

Project:

# Engineering Students' Salary Prediction

---

By Ratana Sovann

28/03/2024

36106 - Machine Learning Algorithms and Applications  
Master of Data Science and Innovation  
University of Technology of Sydney

## Table of Contents

<b>1. Executive Summary</b>	<b>2</b>
<b>2. Business Understanding</b>	<b>3</b>
a. Business Use Cases	3
<b>3. Data Understanding</b>	<b>4</b>
<b>4. Data Preparation</b>	<b>5</b>
<b>5. Modeling</b>	<b>6</b>
a. Approach 1	6
b. Approach 2	6
c. Approach 3	6
<b>6. Evaluation</b>	<b>8</b>
a. Evaluation Metrics	8
b. Results and Analysis	8
c. Business Impact and Benefits	8
d. Data Privacy and Ethical Concerns	9
<b>7. Deployment</b>	<b>10</b>
<b>8. Conclusion</b>	<b>11</b>
<b>9. References</b>	<b>12</b>

## 1. Executive Summary

Engineering students often face uncertainty regarding their career path and salary prospects after graduation. Moreover, the labor market for engineering professionals is highly dynamic as technological advancement continues to progress. Educational institutions and government agencies invest significant resources in engineering education and training programs. By understanding factors that contribute to higher salaries, institutions can tailor to meet industry demand and optimize their resources.

Traditional regression studies are often conducted within a specific timeframe where exogenous variables are controlled for, with a limited dataset. These studies that guided past decision-makers are met with temporal limitations. In the modern world, the labor market continuously evolves, making the results and insights of past studies less relevant. As such, a new method is needed to navigate the dynamic landscape of the modern era. The proposed machine learning models can use real-time data to help guide decision-making, surpassing the temporal limitations of the past. We will compare the results from a linear model with a more complex model such as Random Forest, analyzing closely the models' feature importance and their implications on the issue.



## 2. Business Understanding

### a. Business Use Cases

- The result of this experiment could help aspiring students make the most informed choice when deciding the trajectory of their careers.
- Educational institutions and career counselors can also gain insight into the current trends and market needs in the industry, enabling them to design courses that cater to those needs.

### b. Key Objectives

- A quantifiable metric to measure and interpret the impact of model features importance.
- A visual to help illustrate the impact of selected features on salary.
- **Technical Objective:** A model Root-Mean-Squared that exceeds the baseline model by 30%.

Addressing Stakeholder's requirement with Machine Learning:

- **Engineering Students:** often seek information about the earning potential associated with different engineering disciplines to make an informed decision about their academic and professional pursuits. Machine learning models can be trained to predict salary expectations with input features such as academic performance and specializations.
- **Educational Institutions:** aims to provide relevant and market-driven education programs that require insights into salary trends and industry demand to tailor their curriculum and allocate resources effectively. By leveraging the continuous improvement aspect of machine learning models, stakeholders can gain insight into continually refined factors influencing salary expectations.



### 3. Data Understanding

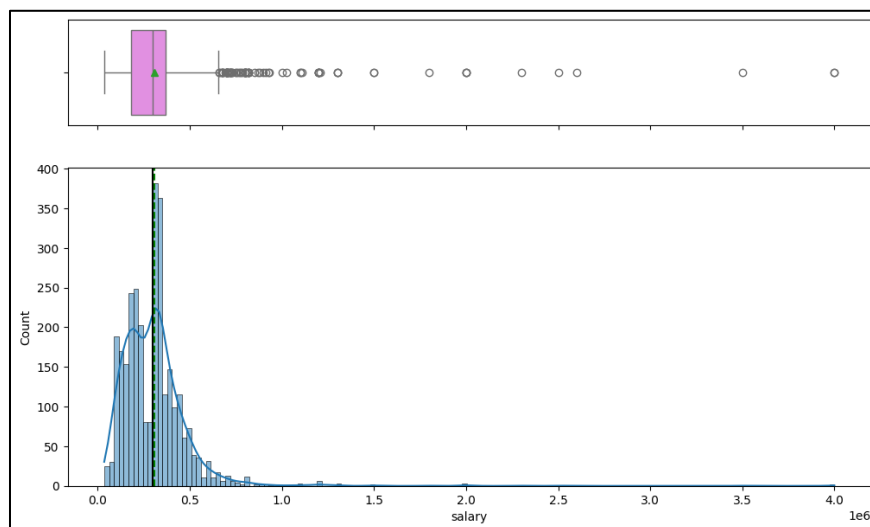
This dataset is provided by the stakeholder with limited information regarding its source and how it is collected.

