B565: Data Mining Project

Integrated Job Post Verification and Personalized Job Recommendation System

Team Members

Sumeet Suvarna

Himanshi Raturi

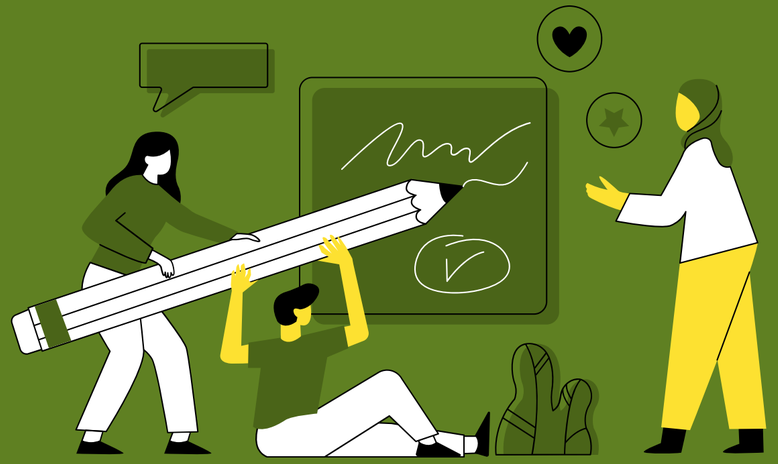
Shubhangi Dabral

Table of Contents

Problem Statement

- Inaccurate Job Listings: Job market flooded with inaccurate job postings, posing a challenge for seekers to distinguish genuine opportunities.
- Untapped Job Data Potential : Rise of online job platforms fuels skepticism among job seekers, eroding trust in the reliability of posted opportunities.
- Impersonal Searches: Job seekers struggle to find personalized matches, resulting in time-consuming and inefficient job searches.

Methodology



Data Collection

Data Preprocessing

Model Creation

Model Evaluation

Result

Step 1

Step 2

Step 3

Step 4

Step 5

Dataset

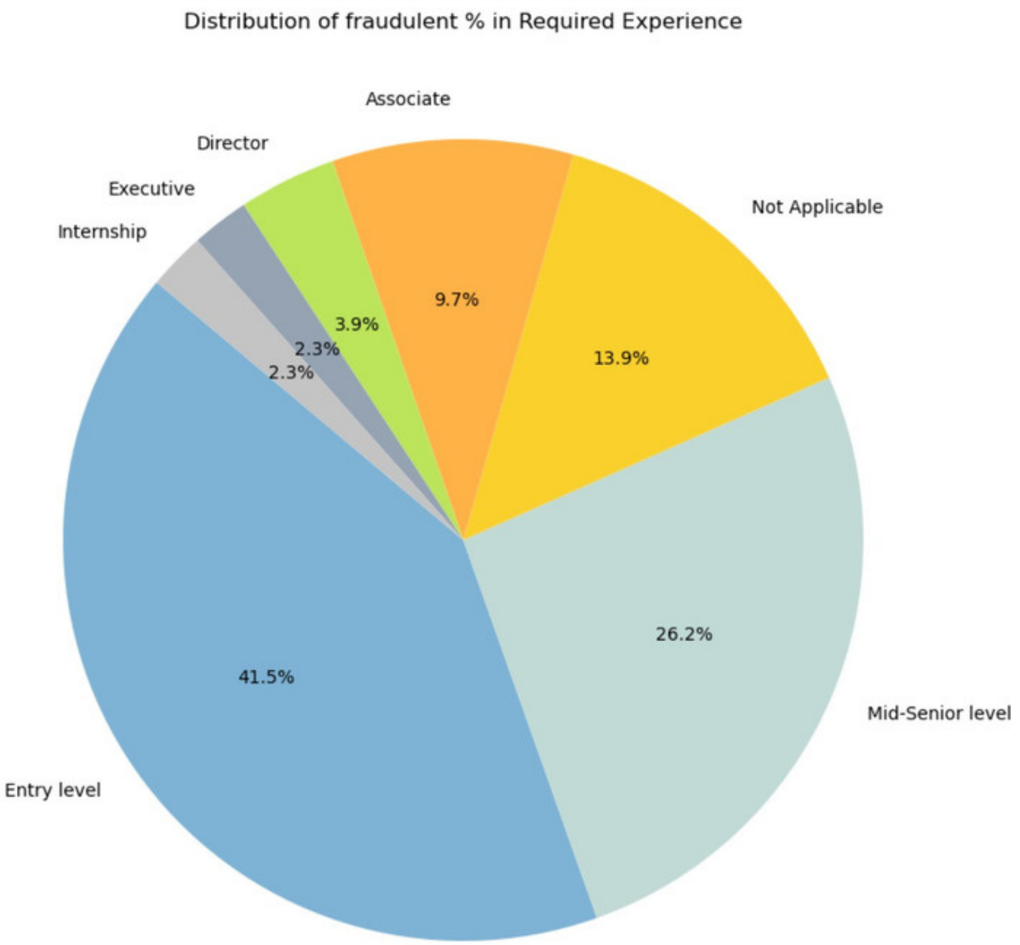
- Dataset collected from Kaggle
- Dataset contains 18K job descriptions and 18 attributes
- It consists of both textual information and meta-information about the jobs
- Sample Data:

