Assignment 1  
Regression Models

short line

Ngoc Quang Vinh Pham

Student ID: 25100660  
February 29th, 2024

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Executive Summary 2**](#_xvtyf97iwmd0)

[**2. Business Understanding 3**](#_if9iwijz2pg3)

[a. Business Use Cases 3](#_bnjfuh5ewogw)

[**3. Data Understanding 4**](#_yg03y3w5hig9)

[**4. Data Preparation 5**](#_tj77rrmixzhi)

[**5. Modeling 6**](#_6vwh14nc44ga)

[a. Approach 1 6](#_uv8sn0ik6017)

[b. Approach 2 6](#_wtwxuxemqtvc)

[c. Approach 3 6](#_nrl8g3ri789x)

[**6. Evaluation 8**](#_mngiyx5tet6w)

[a. Evaluation Metrics 8](#_qx20m6hrk8uo)

[b. Results and Analysis 8](#_25v57zsj5m7g)

[c. Business Impact and Benefits 8](#_hkob76wu4d6q)

[d. Data Privacy and Ethical Concerns 9](#_uf1z6gbsejg6)

[**7. Deployment 10**](#_8rqux13i2o01)

[**8. Conclusion 11**](#_9hr11g79asdx)

[**9. References 12**](#_ma3y8ytjvp1y)

# Business Understanding

## Business Use Cases

The project focuses on evaluating engineering students’ academic performance, personal traits, and their college characteristics in order to make predictions of their salaries in the industry. For that, the application of the predictive model can be seen in a variety of industries and scenarios.

For examples:

**Use case 1** - College Graduates and Students can leverage our model to identify factors affecting their initial salaries, aiding in informed career and skill development decisions.

**Use case 2** – Universities and Educational Institutions can use predictions to guide students on salary expectations and refine programs for better employability and salary outcomes.

**Use case 3** – HR Departments and Recruitment Services can use the model to establish fair salary ranges for recent graduates, aligning offers with industry standards and streamlining recruitment.

This initiative is driven by the challenges of understanding wage determinants and the need for transparency in salary discussions. It capitalizes on machine learning to analyze complex data relationships, offering insights previously difficult to access. With validated historical data, developing algorithms to predict future trends becomes a reliable method for salary assessment. In a world prone to bias and self-interest, machine learning offers a way to ensure fairness in negotiations, providing unbiased metrics that benefit all parties involved.

1. Key Objectives

**The objectives (goals) of the projects are:**

* Enhance predictive accuracy by developing a machine learning model focused on minimizing RMSE, ensuring precise salary forecasts.
* To guarantee that the model promotes equal compensation practices by avoiding increasing biases, especially those related to gender or socioeconomic status.
* To provide stakeholders practical information on the main variables affecting salary levels, allowing data-driven decisions to be made about hiring, career planning, and policy formulation.

**Stakeholders:**

* **HR Departments and Recruitment Services:** aiming for fair and efficient hiring, using statistical analysis to guide decisions.
* **Universities and Educational Institutions:** seeking to boost their graduates' employability and salary prospects while guiding students towards suitable careers based on their profiles.
* **Students and Pupils:** Students can be informed and take the initiative to develop skills and plan for their career through the insights of the models.

These criteria can be addressed by the below methodologies:

* Refine regression models to improve salary prediction accuracy by evaluating a wide range of features.
* Conduct an in-depth analysis of key attributes to enhance stakeholders' understanding of salary determinants and support informed decision-making.
* Integrate fairness and bias assessments in the modeling process to ensure equitable outcomes for all demographic groups.

Machine learning algorithms will be utilized to balance ethical considerations with predictive precision, aiming for inclusive benefits.

short dash

# Data Understanding

The project's dataset, with details on engineering students' profiles, academics, and specializations, facilitates the development of a model to predict starting salaries using historical data.

**Data Sources and Collection Methods**

The dataset provided for the assessment consists of Training, Validation, and Test Sets, compiled from educational institutions, employment agencies, and surveys on engineering graduates' salaries. It includes data on personal qualities and academic achievements through categories and numerical factors. While the data collection method aims for a comprehensive overview, potential limitations include self-reporting biases in salary data and representation across different engineering fields and regions.

