Predicting Real or Fake Job Postings Using Machine Learning

Eshan Bhatnagar
Fengshou Liang
Jane Su
December 12, 2023

Berkeley
UNIVERSITY OF CALIFORNIA

Introduction

- Overwhelming number of job postings on different platforms
- Fake job postings created a misleading impression of an inflated job market
- Lead to frustration and mistrust among job seekers, damaging their perception of the hiring process
- Damage the reputation of job search platforms and the companies using them
- The successful identification of these fake job posting can be helpful to improve the overall integrity of the job market

Description of Data

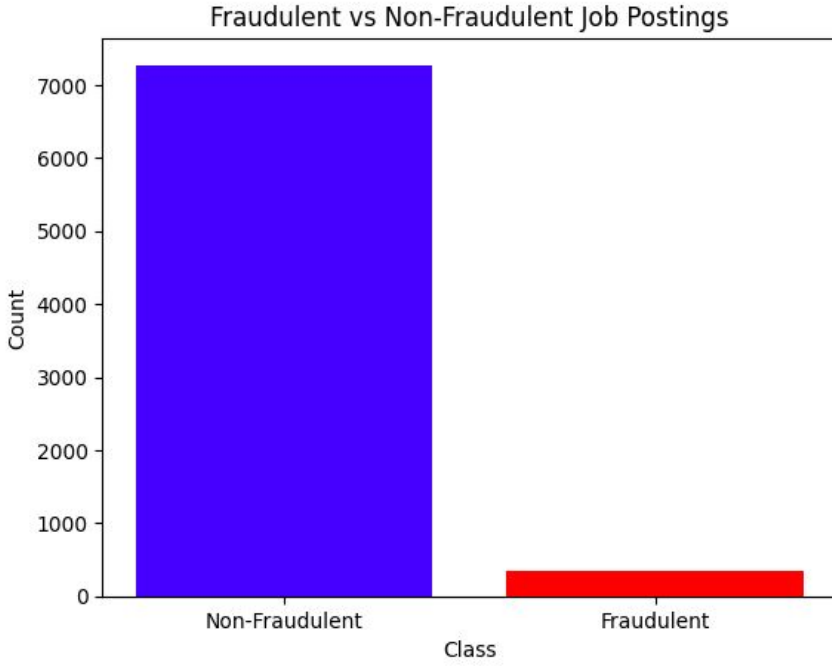
- Identify fake job description based on textual information
- Fake job posting dataset from Kaggle
- Containing 18k job postings
- Columns include: Job ID, Title, Location, Department, Salary Range, Company Profile, Description, Requirements, Benefits, Telecommuting, Benefits, Company Logo, Questions, Employment Type, Required Experience, Required Education, Industry, Function, and Fraudulent (target variable)

Data Cleaning and Feature Selection

- Dropped first few text-based columns since they were too descriptive and varied for initial logistic regression and random forest models
- Dropped all rows with null values
- Resulting features are ‘telecommuting’, ‘has_company_logo’, ‘has_questions’, ‘employment_type’, ‘required_experience’, ‘required_education’, ‘industry’, and ‘function’

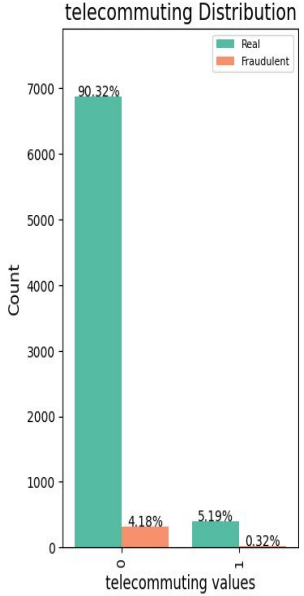
Imbalanced Data

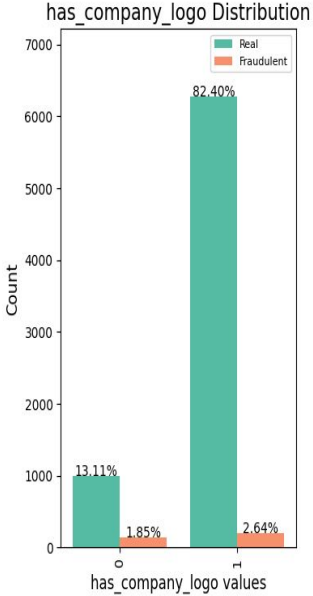
- Only 4.7% of the data is labeled ‘Fraudulent, indicating imbalanced distribution
- Accuracy alone is not sufficient to evaluate models, precision and recall also needed
- Additional solutions to balance the data were also implemented

