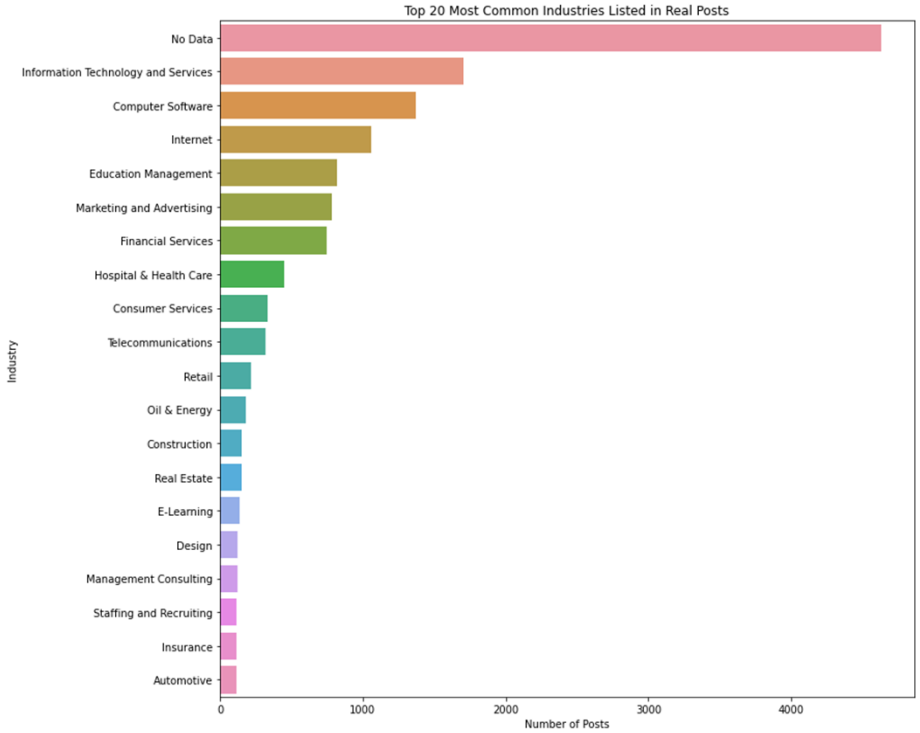
- No Data: 7050
- Mid-Senior level: 3809
- Entry level: 2697
- Associate: 2297
- Not Applicable: 1116
- Director: 389
- Internship: 381
- Executive: 141

