



Real/Fake Job-Posting Prediction

EGR 598: Final Project

- **Apoorva Uplap**
- **Rohan Makarand Khare**
- **Saketh Audipudi**
- **Holmes Joseph**

Problem and Motivation

- This project addresses the rising concern of fraudulent job postings on online platforms through a machine learning approach.
- By analyzing a dataset and employing machine learning and natural language processing techniques, we aim to develop a robust model for distinguishing between real and fake job postings.

DATA SOURCE: From Where

- The dataset is called “Employment Scam Aegean Dataset (EMSCAD)”
- The dataset is compiled and published by the University of Aegean, Mytilene, Greece.
- The dataset has 17,880 total entries of which 866 are of fraudulent job postings.

Reference - Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. *Future Internet*, 9(1), 6.

<http://emscad.samos.aegean.gr/>

DATA SOURCE: How Reliable

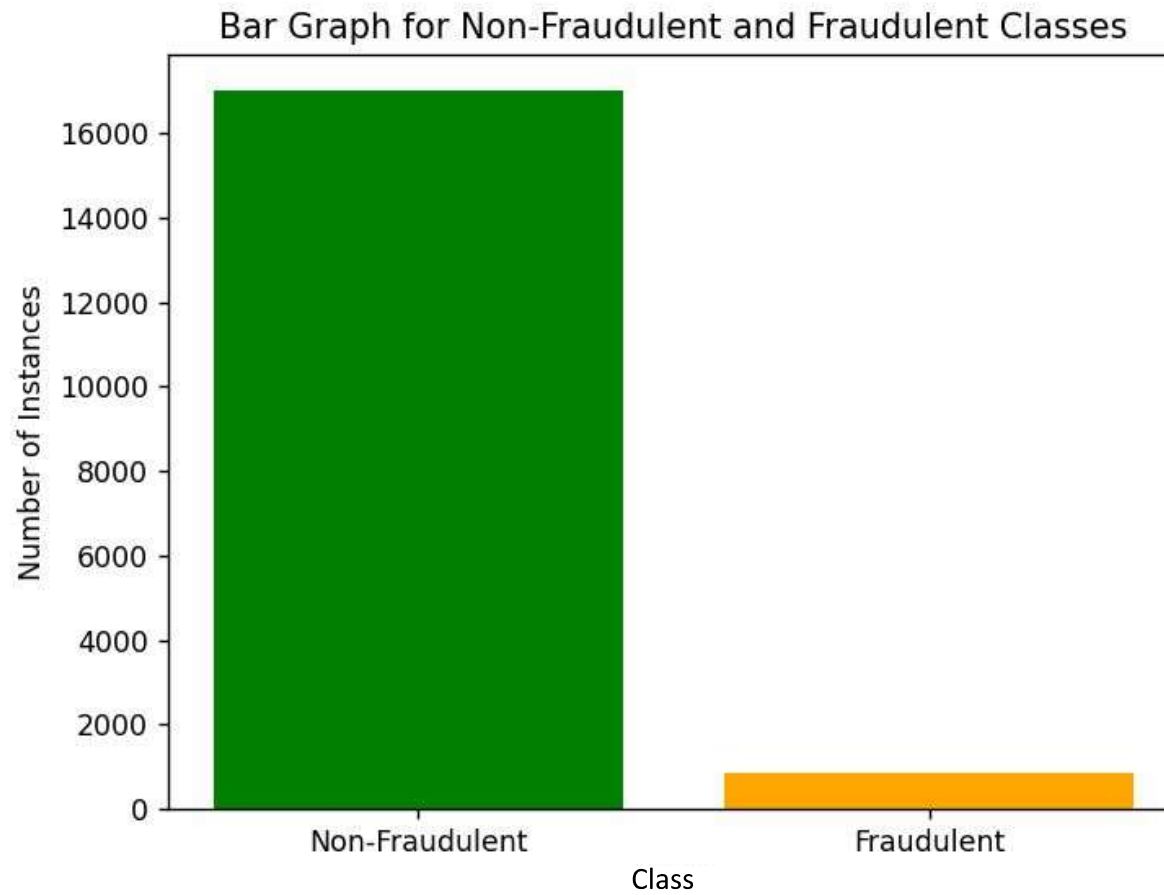
- The dataset has been compiled by employees and the annotation process pertained to out-of-band quality assurance procedures.
- The criteria for the classification were based on client's suspicious activity on the system, false contact or company information, candidate complaints and periodic meticulous analysis of the clientele.

DATA SOURCE: Attributes

- Each record in the dataset is represented as a set of structured and unstructured data. It is formally described by a set of fields $F = \{F1, \dots, Fn\}_{n=16}$, and a binary class field $C\{+, -\}$ indicating whether the current job ad entry is fraudulent or not.

No. of Columns	Attribute type	Column Names
12	String	title, location, department, company_profile, description, requirements, benefits, employment_type, required_experience, required_education, industry, function, salary_range
4	Nominal {0,1}	telecommuting, has_company_logo, has_questions, fraudulent

DATA CHARACTERISTICS



DATA CHARACTERISTICS

Observations:

- The data is highly imbalanced, EMSCAD contains 17,014 legitimate and 866 fraudulent job ads.
- There are only 3 numeric columns and the rest contain strings which would require further manipulation for model fitting.

```
> summary(job_data)
```