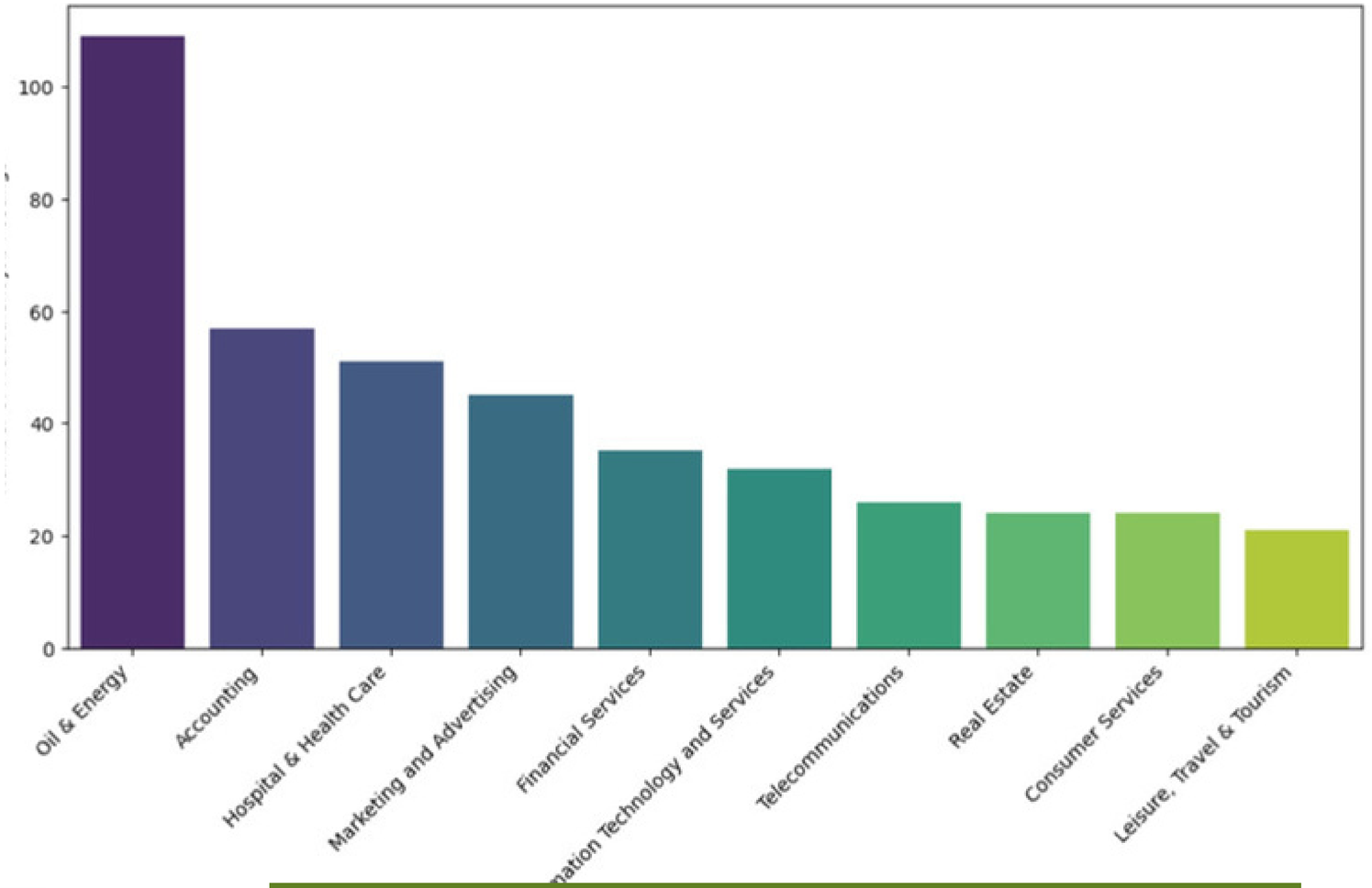
	job_id	title	location	department	salary_range	company_profile	description	requirements	benefits
0	1	Marketing Intern	US, NY, New York	Marketing	NaN	We're Food52, and we've created a groundbreaki...	Food52, a fast-growing, James Beard Award-winn...	Experience with content management systems a m...	NaN
1	2	Customer Service - Cloud Video Production	NZ, , Auckland	Success	NaN	90 Seconds, the worlds Cloud Video Production ...	Organised - Focused - Vibrant - AwesomeIDo you...	What we expect from you:Your key responsibilit...	What you will get from usThrough being part of...
2	3	Commissioning Machinery Assistant	US, IA, Wever	NaN	NaN	Valor Services provides Workforce	Our client, located in Houston, is actively	Implement pre-commissioning and	NaN