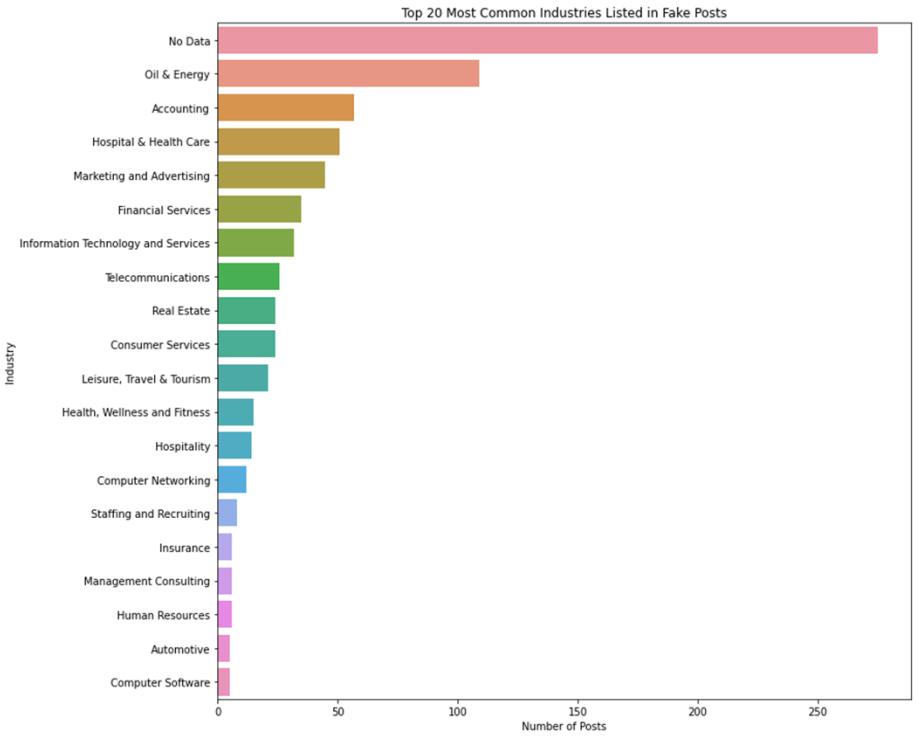


REQUIRED EDUCATION

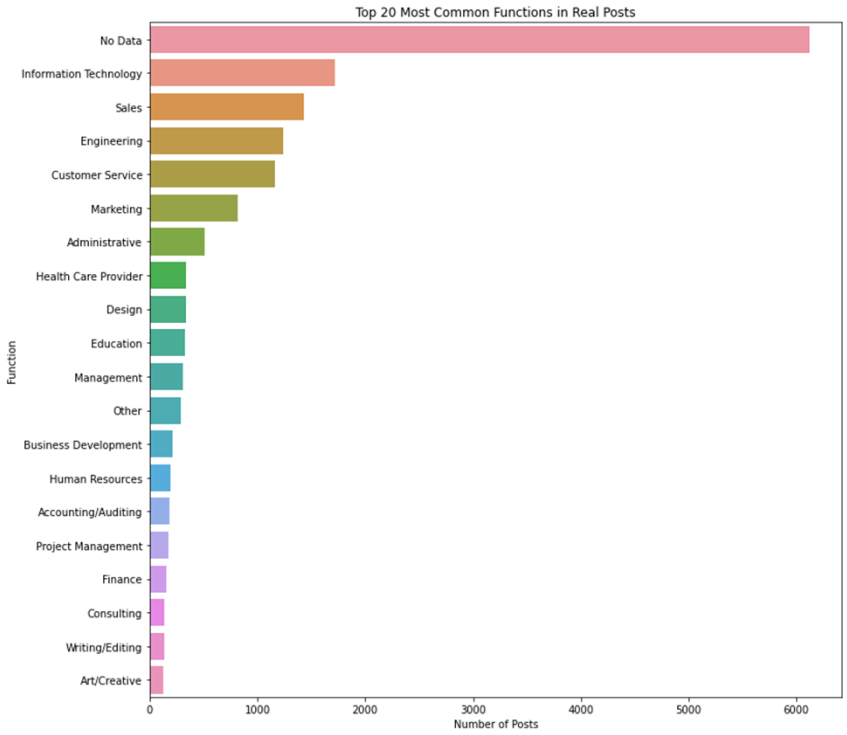
- No Data: 8105
- Bachelor's Degree: 5145
- High School or equivalent: 2080
- Unspecified: 1397
- Master's Degree: 416
- Associate Degree: 274
- Certification: 170
- Some College Coursework Completed: 102
- Professional: 74
- Vocational: 49
- Some High School Coursework: 27
- Doctorate: 26
- Vocational - HS Diploma: 9
- Vocational - Degree: 6