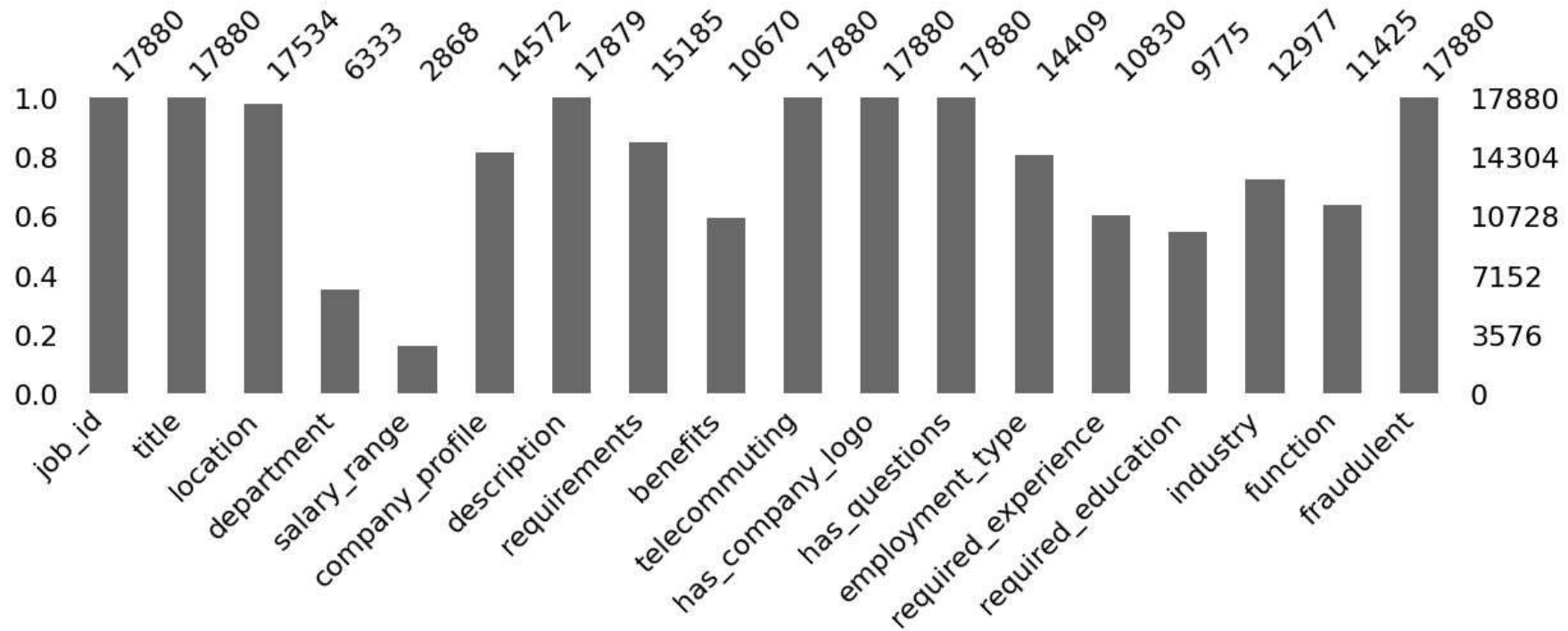
job_id	title	location	department	salary_range	company_profile	description	requirements	benefits
Min. : 1	Length:17880	Length:17880	Length:17880	Length:17880	Length:17880	Length:17880	Length:17880	Length:17880
1st Qu.: 4471	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character	Class :character
Median : 8940	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character
Mean : 8940								
3rd Qu.:13410								
Max. :17880								

telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function.	fraudulent
Min. :0.0000	Min. :0.0000	Min. :0.0000	Length:17880	Length:17880	Length:17880	Length:17880	Length:17880	Min. :0.00000
1st Qu.:0.0000	1st Qu.:1.0000	1st Qu.:0.0000	Class :character	Class :character	Class :character	Class :character	Class :character	1st Qu.:0.00000
Median :0.0000	Median :1.0000	Median :0.0000	Mode :character	Mode :character	Mode :character	Mode :character	Mode :character	Median :0.00000
Mean :0.0429	Mean :0.7953	Mean :0.4917						Mean :0.04843
3rd Qu.:0.0000	3rd Qu.:1.0000	3rd Qu.:1.0000						3rd Qu.:0.00000
Max. :1.0000	Max. :1.0000	Max. :1.0000						Max. :1.00000

```
> |
```

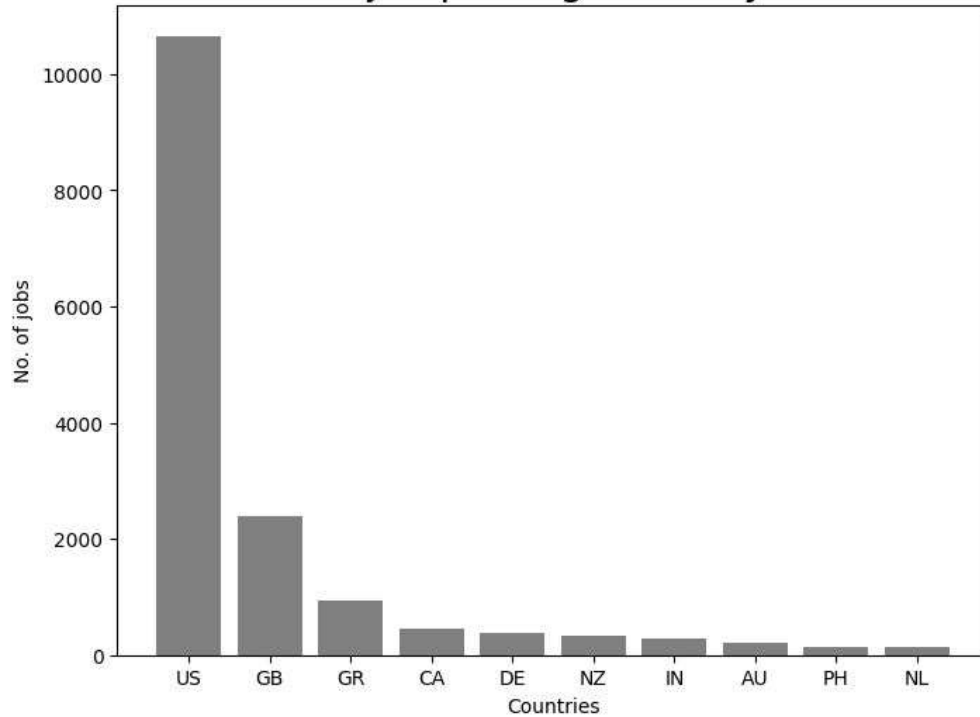
DATA CHARACTERISTICS

Missing Data:

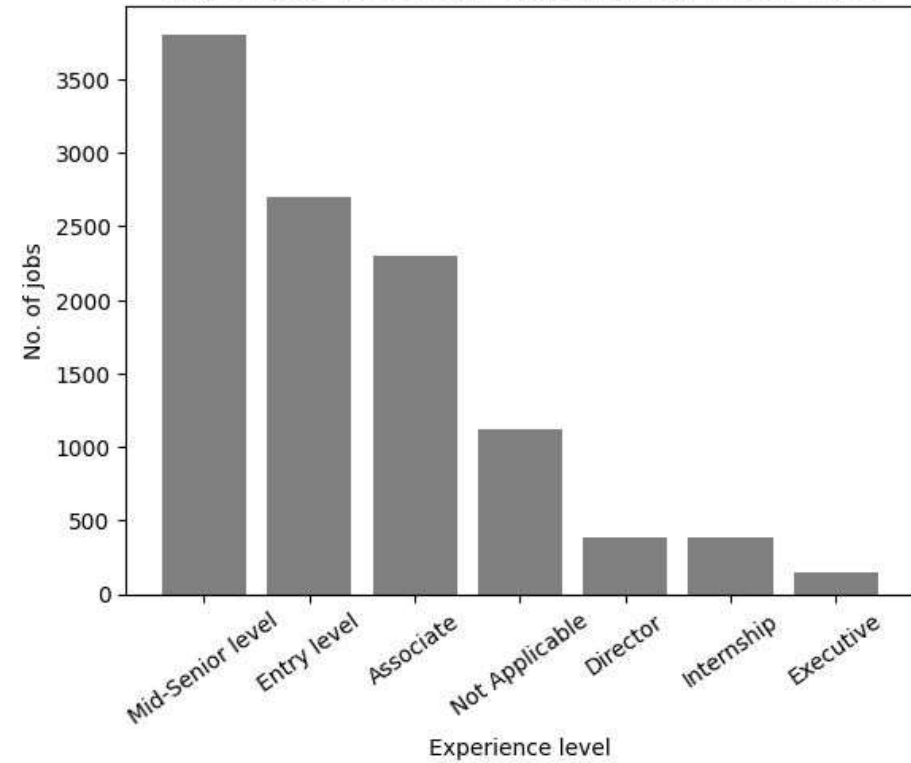


DATA CHARACTERISTICS

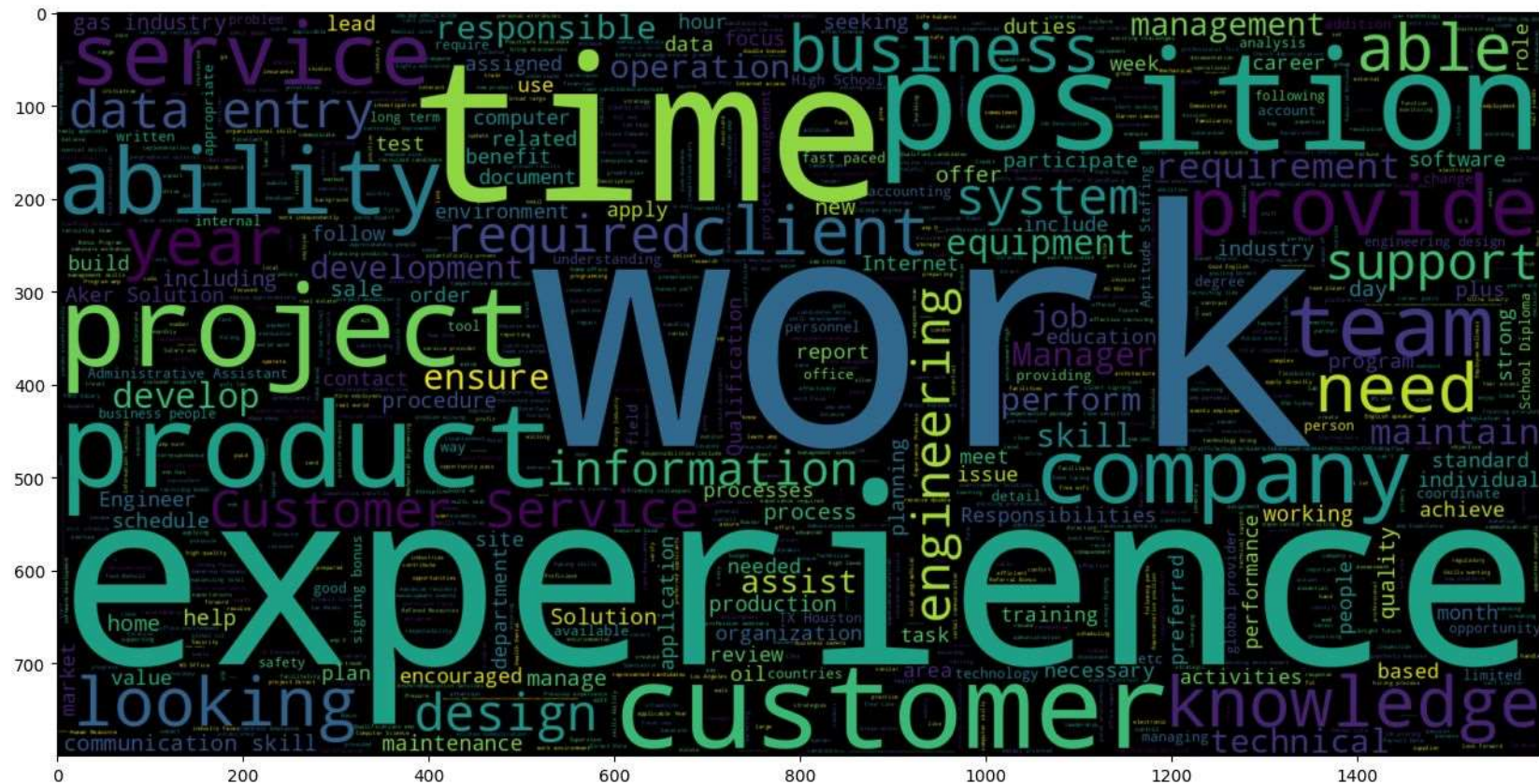
No. of job postings country wise



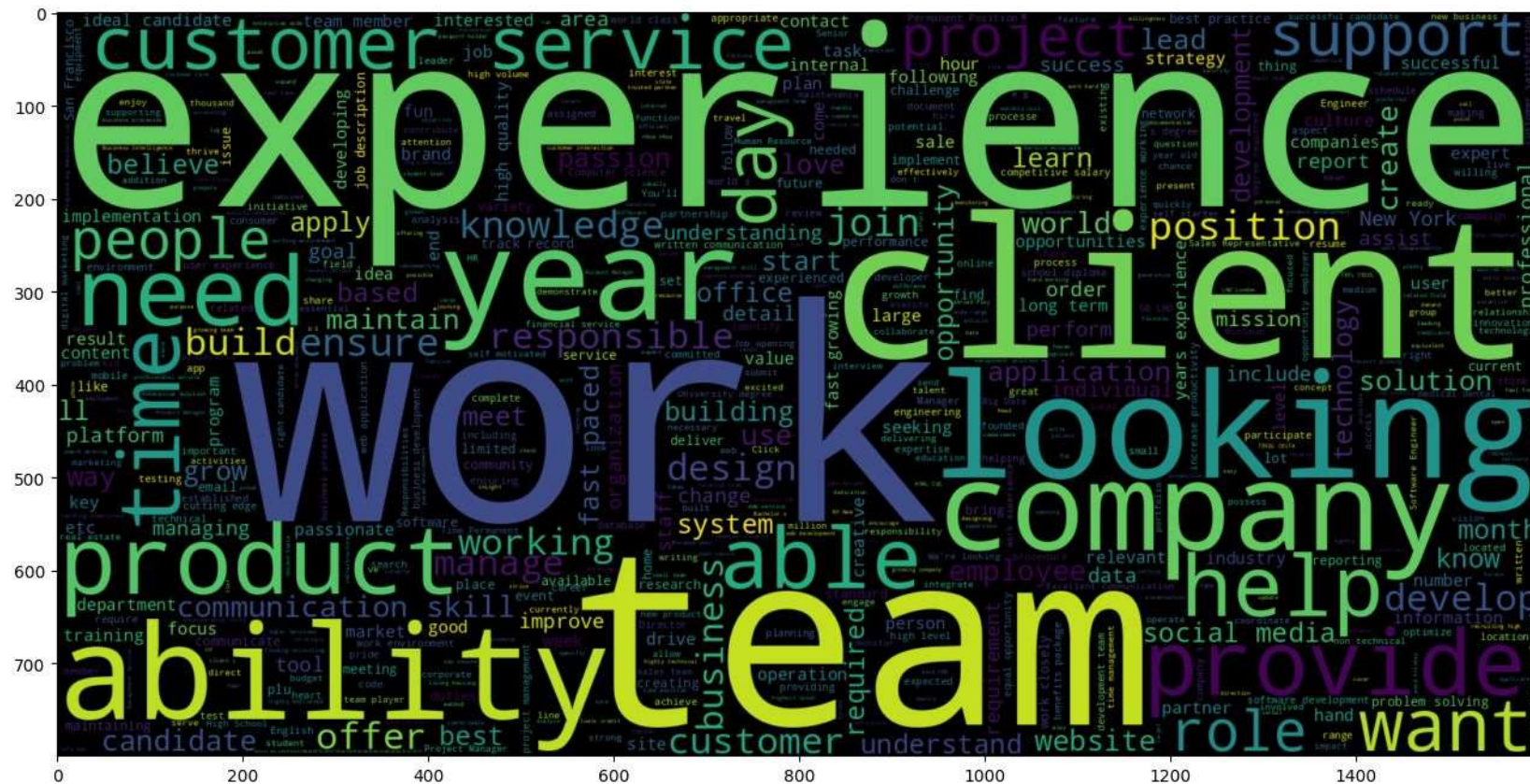
No. of job postings experience level wise