**Data Limitation**

* **Insufficient data:** With only around 2998 records, the dataset's small size could hinder model training and risk overfitting.
* **Outdated Information:** Rapid changes in the job market and education necessitate frequent dataset updates.
* **Bias Possibility:** Individuals might report salaries that are bias, with the problem that recent graduates usually round up their pays. This can be accounted as the integrity of data.
* **Representativeness:** The dataset may not equally cover all engineering disciplines, leading to biased predictions towards more represented areas.
* **Missing and duplicated data:** No missing or duplicated data issues are noted, with minor '0' values fixable by back-fill methods. However, real-world data often presents these challenges, requiring imputation techniques that might introduce biases.

**Variables and Features Significance**

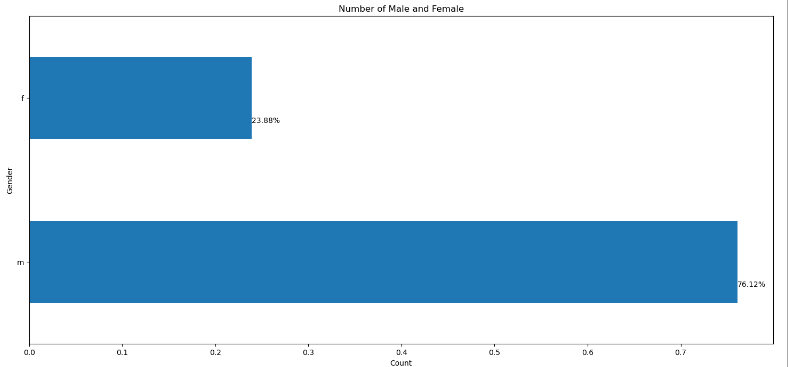
* **De-identification data** (ID, Gender, DOB): Determine the identity information of the candidate
* **Academic Performance** (10percentage, 12graduation, 12percentage, CollegeGPA)**:** Reflects a student’s academic background, which may be related to their knowledge, skills and earning potential.
* **University Characteristics** (CollegeID,CollegeTier, CollegeCityID, CollegeCityTier)**:** Show the college characteristics that may have an impact on salary outcomes by indicating the quality of education and networking possibility.
* **Field of Study data** (Degree, Specialization)**:** This data is essential for comprehending the differences in pay between engineering specialties.
* **Personal Traits** (conscientiousness, agreeableness, extraversion, neuroticism, openess\_to\_experience)**:** These show the score describe the soft skills of the students whose influence maybe less significant than that of academic and professional talents but may nonetheless have an impact on a candidate’s work performance and income.
* **Scores** (English, Logical, Quant, Domain, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, TelecomEngg, CivilEngg, CollegeGPA)**:** Provide certain information and skill sets that may be correlated with pay and work performance.
* **Graduation Year** - Year of graduation (bachelor’s degree): Assist in taking into consideration how the market changes affect pay levels over time.

**Exploratory Data Analysis**

The training dataset contains 2998 rows of records with 30 columns of features and 1 column is the target variable ‘salary’. Thankfully, this dataset doesn't include any null or duplicate values but only some Zero value which can be fill by back fill method that would affect the performance of the models, so we can quickly proceed beyond the cleaning step.



Three columns of qualitative data—"gender," "degree," and "specialization"—with 27 columns of numerical values comprise the characteristics' total of 30 columns of attributes. From that, we can present the distribution of male and female engineering students in the dataset.



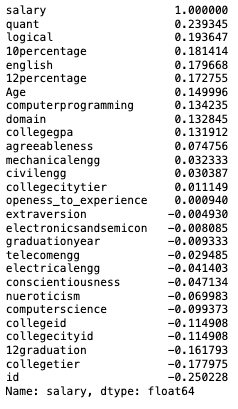
It is clear from the gender differences in your dataset that there is a substantial gender imbalance in the engineering profession, with 76.12% of engineering students being male and only 23.88% being female. Due to a number of socioeconomic, cultural, and educational variables, female involvement in STEM (Science, Technology, Engineering, and Mathematics) industries has historically been lower. This imbalance is symptomatic of a global trend in these subjects.

