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: They can use our model to understand the specific subjects or attributes might impact their starting salaries and be able to examine themselves which help them to be informative and make the right decision about their career planning and skills enhancement.

**Use case 2** – Universities and educational institutions: Based on their existing and extracurricular profiles, these organizations may use forecasts to advise their students about possible income expectations. It helps in refining educational programs to improve employability and pay results, at the same time helping to define realistic professional objectives.

**Use case 3** – HR Departments and Recruitment Services: By using the model, they may improve offers and negotiations by estimating reasonable wage ranges for recent grads. This tool facilitates the recruiting process by bringing offers and expectations into line with the standards of the industry.

The complexity of the factors determining wage results and the need for willingness in salary negotiations are the main obstacles motivating this initiative. The chance is in using machine learning to examine these complex relationships and provide stakeholders insights that were hard to measure before. Data never lies. As long as the historical information is validated and trustful, building an algorithm for formulate a trend that could use for predicting future data is the best thing we can do when it comes to assessing salaries. In this world where everybody placed their benefits above everything else, evaluations can become biased, leaving room for individuals to be exploited. Therefore, I see machine learning as a method to bring balance to negotiations, providing metrics that serve as fair tools benefiting both parties involved in the agreement.

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:** These stakeholders are interested in recruiting procedures that are fair and efficient. From that, they can make statistical decisions in hiring process.
* **Universities and Educational Institutions:** They are seeking to improve the employability and salaries’ range of their graduates. Furthermore, they’ll be able to orientate their students to their suitable careers based on their traits and academic performances.
* **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

As in the project description, the dataset for this project presents information on engineering students, including their personal and academic performances, as well as their career outcomes after graduation. Therefore, we will be able to identify historical students’ personality scores and academical scores as well as their specialization and college information to distinguish their difference in the starting salaries. This data is fundamental for building a predictive model that evaluate the future engineering graduates’ beginning earnings.

**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:** The amount of data is noticeably small (only about 600 records). This can be a major drawback since training a complicated model to increase the prediction score requires a large amount of data. We may need to consider overfitting issues due to the limited amount of data.
* **Outdated Information:** The dataset may require frequent revisions to remain current as the job market and educational requirements change rapidly nowadays.
* **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:** Predictions maybe biased toward more represented areas because not all engineering disciplines may be equally covered by the dataset.
* **Missing and duplicated data:** Luckily, there is no record of missing value of data or any duplicate in this dataset, only a few ‘0’ value which can be replace by back-fill method with no significant impact to the performance of the model. Nevertheless, in the real-world data, this problem will happen spontaneously and require imputation techniques that may further introduce additional biases.

**Variables and Features Significance**

1. **De-identification data:** Determine the identity information of the candidate

- ID: A unique ID to identify a candidate

- Gender: Candidate’s gender

- DOB: Date of birth of the candidate

- CollegeGPA: Aggregate GPA at graduation

2. **Academic Performance:** Reflects a student’s academic background, which may be related to their knowledge, skills and earning potential.

- 10percentage: Overall marks obtained in Year 10

- 12graduation: Year of high school graduation (Year 12)

- 12percentage: Overall marks obtained in Year 12

3. **University Characteristics:** Show the college characteristics that may have an impact on salary outcomes by indicating the quality of education and networking possibility.

- CollegeID: Unique ID identifying the university/college which the candidate attended for her/his undergraduate

- CollegeTier: Each college has been annotated as 1 or 2. The annotations have been computed from the average scores obtained by the students in the college/university. Colleges with an average score above a threshold are tagged as 1 and others as 2.

- CollegeCityID: A unique ID to identify the city in which the college is located in.

- CollegeCityTier: The tier of the city in which the college is located in. This is annotated based on the population of the cities.

4. **Field of Study data:** This data is essential for comprehending the differences in pay between engineering specialties.

- Degree: Degree obtained/pursued by the candidate

- Specialization: Specialization pursued by the candidate

5. **Personal Traits:** 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.