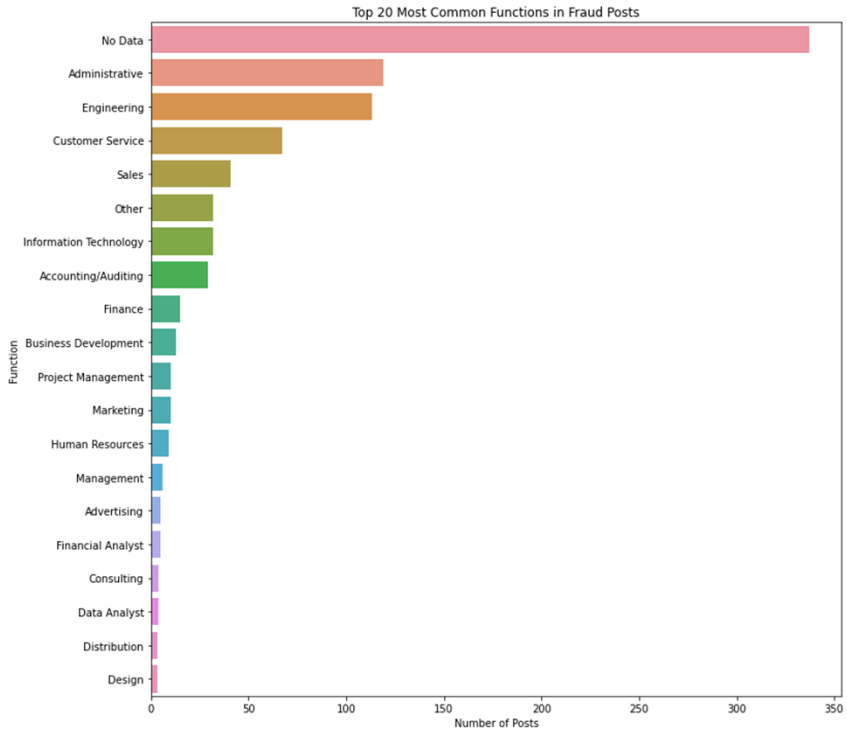
TOP 20 MOST COMMON INDUSTRIES

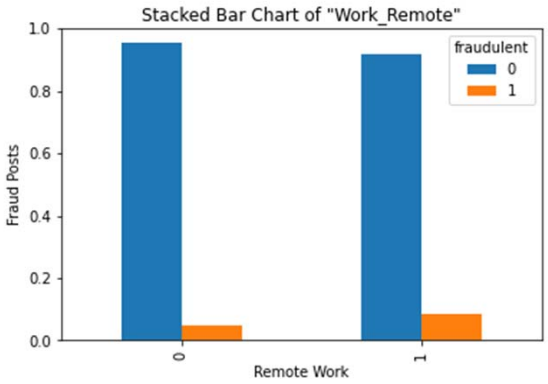


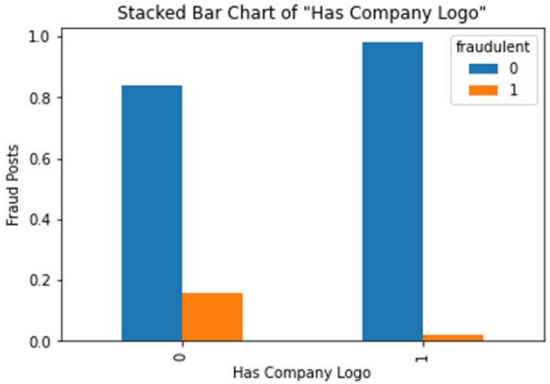


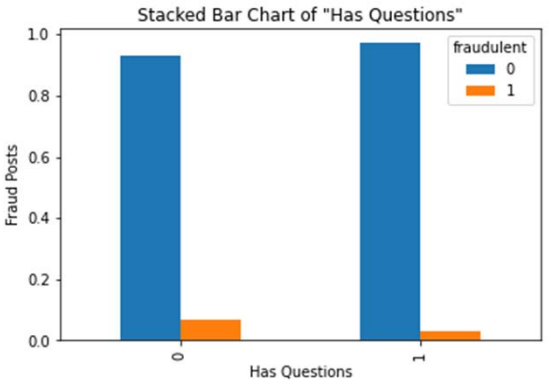
TOP 20 MOST COMMON FUNCTIONS











BINARY VARIABLES (ALL JOB POSTINGS)

- Work Remotely
- Has Company Logo
- Has Questions

OBSERVATIONS FROM THE GRAPHS

Employment Type

- Fewest fraudulent job postings are found for 'Temporary'
- Most fraudulent postings appear to be found for 'No Data', 'Other', and 'Part-time'

Required Experience

- Largest amount of fraud postings are for 'Entry Level', 'Executive', 'No Data' or 'Not Applicable'

Required Education

- No fraudulent postings for any vocational category
- Most fraudulent postings for 'Some High School Coursework' (20 out of 27)
- Large number of fraudulent postings also found for 'Certification', 'Hight School or equivalent', and 'Master's Degree'

OBSERVATIONS FROM THE GRAPHS

Most Common Label

- The lack of an industry or function listed in the job posting was the most frequent among all graphs

Other Top Industries

- Fake Posts: Oil & Energy, Accounting, Hospital & Healthcare, Marketing and Advertising
- Real Posts: Information Technology and Services, Computer Software, Internet, Education Management

Other Top Functions

- Fake Posts: Administrative, Engineering, Customer Service, Sales
- Real Posts: Information Technology, Sales, Engineering, Customer Service, Marketing

Work Remotely

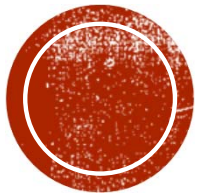
- More fraudulent posts when job posting offered remote work