**Correlation Matrix**

A green square with black line

Description automatically generated

The correlation heatmap is a valuable tool for examining variable relationships in your dataset, displaying correlation coefficients between pairs to show their interconnections. We focus on identifying attributes most strongly correlated with the target value 'salary'. So, I have print out the correlation score between each feature and the output variable:



I discovered that features such as 'quant', 'logical', '10percentage', 'english', and 'age' have a strong correlation with 'salary', alongside attributes like '12percentage', 'mechanicalengg', 'collegegpa', and 'domain', which intuitively align with valuable skills in engineering roles. 'Collegetier' shows a significant negative correlation with salary, suggesting lower salaries for candidates from lower-tier colleges. To address this, I plan to adjust the 'collegetier' value, changing '2' to '0.5' to enhance the data's alignment with this pattern and potentially improve model performance. This adjustment reflects the impact of educational institution reputation on career outcomes. Now let’s observe how salaries vary across different categories:

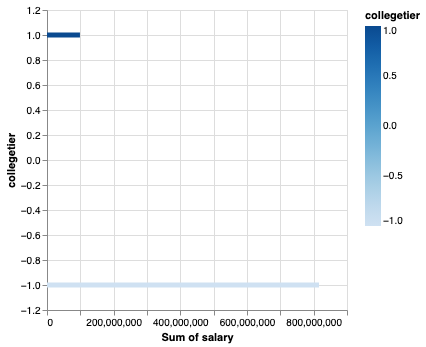
Between Degrees:

A graph with numbers and lines

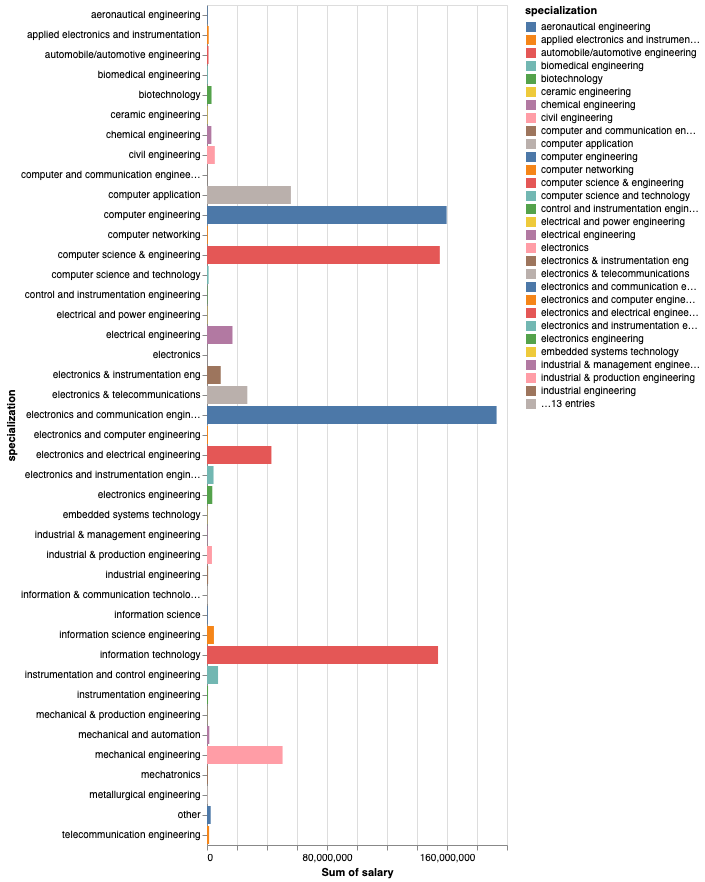
Description automatically generated

The majority of salaries belong to those with the B.Tech/B.E., or Bachelor of Technology with about $900,000,000, as this bar graph illustrates. The Master of Computer Application, or MCA, comes next with only about $60,000,000. Lastly is the two M.Tech./M.E or Master of Engineering and M.Sc. (Tech.) as Master of Science (Technology) respectively.

Between 2 different college-tiers (where ‘1’ indicate higher rank college)



Between specializations:



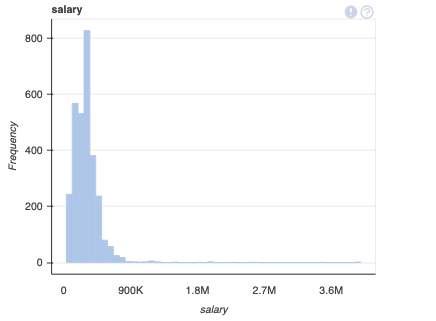
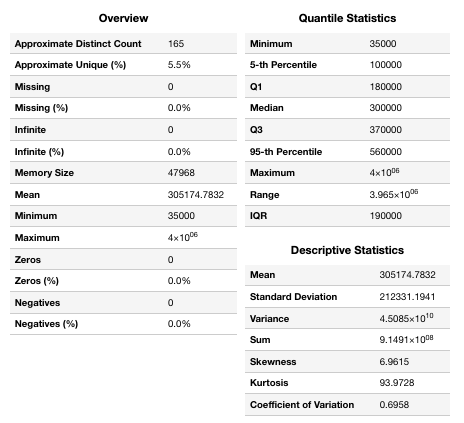
Salaries across 42 engineering specializations vary from $40,000 to $193,070,000, with fields like computer engineering and information technology earning significantly higher averages than others. In contrast, areas such as instrumentation engineering and mechatronics typically see lower average incomes. This research provides valuable insights into the pay disparities among engineering fields, aiding professionals, students, and policymakers in education and career planning.

Nevertheless, there is little to no correlation between these three traits and the target variable, which might have a negative impact on the model, so we won’t apply encoding method to these categorical data. The results section will provide the outcome of the negative impact.

**Numerical Data Visualization**

To visualize salary distributions, skewness, and outliers, we use the dataprep library for plotting statistics, histograms, and boxplots, along with scatterplots to deeply understand target value distributions across features.

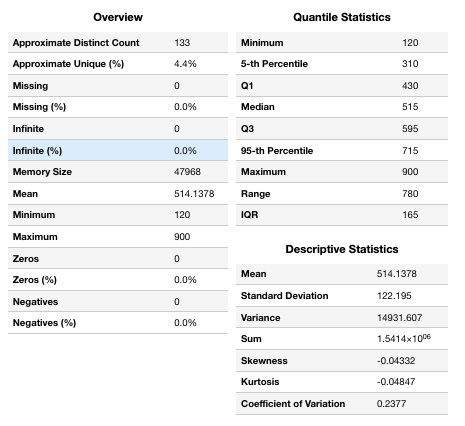
**Distribution and Statistics of target variable ‘salary’:**

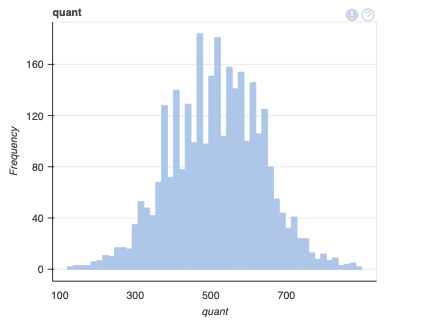


The pay distribution is biased to the right and is regularly distributed, with the bulk of salaries falling between $60,000 and $500,000. The average pay is around $305,000 when the standard deviation is about $212500.

