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:**

* **Predictive Accuracy Optimization**: Using a collection of specified features as a basis, create a machine learning model with predictive accuracy that minimizes the Root Mean Squared Error (RMSE), a performance measure used to assess performance.
* **Equality and Equity**: To guarantee that the model promotes equal compensation practices by avoiding increasing biases, especially those related to gender or socioeconomic status.
* **Insightful Investigation**: 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:

* Optimizing regression models to make precise salary evaluations by taking into account a variety range of features to capture the complexity of the important features.
* Bringing a detailed analysis set of important attributes, giving stakeholders a clear understanding of the factors that most affect the salary prediction outcomes and facilitating targeted interventions and decisions.
* Incorporating fairness and bias examines into the modeling process to identify and differentiate any biases and ensure that the evaluations fairly serve the interests of a diverse population

In order to ensure that the advantages are widespread and inclusive, we are going to apply machine learning algorithms into this research to achieve an equilibrium between the subject of integrity and prediction accuracy.

short dash

# Data Understanding

The project dataset contains information on engineering students' personal, academic performances, specializations, and post-graduation outcomes, enabling the analysis of factors influencing starting salaries. This foundational data supports the creation of a model to predict future graduates' initial earnings based on historical insights.

**Data Sources and Collection Methods**

The dataset has been given by the assessment dataset with 3 different set: Training Set, Validation Set and Test Set. However, based on the structure of the data, we can conclude that it has been compiled from various educational institutions, employment agencies, and possibly surveys to obtain information on engineer graduates’ salaries. It reflects both personal qualities and academic accomplishments through a combination of categories and numerical factors. Although the technique of data collection that we anticipate guarantees a thorough overview, it’s important to be aware of potential limitations, including self-reporting biases in income data and the likelihood that the sample is representative across various engineering field and geographical areas.

**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**

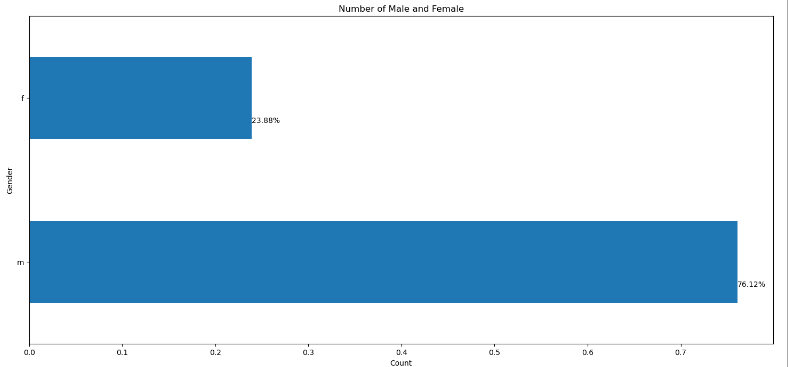
The dataset contains 599 rows of records with 30 columns of features and 1 column is the target variable ‘salary’. Thankfully, this dataset doesn't include any null or duplicate values but only some Zero value which can be fill by back fill method that would affect the performance of the models, so we can quickly proceed beyond the cleaning step.



I made a column called "Age" from "DOB" to extract the age range of the students and see the variation in pay between different ages.



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.



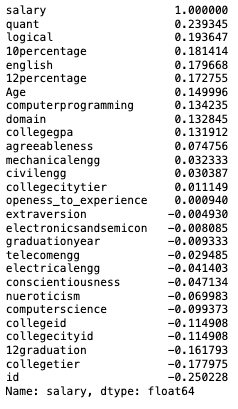
It is clear from the gender differences in your dataset that there is a substantial gender imbalance in the engineering profession, with 76.12% of engineering students being male and only 23.88% being female. Due to a number of socioeconomic, cultural, and educational variables, female involvement in 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

Above is a great visual tool for exploring the relationships between different variables in your dataset. In a correlation heatmap, each cell shows the correlation coefficient between two variables, indicating how much one variable is related to another. However, we are considering which attributes are most like correlate with the target value ‘salary’. So, I have print out the correlation score between each feature and the output variable:



After presenting the correlation between attributes, I found out that some features are highly correlate with the target feature ‘salary’. The data show that 'quant', 'logical', '10percentage', 'english', ‘age’ is highly correlative with 'salary'. Following are the attributes: '12percentage', 'mechanicalengg', 'collegegpa', 'domain'

* Intuitive sense that these skills are valuable in many engineering roles.