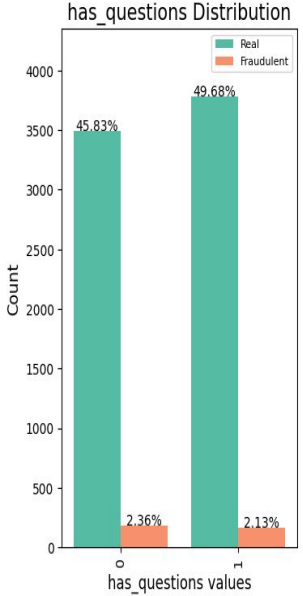


Exploratory Data Analysis

- Majority of telecommuting values for both real and fraudulent postings are 0, we dropped this as a feature
- Majority of fraudulent and real postings have company logo although the proportion for real postings is greater
- Distribution for having questions is roughly equal for both real and fraudulent postings

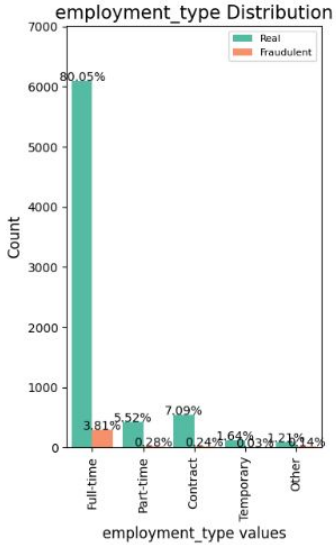


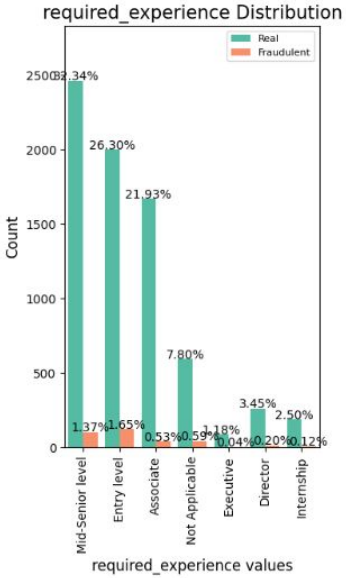


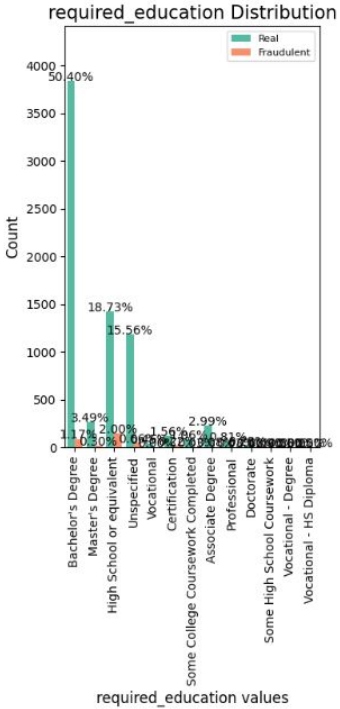


Exploratory Data Analysis

- Majority of distribution for employment type is full-time for both real and fraudulent postings
  - Contract has the second highest count for real postings, while part-time has second highest count for fraudulent postings
- Majority of required experience for real postings is mid-senior level, for fraudulent it's entry level
- Majority of education level for real postings is Bachelor's degree, for fraudulent it's high school
- Biggest discrepancy between real and fake postings is in education level and required experience



