Predicting Fraudulent Job Postings

Presented by: Khashayar Behyar (Group 9)
Janzen Liu (Group 9)
Ikyu Park (Group 9)
Anthony Ransom (Group 9)

Date: August 2022

Background & Purpose

This Project was created knowing that:

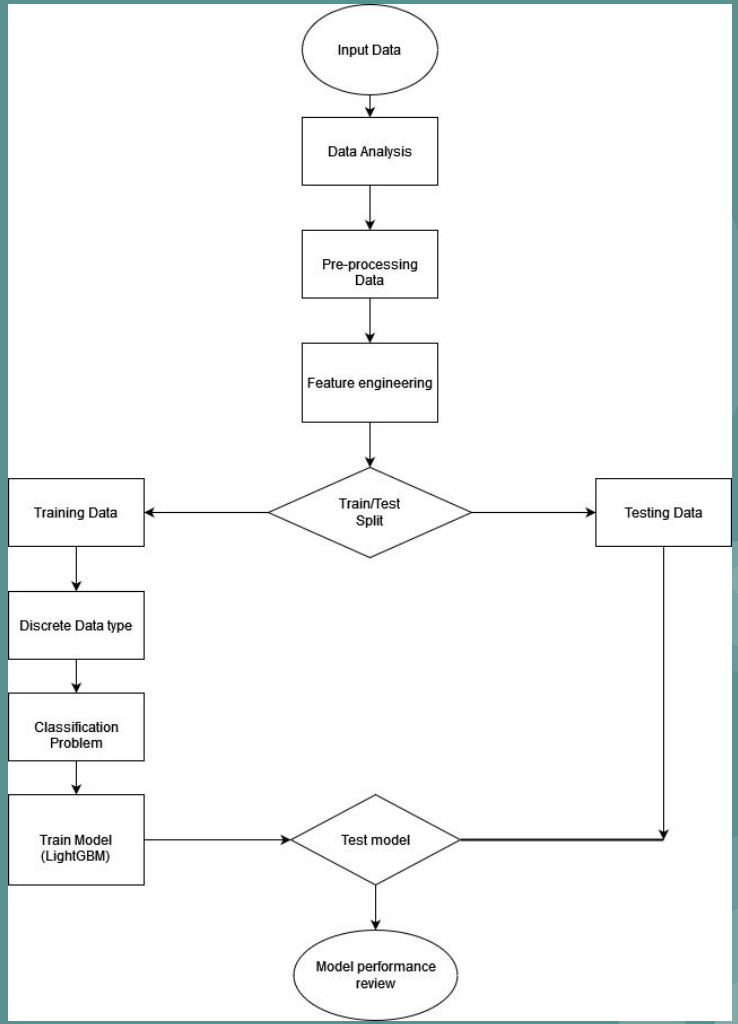
- Fraudsters and scammers use of phishing techniques within job posting to gain access to private information
- Unsuspecting job seekers provide personal information believes job posts on popular job boards are safe without knowing the dangers

The purpose of this project is to:

- Create a machine learning model to integrate with job board systems
- Accurately determine if a newly created job posting is fraudulent
- Prevent identity theft

Process

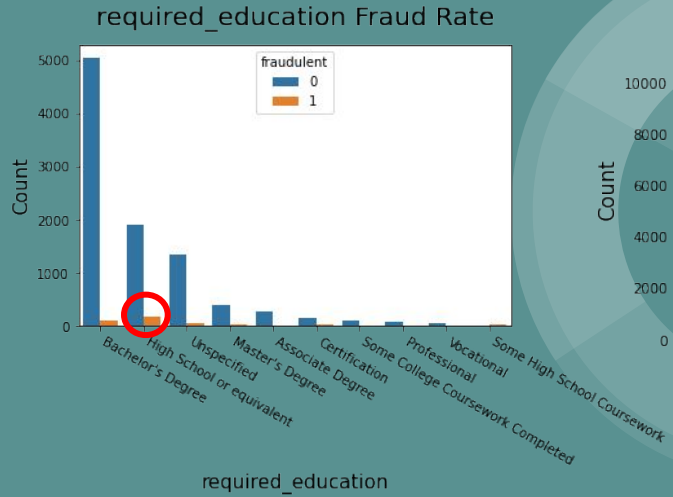
1. Data Analysis/Exploration
2. Data preprocessing
3. Feature engineering
4. Split data into training and testing datasets
5. Trained the machine learning model
6. Tested machine learning model
7. Review the model performance and tuned the model accordingly
8. Use model to score new data

