2023.05.09

1) PROBLEM OVERVIEW & MOTIVATION

I. What problem are you solving?

I am trying to solve how to predict salaries for data science jobs worldwide. In general, I wanted to gain a deeper understanding of the indicators of varying levels of salaries. There is no standardized expectation for data science salaries worldwide, but for example, based on an individual's experience level, geographical location of residence, employment type, company location, and accepted remoteness of their work to name a few, I strongly believe I could develop a machine learning model to help individuals get a stance on the quantification of their compensation about their background. This would help them get a better feel of how to improve it by assessing the various coefficients and indicators.

II. Why?

As I am an undergraduate senior and an aspiring data scientist too, I came into this project with a shared desire to learn how to excel financially in the field of data science. As the field of Data Science is rapidly developing and changing, alongside an already turbulent economy, entry-level workers entering data science roles and workers pushing through their career journeys need as much support and insight as possible to meet their financial, career, and other important goals. I wanted to know specific indicators I could focus on to help improve my salaries and careers, which my dataset and model provided me the freedom to explore.

2) THE DATA

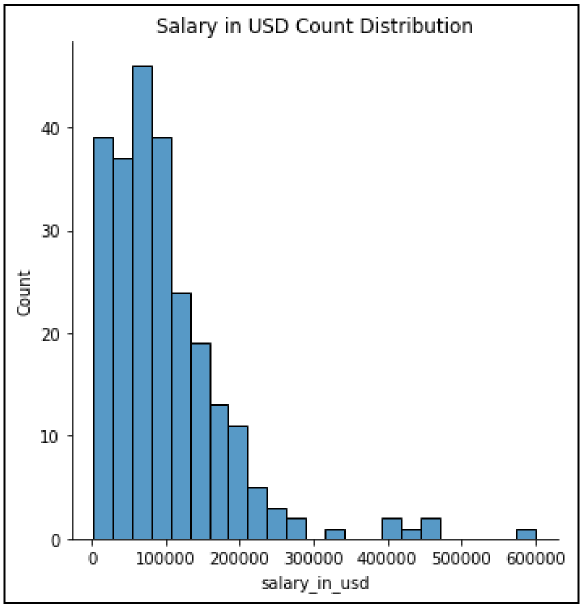
I. What is the nature of the data used?

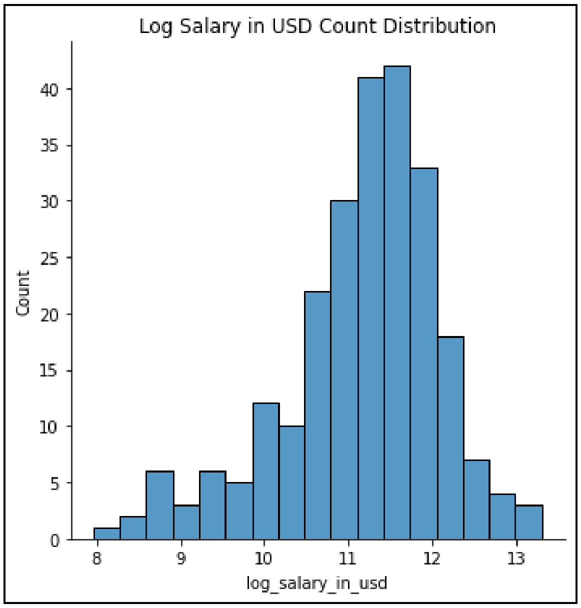
The dataset I used provided a sleuth of categorical indicators that all provided unique insight into how each is uniquely correlated to a data scientist's compensation. The data is all on recent times of 2020 and 2021, with varying levels of experience ranging from entry-level to executive level, all types of employment to suit an individual's personal needs, specified job title, individual's current country of residence, remote ratio of their job, current company's country to residence, and how large their company is.

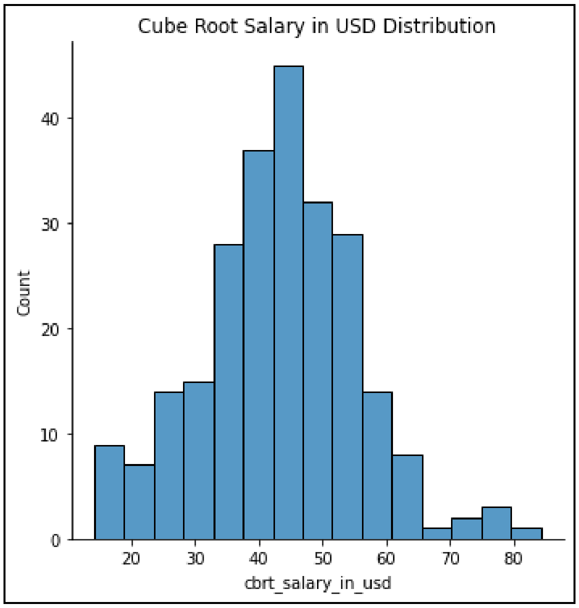
II. How did you collect and process the data?