WordCloud visualization of frequent words in fraudulent job ads:



WordCloud visualization of frequent words in actual job ads:



MODELLING METHODS

- Exploratory Data Analysis is done for the dataset.
- We have built multiple models and compared the performance of each later.
- In a few models, we have taken just the numeric columns and built classifiers.
- In the rest, we have manipulated the text columns and classified accordingly.
- There are entries which have words or even sentences which must be converted to a numerical form for ease of calculation.
- For models like RandomForest, Logistic Regression etc. we skipped the columns which have large entries and only considered columns which have singular entries. These are then scaled uniformly using standard scalar.
- 80-20 split was used for modelling.

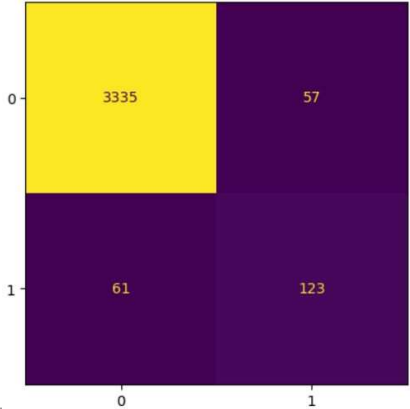
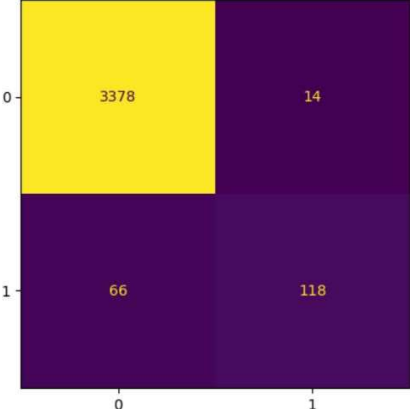
MODELLING METHODS - ENCODED DATASET

	department	salary_range	telecommuting	has_company_logo	has_questions	employment_type	required_experience	required_education	industry	function
0	0	0	0	1	0	0	0	-1	-1	0
1	1	0	0	1	0	1	1	-1	0	1
2	-1	0	0	1	0	-1	-1	-1	-1	-1
3	2	0	0	1	0	1	2	0	1	2
4	-1	0	0	1	1	1	2	0	2	3
...
17875	2	0	0	1	1	1	2	-1	1	2
17876	77	0	0	1	1	1	2	0	8	26
17877	-1	0	0	0	0	1	-1	-1	-1	-1
17878	-1	0	0	0	1	3	1	8	82	9
17879	15	0	0	1	1	1	2	-1	1	7

MODELLING METHODS

1. **Decision Tree:** A decision tree is a flowchart-like structure where each node represents a decision based on a feature, aiming to split data into subsets and ultimately predict the target variable.
2. **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes, providing better accuracy and robustness through the combination of individual tree predictions.

MODELLING RESULTS

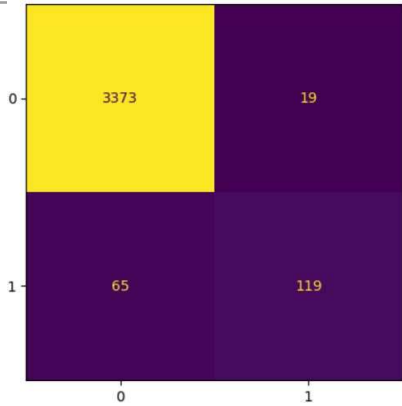
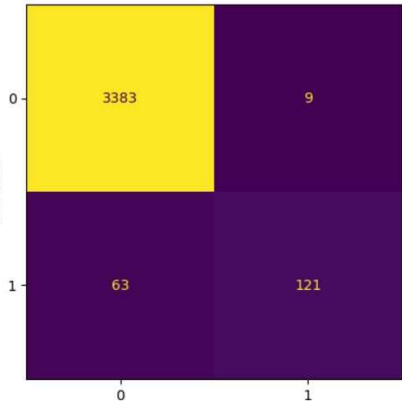
METHOD(CLASSIFICATION)	ACCURACY	BALANCED ACCU	CONFUSION MATRIX
DECISION TREE	96.70	82.58	
RANDOM FOREST	97.76	81.86	



MODELLING METHODS

1. **XGBoost:** XGBoost is an efficient and scalable gradient boosting framework that uses decision trees as base learners, employing a regularization term to control model complexity and enhance predictive performance.
2. **LightGBM:** LightGBM is a gradient boosting framework designed for speed and efficiency, using a tree-based learning algorithm and histogram-based approach to split data, making it particularly effective for large datasets.