**Distribution and Descriptive Statistics of Important Features**

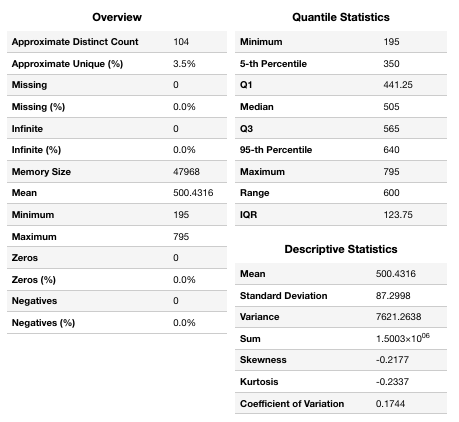
**‘Quant’**

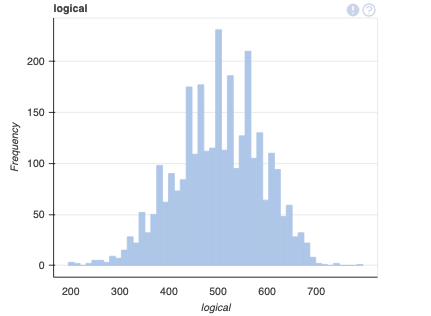
****

 A graph with blue dots

Description automatically generated

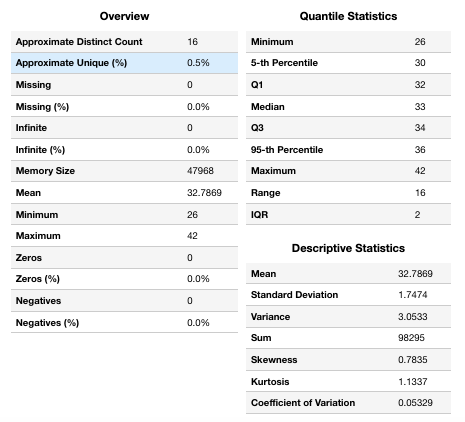
**‘Logical’**

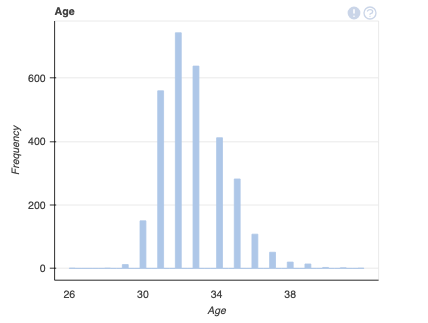
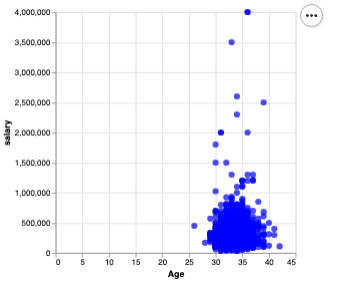
****

 A graph with blue dots

Description automatically generated

**‘Age’**

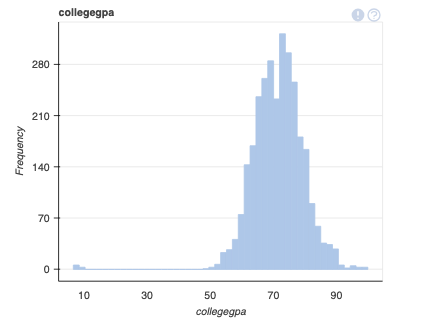
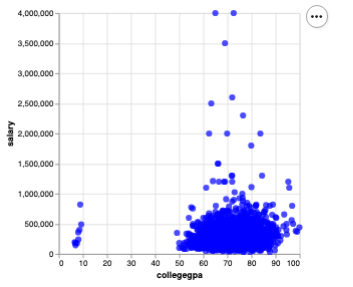
****