The 'collegetier' metrics show significant negative correlation with salary, indicating that candidates from colleges with a lower tier (higher numerical value) tend to have lower salaries.

* This could reflect the perceived quality or reputation of the educational institution affecting career prospects. So, I tend to change the value of the ‘collegetier’ to increase the weights of the data point ‘1’ which indicate the higher score rank college.

Since the 'collegetier' attributes is negatively correlated with the target output because of the students from a lower tier institution (higher value) usually have a higher salary than the one from higher tier college. In order to tuning the data to fit the algorithms, I tend to change the value ‘2’ in the collegetier to ‘0.5’ to see if the performance is improved or not.

No let’s observe how salaries vary across different categories:

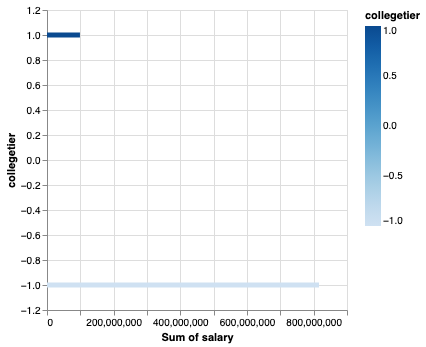
Between Degrees:

A graph with numbers and lines

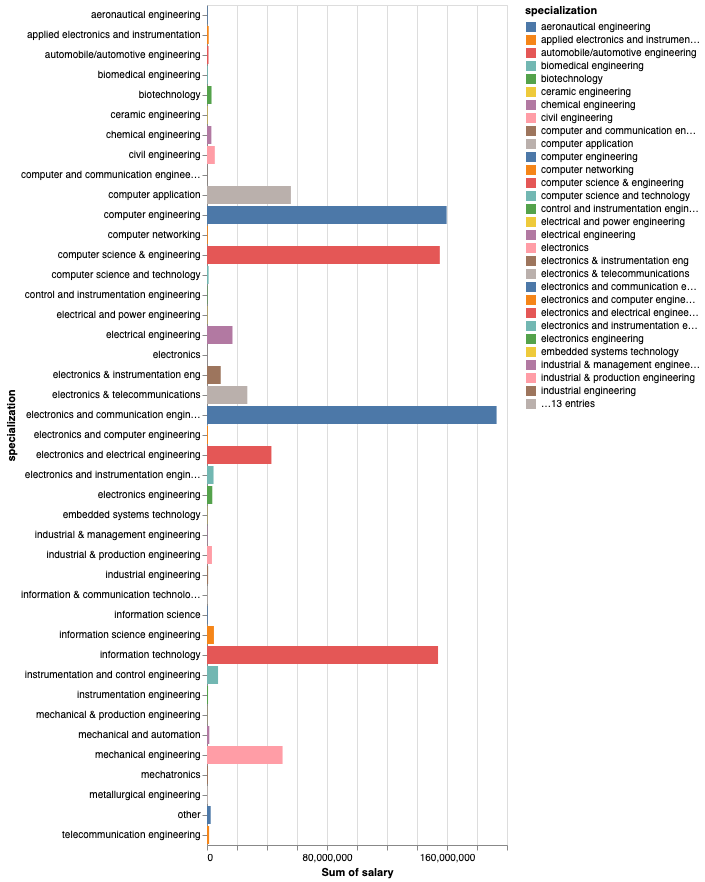
Description automatically generated

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:



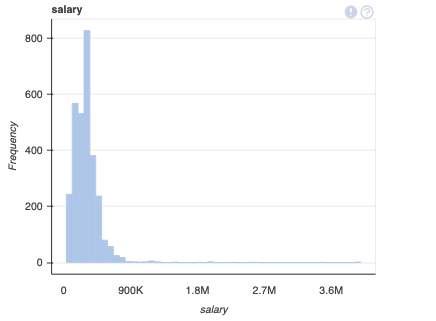
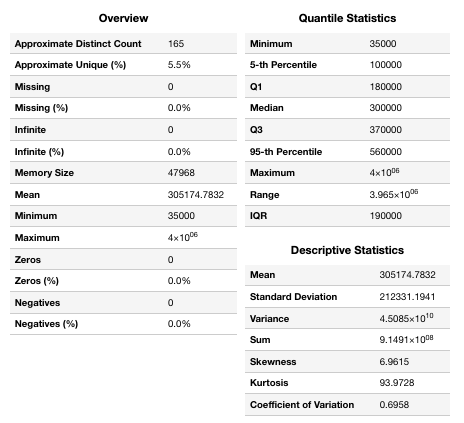
The salaries vary across 42 different engineering specializations, ranging from $40,000 to $193,070,000. The average salary for engineering specialties including computer engineering, computer science & engineering, electronics and communication engineering, and information technology is much greater than for the other specializations. However, as compared to the other engineering specialties, specializations like instrumentation engineering, mechanical and automation, and mechatronics often have lower average incomes.

All things considered, this research offers insightful information on the pay range for different engineering specializations, which may help professionals, students, and policymakers make wise decisions about their education, careers, and workforce planning.

**Numerical Data Visualization**

In order to visualize the distribution of salaries and other continuous variables, identifying skewness or outliers, we are using dataprep library to plot the statistics, histograms, and boxplots to visualize the quantile statistics, descriptive statistics and distributions, outliers of the data, combining with scatterplot to understand deeper the distribution of the target values across different features.

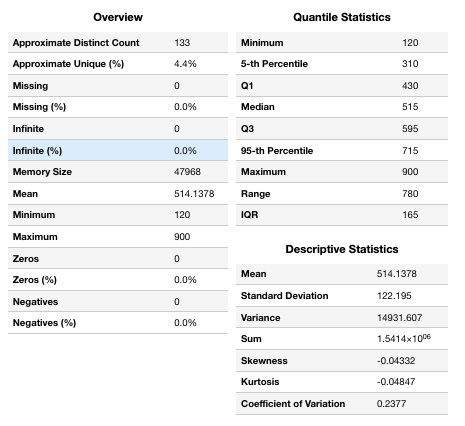
**Distribution and Statistics of target variable ‘salary’:**

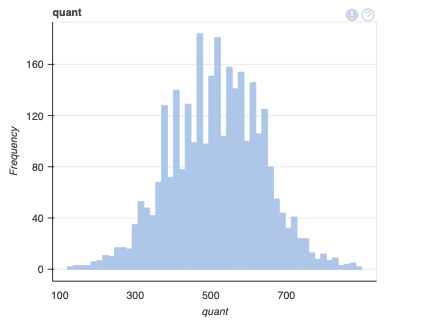


The pay distribution is biased to the right and is regularly distributed, with the bulk of salaries falling between $60,000 and $500,000. The average pay is around $305,000 when the standard deviation is about $212500.