MODELLING RESULTS


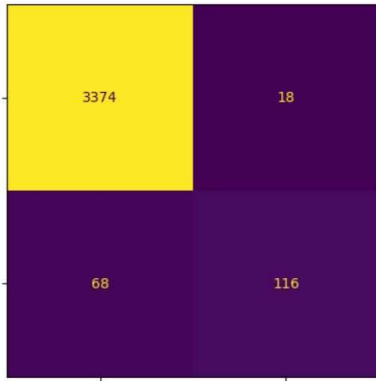
METHOD(CLASSIFICATION)	ACCURACY	BALANCED ACCURACY	CONFUSION MATRIX									
XGB	97.65	82.06	 <table><tr><td>0</td><td>3373</td><td>19</td></tr><tr><td>1</td><td>65</td><td>119</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3373	19	1	65	119		0	1
0	3373	19										
1	65	119										
	0	1										
LGBM	97.99	82.75	 <table><tr><td>0</td><td>3383</td><td>9</td></tr><tr><td>1</td><td>63</td><td>121</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3383	9	1	63	121		0	1
0	3383	9										
1	63	121										
	0	1										



MODELLING METHODS

1. **BaggingClassifier:** BaggingClassifier is an ensemble method that combines multiple models, typically decision trees, by training each on different subsets of the training data to improve overall classification accuracy and reduce overfitting.
2. **ExtraTree:** ExtraTree, or Extremely Randomized Trees, is an ensemble learning method similar to Random Forest, but it builds decision trees with random splits, making it computationally more efficient while maintaining strong predictive performance.

MODELLING RESULTS

METHOD(CLASSIFICATION)	ACCURACY	BALANCED ACCURACY	CONFUSION									
BAGGING	97	81	 <table><tr><td>0</td><td>3368</td><td>24</td></tr><tr><td>1</td><td>69</td><td>115</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3368	24	1	69	115		0	1
0	3368	24										
1	69	115										
	0	1										
ExtraTrees	97.60	81.26	 <table><tr><td>0</td><td>3374</td><td>18</td></tr><tr><td>1</td><td>68</td><td>116</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3374	18	1	68	116		0	1
0	3374	18										
1	68	116										
	0	1										



MODELLING METHODS

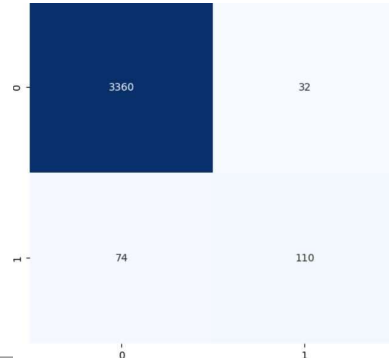
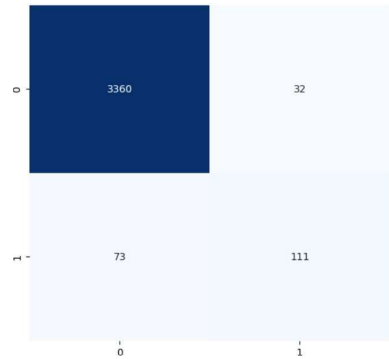
1. Label Propagation:

Label Propagation is a semi-supervised learning algorithm that assigns labels to data points based on the labels of their neighbors, propagating information through the underlying structure of the data graph. It operates by iteratively updating labels, considering the majority class of neighboring points, making it effective for datasets with limited labeled samples.

2. Label Spreading:

Label Spreading is an extension of Label Propagation that introduces a softening factor, allowing labels to be distributed probabilistically among neighbors, providing a more nuanced and flexible approach to semi-supervised classification. By incorporating label spreading, this algorithm accommodates uncertainty and yields more nuanced predictions, particularly beneficial when dealing with noisy or ambiguous data.

MODELLING RESULTS

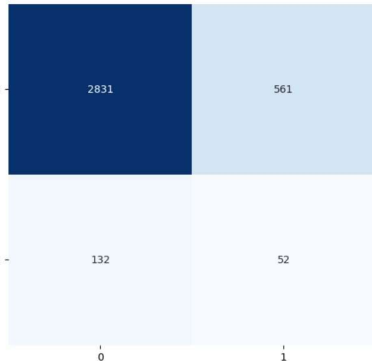
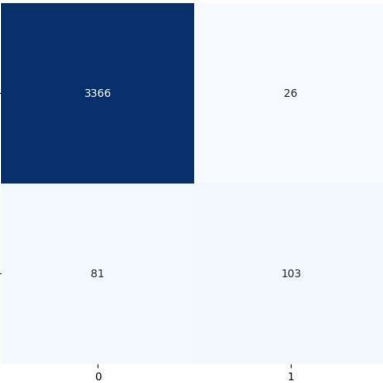
METHOD(CLASSIFICATION)	ACCURACY	BALANCED ACCURACY	CONFUSION												
LABEL PROPAGATION	97	79	 <table><tr><td>0</td><td>3360</td><td>32</td></tr><tr><td>1</td><td>74</td><td>110</td></tr><tr><td>0</td><td></td><td>1</td></tr><tr><td>1</td><td></td><td></td></tr></table>	0	3360	32	1	74	110	0		1	1		
0	3360	32													
1	74	110													
0		1													
1															
LABEL SPREADING	97	80	 <table><tr><td>0</td><td>3360</td><td>32</td></tr><tr><td>1</td><td>73</td><td>111</td></tr><tr><td>0</td><td></td><td>1</td></tr><tr><td>1</td><td></td><td></td></tr></table>	0	3360	32	1	73	111	0		1	1		
0	3360	32													
1	73	111													
0		1													
1															



MODELLING METHODS

1. **Nearest Centroid:** Nearest Centroid classifies samples based on the class of the nearest centroid, calculated as the mean of each class's feature values, making it a simple and efficient algorithm for classification tasks.
2. **KNN (K-Nearest Neighbors):** KNN is a non-parametric algorithm that classifies a data point based on the majority class of its k-nearest neighbors, making it effective for both classification and regression tasks.

MODELLING RESULTS

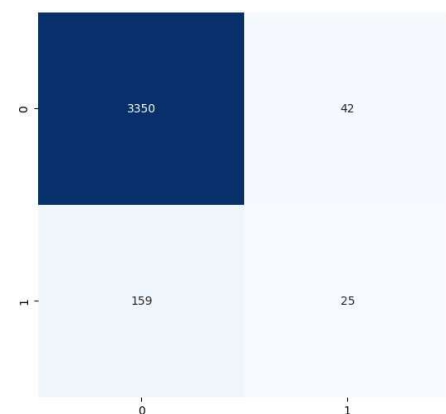
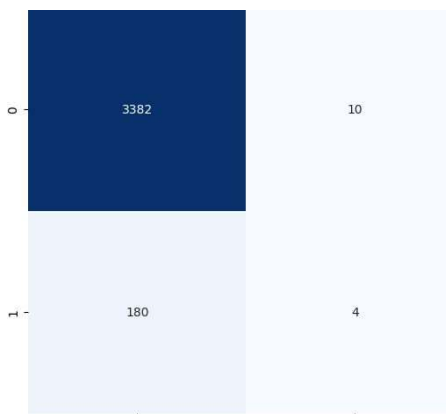
METHOD(CLASSIFICATION)	ACCURACY	BALANCED ACCURACY	CONFUSION									
NEAREST CENTROID	81	56	 <table><tr><td>0</td><td>2831</td><td>561</td></tr><tr><td>1</td><td>132</td><td>52</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	2831	561	1	132	52		0	1
0	2831	561										
1	132	52										
	0	1										
K NEAREST NEIGHBORS	97	56	 <table><tr><td>0</td><td>3366</td><td>26</td></tr><tr><td>1</td><td>81</td><td>103</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3366	26	1	81	103		0	1
0	3366	26										
1	81	103										
	0	1										



MODELLING METHODS

1. **Perceptron:** The Perceptron is a linear binary classification algorithm that learns a hyperplane to separate classes, adjusting its weights based on misclassifications to converge towards an optimal decision boundary.
2. **PassiveAggressive Classifier:** The PassiveAggressive Classifier is an online learning algorithm for binary classification that adapts its model incrementally, updating weights to minimize loss when misclassifications occur in streaming data.

