Understanding Employment Trends in NYC Job Postings

Shimil Shijo [1] Nikhil Patil [2] Srinivasa Reddy Gurram [3]

University of Michigan, Dearborn

Introduction

This project examines various aspects of New York City's job market, including salary differences, competitive job postings, residency requirements, and demographic impacts, using data science techniques. It aims to provide a clear picture of how these factors influence job accessibility for different groups within the city.

The goal is to help employers and job seekers better understand the New York City employment ecosystem. By highlighting key aspects, the study aims to enhance awareness of job market dynamics, enabling more informed decisions for hiring practices and job searches, fostering a more efficient and equitable job market.

Dataset

The **NYC Jobs dataset** comprehensively provides details related to job postings across different agencies in New York City.This dataset is useful for exploring trends in public sector employment, analyzing salaries, identifying skills in demand, and studying job-related data patterns. It contains 5411 rows and 30 fields.

An overview of the NYC Jobs dataset is given below.

cc

Feature	Description
Job ID	Unique job posting.
Agency	Government offering job.
Posting Type	Internal or external.
Business Title	The title of the job.
Civil Service Title	Yes or No.
Salary Range (From)	Minimum salary.
Salary Range (To)	Maximum salary.
Salary Frequency	Annual and Hourly.
Work Location	Address or borough).
Division/Work Unit	Organizational unit or division.
Job Category	IT or Healthcare.
Full-Time/Part-Time	Full-Time or Part-Time.
Career Level	Entry-Level, Manager or Experienced
Civil Service Exams	Yes or No
Posting Date	Job posted date.
Application Deadline	Last day to apply.
Job Description	Detailed description of job.
Preferred Skills	Specific skills.
Minimum Qualifications	Minimum qualifications required for the position.

Table 1. NYC Jobs Dataset Description

Research Topics

The primary objective of this project is to explore and provide insights into the following research questions:

1.Do different populations within similar job titles have observable salary differences?

2.What are the main wage distinctions between competitive job postings and those that are not?

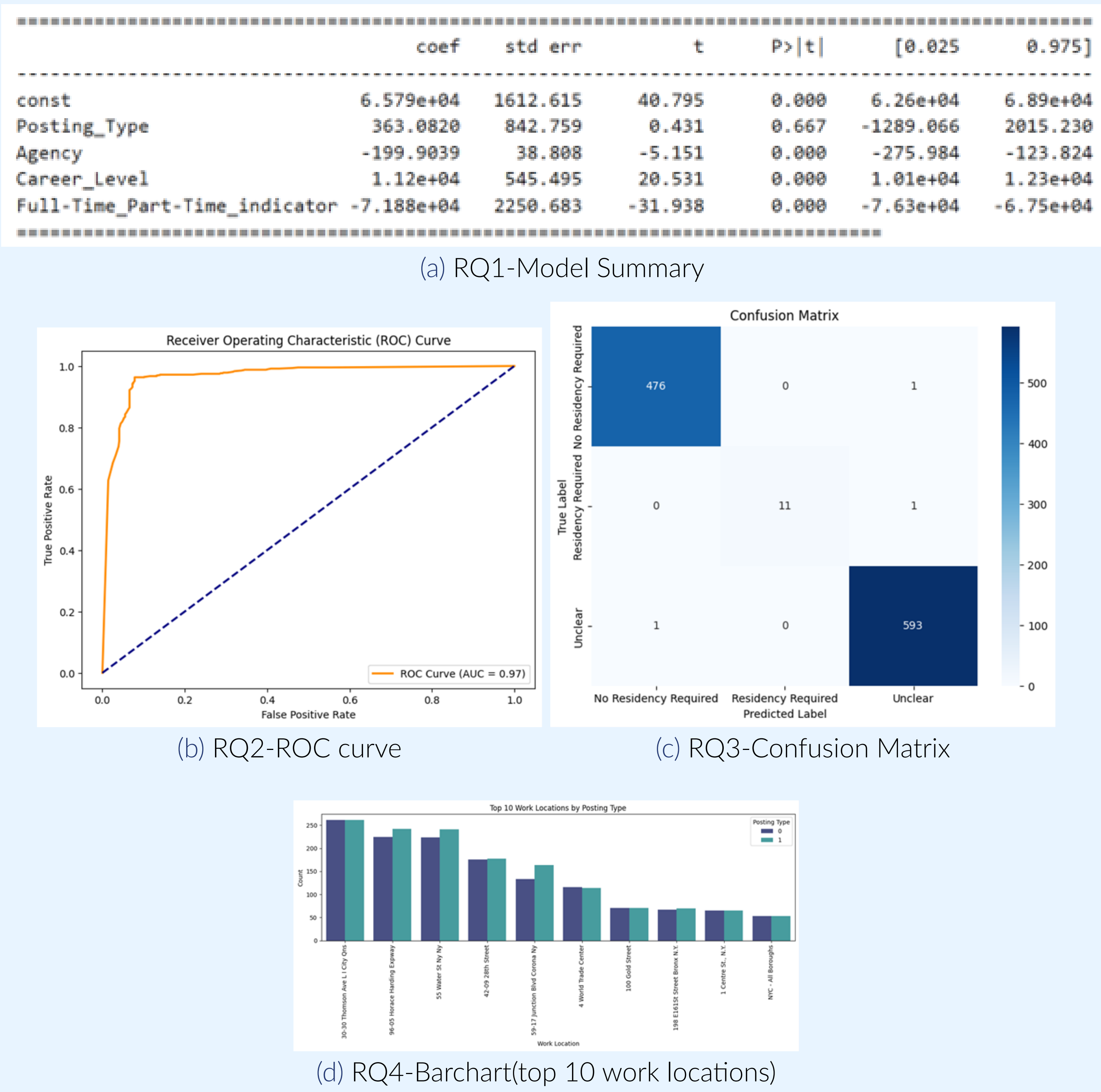
3.Do residency requirements make some organizations more likely to have open positions?