- conscientiousness: Scores in one of the sections of personality test

- agreeableness: Scores in one of the sections of personality test

- extraversion: Scores in one of the sections of personality test

- nueroticism: Scores in one of the sections of personality test

- openess\_to\_experience: Scores in one of the sections of personality test

6**. Scores:** Provide certain information and skill sets that may be correlated with pay and work performance.

- English: Scores in English section

- Logical: Score in Logical ability section

- Quant: Score in Quantitative ability section

- Domain: Scores in domain module

- ComputerProgramming: Score in Computer programming section

- ElectronicsAndSemicon: Score in Electronics & Semiconductor Engineering section

- ComputerScience: Score in Computer Science section

- MechanicalEngg: Score in Mechanical Engineering section

- ElectricalEngg: Score in Electrical Engineering section

- TelecomEngg: Score in Telecommunication Engineering section

- CivilEngg: Score in Civil Engineering section

7**. 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**

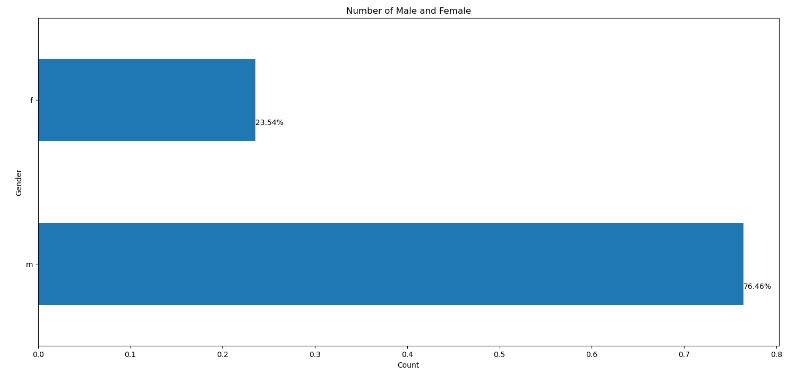
The dataset contains 599 rows of records with 30 columns of features and 1 column is the target variable ‘salary’. Thankfully, this dataset doesn't include any null or duplicate values but only some Zero value which can be fill by back fill method that would affect the performance of the models, so we 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 76.46% of engineering students being male and only 23.54% 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:

A screenshot of a computer code

Description automatically generated

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' 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 ‘-1’ 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 $170,000,000, as this bar graph illustrates. The Master of Computer Application, or MCA, comes next with only about $10,000,000. 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)

A graph with numbers and lines

Description automatically generated

Between specializations:

A graph of various colored bars

Description automatically generated with medium confidence

The salaries vary across 27 different engineering specializations, ranging from $215,000 to $38,290,000. 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.

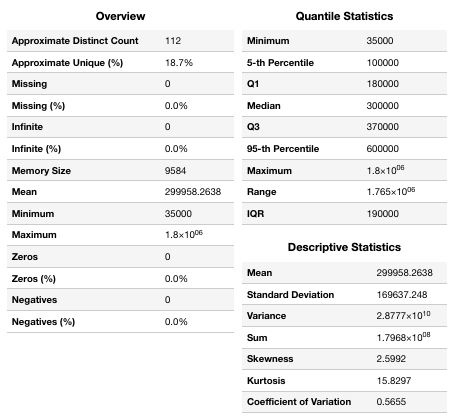
**Numerical Data Visualization**

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’:

A graph of a salary

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.

short dash

# Modeling

* Describe the machine learning algorithms used for modeling.
* Discuss the rationale behind selecting these algorithms.
* Explain the parameter tuning and model selection process.

Instructions: Describe the machine learning algorithms used for modeling, providing a rationale for their selection based on the project goals. Explain the process of parameter tuning and model selection. Include details about the algorithms' implementation and any considerations made during the modeling phase.

short dash

# Evaluation

## Results and Analysis

* Present the results of the model evaluation, including accuracy, precision, recall, F1-score, etc.
* Analyze and compare the performance of each model.
* Discuss the key insights gained during the experimentation phases.

Instructions: Present the results of the model evaluation, including accuracy, precision, recall, F1-score, or any other relevant metrics. Analyze and compare the performance of each model, highlighting the key insights gained during the experimentation phases. Discuss the implications of these insights on the project's goals and potential areas for further improvement.

## 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.

## 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.

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