MODELLING RESULTS

METHOD(CLASSIFICATION)	ACCURACY	BALANCED ACCURACY	CONFUSION									
PERCEPTRON	94	56	 <table><tr><td>0</td><td>3350</td><td>42</td></tr><tr><td>1</td><td>159</td><td>25</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3350	42	1	159	25		0	1
0	3350	42										
1	159	25										
	0	1										
PASSIVE AGGRESSIVE	95	56	 <table><tr><td>0</td><td>3382</td><td>10</td></tr><tr><td>1</td><td>180</td><td>4</td></tr><tr><td></td><td>0</td><td>1</td></tr></table>	0	3382	10	1	180	4		0	1
0	3382	10										
1	180	4										
	0	1										

MODELLING METHODS

Support Vector Machine: Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. It is highly effective in high-dimensional spaces and is suited for cases when the data is not linearly separable.

SVM works by finding the hyperplane that best separates the data into different classes using support vectors (data points that lie closest to the decision boundary) as a minimization metric.

MODELLING METHODS - SVM

The text fields were now taken and everything else was dropped. All the text columns were combined into a single column. A sample text field looked like this:

BI Developer IL, TA, Tel Aviv The only thing we love more than our data is our team. We're a group of developers, designers, mathematicians, data scientists, researchers and marketers that work relentlessly to measure online behavior worldwide and to generate marketing insights. Together, we are shaping the future of web measurement and competitive intelligence. SimilarWeb is a technology-driven Big Data internet company. Our products are based on unique, proprietary technology and data, and use sophisticated algorithms to analyze every website on the internet. We pride ourselves on providing users with comprehensive and beneficial information, giving them valuable insights about their competitor's traffic. As a BI developer, the ideal candidate will have a very good Java background. We are looking for someone who is highly passionate about data, databases and data analysis. We will prefer someone who is eager to learn and keeps his knowledge up to date with new methodologies, best practices and technologies. The day to day will be working closely with our business analysts, developing our BI platform and integrate data for internal and external sources. Skills and Requirements: Minimum 2 years' Experience of Java Development Experience with Databases Love for Data Business Oriented Experience in BI Development - Advantage Highly passionate about architecture and server code design TDD and DDD Strong willingness to learn

MODELLING METHODS - SVM

We created a corpus of high frequency words occurring in actual and fake job postings in R, added the fraudulent dependent column and trained a SVM model on a subset of the data due to high imbalance with 70-30 train_test_split.

```
> summary(sparse_df)
```

administr	affili	amp	andor	assist
communiti	content	curat	current	daytoday
execut	exist	fastgrow	file	food
mail	manag	market	new	number
prepar	program	provis	screen	site
various	websit	work	york	X100
and	anywher	aspect	auckland	beauti
cloud	commiss	communic	compani	cool
deliveri	drive	easi	enabl	enter
focus	get	global	googl	grow
huge	includ	innovat	issu	its
need	network	next.	opportun	organis
process	produc	product	profession	project
rate	regular	remov	repres	right
skill	speed	stage	success	talent
video	world	X.we	activ	assistant
environment	experienc	health	houston	ideal

MODELLING RESULTS - SVM

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	1677	29
1	35	47

Accuracy : 0.9642

95% CI : (0.9545, 0.9723)

No Information Rate : 0.9575

P-Value [Acc > NIR] : 0.08622

Kappa : 0.5762

McNemar's Test P-Value : 0.53197

Sensitivity : 0.9796

Specificity : 0.6184

Pos Pred Value : 0.9830

Neg Pred Value : 0.5732

Prevalence : 0.9575

Detection Rate : 0.9379

Detection Prevalence : 0.9541

Balanced Accuracy : 0.7990

'Positive' Class : 0

- The model performed well on the test set.
- As the data is imbalanced, many fake job postings were misclassified.
- From the confusion matrix,
TPR = 0.573
FPR = 0.021
FNR = 0.381
TNR = 0.983
F1 score = 0.594

LSTM and Text Generation

- This code is a comprehensive script that uses a Long Short-Term Memory (LSTM) neural network to generate text for fake job postings. The LSTM is a part of a larger workflow which involves data loading, preprocessing, text generation, and the creation of new job postings.
- Python script is designed to generate text using a machine learning model, specifically for the purpose of creating content related to job postings.
- It works by augmenting a seed text with words predicted by a model and also inserts random categories related to job postings.

GOAL

- To create an entry for the job postings dataset using LSTM text generation for the Verbose columns in the dataset, and randomly fill in the various categorical and other data types.
- The goal is to use LSTM to generate “Fake Job Postings” and evaluate its performance.
- To achieve this we will Train a model on the corpus of the job description column since it is the most dense.
- Ideally we would train models for each of the text columns ie; Company Profile, Job Description, Requirements and Benefits. To give a tailored output.
- However this is computationally expensive and we can achieve the same goal by choosing a proper dataset for a single model.

Data Preparation and Preprocessing

- Loading Data: We first load a CSV file containing job postings.
- Data Filtering: It filters the dataset to select non-fraudulent job postings and shuffles them.
- Text Tokenization: The job descriptions are tokenized, converting the text into sequences of integers.
- Padding Sequences: These sequences are then padded to a uniform length.
- Preparing Inputs and Labels: The sequences are split into inputs and labels for training the LSTM model.
- One-Hot Encoding: The labels are one-hot encoded, a common practice in multi-class classification tasks.

LSTM Model Specifications

Model Architecture: The model is a Sequential model from Keras, which is defined by a linear stack of layers.

Embedding Layer: The first layer is an Embedding layer, which transforms the input sequence into dense vectors of fixed size (100 in this case). The number of possible words is used as the input dimension.

LSTM Layers: There are three LSTM layers.

1.The first LSTM layer has 200 units and returns sequences, meaning it outputs the full sequence of outputs for each input sequence (useful for stacking LSTM layers).