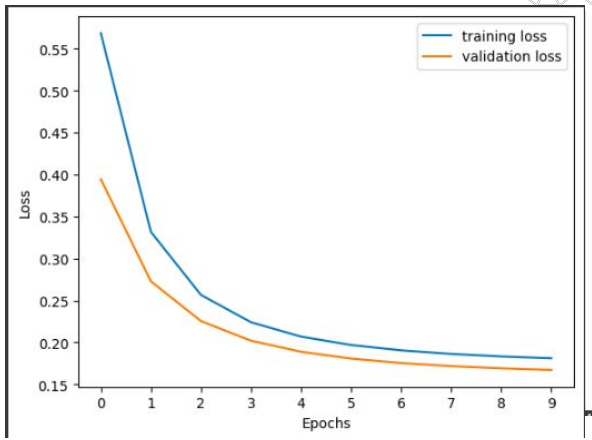


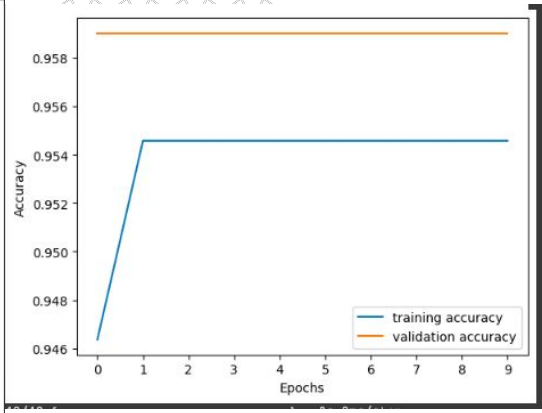


Berkeley
UNIVERSITY OF CALIFORNIA

Logistic Regression

- Baseline model accuracy: 0.955089
- Used one hot encoding for the features
- Used logistic regression model with learning rate of 0.01, sigmoid activation function, and binary cross entropy
- Performed slightly better than baseline for training and test accuracy, but validation accuracy indicates overfitting, and AUC score indicates low precision and recall performance






```
Test accuracy:  0.9553512930870056
              precision    recall  f1-score   support

           0       0.96      1.00      0.98      1455
           1       0.00      0.00      0.00        68

    accuracy                           0.96      1523
   macro avg       0.48      0.50      0.49      1523
weighted avg       0.91      0.96      0.93      1523

AUC score : 0.5
```

Random Forest

- Balanced data using Synthetic Minority Over-Sampling Technique (SMOTE) to oversample minority class (fraudulent job postings)
- Used grid search and random search for hyperparameter tuning
  - Grid search places hyperparameters in a matrix and exhaustively trains every combination of hyperparameter values to select best model
  - Random search randomly samples from hyperparameter grid
- Hyperparameters used:
  - Max depth: Maximum level of each tree
  - Number of estimators: Number of decision trees
  - Max features: Number of features at each split
  - Min samples leaf: Number of samples required at each leaf node
  - Min samples split: Number of samples required to split internal node in each tree

Random Forest

	Random Search	Grid Search
Max Depth	50	None
Max Features	2	1
Min Samples Leaf	1	1
Min Samples Split	4	3
Number of Estimators	300	200

Random Search Results

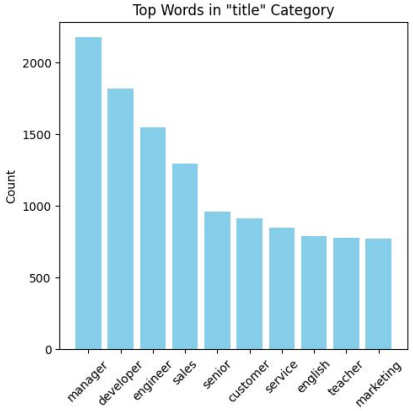
```
Train accuracy:  0.9793725827245 38
Train precision:  0.9768773865082733
Train recall:  0.9822952218430034
Validation accuracy:  0.9634879725085911
Validation precision:  0.9560723514211886
Validation recall:  0.9702797202797203
Test accuracy:  0.9701030927835051
Test precision:  0.9592944369063772
Test recall:  0.981263011797363
              precision    recall  f1-score   support

           0       0.98      0.96      0.97      1469
           1       0.96      0.98      0.97      1441

    accuracy                           0.97      2910
   macro avg       0.97      0.97      0.97      2910
weighted avg       0.97      0.97      0.97      2910

AUC score : 0.97021
```