Data Preprocessing

- **Data Cleaning and Text Processing**
  - Handled missing values
  - Location and Salary Column Data Refinement
  - Dropped null rows for ‘description’ column
  - Comprehensive Text Representation
  - Standardization for Analysis
  - Optimized Textual Information

EDA





Visualizing the distribution of fraudulent job postings based on 'Required Experience'

Visualize the relationship between different industries and number of fraudulent job postings

Logistic Regression

- Reasons for choosing Logistic Regression:
  a. Good baseline model for binary classification problems
  b. Handle mix of numerical & categorical variables

- Utilized Scikit-Learn (Sklearn) for Logistic Regression

- Under data processing and feature engineering for logistic regression, we dropped irrelevant columns, applied TF-IDF vectorization for text columns and performed one-hot encoding on categorical columns

- Assessed model performance and got the following results:
  a. accuracy: 96.085%
  b. ROC-AUC: 0.954

KNN

- Reasons for choosing KNN Classifier:
  a. Less sensitive to irrelevant or redundant features.
  b. Well-suited for datasets of moderate size, balancing computational efficiency and accuracy.

- For KNN, performed label encoding to certain categorical columns like ‘location’, ‘department’, etc using LabelEncoder and performed one-hot encoding to the entire dataset using pd.get_dummies

- Chose K-Nearest Neighbors (KNN) Classifier with 5 neighbors (**n_neighbors=5**)

- Assessed model performance and got the following results:
  a. accuracy: 96.308%
  b. ROC-AUC: 0.716

Random Forest

- Reason for selecting Random Forest:
  a. Ensemble method to handle a mix of data types and less likely to overfit
  b. Good for capturing non-linear relationships and interactions between features.

- Utlized Random Forest Classifier from Scikit-Learn

- Utilized 100 decision trees (**n_estimators=100**) for ensemble learning.

- Assessed model performance and got the following results:
  a. accuracy: 98.022%
  b. ROC-AUC: 0.834

Neural Network

- Reasons for choosing Neural Network:
  1. DL approaches very effective for text data, capturing complex patterns and relationships.

