TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHTN

ĐHQG-HCM

PROJECT 03: Linear Regression

Applied mathematics and statistics for information technology

I. Student information and complete progress

a) Student information

Name	**:**	**Bùi Nguyễn Nhật Minh**
Student ID	:	21127105
Class	:	21CLC08

b) Complete progress

TASK	COMPLETE
TASK 1A	100%
TASK 1B	100%
TASK 1C	100%
TASK 1D	100%

II. Project Introduction

The goal of this project is to investigate the determining factors behind the salary and employment prospects of engineers immediately after graduation. Factors such as academic performance at different educational levels, candidate skills, the connection between universities and industrial/technological hubs, degree qualifications, and the market conditions for specific industries all play a role in shaping these outcomes.

The dataset utilized for this project was collected in India, a country with over 6000 technical training institutions and approximately 2.9 million students enrolled. On average, around 1.5 million engineering/technical students graduate each year. However, due to the lack of necessary skills, less than 20% of them find employment that matches their expertise. This dataset not only aids in building a predictive salary tool but also provides insights into the factors impacting salaries and job roles in the labor market.

In this project, the requirement is to build a model to predict the salary of engineers using a linear regression model.

III. Libraries and functions

1) Libraries

matplotlib.pylot	Used for working with and processing tabular data (DataFrames). Supports reading, writing, cleaning, transforming, and analyzing data.
pandas	Used for working with and processing tabular data (DataFrames). Supports reading, writing, cleaning, transforming, and analyzing data.
numpy	Utilized for numerical computations and handling powerful multidimensional arrays. Increases efficiency in mathematical operations.
seaborn	Utilized for numerical computations and handling powerful multidimensional arrays. Increases efficiency in mathematical operations.
sklearn.model selection	Used to perform model selection-related tasks, such as data splitting for training and testing, cross-validation, and hyperparameter tuning.

2) Functions

fit	• Description: This function is part of a linear regression class and performs the fitting process of the linear regression model It computes the model weights using the formula: $w = (X^TX)^{-1}X^Ty$, here X is the input feature matrix and y is the target vector. • Parameters: ○ self: Instance of the linear regression class. ○ X: Input feature matrix. ○ y: Target vector. • Returns: The fitted instance of the linear regression class.
get_params	• Description: This function returns the learned model parameters (weights) after the linear regression model has been fitted. • Parameters: ○ self: Instance of the linear regression class. • Returns: The learned model parameters (weights).

predict	• Description: This function makes predictions using the linear regression model. It multiplies the learned model weights with the input feature matrix to make predictions. • Parameters: ○ self: Instance of the linear regression class. ○ X: Input feature matrix for which predictions are to be made. • Returns: An array of predicted values.
mae	• Description: This function calculates the Mean Absolute Error (MAE) between the actual target values (y) and the predicted values (y hat). • Parameters: ○ y: Actual target values. ○ y_hat: Predicted values. • Returns: The calculated MAE.
preprocess	• Description: This function preprocesses the input data by adding a column of ones to the left of the input matrix. It is often used to include a bias term in linear regression models. • Parameters: ○ x: Input data matrix. • Returns: The preprocessed input data matrix with an additional column of ones.

IV. Project result

1. Task 1a

Regression formula:

$$\begin{aligned} Salary = {} & 49248.089 - 23183.329 \times Gender \\ & + 702.766 \times 10percentage \\ & + 1259.018 \times 12percentage \\ & - 99570.608 \times CollegeTier + 18369.962 \times Degree \\ & + 1297.532 \times collegeGPA \\ & - 8836.727 \times CollegeCityTier + 141.759 \times English \\ & + 145.742 \times Logical + 114.643 \times Quant \\ & + 34955.750 \times Domain \end{aligned}$$

MAE on test.csv : 105052.529

2. Task 1b

No	Model with a feature	MAE
1	nueroticism	123473.399
2	agreeableness	123706.054
3	extraversion	123809.926
4	openess_to_experience	123818.333
5	conscientiousness	124182.563