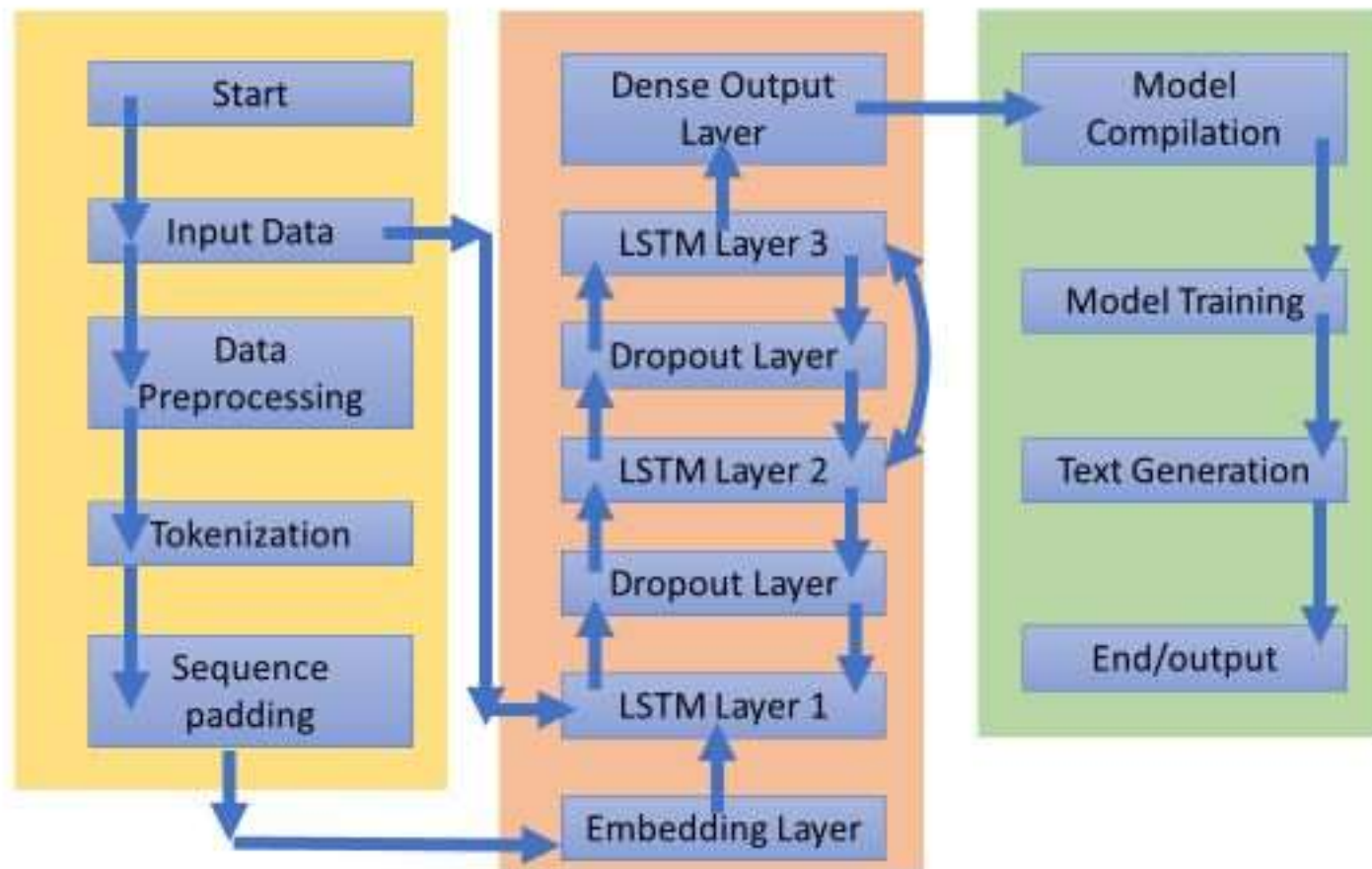
2.The second LSTM layer has 150 units and also returns sequences.

3.The third LSTM layer has 100 units and does not return sequences, meaning it outputs only the last output in the output sequence.

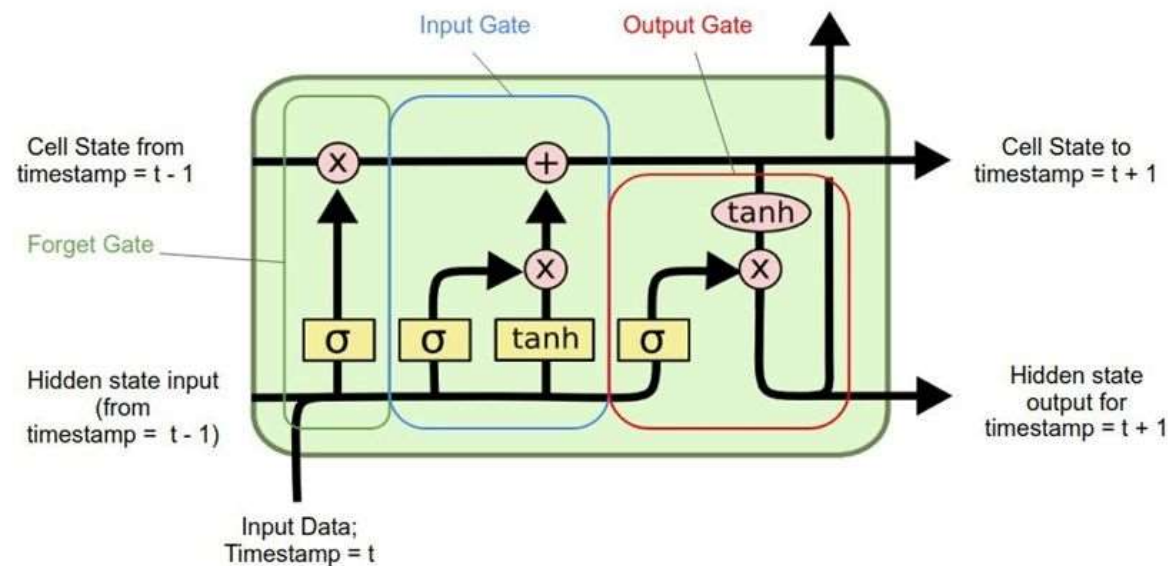
Dropout Layers: After the first and second LSTM layers, there are Dropout layers with a rate of 0.2, used to prevent overfitting by randomly setting input units to 0 at each update during training time.

Dense Layer: The final layer is a Dense layer with a softmax activation function, used for multi-class classification. The size of this layer is equal to the number of unique words.