4.Does the demography that is drawn to external postings differ from that of internal postings?

Methodology

1. Data Pre-processing - Performed preprocessing tasks such as handling missing value, removing unnecessary fields,renaming fields, cleaning inconsistencies, scaling data and encoding categorical variables for each research question and ensuring appropriate data types to prepare the dataset for effective analysis.
2. Feature Engineering - transformed preprocessed data into meaningful features by selecting, modifying, and creating new features to improve the performance of machine learning models.
3. Model Building
   - RQ1- Used linear regression to identify variations in salary for similar job roles and apllied K-Means clustering to group the populations by job and demographics.
   - RQ2- Random Forest Classification to categorize postings, with Comparative Analysis to highlight wage differ- ences and used descriptive statistics and visualizations to compare average salaries between these two roles.
   - RQ3- Used TF-IDF for text vectorization and build a Logistic Regression pipeline with L2 regularization.Hyperparameter Tuning is performed to optimize Logistic Regression using GridSearchCV. Also, 5-fold cross-validation has been done for robust accuracy.
   - RQ4- Focused on Exploratory Data Analysis(EDA) to examine distributions and relationships between variables like Posting Type, Average Salary, and Work Location. Data grouping has been done to explore relationships by categories like Posting Type and Career Level.
4. Model Evaluation
   - RQ1 – OLS Regression Model Summary is used for model evaluation. Also a linear regression equation is formed with depandet and independent variables.
   - RQ2 – The evaluation metrices used are Classification Report, Confusion Matrix, ROC Curve, Precision-Recall Curve, Descriptive Salary Statistics and Histogram of Predicted Probabilities.
   - RQ3 – Evaluated with classification report and Confusion Matrix.

Results



```
==========================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                        6.579e+04   1612.615     40.795      0.000    6.26e+04    6.89e+04
Posting_Type                  363.0820    842.759      0.431      0.667   -1289.066    2015.230
Agency                       -199.9039     38.808     -5.151      0.000    -275.984    -123.824
Career_Level                  1.12e+04    545.495     20.531      0.000    1.01e+04    1.23e+04
Full-Time_Part-Time_indicator -7.188e+04  2250.683    -31.938      0.000   -7.63e+04   -6.75e+04
==============================================================================
```

(a) RQ1-Model Summary

(b) RQ2-ROC curve

(c) RQ3-Confusion Matrix

(d) RQ4-Barchart(top 10 work locations)

Figure 1. 2x2 Grid of Images

Machine Learning Models

Machine learning models used to train the data:

1. Linear Regression

Agency: The negative coefficient for Agency (-199.90) indicates significant salary differences across agencies, even within similar job titles. Career Level: A positive coefficient for Career Level (+11,199.68) highlights that higher organizational levels are associated with significantly higher salaries in similar roles. Full-Time vs. Part-Time: The negative coefficient for the Full-Time/Part-Time indicator (-71,883.26) demonstrates that part-time employees earn substantially less than their full-time counterparts in equivalent job titles. KMeans Clustering

2. KMeans Clustering

For the Elbow method we have taken k in the range of 1-15. While initializing we took n init=20 and random state=49. For Silhouette Analysis we have taken k in the range of 2-15. While initializing we took n init=20 and random state=49

3. Random Forest Classification

We took n estimators which are called Number of trees in the forest typical ranging between 10–2000. The maximum depth of the trees are typical ranged between None (unlimited) or specific values like 10–100. The minimum number of samples required to split an internal node. Their range falls between 2–20. The minimum number of samples required to be at a leaf node and their range fals between 1–10.

4. Logistic Regression

A type of regression analysis used for predicting binary outcomes. It uses a logistic function to model the probability of an event occurring.

Conclusion & Limitation

The comprehensive review of the NYC Jobs dataset shows complex relationships in the job environment of the city. The study revealed notable salary differences among agencies, job categories, and career levels, underscoring ongoing pay inequalities. Despite competitive job advertisements show more uniform pay, noncompetitive positions show more variation, which may be a reflection of different function complexity and experience levels. Strong similarities between internal and external employment possibilities in terms of career level distributions and work location patterns were also shown by the study, which also shed light on subtle trends in job listings.

The results emphasize the significance of understanding complex labor market systems and the necessity of ongoing research into fair employment practices. By using rigorous analytical methods including regression, classification, and comparison analysis, the study provides insightful information about the complex variables influencing the labor market in New York City. It indicates possible areas for resolving workforce inequalities and offers a data-driven perspective on job trends.

Certain research questions such as clustering could not be adequately addressed due to inherent constraints and limitations in the structure, scope, and characteristics of the dataset. But it is addressed by linear regression.

References

1. Eric Kober. "New York City Jobs Data Shine Light on NYC's Economic Strengths and Weaknesses". In: Manhattan Institute (2024).

2. Rahul Jain Thomas P. DiNapoli. "New York City's Uneven Recovery: An Analysis of Labor Force Trends". In: New York State Office of the State Comptroller (2023).

3. Jennifer Gravel Carl Weisbrod Purnima Kapur. "Employment Patterns in New York City". In: New York City Department of City Planning (2016).