I **acquired** the data for my analysis on data science job compensation from Kaggle, where I found a pre-existing dataset that closely aligned with my research topic. I prioritized finding a more recent dataset since the area of this data is constantly and rapidly evolving. The dataset contained the aforementioned regression variables that I believed would be pertinent to my analysis.
I **processed** the data by using .ipynb Colaboratory notebooks to write code that cleaned the data. The tools I used to process my data are Python, Pandas, and Numpy. I created data frames and used various functions in Pandas to organize my code. I deleted/dropped unnecessary columns. Furthermore, I one-hot encoded my categorical variables and removed columns after using "dummy variables." I plotted my dependent variable, "salary_in_usd" (left-side plot), which shows a right-skewed distribution. I then processed this variable by taking the log("salary_in_usd") which resulted in a

relatively better distribution (right-side plot). However, I gained a much better distribution after taking the cube root of “salary_in_usd” (bottom plot) which is the best distribution of the three.


Salary in USD Count Distribution


Log Salary in USD Count Distribution


Cube Root Salary in USD Distribution

3) ANALYTIC TECHNIQUES & MODELS

I. What methods did you use?

My model mainly utilizes five libraries and/or corresponding methods.

A. **Pandas** - Pandas was the bread and butter of my notebook for manipulation and analysis of the dataset. I used it to read my dataset as a csv, add and drop features, create dummy variables, and more.

B. **Sklearn** - I implemented quintessential machine learning tasks such as splitting the dataset into training and testing sets using sci-kit-learn's train_test_split(). I am aiming for a 70% training: 30% testing split and accomplished exactly this.

C. **Statsmodel API** - I utilized the OLS function of the Statsmodel API to create a multivariable linear regression model. Because I'm working with such a large dataset with many different features (30+), this helped me analyze everything, ranging from the coefficients to the p-values, to the R^2 values. I later used my OLS models to predict the test sets as well as salaries for my randomized data scientists.

D. **Out of Sample R^2 (OSR2)** - I implemented the OSR2 function to test whether the model has any sort of out-of-sample predictability.

E. **Other minor libraries and functions (Numpy, Matplotlib, Seaborn)** - I also implemented a few other libraries and methods for minor tasks like data cleaning and generating visualization plots.

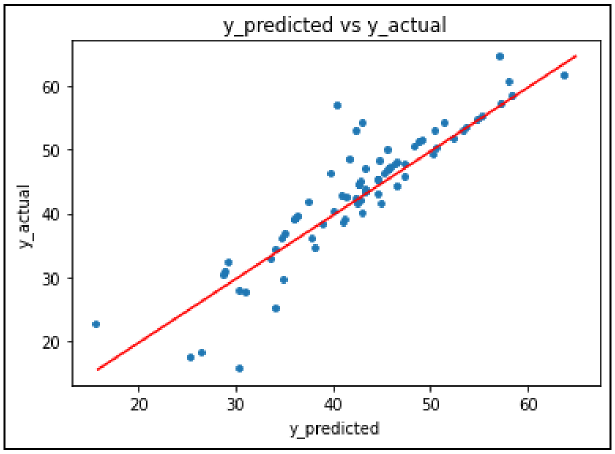
II. What are the results?