**‘CollegeGPA’**

**A screenshot of a graph

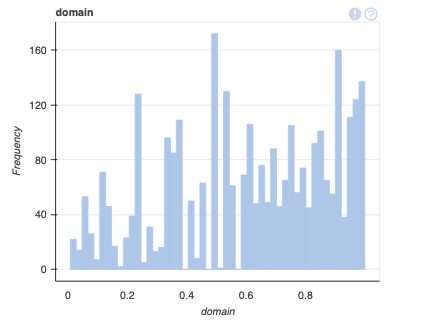
Description automatically generated**

**‘Domain’**

**A screenshot of a table

Description automatically generated**

 A graph with blue dots

Description automatically generated

**‘10percentage’**

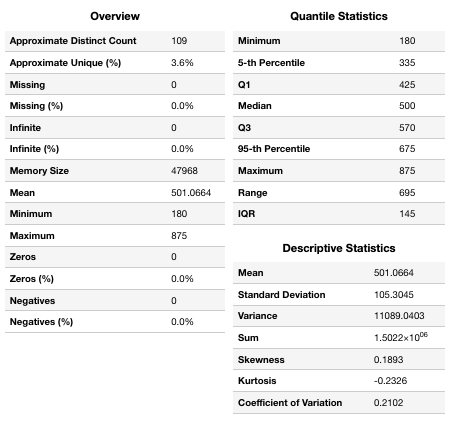
**A screenshot of a table

Description automatically generated**

**A graph of a number of percentages

Description automatically generated**

**‘English’**

****

**A graph of a number of english

Description automatically generated**

**‘12percentage’**

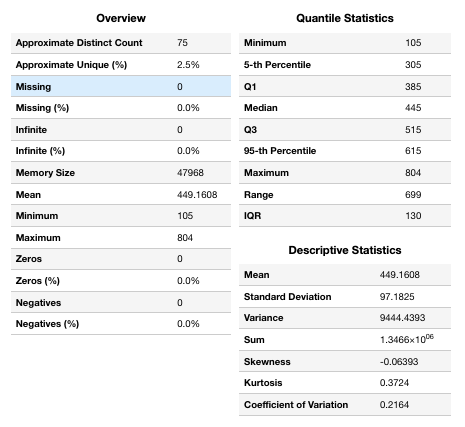
**A screenshot of a table

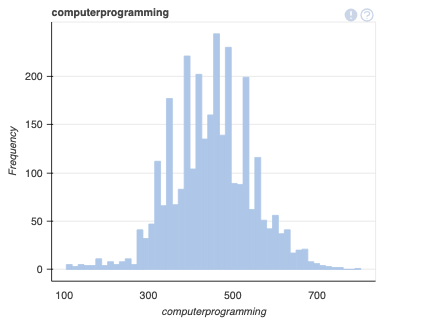
Description automatically generated**

**A graph of a graph

Description automatically generated**

**‘Computerprogramming’**

****

****

**‘Agreeableness’**

**A screenshot of a table

Description automatically generated**

**A graph of a graph

Description automatically generated**

The following summary is shown by the histograms of several student attributes:

* Students' skills in quantitative and logical reasoning are typically average, as shown by their scores, which fall in the center of the range.
* Most students are in their early 20s, which is normal for those planning to attend college.
* The majority of students have ordinary scores, with very few having very high or low GPAs.
* Specialization areas vary, with no one domain predominating.
* Percentages of students in the 10th and 12th grades: Students' performance in the 10th grade was superior to their more inconsistent performance in the 12th grade.
* English Proficiency: Most students speak the language at a moderate to high level.
* Students are split into two groups: those with strong programming abilities and those with weaker abilities.
* Lower scores for agreeableness suggest a more competitive personality characteristic among students.

Overall, these distributions provide an overview of the skill sets and personality attributes of the student population, demonstrating a competitive spirit, various levels of proficiency in specific subjects, and average academic and logical ability.

short dash

# Data Preparation

**Data Cleaning**

* Handling Missing Values: There is no null value in the dataset, only ‘Zero’ value and ‘-1’ value that can be dealt by replacing with nan value and using back fill null value method. Depend on the proportion of missingness, we would consider drop the whole column from the dataset for the predictive model (civilengg and mechanicalengg columns)

Ex: A close-up of a sign

Description automatically generated

* No record of duplicated value.

**Feature Engineering**

* Feature Selection: We must choose the important features that are highly correlated with the output variable ‘salary’. After the Exploratory Data Analysis, the features that are significant and reliable for the models are ‘quant’, ‘logical’, ‘10percentage’, ‘english’, ‘12percentage’, ‘collegegpa’, ‘domain’, ‘collegetier’, ‘computerprogramming’, ‘agreeableness’, ‘Age’.
* Feature Creation: We have created ‘Age’ feature from ‘DOB’ and would be beneficial for the visualization part to see difference of salary between ages.



**Data Preprocessing**

* Data Encoding: Categorical data are mostly uncorrelated with the output value, which should only be used for data visualization and analysis.
* Feature Scaling: We scaled the data using MinMaxScaler from the sklearn.preprocessing module to generalize the weight of each feature in the predictive models.

**Data Splitting**

The project provided 3 different datasets:

* Training Dataset: 2998 records
* Validation Dataset: 599 records
* Testing Dataset: 599 records

Given that the validation and testing data each represent approximately 20% of the dataset in comparison to the training data, it's important to split the data accurately to ensure that the model is trained, validated, and tested properly. We just need to apply the previous data manipulation step to the validation and the test set and split the column ‘salary’ of 3 dataframes into y\_train, y\_valid and y\_test and the dataframes containing all the chosen features into X\_train, X\_valid, X \_test.

short dash

# Modeling

For this assignment, we'll test the dataset using five machine learning models: Multivariate Linear Regression, Ridge, Lasso, ElasticNet Regression, and k-Nearest Neighbors (kNN). Initially, we'll establish a baseline model, gauging its performance by the Root Mean Squared Error (RMSE) between the sum of the target variables' mean and the actual outputs. Here is the performance of the baseline model:

RMSE value of baseline model performance on training set: 212295.77905147275

RMSE value of baseline model performance on validation set: 286019.14519087254

RMSE value of baseline model performance on test set: 169575.84318922673

The breakdown of each algorithm and the rationale for its selection are:

**Multivariate Linear Regression:** Maps the relationship between multiple independent variables and a continuous dependent variable. It's used as a baseline for its simplicity and clarity in showing how variables affect salary, setting a performance benchmark.

**Ridge Regression:** Incorporates L2 regularization to mitigate overfitting while improving model performance by addressing multicollinearity and overfitting risks with many features.