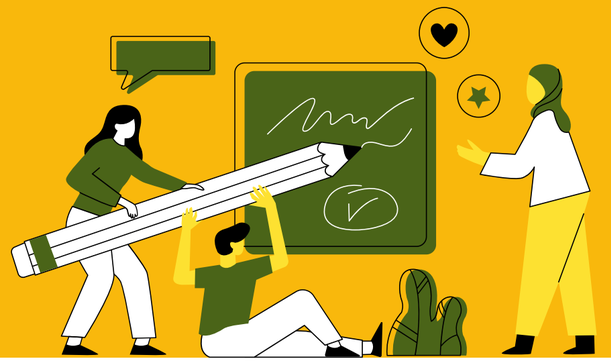
- MLPClassifier, short for Multi-Layer Perceptron Classifier, is an artificial neural network primarily used for classification tasks. It is implemented through Python's scikit-learn package.

- Due to the substantial presence of textual data in our dataset, we opted for the MLPClassifier model to comprehend intricate relationships and patterns inherent in this complex form of data.

- Preceding the model implementation, we conducted text cleaning utilizing NLP techniques and subsequently applied the MLPClassifier.

- Assessed model performance and got the following results:
  1. accuracy: 98.320%
  2. ROC-AUC: 0.879

Recommendation For The Jobs



Search Query Analysis:

- In our search query analysis, we employed Natural Language Processing (NLP) techniques to comprehend the user's intent.
- Following this, we extracted rows from the dataset that matched the user's search query using cosine similarity.
- The final result presented corresponds to the row with the highest cosine similarity to the user's input.
- This approach ensures that the system returns the most relevant result based on the similarity between the search query and the dataset entries.

Input query

```
# Example usage
user_query = "Looking for job in marketing"
result_indices = search_for_jobs(user_query, X, vectorizer, job_data)
```

Output

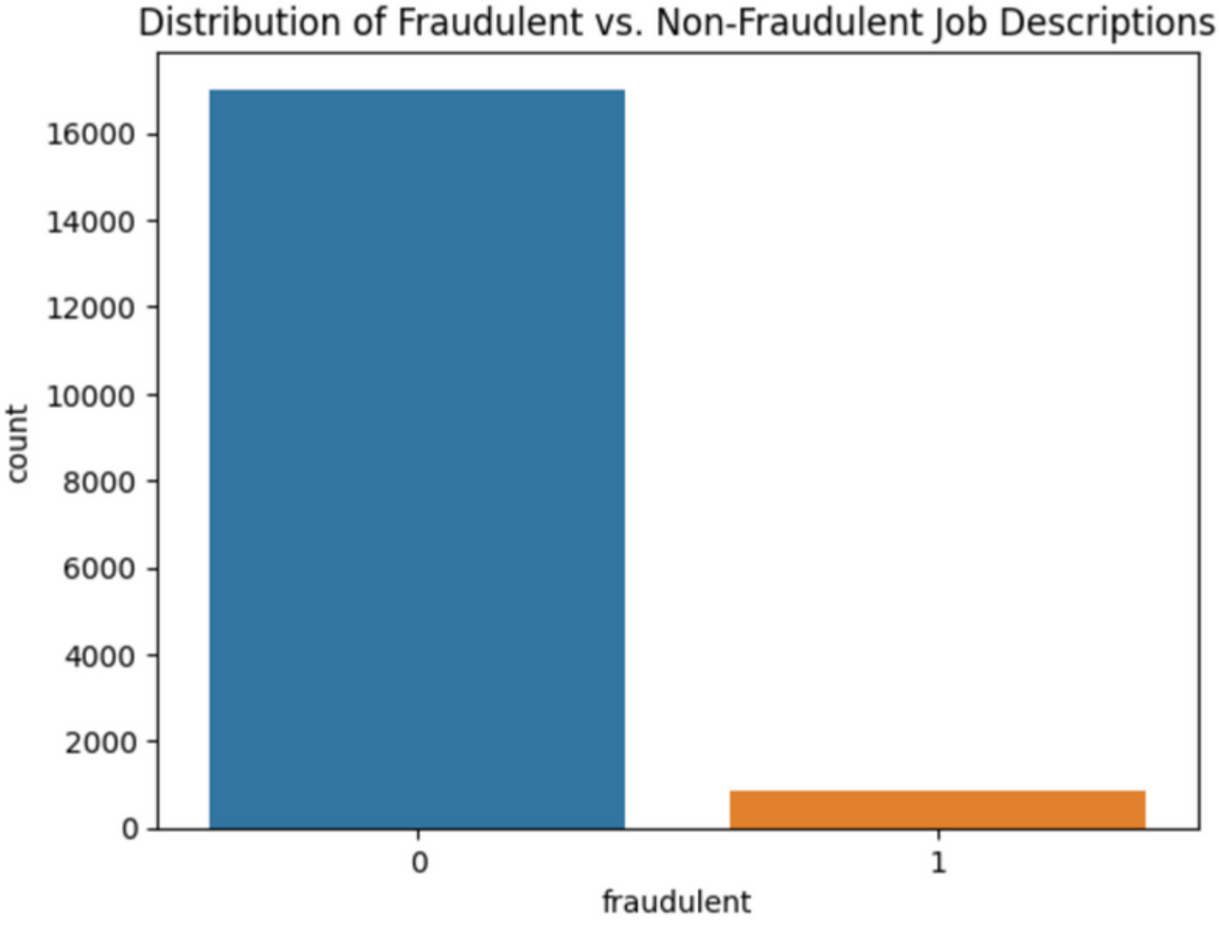
	title	department	description	industry	function
95	Senior Marketing Manager	Marketing	Senior Marketing ManagerOur photography and vi...	Internet	Marketing
12180	Marketing Consultants, Contractors and Freelan...	Marketing	Marketing Consultants, Contractors and Freelan...	Marketing and Advertising	Marketing
3462	Online Marketing Manager	Marketing	CVR Marketing Job Description:Online Marketing...		
15000	Marketing Manager		JOB SUMMARY:The Marketing Manager will be resp...	Information Technology and Services	Marketing
5637	Marketing Coordinator	Marketing	Job Summary:Under general supervision, provide...	Investment Management	Marketing
9011	Marketing Manager		ProServices is looking for a passionate market...	Consumer Goods	Sales
15365	Marketing Manager	Marketing	Role summary:Responsible for developing and ma...	Telecommunications	Marketing
6460	Marketing Assistant		We are looking for a junior/entry level market...		Marketing
14922	Internet Marketing Manager/Internet Marketing ...	Marketing	Positionly Inc. is a company that provides the...	Computer Software	Marketing
8526	Admin / Marketing Assistant		We are looking for a junior/entry level admin ...		Marketing