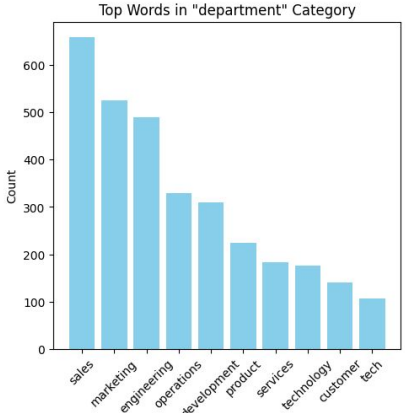
Grid Search Results

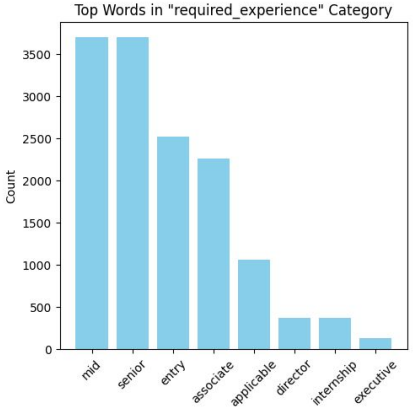
```
Train accuracy:  0.9810915341641598
Train precision:  0.9917175239755885
Train recall:  0.9705631399317406
Validation accuracy:  0.9656357388316151
Validation precision:  0.9707964601769912
Validation recall:  0.958916083916084
Test accuracy:  0.9721649484536082
Test precision:  0.9715672676837726
Test recall:  0.9722414989590562
              precision    recall  f1-score   support

           0       0.97      0.97      0.97      1469
           1       0.97      0.97      0.97      1441

    accuracy                           0.97      2910
   macro avg       0.97      0.97      0.97      2910
weighted avg       0.97      0.97      0.97      2910

AUC score : 0.97217
```