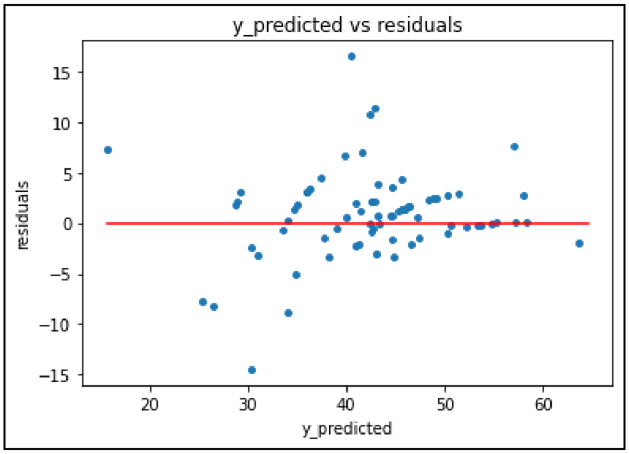
At first, I didn't take the cube root of my prediction variable, which resulted in very poor results. The OSR2 after feature selection was -1.3847 (and the model.summary() output had a R-squared of 0.816 and Adj. R-squared of 0.792). This initial model also predicted negative salaries in my proof-of-concept randomized DataFrame dataset of hypothetical jobs, which didn't make sense in the problem's context. Therefore, I went back to the drawing board to reevaluate my dataset processing. After visualizing/plotting the log(salary_in_usd) distribution, I created a model that predicts these values perform significantly better than my initial model. The updated model's OSR2 after feature selection was roughly 0.5810 (R-squared of 0.920 and Adj. R-squared of 0.903). Furthermore, my log model didn't predict negative log(salary_in_usd) values for the proof-of-concept dataset.

FINAL & BEST MODEL: For my cube root model, after feature selection, it had the best OSR2 value of 0.8035 (R-squared of 0.966 and Adj. R-squared of 0.959) and doesn't predict negative values for my proof-of-concept dataset as Ill. Overall, the results of the cube root model can be seen below in the next section's plots.

III. How confident are you in your results?

Originally, I was not confident at all in my results. Due to predicting negative salaries, a low negative OSR2 value (< -1), and a lolr R-squared (~0.8), I felt like there was room for improvement and there may not be enough applicability for my model in a real-world context. Holver, after performing the log transformation, later going with the cube root transformation, and analyzing the results explained in part II (higher R-squared, OSR2, and all positive salaries), I felt a lot more confident in them and I'm excited to see that my model may have potential when applied to real-world data, as indicated in my predictions on my testing data with lolr residuals this time. The improvement seen from the progression of my model implementations provides a greater sense of confidence in my final (cube root) model.





IV. How might you extend your analysis in the future?

In an ideal world where I have a much more expansive and stratified dataset, it could've been possible to develop separate regression models for different subpopulations or to include interaction terms in the model to capture the varying effects of predictors across subpopulations. For example, if my dataset had more specificity in the category of experience level such as listing years of experience as opposed to broad classifications for entry-level through to executive level, my model's predictive capabilities would be much more granular. Additionally, it may have been useful to perform subgroup analyses to assess the performance of the model across different subpopulations of interest subsequently. I also hope to use my model to predict salaries specifically for UC Berkeley students (new grads) such as myself to give others an analytical tool to help them better understand what to expect in their job search.

4) OVERALL IMPACT

I. What is the (potential) impact of your work regarding the problem that you are trying to solve?

my work could provide valuable insights and recommendations to new graduates who are seeking work on what factors are most important to consider when they are choosing the type of data science job they want to pursue or even help them negotiate their salary with the company they are working for. Especially for people in underrepresented backgrounds, being more knowledgeable about industry salaries could help them advocate and fight for higher salaries versus not getting paid as much as their industry peers. Employers may also find this report useful by helping them better understand the factors that drive salary expectations. This would help employers attract high-quality employees which in turn could lead to greater productivity, profitability, and innovation.

II. How might you expand the scope of your analysis to improve its impact even more?

One possible way I could improve the impact of my analysis is by incorporating more data sources from other public data sets that include information on the labor market, salaries listed on job postings, and the demand for different types of data science jobs. I could also collect my data through surveys from both employers and employees to get a better understanding of their experiences with data science jobs and possibly include more qualitative data into my model like work-life balance and job satisfaction to see if these new variables have an impact on predicted salaries. Furthermore, I could incorporate data on other employee benefits and compensation methods that are also important, such as bonuses, stock options, RSUs, health insurance, paid leave, housing support, and more.

III. Does the impact of my model vary across different subpopulations of interest? (Note that an analysis of the potential impact of the model might also include possible negative consequences.)

Yes, the impact of the regression model that predicts data science job salaries could vary across different subpopulations of interest. For example, different subpopulations may have different levels of education, work experience, job titles, locations, and other factors that could impact their salaries. Therefore, a regression model that works Ill for one subpopulation may not work as Ill for another.

REFERENCES

1. https://www.kaggle.com/datasets/saurabhshahane/data-science-jobs-salaries