Has Company Logo

- More fraudulent posts when job posting did not include a company logo

Has Questions

- More fraudulent posts when job posting did not contain screening questions



ONE-HOT-ENCODING AND SPLITTING THE DATASET

ONE-HOT-ENCODE CATEGORICAL VARIABLES

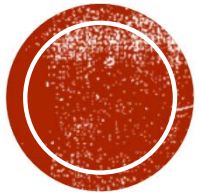
- Columns to be one-hot-encoded:
  - Employment Type
  - Required Experience
  - Required Education
  - Industry
  - Function
- Original data frame had 12 columns
- One-hot-encoding expanded it to 205 columns

TRAIN/VAL/TEST SPLIT

- Data was already split into train/val/test to calculate salary means
- Split was done again now that all data is processed satisfactorily
- Train/Test and Train/Val were split based on 80/20 ratio
- Same random state was used for all splits
- Train_x had 10,889 real posts and 554 fraud posts

- Shape of each data frame:
  - Original dataframe shape: (17880, 205)
  - train_x shape: (11443, 205)
  - val_x shape: (2861, 205)
  - test_x shape: (3576, 205)

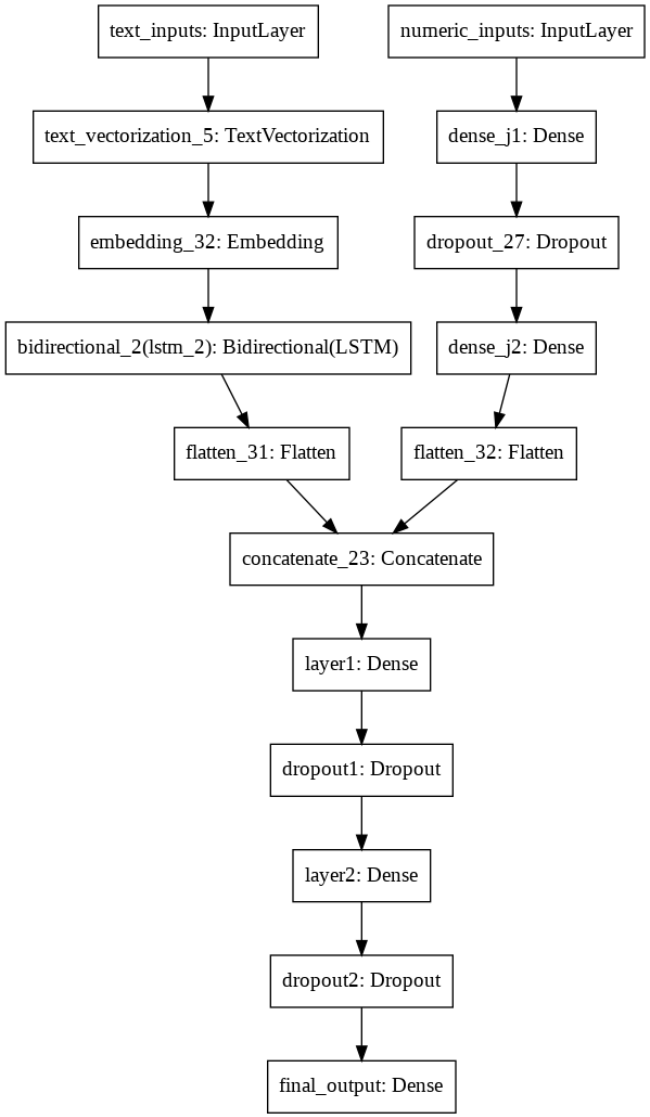
- There are 5 industry levels where training data only has real postings:
  - Libraries, Military, Package/Freight Delivery, Shipbuilding, Wine and Spirits
  - Models will not learn to predict fraudulent postings for these levels



DEEP LEARNING MODELS

MODELS AND METRICS

- Models:
  - RNN with TextVectorization, Embedding, and Bidirectional layers
  - BERT with modified preprocessing function
- Metrics:
  - False Positives (fp)
  - False Negatives (fn)
  - Binary Accuracy (accuracy)
  - AUC (auc)
- Loss Function:
  - Binary Crossentropy
- Optimizer:
  - Adam (RNN)
  - AdamW (BERT)



RNN MODEL ARCHITECTURE

Text and numeric data

TRAINING AND TESTING WITH ALL VARIABLES

- Inputs (train / validation):
  - Full text vectorized to 128 tokens
  - Numeric variables
- Inputs (train+validation / test):
  - Full text combining with train and validation combined
  - Numeric variables with train and validation combined
- Ran models for 5 epochs each