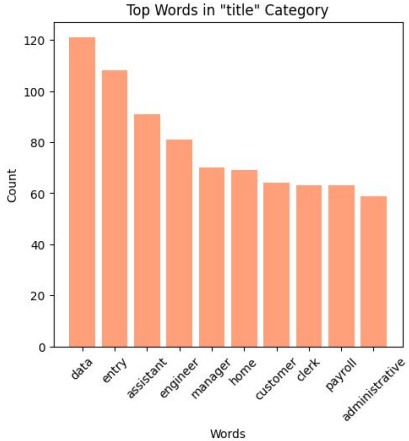
Berkeley
UNIVERSITY OF CALIFORNIA

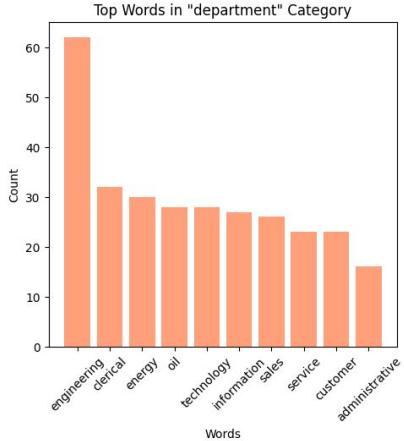
Text Data Analysis

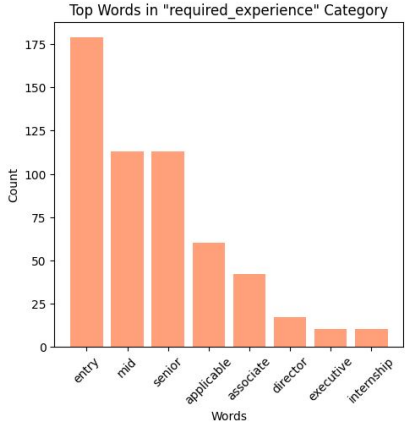




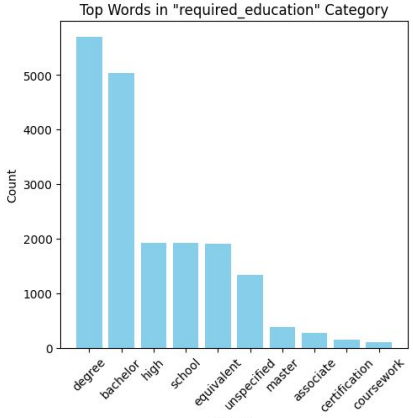


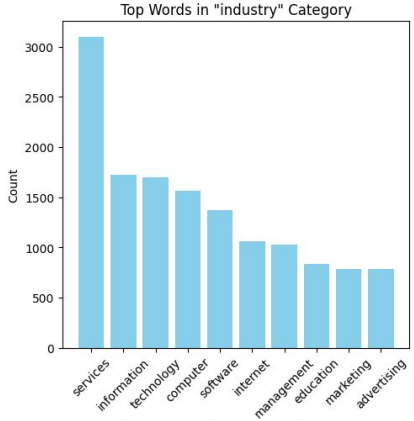


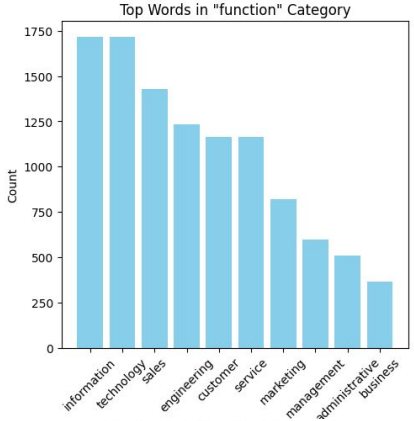


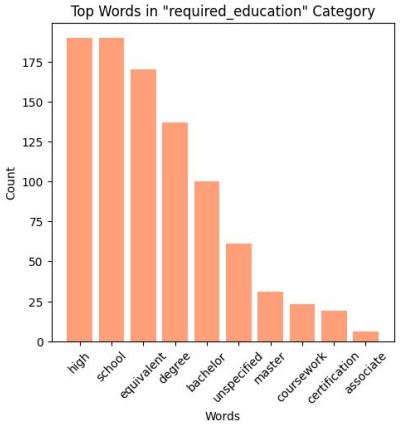


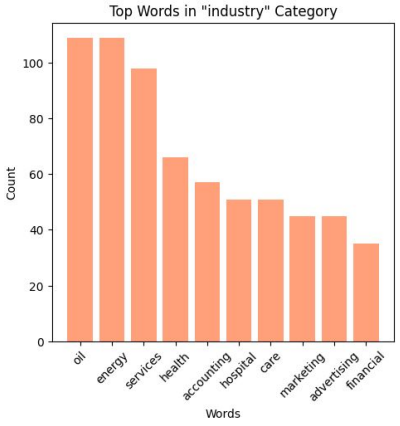
Text Data Analysis

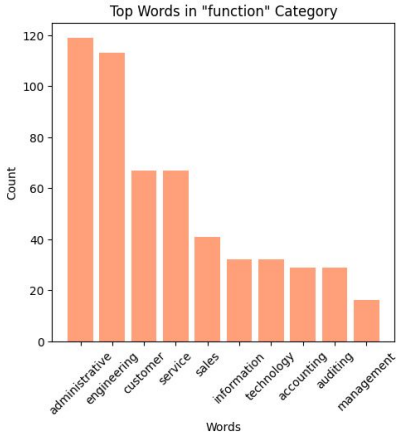








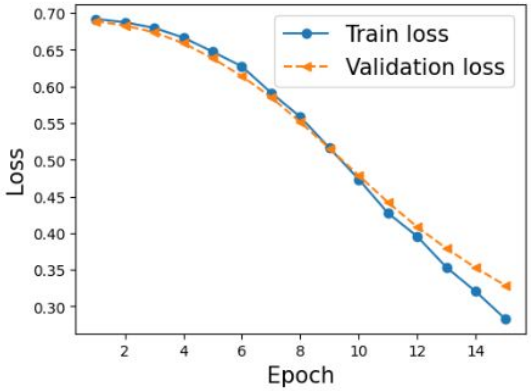


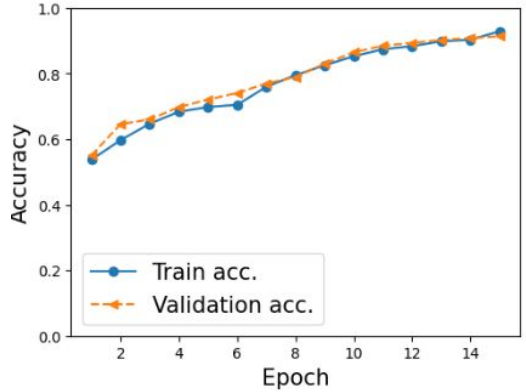


Deep Learning Model

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, None, 20)          378540

 global_average_pooling1d (  (None, 20)                0
 GlobalAveragePooling1D)

 dense (Dense)               (None, 16)                336

 dropout (Dropout)           (None, 16)                0

 dense_1 (Dense)             (None, 1)                 17

=================================================================
Total params: 378893 (1.45 MB)
Trainable params: 378893 (1.45 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