**Distribution and Descriptive Statistics of Important Features**

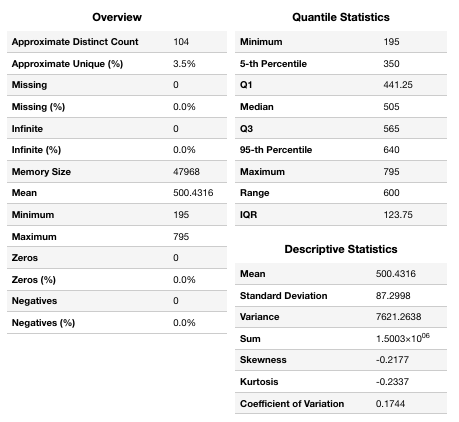
**‘Quant’**

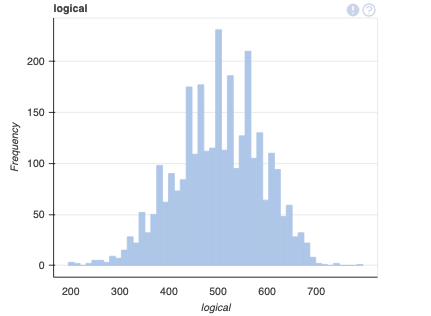
****

 A graph with blue dots

Description automatically generated

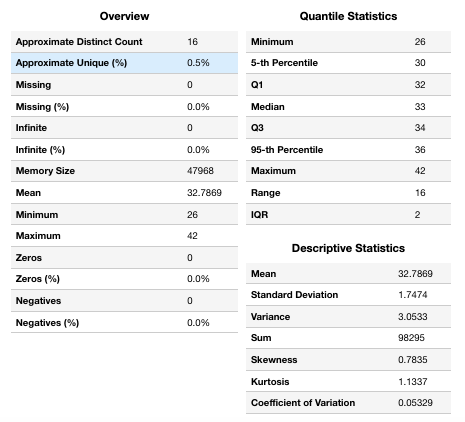
**‘Logical’**

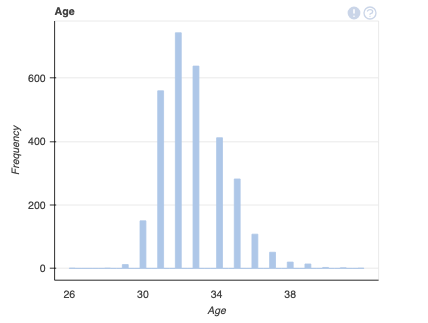
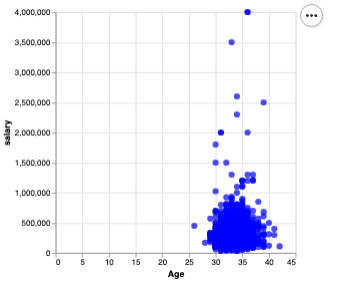
****

 A graph with blue dots

Description automatically generated

**‘Age’**

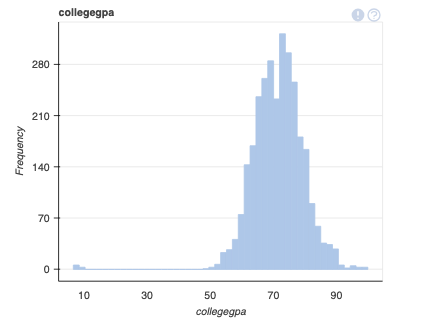
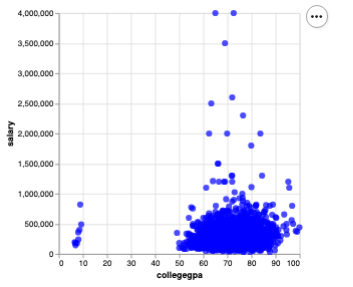
****

