Is Data Science Still the Sexiest Job of the 21st Century?

Michael Zelaya & Dr. Robert Kelley

Bellarmine University Data Science Program

Abstract

The project aimed to develop a model that predicted salaries in the data industry based on variables like experience level, employment type, work setting, company size, and country category (U.S. or not), and data was cleaned and analyzed using Python and Tableau to achieve this. The accuracy score of the preliminary results for both linear regression and K-nearest neighbor was around 33%, with k = 44 out of 50 having the best RMSE. The validation data for the K-nearest neighbor improved the accuracy to 53%—the project goal was to provide valuable insights into the data field job market.

Introduction

His long-term semester project focused on salary trends in data-related careers. He chose the dataset from Kaggle because he was interested in further exploring and accurately predicting salaries in the data job industry. The cleaned dataset had 9,356 rows with eight columns, mostly categorical data, while two columns had numerical data. The variables in the dataset included the work year, job category, salary (USD), experience level, employment type, work setting, company size, and country category (U.S. or not). The poster provides more detailed information regarding and understanding the dataset through various visualizations and graphs.

Objectives

- He focused on determining a machine learning (ML) model that could most accurately predict a job salary based on previous data collected in the data field.
- As a baseline for evaluating predictive models, he used linear regression to compare the accuracy scores of the K-nearest neighbor, Random Forest, and Gradient-Boosting regression algorithms.
- Ultimately, the main objective of his entire project was to determine which algorithm had the best accuracy score to predict salaries in the data industry.

Materials and Methods

He completed the project using the Python programming language in Jupyter Notebook from Anaconda-Navigator. As for libraries in Python, he implemented numpy, pandas, matplotlib, seaborn (Fig. 1, 2, and 4), and sklearn within his project. He chose these libraries because these packages helped him clean the data, perform explanatory data analysis, and ultimately perform machine learning predictive modeling analysis. As for graphs, he also used Tableau to help him best with the EDA process (Fig. 3). First, he created a lot of charts and visualizations to help him better understand the dataset. After dropping repeating or unnecessary columns, he converted the categorical columns to numerical variables using one-hot coding to allow the ML models to run. Then, he split the data into training and test sets to predict the best accuracy score.

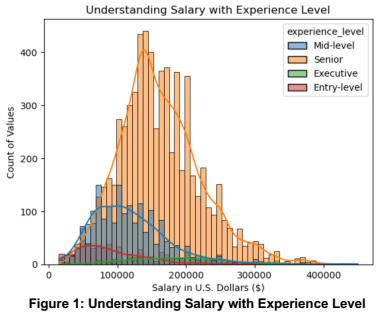

Figure 1: Understanding Salary with Experience Level

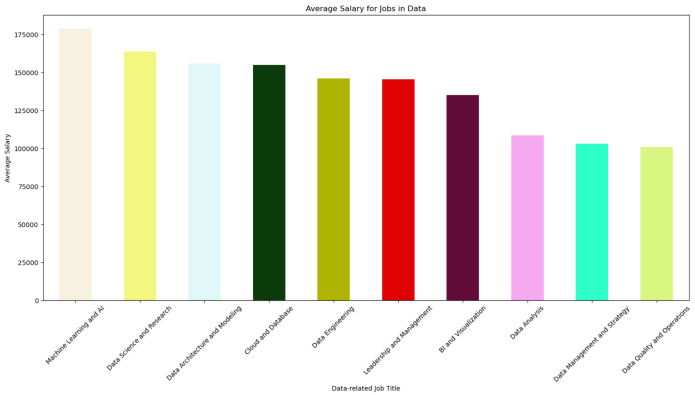

Figure 2: Understanding Average Salary in the Data Industry

Results

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

Figure 4: Reduction in RMSE as K increased

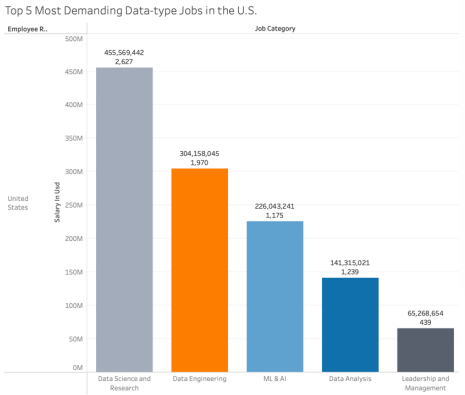

Figure 3: Top 5 Most Demanding Data Roles in the USA

Conclusion/Future Work

In summary, his 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 was not what he hoped for, he believed it could improve significantly. He wanted 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