METRICS RESULTS AND OBSERVATIONS

Training

- AUC: 0.5109 (increase)
- False Negatives: 515.0 (varied)
- False Positives: 492.0 (decrease)

Validation

- AUC: 0.5054 (varied)
- False Negatives: 137.0 (increase)
- False Positives: 10.0 (decrease)

Problems with results

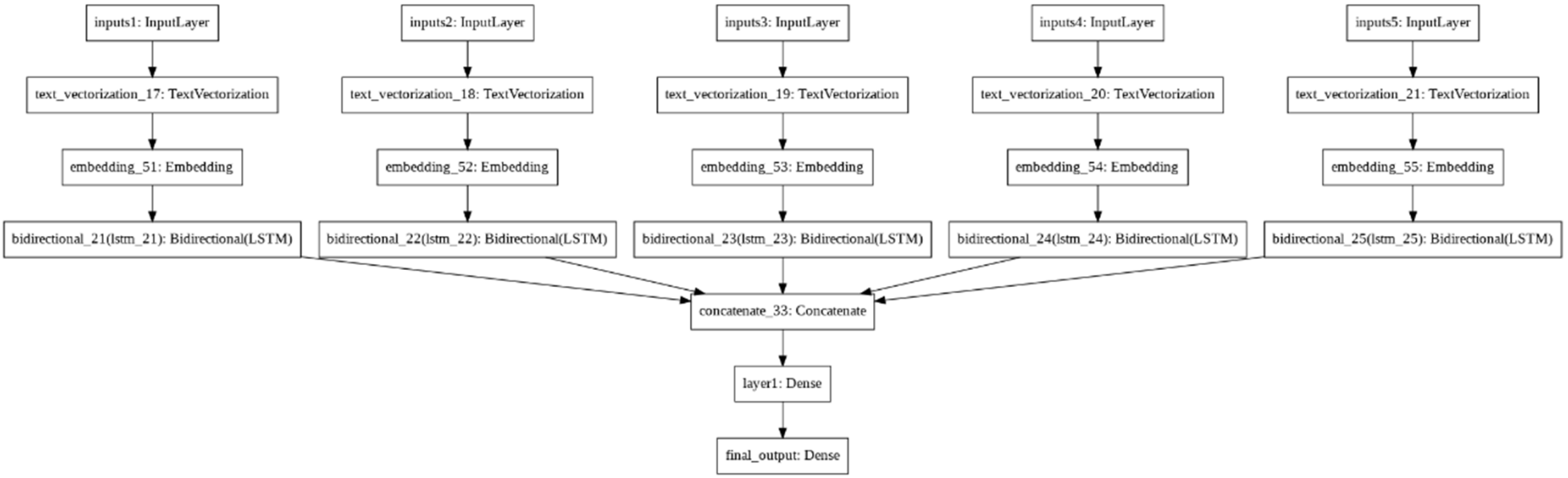
- Data imbalance caused the initial models (text + numeric data) to have high accuracy but a low AUC score of 50%
- Under-sampling the real postings in the training data did not significantly improve results
- Second RNN model will only use the text columns to see if the AUC score improves
  - TextVectorization, Embedding, and Bidirectional(LSTM) layers applied to each text column

TEXT-ONLY DATA COLUMNS

- The text columns were split into 5 groups:
  - Title, Location, Description
  - Department, Employment type, Required Experience, Required Education, Industry, Function
  - Company Profile
  - Requirements
  - Benefits
- The text-only dataset was split again into train/validation/test and each column group was saved into its own Numpy array
- Each text group was vectorized into 128 tokens
  - This amount matches the default number of tokens with BERT preprocessing models
  - Next slide shows top 20 and bottom 20 tokens in each text group