Model Evaluation

We selected the ROC-AUC Score as our preferred model evaluation metric for the following reasons:

1. **Binary Classification Model:**
    - Our model addresses a binary classification task, specifically distinguishing between fraudulent and non-fraudulent job postings.

2. **Imbalanced Class Distribution:**
    - Given the imbalanced nature of the dataset, with one class substantially outnumbering the other, the ROC-AUC Score is chosen to offer a more nuanced and comprehensive assessment of the model's performance.



Result

MODEL	ACCURACY	ROC-AUC SCORE
LOGISTIC REGRESSION	96.085%	0.954
KNN	96.308%	0.716
RANDOM FOREST	98.022%	0.834
MLPCLASSIFIER	98.320%	0.879

Conclusion

- Recognized the pivotal role of data processing and feature engineering steps in ML models, witnessing significant variations in model accuracy

- Successfully implemented ML models, including Logistic Regression, K-Nearest Neighbors, Random Forest, and MLPClassifier where Logistic Regression achieved outstanding test roc-auc of 88% and accuracy of 98.32% aligning with the original project goal

- Search Query Analysis for job recommendation successfully leverages NLP techniques and cosine similarity to offer personalized and relevant job recommendations

References

- https://www.researchgate.net/publication/319990923_Review_of_Data_Preprocessing_Techniques_in_Data_Mining
- https://www.researchgate.net/publication/225924875_Data_Mining_Methods_for_Recommender_Systems
- https://www.sciencedirect.com/science/article/pii/S2666412721000489
- https://gdeepak.com/thesisme/Applying%20Data%20Mining%20For%20Job%20Recommendations.pdf
- https://link.springer.com/article/10.1007/s00146-022-01469-0
- https://www.researchgate.net/publication/349884280_A_Comparative_Study_on_Fake_Job_Post_Prediction_Using_Different_Data_mining_Techniques

THANK YOU