**‘CollegeGPA’**

**A screenshot of a graph

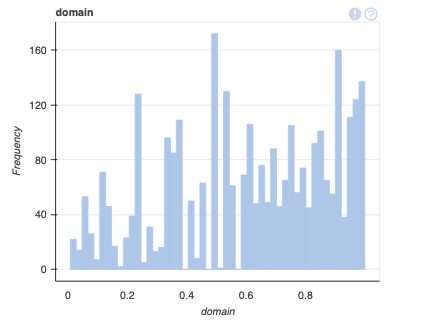
Description automatically generated**

**‘Domain’**

**A screenshot of a table

Description automatically generated**

 A graph with blue dots

Description automatically generated

**‘10percentage’**

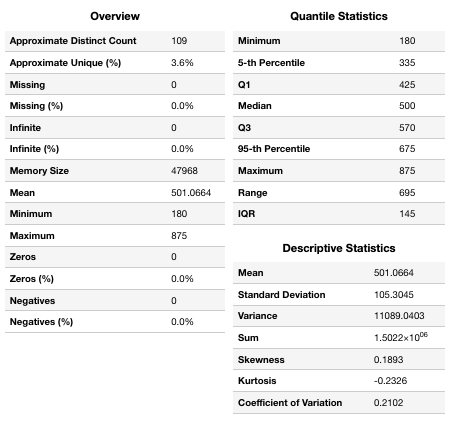
**A screenshot of a table

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’**

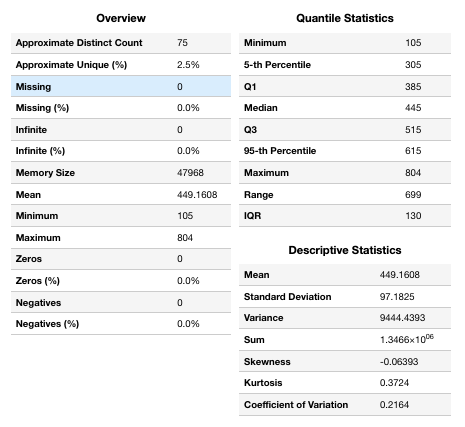
**A screenshot of a table

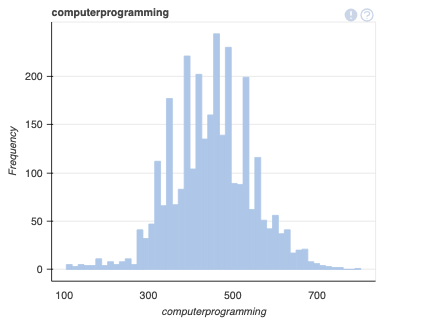
Description automatically generated**

**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**

short dash

# Data Preparation

* Describe the steps taken to prepare the data for modeling.
* Discuss the data cleaning, preprocessing, and feature engineering techniques applied.
* Document any handling of missing values, outliers, or imbalanced data.

Instructions: Describe the data preparation steps taken before modeling. Include details about data cleaning, preprocessing, and feature engineering techniques applied. Explain how missing values, outliers, or imbalanced data were handled and any transformations performed on the dataset.

**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

In this assignment, we need to run multiple experiments on the dataset with 5 different machine learning models: Multivariate Linear Regression, Ridge Regression, Lasso Regression, ElasticNet Regression, and k-Nearest Neighbors (kNN). First and foremost, we have to create a baseline model that the performance of this model is the Root Mean Squared Error of the difference between sum of the mean of target variables and the real output variables that are provided.

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:** A straightforward linear method for simulating the connection between a continuous dependent variable and many independent variables.

Rationale: It offers a clear description of how each aspect affects pay as a baseline model. Setting a performance standard for models with greater complexity might be beneficial.

**Ridge Regression:** we are obligated to generate a regression model with the regularization terms (L2 regularization) in the cost function to reduce overfitting and enhance the model performance at the same time.

Rationale: Chosen to minimize the possibility of multicollinearity across features and to avoid overfitting, which is prone to occur when there are a lot of features.

**Lasso Regression:** Similar to the Ridge Regression model, The Lasso regression applies the regularization terms (L1 regularization) which is the absolute value of the weights that can shrink some unimportant coefficients to zero, effectively performing feature selection.

Rationale: Selected due to its capacity to generate more comprehensible models by removing uninformative elements, which is beneficial in comprehending the most significant pay predictors.

**ElasticNet Regression:** Combining both L1 and L2 Regularizations, offering a balance between Ridge and Lasso Regression’s properties.

Rationale: Chosen for its adaptability to a range of data situations and its capacity to optimize model complexity and performance by combining the shrinkage approach of L2 regularization with feature reduction of L1 regularization.

**k-Nearest Neighbors (kNN):** A non-parametric approach that anticipates the output variable base on the ‘k’ most similar instances in the training set.

Rationale: Offering an alternative method to linear models by capturing nonlinear interactions between the target and features. The purpose of choosing it is to assess performance against linear models and determine whether pay is commensurate with feature space similarity.

**Parameter Tuning and Model Selection Process**

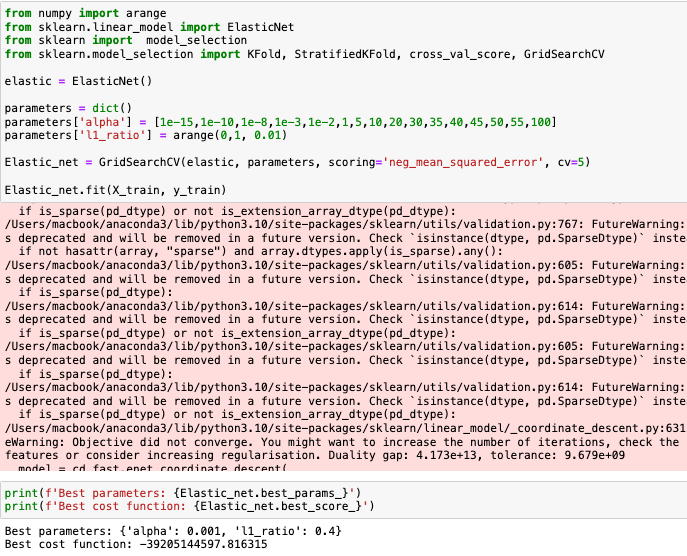
1. Parameter Tuning

* 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 testing set with the lowest RMSE demonstrated the best performance of the k-Nearest Neighbors (kNN) model, demonstrating the model's effectiveness in identifying complicated trends in the data. In spite of its simplicity, kNN performed better than more complex regression models, indicating that distance in feature space has a substantial correlation with pay similarity in terms of salary prediction.