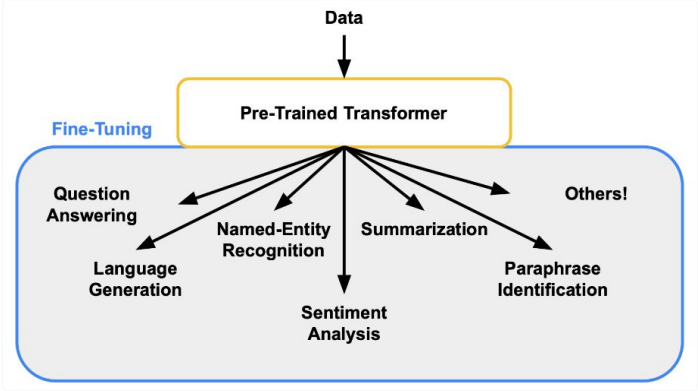
After training this model for 15 epochs, evaluation on the validation data shows an accuracy of: 90.78 %.
Test Acc. 87.57%, AUC. 95.05%





Transformers (BERT)

The Transformer, exemplified by models like BERT (Bidirectional Encoder Representations from Transformers), is a powerful machine learning architecture that excels in natural language processing tasks. It employs attention mechanisms to analyze and capture contextual relationships between words bidirectionally, enabling it to understand and generate human-like language representations.



	precision	recall	f1-score
0	1.00	1.00	1.00
1	1.00	0.95	0.98
accuracy			1.00
macro avg	1.00	0.98	0.99
weighted avg	1.00	1.00	1.00

AUC score : 0.97674

Limitations

- Imbalance of data and limited training data after the cleaning process: with only 800 instances of fake job postings out of 18,000, may lead to models being biased towards the majority class.
    - This imbalance could result in suboptimal performance in detecting fraudulent job postings. The reduction in the volume of training data after data cleaning raises concerns about the model's robustness.
- Generalization of models: Models trained on this dataset may struggle to generalize well to new, unseen data
    - Considering the ever-evolving nature of fraudulent tactics, models trained on historical data may not capture novel strategies used in fake job postings. Regularly updating the model with new data is essential to adapt to emerging trends
- The inherent complexity of language in job descriptions
    - Nuanced language and contextual understanding may prove challenging for models, leading to potential misinterpretation of job descriptions

Limitations

- The use of advanced machine learning models such as BERT transformers and neural networks introduces challenges related to transparency and interpretability
- Incorrectly classifying a legitimate job posting as fake or vice versa may have profound consequences for individuals and organizations
    - It is imperative to continually assess the social implications of the model and actively work towards minimizing biases
- Models, including BERT transformers and neural networks, are susceptible to changes in the data distribution over time.
    - Monitoring for shifts in the job market or language usage patterns is crucial to maintain the model's effectiveness.

Conclusion

- After data cleaning, we discovered that fake job postings tend to include bad details on required education, job descriptions, employment type, telecommuting, etc.
- We employed a number of ML models (base, logistic, random forest) and eventually deep learning model, specifically leveraging BERT on both textual and numeric data, which yielded the highest AUC of 0.97674 and the lowest loss among all the models
- Future efforts should be directed towards addressing class imbalance, obtaining more diverse and representative data, and enhancing the model's adaptability to evolving deceptive tactics.

Contributions

- Eshan
  - Feature selection and EDA for logistic regression and random forest models
  - Evaluated baseline accuracy
  - Implemented logistic regression and random forest models, including hyperparameter tuning and data rebalancing
- Fengshou
  - Text Analysis
  - Deep learning model on text data
  - Application of BERT on text data
  - Application of BERT on text +numeric data
- Jane
  - Data cleaning, description, and analysis
  - Base model and logistic model
  - Introduction, ethical considerations and limitations, and conclusions

References

1. https://medium.com/thecyphy/handling-imbalanced-datasets-with-imblearn-library-df5e58b968f4#:~:text=Imblearn%20library%20is%20specifically%20designed,the%20imbalance%20from%20the%20dataset
2. https://www.kdnuggets.com/2022/10/hyperparameter-tuning-grid-search-random-search-python.html