Is Data Science still the sexist job of the 21st Century?

Michael Zelaya & Dr. Robert Kelley

Bellarmine University Data Science Program

Abstract

This semester-long project focuses on analyzing, cleaning, and predicting various machine learning models like linear regression, K-nearest neighbor, Random Forest, and Gradient Boosting for salaries in the Data Industry. The accuracy score of the preliminary results for both linear regression and K-nearest neighbor had an accuracy score of around 33%. However, after splitting the data again in half for my validation data, the accuracy score for the K-nearest neighbor model improved by about 53%. Although the accuracy score is not what I hoped for, I believe it can improve significantly.

Introduction

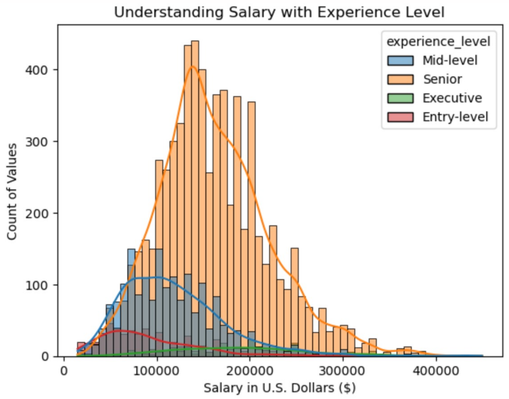
My long-term semester project focuses on salary trends in data-related careers. I chose this dataset from Kaggle because I'm interested in further exploring and accurately predicting salaries in the data job industry. The cleaned dataset has 9,356 rows with eight columns, mostly categorical data, while two columns have numerical data. The variables in the dataset include the work year, job category, salary (USD), experience level, employment type, work setting, company size, and country category (U.S. or not). In this poster, you can find more detailed information regarding and understanding the dataset from the various visualizations and graphs.

Objectives

The focus of my semester-long project was to determine a machine learning (ML) model that can most accurately predict a job salary based on previous data collected in the data field. At first, as a baseline for evaluating predictive models, I used linear regression to compare the accuracy scores with the other three predictive models I will also use. The other three predictive models I implemented in my project were the K-nearest neighbor, Random Forest, and Gradient-Boosting regression algorithms. Ultimately, the main objective of my entire project was to determine which algorithm had the best accuracy score to predict salaries in the data industry.