THE MOST AND LEAST FREQUENT TOKENS

- Title + Location + Description
  - Top 20 words in vocab: ['', '[UNK]', 'and', 'the', 'to', 'of', 'a', 'in', 'for', 'with', 'our', 'is', 'you', 'are', 'will', 'be', 'as', 'we', 'on', 'team']
  - Last 20 words in vocab: ['media', 'into', 'growing', 'knowledge', 'engineer', 'build', 'systems', 'required', 'office', 'information', 'if', 'has', 'training', 'one', 'lead', 'social', 'do', 'communication', 'ca', 'years']
- Department + Employment type + Required Experience + Required Education + Industry + Function
  - Top 20 words in vocab: ['', '[UNK]', 'data', 'no', 'fulltime', 'level', 'degree', 'bachelors', 'midsenior', 'technology', 'information', 'services', 'and', 'entry', 'associate', 'marketing', 'sales', 'or', 'school', 'high']
  - Last 20 words in vocab: ['security', 'building', 'materials', 'general', 'communications', 'travel', 'nonprofit', 'tourism', 'leisure', 'social', 'entertainment', 'electricalelectronic', 'client', 'beverages', 'support', 'goods', 'creative', 'team', 'cosmetics', 'vocational']
- Company Profile
  - Top 20 words in vocab: ['', '[UNK]', 'and', 'the', 'to', 'of', 'a', 'in', 'we', 'our', 'is', 'for', 'with', 'are', 'that', 'you', 'on', 'as', 'their', 'have']
  - Last 20 words in vocab: ['growing', 'we're', 'online', 'what', 'marketing', 'home', 'across', 'mission', 'jobs', 'believe', 'job', 'improve', 'candidates', 'many', 'high', 'up', 'platform', 'focus', 'been', 'creative']
- Requirements
  - Top 20 words in vocab: ['', '[UNK]', 'and', 'to', 'of', 'in', 'a', 'the', 'with', 'experience', 'or', 'skills', 'for', 'ability', 'be', 'work', 'you', 'is', 'as', 'years']
  - Last 20 words in vocab: ['support', 'level', 'if', 'technology', 'about', 'microsoft', 'company', 'detail', 'proven', 'technologies', 'attention', 'applications', 'information', 'who', 'position', 'writing', 'mobile', 'school', 'role', '–']
- Benefits
  - Top 20 words in vocab: ['', '[UNK]', 'and', 'to', 'the', 'a', 'of', 'in', 'we', 'with', 'you', 'for', 'our', 'work', 'is', 'your', 'benefits', 'on', 'an', 'as']
  - Last 20 words in vocab: ['bonus', '–', 'position', 'their', 'startup', 'do', 'per', 'international', 'want', 'offers', 'personal', 'excellent', 'exciting', 'solutions', 'creative', 'sick', 'world', 'sales', 'provide', 'day']



TEXT-ONLY RNN MODEL ARCHITECTURE

- Each column group is fed as input to the model, passing through three layers – TextVectorization, Embedding, and Bidirectional LSTM
- All inputs are combined and passed into a Dense Layer
- Final output layer has 1 neuron with sigmoid activation function

METRICS RESULTS

Training

- AUC: 0.7930 (increase)
- False Negatives: 554 (increase)
- False Positives: 0.0 (decrease)

Training + Validation

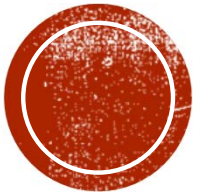
- AUC: 0.8009 (increase)
- False Negatives: 693.0 (decrease)
- False Positives: 0.0 (decrease)

Validation

- AUC: 0.7972 (increase)
- False Negatives: 139.0 (no change)
- False Positives: 0.0 (no change)

Testing

- AUC: 0.7864 (increase)
- False Negatives: 173.0 (no change)
- False Positives: 0.0 (no change)

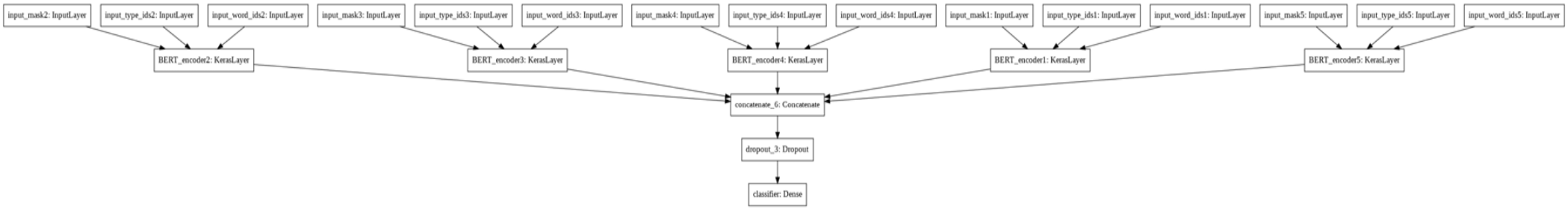


NLP MODEL: BERT

BERT MODELS

	H=128	H=256	H=512	H=768
L=2	2/128 (BERT-Tiny)	2/256	2/512	2/768
L=4	4/128	4/256 (BERT-Mini)	4/512 (BERT-Small)	4/768
L=6	6/128	6/256	6/512	6/768
L=8	8/128	8/256	8/512 (BERT-Medium)	8/768
L=10	10/128	10/256	10/512	10/768
L=12	12/128	12/256	12/512	12/768 (BERT-Base)