Theory: The best model is the one that uses the nueroticism feature, the reason being that this feature affects salary. “Neurotic people tend to be anxious, negative, and get stressed out more easily. Although previous research has found that neuroticism is not correlated with job performance, the current study found that a one standard deviation increase in neuroticism is associated with a 5–9% decrease in annual wages” (Min, 2015)

Regression formula:

$$Salary = 304647.552 - 16021.493\ \times nueroticism$$

MAE on test.csv: 119361.917

3. Task 1c

No	Model with a feature	MAE
1	Quant	117353.838
2	Logical	119932.503
3	English	120728.603

Theory: The best model is the one that uses the Quant feature, the reason being that this feature affects salary because without quantitative and analytical skills, assessments and decisions can be subjective and more likely to have errors or unintended outcomes (TEXAS A&M UNIVERSITY CORPUS CHRISTI, 2022). Quantitative skills are like a currency in today's data-driven economy, where professionals who can translate numbers into insights hold the key to unlocking higher earning potential.

Regression formula:

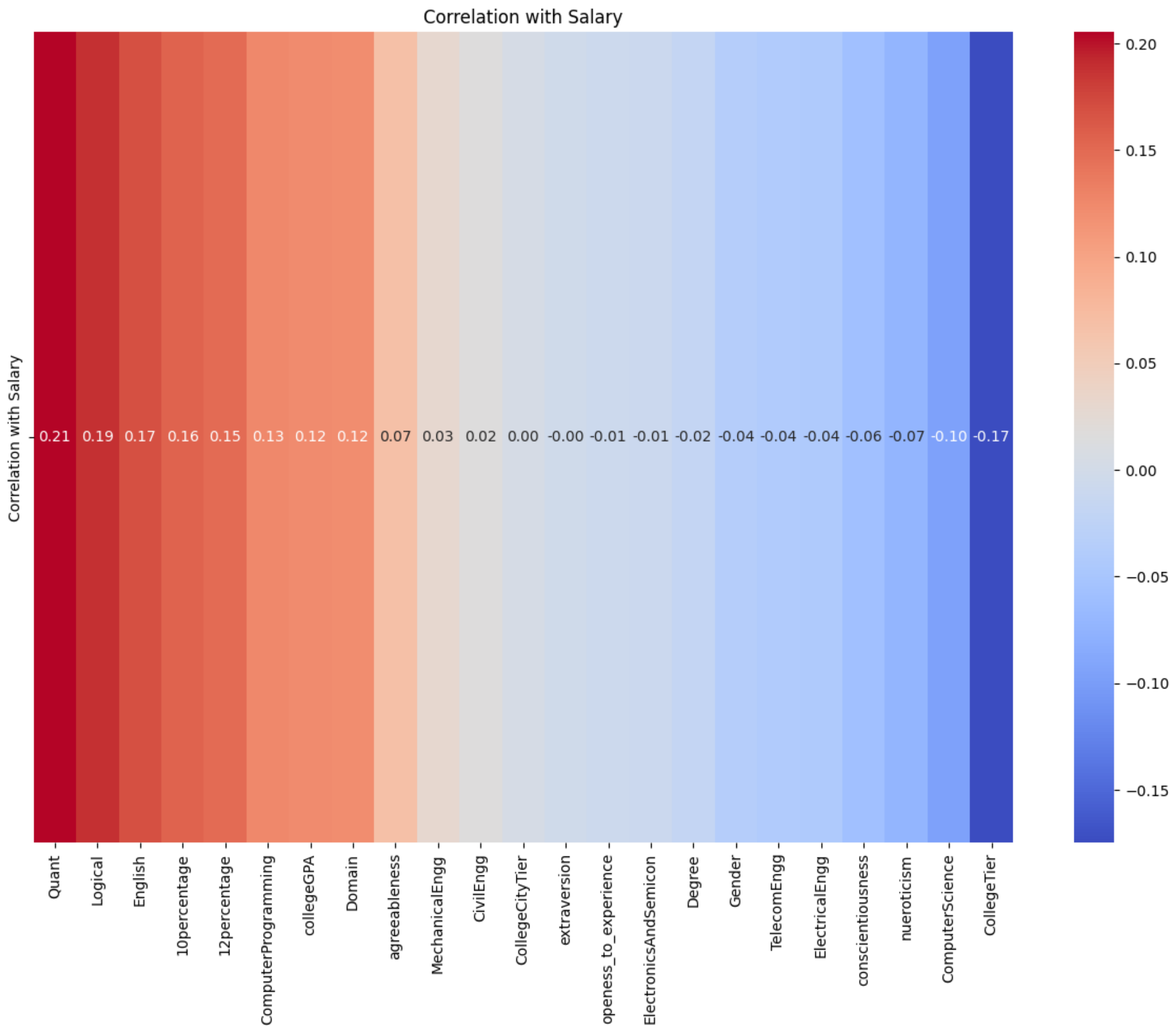
$$Salary = 117759.729 + 368.852 \times Quant$$