Regularization prevents overfitting, which significantly improves model performance. Linear models showed similar performance measures, with Ridge Regression slightly outperforming others on the testing set. The little advances, meanwhile, suggest that either more sophisticated feature engineering or the investigation of different modeling approaches are required.

Due to the challenge of balancing the L1 and L2 penalties in this specific scenario, the ElasticNet model did not perform as well as anticipated. This result emphasizes the significance of selecting models based on validation performance and fine-tuning hyperparameters instead of supposing that a more complicated model would inevitably perform better.

**Key Insights**

Model Complexity vs Performance: The different models' performances highlight the fact that more complicated models do not necessarily produce more accurate predictions. Rather, it is essential to comprehend the fundamental features of the data and select a model appropriately.

Feature Importance: The minor improvements with regularization demonstrate that all features—possibly not equally—contribute to salary prediction. This illustrates a need for more research on the characteristics that most accurately predict salary.

Generalization Ability: Each model's capacity to generalize is indicated by the variation in RMSE across its training, validation, and testing sets. Particularly, the kNN model's outstanding performance on the testing set raises the possibility that it is identifying underlying patterns in the data that are similar across multiple subsets.

## Business Impact and Benefits

* Assess the impact and benefits of the final model on the business use cases.
* Discuss how the model contributes to solving the identified challenges or exploiting opportunities.
* Quantify the improvements achieved and the potential value generated.

Instructions: Assess and discuss the impact and benefits of the final model on the identified business use cases. Explain how the model contributes to solving the identified challenges or exploiting opportunities. Quantify the improvements achieved and discuss the potential value generated by the model.

As of the final performance based on the testing set, the k-Nearest Neighbor is the most effective model for this project which can bring a lot of impact to the business use cases:

* **Use case 1 - College Graduates and Students**: By using these insights, students can leverage our model to identify factors affecting their initial salaries which is ‘quant’, ‘logical’, ‘10percentage’, ‘english’, ‘12percentage’, ‘collegegpa’, ‘domain’, ‘computerprogramming’, ‘agreeableness, these scores help students realize what aspects should they need to improve, aiding in informed career and skill development decisions. Moreover, the majority of salary distribution fall on these specializations: computer engineering, computer science & engineering, electronics and communication engineering, and information technology, etc. which can guide college graduates on the right direction for their career paths.
* **Use case 2 - HR Departments and Recruitment Services**: Organizations may increase offer acceptance rates and employee satisfaction by making better educated offers to candidates by properly forecasting a settlement range based on a variety of criteria. Nevertheless, the predictive model is only giving the employers more options and information about the interviewee, not the decision-making factor since the employers still have to assess a lot of aspects for recruitment purposes. By identifying trends and skill shortages in the labor market, firms and policymakers may make more informed decisions on workforce development and education policy. This is made possible by having a better understanding of the variables that lead to higher salaries.
* **Use case 3 - Universities and Educational Institutions:** By using these insights, educational institutions may help students choose to use the specializations above to apply the right guidance to the right cluster of students that best suitable for their traits and educational performances.

## Data Privacy and Ethical Concerns

* Assess the data privacy implications of the project.
* Discuss any ethical concerns related to data collection, usage, or model deployment.
* Address steps taken to ensure data privacy and ethical considerations.

Instructions: Assess the data privacy implications of the project, considering any sensitive information or privacy concerns related to data collection, usage, or model deployment. Discuss any ethical concerns and considerations. Address the steps taken to ensure data privacy and mitigate 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

* Summarize the key findings, insights, and outcomes of the project.
* Reflect on the project's success in achieving its goals and meeting stakeholders' requirements.
* Discuss any future work, recommendations, or next steps based on the project's outcomes.

Instructions: Summarize the key findings, insights, and outcomes of the project. Reflect on the project's success in achieving its goals and meeting stakeholders' requirements. Discuss any future work, recommendations, or next steps based on the project's outcomes.

short dash

# References

* Include a list of references used throughout the project report.

Instructions: Include a list of references used throughout the project report, following the appropriate citation style.

short dash