#### a. Data Dictionary

- **ID**: A unique ID to identify a candidate
- **Salary**: Annual salary offered to the candidate
- **Gender**: Candidate's gender
- **DOB**: Date of birth of the candidate
- **10percentage**: Overall marks obtained in Year 10
- **12graduation**: Year of high school graduation (Year 12)
- **12percentage**: Overall marks obtained in Year 12
- **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.
- **Degree**: Degree obtained/pursued by the candidate
- **Specialization**: Specialization pursued by the candidate
- **CollegeGPA**: Aggregate GPA at graduation
- **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.
- **GraduationYear**: Year of graduation (Bachelor's degree)
- **English**: Scores in English section
- **Logical**: Score in Logical ability section

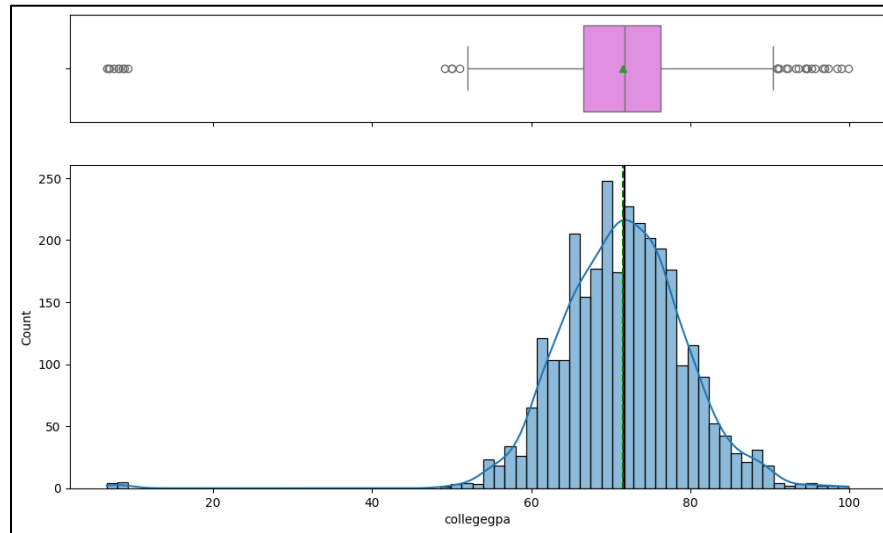
- **Quant:** Score in Quantitative ability section
- **Domain:** Scores in domain module
- **ComputerProgramming:** Score in Computer programming section
- **ElectronicsAndSemicon:** Score in Electronics & Semiconductor Engineering section
- **ComputerScience:** Score in Computer Science section
- **MechanicalEngg:** Score in Mechanical Engineering section
- **ElectricalEngg:** Score in Electrical Engineering section
- **TelecomEngg:** Score in Telecommunication Engineering section
- **CivilEngg:** Score in Civil Engineering section
- **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