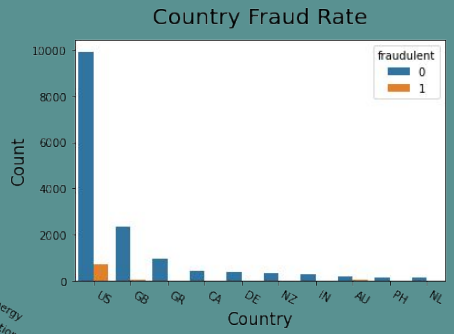


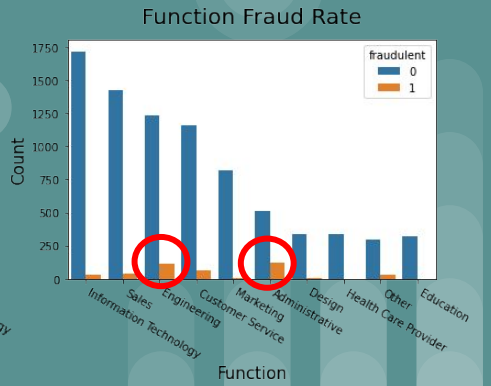
Data Analysis - Data Overview

- Dataset obtained from Kaggle
- Database Integration with Postgres & AWS RDS
- 17838 data points with 17 features and 1 target
- Identified target Feature in dataset marked as “Fraudulent”
- Target Feature data is skewed within dataset

Data Analysis/Exploration – Feature to Target Relationship







Data Preprocessing

- Job ID is an identification column and has no relation to the target Feature and was dropped
- Title contained majority unique values with a small word count and was dropped
- Salary_range has majority null values and was dropped - cutoff was 80% null values
- States and cities removed from dataset keeping only Country due to numerous incorrect spelling and entries with multiple cities and missing entries
- Null values changed to "Not Specified" for remaining features due to the already limited number of data points

	Percent Unique	Average Text Length
job_id	100.00	4.4
description	82.75	1209.0
requirements	66.92	585.6
title	62.80	28.5
benefits	34.69	207.1

	Percent Null
salary_range	83.959732
department	64.580537
required_education	45.329978
benefits	40.324385
required_experience	39.429530

Country

```
0       US; NY; New York
1         NZ; ; Auckland
2          US; IA; Wever
3     US; DC; Washington
4     US; FL; Fort Worth
Name: location, dtype: object
```

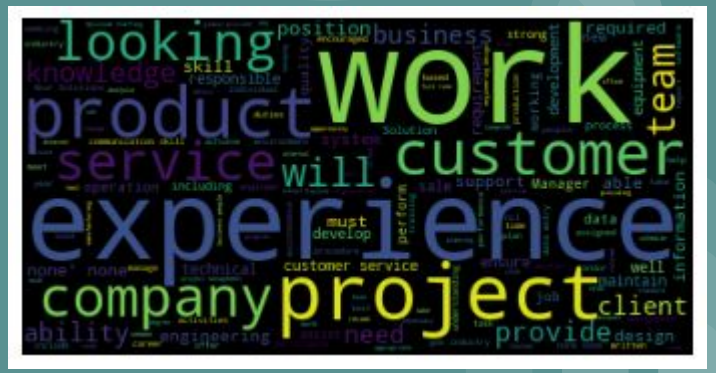
```
0    US
1    NZ
2    US
3    US
4    US
Name: Country, dtype: object
```

Feature Engineering - Natural Language Processing

- Natural Language Processing (NLP) used to process features with large amounts of text in each job posting
- Removal of punctuation and stopwords (“the”, “a”, “an”, “so”, “what”)
- NLP was used to identify the words that are most relevant to determine if a job posting is fraudulent

Non-Fraudulent Top Words

Fraudulent Top Words



Feature Engineering - Ordinal Data Types

- Ordinal Data types is data that has a natural/set order
  - Example: Internship -> entry level -> Associate -> Director -> Executive
- Label Encoder was used to convert these features to numeric form for the machine learning model

	employment_type	required_experience	required_education
0	other	internship	not specified
1	full-time	not applicable	not specified
2	not specified	not specified	not specified
3	full-time	mid-senior level	bachelor's degree
4	full-time	mid-senior level	bachelor's degree

	employment_type	required_experience	required_education
0	3	4	6
1	1	6	6
2	2	7	6
3	1	5	1
4	1	5	1

Feature Engineering - Nominal Data Types

- Nominal data is data grouped into categories with no meaningful order between the categories
    - Example: Computer Software industry vs Hospital & Health Care industry
- Target Encoding was used in order to convert these features to numerical form based on the mean of the target feature (Fraudulent) to the count for each category

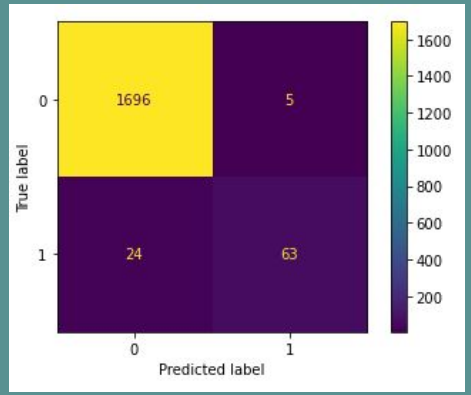
	department	industry	function	Country
0	marketing	not specified	marketing	us
1	success	marketing and advertising	customer service	nz
2	not specified	not specified	not specified	us
3	sales	computer software	sales	us
4	not specified	hospital & health care	health care provider	us