Image found at: https://huggingface.co/google/bert_uncased_L-4_H-512_A-8



TEXT-ONLY BERT MODEL ARCHITECTURE

- BERT-Small (L=4, H=512, A=4)
- Token Length: 128
- Each input passes through a Tokenizer layer and a Packing layer for preprocessing
- Preprocessing inputs to each encoder: input_mask, input_word_ids, and input_type_ids

METRICS RESULTS

Training

- AUC: 0.8055
- False Negatives: 520.0
- False Positives: 27.0

Validation

- AUC: 0.9625
- False Negatives: 109.0
- False Positives: 1.0

Testing

- AUC: 0.9637
- False Negatives: 119.0
- False Positives: 4.0

COMPARING THE MODELS AND RESULTS

COMPARING THE TEXT-ONLY MODELS

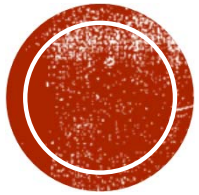
RNN

- Epochs: 5
- Fixed learning rate with Adam optimizer
- Final Metrics Results:
  - AUC: 0.7864
  - False Negatives: 173.0
  - False Positives: 0.0

Small-BERT

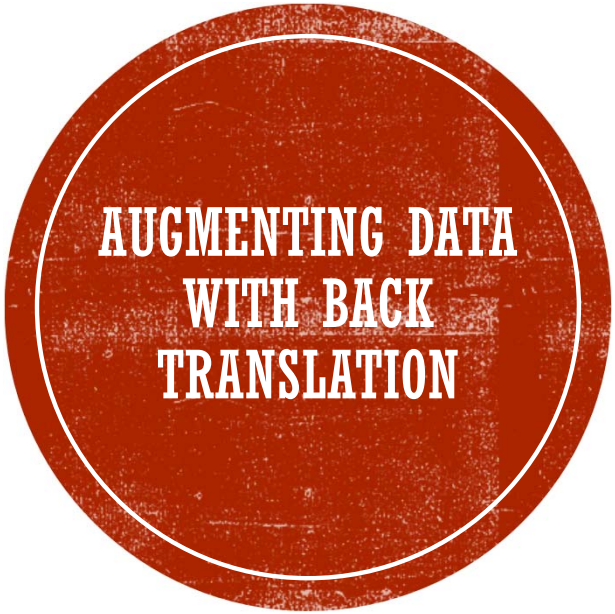
- Epochs: 1
- Warm-up learning rate scheduler with AdamW optimizer
- Final Metrics Results:
  - AUC: 0.9637
  - False Negatives: 119.0
  - False Positives: 4.0

Small-BERT is the better model to predict fraudulent job postings from a highly imbalanced dataset using only the text components of the posting



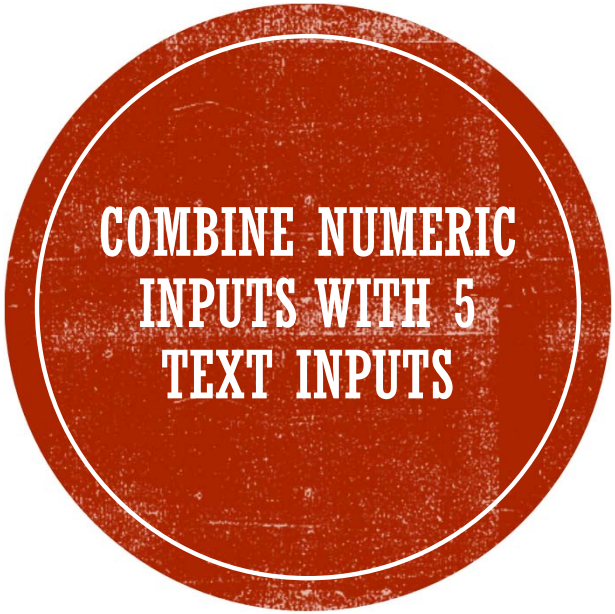
EXPANDING THE PROJECT

Ideas to apply with more time and/or resources



- What is Back Translation?
  - Translating a text from a source language and back via a second language
- Python Package: BackTranslation on PyPI
  - Implements googletrans
- Languages to be used: French, Spanish, Chinese, Russian
- Back translation is only applied on fraudulent data posts
- Steps:
  - Create data frame to save all augmented fraudulent posts (i.e., df_augmented)
  - Apply translate() function from BackTranslation package to the original 866 fraudulent posts in English using one of the above languages
  - Append results to df_augmented
  - Repeat steps 2-3 with a new language, applying translate() to all posts found in df_augmented
  - When finished, append df_aug to original data frame
- Final data frame would have 13,856 fraudulent posts (approx. 45% of the total data)