**Lasso Regression:** Applies L1 regularization, reducing some coefficients to zero for feature selection, thereby eliminating less informative variables and simplifying the model.

**ElasticNet Regression:** Merges L1 and L2 regularization advantages, offering flexibility and optimizing model by balancing feature reduction and shrinkage.

**k-Nearest Neighbors (kNN):** Uses a non-parametric method based on the closest 'k' instances, chosen for its ability to capture nonlinear relationships and compare linear model performance by similarity in the feature space.

**Parameter Tuning and Model Selection Process**

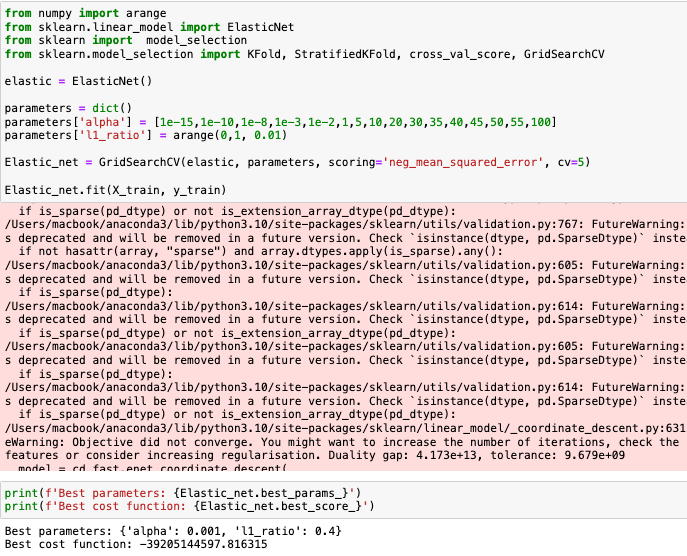
1. **Parameter Tuning**

* Methodology: Using GridSearchCV (scikit-learn) and cross-validation, each model’s parameter values were systematically explored.

Ex\_1: Ridge Regression Parameter Tuning



Ex\_2: ElasticNet Regression Parameter Tuning



* Objective: To find the ideal parameters that reduce the model’s root mean squared error on the validation set and improve generalization to unseen data.

**2. Model Selection**

* Criteria: Predicted on the validation set’s root mean squared error (RMSE), which measures the difference between anticipated and actual salaries.
* Procedure: Every model was assessed on the validation set after being trained on the training set. The best performing model was defined as the on with the lowest RMSE on the validation set.
* Complexity: Established a balance in the trade-off between model performance and complexity, seeking for comprehensible and accurate models.

We want to create a reliable predictive model that reliably predicts engineering students' salary using this logical approach to modeling, parameter manipulation, and model selection. This model will be useful for students, teachers, and employers worldwide.

short dash

# Evaluation

## Results and Analysis

**Results**

* Multivariate Linear Regression

RMSE value of my multivariate linear regression model performance on testing set: 159874.1581536953

* Ridge Regression Model

RMSE value of Ridge linear regression model performance on testing set: 159318.96385930316

* Lasso Regression Model

RMSE value of Lasso linear regression model performance on testing set: 159491.14258425796

* ElasticNet Regression Model

RMSE value of Elastic-Net regression model performance on testing set: 167821.70988653516

* k-Nearest Neighbors Regression Model

RMSE value of K-Nearest neighbors regression model performance on testing set: 154556.51795681656

**Analysis**

The k-Nearest Neighbors (kNN) model had the lowest RMSE on the testing set, outperforming more complex models by effectively capturing complex data patterns. This suggests a strong link between feature space distance and salary similarity. Despite the simplicity of kNN, it surpassed complex regression models, highlighting the importance of feature space in salary prediction.

Regularization improved model performance by preventing overfitting. Linear models had comparable results, with Ridge Regression slightly leading, indicating a need for either enhanced feature engineering or alternative modeling techniques.

ElasticNet model's performance was below expectations due to difficulties in balancing L1 and L2 penalties, underlining the importance of model selection based on validation results and hyperparameter tuning over assuming complexity ensures superiority.

