Data Analyst Job Postings

Adapa Aditya

Aditya.Adapa@coyotes.usd.edu

CSC 457/557 (Data Analysis, Decision Making, and Visualization) Project Report – Fall 2023

ABSTRACT:

The project aims to visually extract essential insights from the Data Analyst Job Postings Dataset. These insights include an understanding of job market trends, skills and qualifications required, salary dynamics, demographic analysis, comparisons among distinct job role names within Data Analysis, exploration of job posting sources, and the implementation of predictive modeling through a machine learning model.

Introduction:

The project initiates a detailed examination and visualization of a dataset that encompasses job postings for Data Analyst positions, sourced from Kaggle. This dataset is a valuable repository, presenting an extensive array of information concerning job requirements, skills, demographic locations, and various other factors pertinent to positions within the field of Data Analysis. The overarching objective of this project is to extract actionable insights from the dataset, with the intention of providing benefits to job seekers, hiring managers, and recruiters navigating the dynamic landscape of the Data Analyst job market.

Methods:

The Project methods include Data Cleaning to identify and handle issues in the dataset, such as missing values, duplicates, and outliers, ensuring that the data is suitable for analysis. Exploratory Data Analysis is used to summarize, visualize, and understand the main

characteristics of the dataset. Linear Regression for predicting a continuous dependent variable based on one or more independent variables. polynomial Regression for extending linear regression to capture more complex relationships between variables., and Random Forest Regression for building an ensemble model for predictive tasks.

Result:

In this project, the Data Analyst Job Postings dataset served as the foundation for a comprehensive examination of factors influencing the salaries of data analysts. The variables scrutinized encompassed company rating, location (city), benefits availability (such as vacation or insurance), and required skills.

Upon analysis, it was observed that the R-squared value of the simple linear model was notably low, suggesting that the correlation between the variables does not exhibit a linear relationship. Consequently, a decision was made to explore a polynomial regression model as an alternative approach.

Despite the implementation of a multinomial regression, which demonstrated some improvement, the findings indicated a lack of substantial correlation between the number of languages mastered and salary levels. The model revealed a weak negative correlation, indicating that companies may prioritize candidates fluent in specific languages over those proficient in multiple computer languages in general.

Further investigations involving both linear regressions and random forests revealed poor performance, suggesting a lack of a robust correlation between the number of benefits and wages. These outcomes collectively contribute to a nuanced understanding of the factors influencing data analyst salaries within the context of the dataset.

Conclusion:

In our study, we examined various factors, including company rating, location (city), benefits such as vacation or insurance availability, and required skills, with a primary focus on understanding whether the number of coding skills mastered significantly predicts annual salary. The underlying hypothesis posited that a higher proficiency in coding skills correlates with increased salary, based on the belief that a broader skill set enhances a candidate's attractiveness to employers and strengthens their negotiating position.

Our analyses encompassed the exploration of correlations between individual factors and salaries, leading to the development of predictive models. Notably, the size of a company did not exhibit a significant correlation with wages. Conversely, location demonstrated a statistically significant impact on average wages, albeit with relatively modest explanatory power. Among programming languages, Python exhibited a positive correlation with salary, while Excel and C showed negative correlations. Consequently, we reject the null hypothesis that the number of coding skills mastered, the location of work, and the employing company insignificantly impact annual salary. Instead, our data supports the alternative hypothesis, suggesting that specific coding skills, location, and certain employee benefits do influence the salary of a data analyst.

However, it is crucial to highlight that the number of coding skills mastered did not strongly correlate with salary levels, indicating that the types of skills may outweigh their quantity in influencing salary outcomes. Despite these findings, our project faced limitations. Firstly, our dataset may not comprehensively capture all factors influencing salaries. Secondly, the linear regression models employed might not effectively capture complex,

potentially non-linear relationships between variables and salaries.

Despite these limitations, our work contributes to a deeper understanding of the factors influencing data analysts' salaries. This knowledge can inform both individuals and organizations, guiding them on what to prioritize in career development and recruitment processes.

References:

1]https://github.com/pranavsai-98/Data-Analyst-Job-EDA/tree/master

2]https://github.com/MohamedMuneerM/naukri-data-analyst-job-posting-analysis#naukri-data-analyst-job-posting-analysis3]

3] https://github.com/pranavsai-98/Data-Analyst-Job-EDA/blob/master/data-analyst-jobs-eda%20(1).ipynb