- After RNN and BERT, I would like to train a GPT-2 model to create a realistic fake job posting
- This idea was inspired by a Coursera Guided Project, 'Generating New Recipes using GPT-2' https://www.coursera.org/projects/generating-new-recipes-python

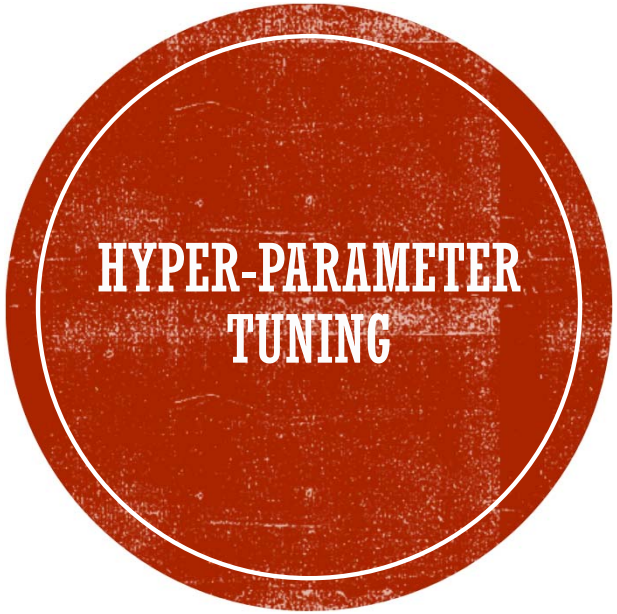


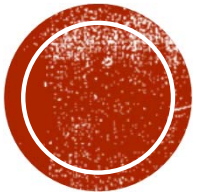
- Both models performed well on the text-only data that was split into 5 separate inputs
- Add the numeric variables (Work Remote, Has Questions, Has Company Profile, Salary Low, Salary High) to these text inputs
- Compare the performance on each model on this combined data versus the text-only data

- This project could only run the data through a Small-BERT model due to a lack of resources in Colab
- I would like to explore whether Classic BERT would provide better AUC scores than Small-BERT and then run the final testing data on the best-performing model

- BERT allows a maximum of 512 tokens
- Current model used 128 because Colab resources were not sufficient for inputs with 512 tokens
- I would like to see whether a model with 512 tokens for each input would improve on the results of this project

- Model Training Times (per epoch):
  - RNN Baseline Model: 25 to 36 minutes
  - Small-BERT: 3.5 to 4 hours
- The models did not overfit during training but also the available resources in Colab were not conducive to extensive parameter tuning
- I would like to see if these results could be further improved by tuning the hyper-parameters





CONCLUSIONS

PROBLEM WITH FALSE NEGATIVES

- The model predicted 109 real job postings in the test data as being fake
- The details (or lack thereof) provided in these real job postings share many characteristics with fraudulent posts

- Advice to Job Posters:
  - Include as much information as possible about the position's requirements and responsibilities so the post is not a "yellow flag" to the job seeker
  - Make sure the company has an active online presence to support its authenticity as a legitimate company and provide working links in the job posting
  - If provided, salary range should be realistic and clear

PROBLEM WITH FALSE POSITIVES

- The model predicted 4 fraudulent postings in the test data as being real
- Authentic-looking fake job postings attract the job seeker to submit a resume, providing the scammer with their personal job history and contact details
- Future requests for additional personal information or money to "speed up the application process" may be more believable to the job seeker if thought to be from a legitimate company, resulting in possible identity or financial theft.

- Advice to Job Seekers:
  - A lack of information in the job posting makes it more likely to be fraudulent
  - An emphasis on remote work and lack of an interview process are more likely to be in fraudulent posts
  - The company should have an online presence
    - Check out their website and the contact details provided
    - Look for past and current activity on their social media account(s)
    - Oftentimes, the company has its own career site with the same job posting – apply there
  - If in doubt, err on the side of caution and do not apply

QUESTIONS?