Process Flow:



- Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) that is specifically designed to handle sequential data, such as time series, speech, and text.
- LSTM networks are capable of learning long-term dependencies in sequential data, which makes them well suited for tasks such as language translation, speech recognition, and time series forecasting.
- LSTM has a chain structure that contains four neural networks and different memory blocks called cells.





job_id	title	location	salary_range	company_profile	description	requirements	benefits	telecommuting	has_company_logo	has_questions	employment_type	required_e
17880	Massage Therapist	US, CO, denver		Data entry Company Director Internship Associa...	Executive Contract Sales/Engineering High Scho...	Options Some College Coursework Completed Publ...	Analyst Hospital & Industrial Automation Healt...	0	1	1	Full-time	
17881	Full-stack JavaScript Software Developer	IN, , Bangalore		Education Engineering PSG Full-time Quality Mu...	Management Job Education Information Technolog...	Part-time Media Professional Executive Product...	Some High Full-time Director School Coursework...	0	0	0	Other	
17882	Drupal Developer - Remote - USA Only	US, MA, Westborough		Certification Not High School or equivalent Ap...	Mid-Senior level High Temporary Part-time Scho...	Computer Software Training Consumer Goods Qual...	Certification Management Team Director Benefit...	1	1	0	Contract	
17883	Software Development Engineer	US, NJ, Jersey City		Distribution Fabrication Company Profile: Cont...	SMRA Training Maritime Part- time Temporary Min...	Associate PLANNING AND GENERAL AFFAIRS Legal R...	Entry level High Environmental Services Intern...	1	0	0	Contract	
17884	Information Technology Technician	GB, , Bury St Edmunds		Part-time Content Programming Other Company In...	Job Data Analyst Meanwood Director Associate B...	Internship Other Business Contract Tech Market...	Insurance Benefits: Mid- Senior level SALES The...	0	1	0	Other	Mid-

Seed text: “Job Description”

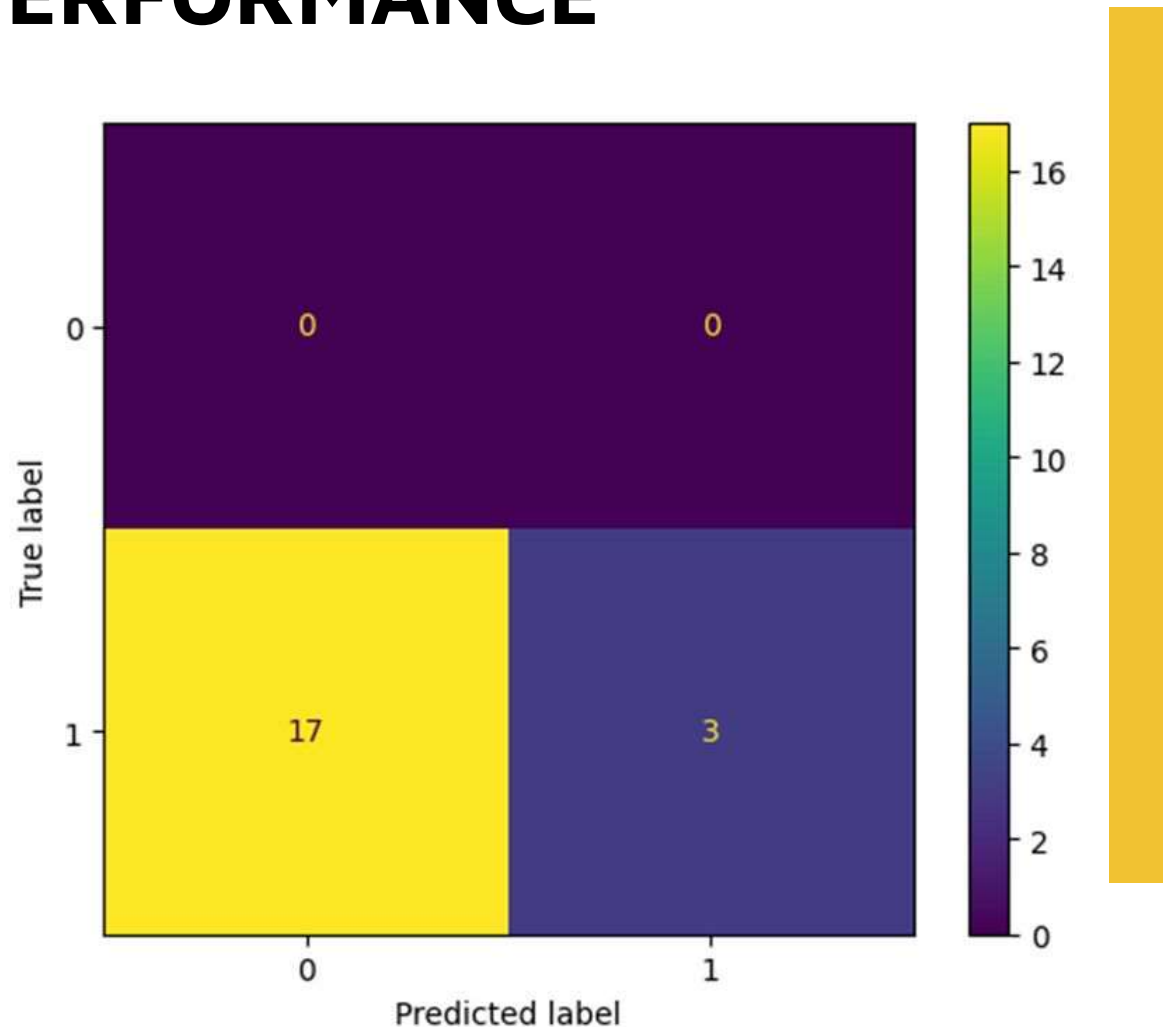
- Company Consumer Goods Full-time Profile: Vocational - Computer Accounting Networking Entry level Degree Printing Motion Pictures and Film Vocational Vocational - Degree Distribution Associate Unspecified Marketing Hospitality Lab Sales HIS Broadcast Electrical/Electronic Engineering Manufacturing Photography Media Part-time Marketing.....

Seed text: “Requirements”

- Requirements: Contract Options Away Art/Creative Temporary Part-time Events Temporary Unspecified Department Technology Innovation Not Entrepreneur Applicable Account/finance Venture Associate Not Applicable Research Capital Success / Sales / Production & Private Human Resources Equity Product Contract Ops

MODEL PERFORMANCE

- In order to evaluate the performance of our model we generated a database of job postings and tested them against the Decision tree classification model we defined.
- Of the 20 fake entries, when tested if they were fraudulent we got:
 - 3 instances were fake and correctly identified (true positives).
 - 17 instances were fake but incorrectly labeled (false negatives).
- This means that our LSTM model performed well as 17 fake entries were able to pass as original ones



DISCUSSION

- We were able to apply different models on the dataset and study the performance of each.
- As the data is highly imbalanced, we have used under-sampling and resampling.
- Working with a dataset comprised of text, numeric fields, and binary columns was a good learning experience.
- We were also able to employ NLP techniques to generate new job postings from the data.

Q/A