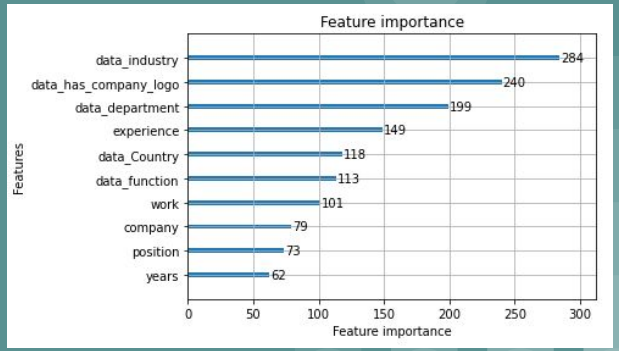
	department	industry	function	Country
0	0.004963	0.056088	0.012048	0.068506
1	0.000871	0.054348	0.054516	0.000000
2	0.045986	0.056088	0.052208	0.068506
3	0.029930	0.003634	0.027929	0.068506
4	0.045986	0.102616	0.002959	0.068506

Machine Learning Model

- Goal is to achieve high accuracy from the machine learning model
- Classification type machine learning model
- LightGBM machine learning model was chosen due to it higher performance to accuracy output
- Accuracy value of 98% meaning that 2 of every 100 job postings analysed are labeled incorrectly and 98 is correctly labeled
- Recall value of 72% means that 28 of every 100 fraudulent job postings analysed in reality are missed by the model and 72 are correctly identified as fraudulent

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1701
1	0.93	0.72	0.81	87
accuracy			0.98	1788
macro avg	0.96	0.86	0.90	1788
weighted avg	0.98	0.98	0.98	1788





Live Tableau Dashboard

FRAUDULENT DATA ANALYSIS & ML MODEL

Number of Fraudulent vs Non-Fraudulent

Fraudulent	
0	17,014
1	866

Location Fraudulent Rate

Word Cloud

FRAUDULENT NON-FRAUDULENT

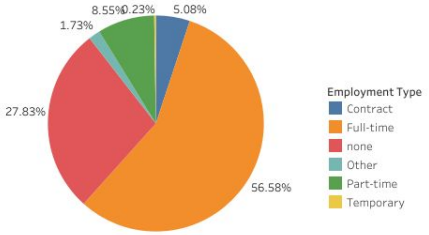
Required Educaiton Fraud Rate

Required Education	
none	451
High School or equivalent	170
Bachelor's Degree	100
Unspecified	61
Master's Degree	31
Some High School Course..	20
Certification	19
Associate Degree	6
Professional	4
Some College Coursework..	3
Doctorate	1
Vocational	0
Vocational - Degree	0

Industry Fraud Rate

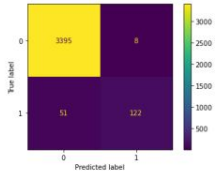
Industry	
none	275
Oil & Energy	109
Accounting	57
Hospital & Health Care	51
Marketing and Advertising	45
Financial Services	35
Information Technology a..	32
Telecommunications	26
Consumer Services	24
Real Estate	24
Leisure Travel & Tourism	21
Health Wellness and Fitne..	15
Hospitality	14
Computer Networking	12

Employment type Fraud Rate

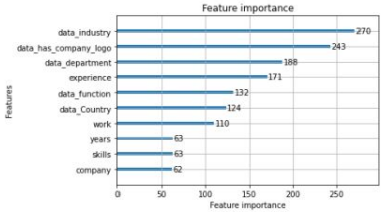


ML MODEL ANALYSIS

CONFUSION MATRIX (LIGHTGBM)



FEATURE IMPORTANCE (LIGHTGBM)



CLASSIFICATION REPORT (LIGHTGBM)

	precision	recall	f1-score	support
0	0.99	1.00	0.99	1701
1	0.93	0.72	0.81	87
accuracy			0.98	1788
macro avg	0.96	0.86	0.90	1788
weighted avg	0.98	0.98	0.98	1788

Conclusion and Improvements

- Fraud through job postings is prevalent which our machine learning model achieved high accuracy in predicting the job postings bad actors are using to gain access to private information.

- A larger dataset can be used to improve the machine learning model and once the model has been implemented in a live system will improve over time as the model learns

- This model could be implemented to analyse each new job posting that is created to determine if a job posting would be used for fraudulent purposes and prevent the job postings from going onto the live public job board

Q&A