**Hypothesis Testing:** Categorical encoding can enhance model performance.

The application of the encoding method to the categorical data shows the negative impact on the model performance which will be shown in these hypothesis testing:







Results:

Without feature encoding: RMSE value of K-Nearest neighbors regression model performance on testing set: 154556.51795681656

Using feature encoding: RMSE value of K-Nearest neighbors regression model performance on testing set: 155232.34548933583

**The hypothesis is rejected.**

**Key Insights**

Model Complexity vs Performance: The performance of various models shows that complexity doesn't guarantee accuracy. Understanding the data's key features and choosing a suitable model is crucial.

Feature Importance: The slight improvements seen with regularization suggest that while not all features are equally important, they do contribute to salary prediction, indicating a need for further investigation into which features most effectively predict salary.

Generalization Ability: The variation in RMSE across training, validation, and testing sets reflects each model's generalization capability. The exceptional performance of the kNN model on the testing set suggests it may be capturing underlying data patterns consistent across different subsets.

## Business Impact and Benefits

Final testing reveals the k-Nearest Neighbor model as highly effective, offering significant implications for various business scenarios:

**College Graduates and Students:** This model provides insights into factors influencing initial salaries, such as 'quant', 'logical', '10percentage', 'english', and others, guiding students on areas for improvement and informed career choices. It also indicates higher salary distributions in fields like computer engineering and IT, directing graduates towards lucrative career paths.

**HR Departments and Recruitment Services:** By forecasting salary ranges based on diverse factors, organizations can enhance offer acceptance rates and employee satisfaction. Although the model offers insights, employers must evaluate other aspects for recruitment. It also assists in identifying market trends and skill gaps, aiding in strategic workforce planning and education policy.

**Universities and Educational Institutions:** Insights from the model enable institutions to guide students towards specializations aligned with their traits and academic performance, optimizing educational outcomes and career readiness.

## Data Privacy and Ethical Concerns

Throughout the whole process, ethical issues and data protection were crucial due to the sensitivity of pay data and its potential for improper utilization:

* Anonymization: A crucial initial step in data management was ensuring that individual identities could not be tracked from the dataset.
* Bias Prevention: Consideration was made to evaluate models for potential biases, especially those that would maintain current wage inequality based on racial or ethnic background, gender, or other non-merit-based characteristics. Models have been modified where necessary to reduce these biases.
* Reliable Application: Giving stakeholders clear understanding about the models' development process, constraints, and suitable use cases assists in preventing incorrect interpretation and inappropriate utilization of the predictions.

The project's results show how machine learning models or in this case is the k-Nearest Neighbor Regression model may be used to make data-driven judgments about hiring and career counseling, but they also serve as an ongoing reinforcement of the necessity for privacy protection and ethical awareness in data science applications.

short dash

# Conclusion

**Key Insights**

* Model Performance: The test set showed that the k-Nearest Neighbors (kNN) model performed better than the others, demonstrating its capacity to recognize complicated patterns in data. By addressing overfitting, regularization in the Ridge and Lasso models provided marginal improvements.
* Importance of Feature Engineering: The project demonstrated how important feature engineering and feature selection are to improving model performance. In addition to reducing overfitting, regularization approaches highlighted the importance of carefully balancing model complexity.
* Data preparation and Ethical Issues: The project's foundation was strict data preparation, which included managing outliers and missing values as well as ethical issues with data privacy and bias reduction.

By providing a data-driven viewpoint on compensation standards and career guidance that matched stakeholder demands, the initiative achieved its goal.

* Investigate more complex models and group techniques.
* To increase model accuracy, add more variables to the feature set.
* As the model grows, it is important to conduct ongoing audits for fairness and bias in the model.
* Testing and deployment in the real world will give light on useful applications and potential areas for development.

In conclusion, the research represents an important breakthrough in the use of machine learning to career counseling, with possible advantages for a range of engineering industry stakeholders. The benefit of the model will be further validated and improved by future developments and practical use.

short dash

# References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer. [A foundational text that covers many of the statistical learning techniques used in the project.]

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10111288/>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830. [For the implementation of machine learning models like Ridge, Lasso, ElasticNet, and kNN.]

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. AI Magazine, 38(3), 50-57. [Discusses the ethical and regulatory considerations relevant to the project, especially around data privacy and algorithmic transparency.]

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7572278/>

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-0285-1>

McKinney, W. (2012). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc. [Provides insights into the data preprocessing steps used in the project.]

Hyndman, R.J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts. [Reference for methodologies related to time series forecasting and model evaluation metrics like RMSE.]

<https://arxiv.org/ftp/arxiv/papers/1603/1603.01857.pdf>

short dash