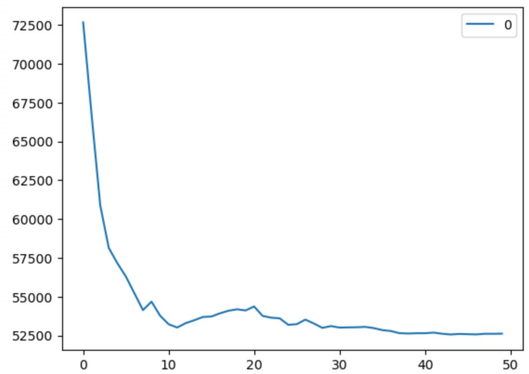
Materials and Methods

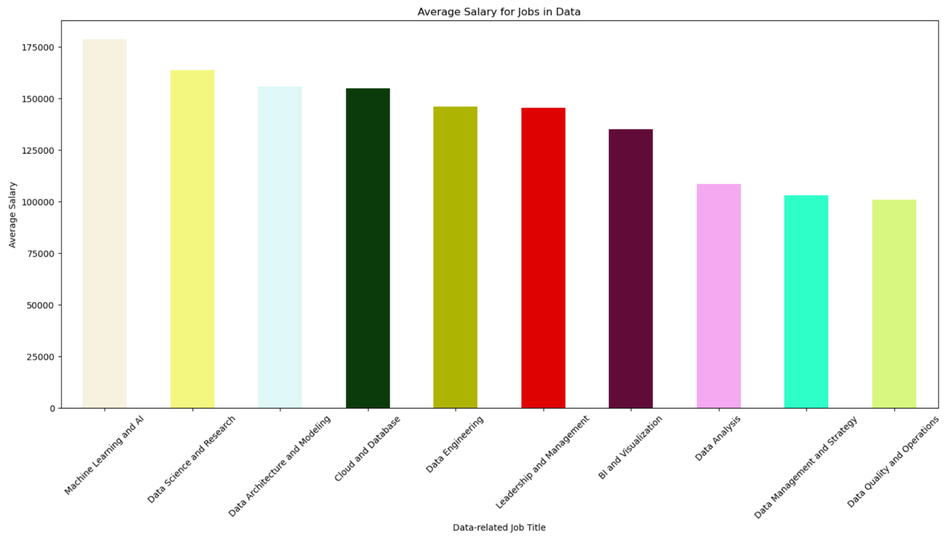
I completed this project using the Python programming language in Jupyter Notebook from Anaconda-Navigator. As for libraries in Python, I implemented numpy, pandas, matplotlib, seaborn, and sklearn within my project. I chose these libraries because these packages would help me clean the data, perform explanatory data analysis (EDA), and ultimately perform machine learning predictive modeling analysis. Lastly, as for visualization tools, I also used Tableau in my tool pack to help me best with the EDA process. First, I created a lot of graphs and visualizations to help better understand the dataset. After dropping repeating or unnecessary columns, I converted the categorical columns to numerical variables using one-hot coding to allow the ML models to run. Then, I split the data into training and test sets to predict the best accuracy score.

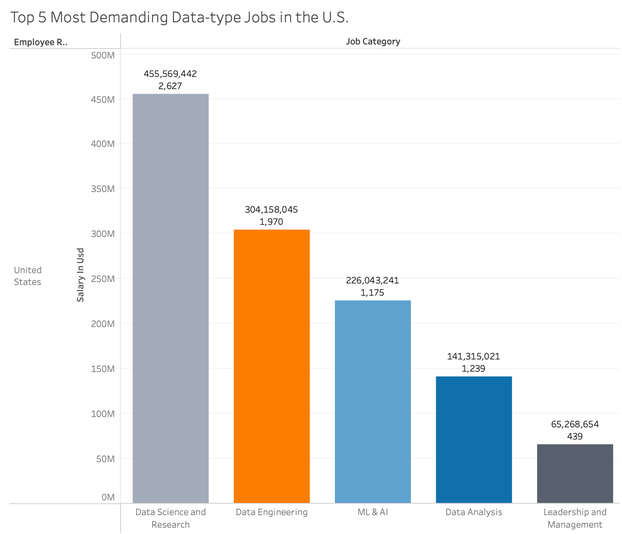


Results

As for results, the preliminary results that I obtained so far in my model for both my linear regression and K-nearest neighbor had an accuracy score of around 33%. However, while splitting the data again in half for my validation data, the accuracy score for the K-nearest neighbor model improved by about 53%. A few problems I encountered in my project was handling the categorical data. Most categorical data had fewer unique variables except salary (USD) and company location. While different job salaries can make sense and vary from person to person, I decided to leave it as it was. As for company location, since most of the data were from the United States, but there were few countries worldwide, I thought it was best to separate it as either the United States or non-United States to make the dummy variables as simple as possible. The results humbled me and made me realize I had to try and experiment in new ways.







Conclusion/Future Work

In summary, my long-term semester project focused on analyzing, cleaning, and predicting various machine learning models like linear regression, K-nearest neighbor, Random Forest, and Gradient Boosting for salaries in the data field. Although the accuracy score is not what I hoped for, I believe it can improve significantly. I would want to work and experiment more by preparing and cleaning the data differently and experimenting with different test sizes.

References

- https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data/data
- https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century
- https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python
- https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/
- https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/

Contact Information

Email:
Mzelaya@bellarmine.edu

rkelley@bellarmine.edu