MAE on test.csv: 108814.059

4. Task 1d

Explain the reason for choosing the design for the model based on the heatmap:

- **High Positive Correlation**: Features with a high positive correlation with salary might be important factors in predicting salary. These features have a positive influence on engineers' salaries. Selecting these features can improve the model's predictive performance.
- **Low Correlation**: Features with correlations close to zero with salary might not contribute significantly to predicting salaries. These features can be omitted to focus on more important ones.
- **Negative Correlation**: Features with a negative correlation with salary could have a negative impact on salary. However, it's important to investigate further to determine whether this is a random relationship or if there is a valid reason. You may need to verify the data for accuracy.



Based on the heatmap chart and combined with the 2 best features from tasks 1b and 1c, I have decided to design 3 models:

1. **Model 1**: I have selected the features from the 1st to the 11th column of the original dataset. Additionally, I have created a new feature named "Quant_nueroticism_product" by multiplying the "Quant" and "nueroticism" columns. The intuition behind this model is to capture the interaction between "Quant" and "nueroticism" and investigate if their combined effect significantly impacts the Salary.
2. **Model 2**: I have chosen a specific set of features including educational metrics such as 10percentage,12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain, ComputerProgramming and openess_to_experience. Furthermore, I have added a feature "nueroticism_cubed" which is the cube of the "nueroticism" column and "Quant_nueroticism" by multiplying "Quant" and "nueroticism”.

3. **Model 3**: I have selected a subset of features that includes 10percentage, English, Logical, Quant, nueroticism. I have added "nueroticism_cubed" which is the cube of "nueroticism," "Quant_cubed" which is the cube of "Quant," "Quant_square" which is the square of "Quant," and "nueroticism_squarer" which is the square of "nueroticism."

No	Model	MAE
1	Model 2	113495.912
2	Model 1	113637.708
3	Model 3	115092.132

Regression formula:

$$\begin{aligned} Salary = {} & 54566.200 - 628.517 \times 10percentage \\ & + 1140.334 \times 12percentage - 100810.331 \times CollegeTier \\ & + 12114.761 \times Degree + 1003.468 \times collegeGPA \\ & - 9897.619 \times CollegeCityTier + 137.011 \times English \\ & + 119.559 \times Logical + 124.330 \times Quant \\ & + 25996.491 \times Domain \\ & + 68.762 \times ComputerProgramming \\ & - 4067.990 \times openess\ to\ experience \\ & + 698.259 \times nueroticism\ cubed \\ & - 12.170 \times Quant\ nueroticism \end{aligned}$$

V. Reference

(2022, November 23). Retrieved from TEXAS A&M UNIVERSITY CORPUS CHRISTI: https://online.tamucc.edu/degrees/business/mba/general/improve-quantitative-skills-mba/

Min, J.-A. (2015, May 13). *The Personality Traits That Increase Your Salary.* Retrieved from Linkedin: https://www.linkedin.com/pulse/personality-traits-increase-your-salary-ji-a-min-masc/

NgocTien0110. (n.d.). *Applied-Mathematics-and-Statistics.* Retrieved from github: https://github.com/NgocTien0110/Applied-Mathematics-and-Statistics/tree/main/project%203