## b. Explorative Data Analysis



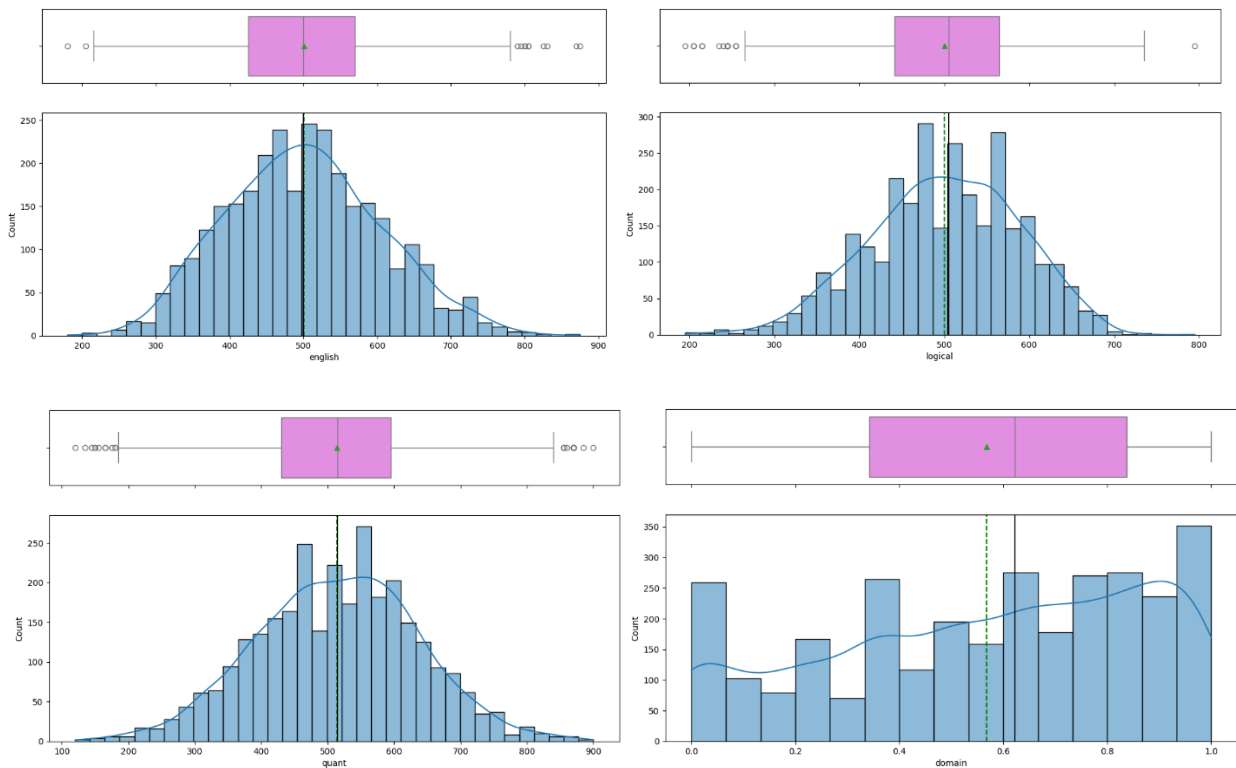
**Fig 2.1. Salary Distribution**

Fig 2.1 shows that salary is skewed to the right with abnormal earnings reported in the data upward of \$ 4,000,000 which is highly unlikely.



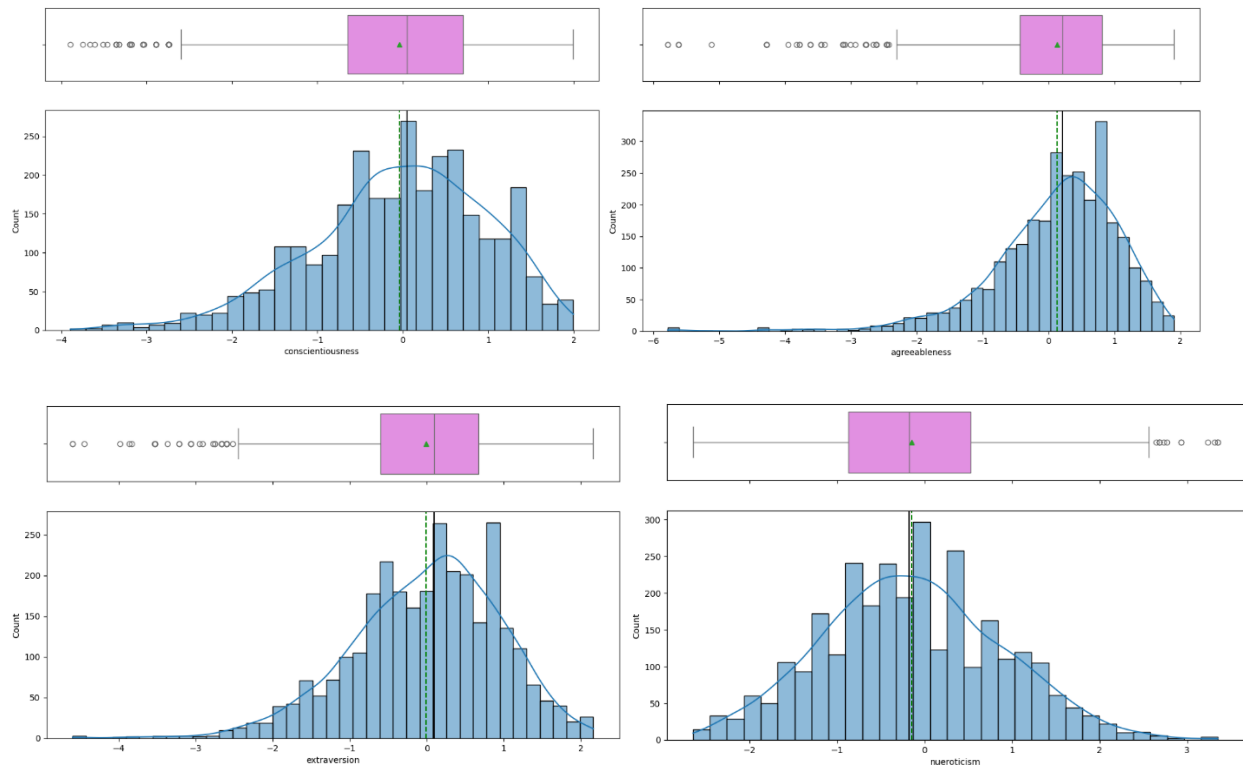
**Fig 2.2. College GPA distribution**

Fig 2.2 College GPA is skewed to the left with many outliers close to 10 reported GPA (100 scale).



**Fig 2.3. English, Logical, Quantitative, and Domain distribution**

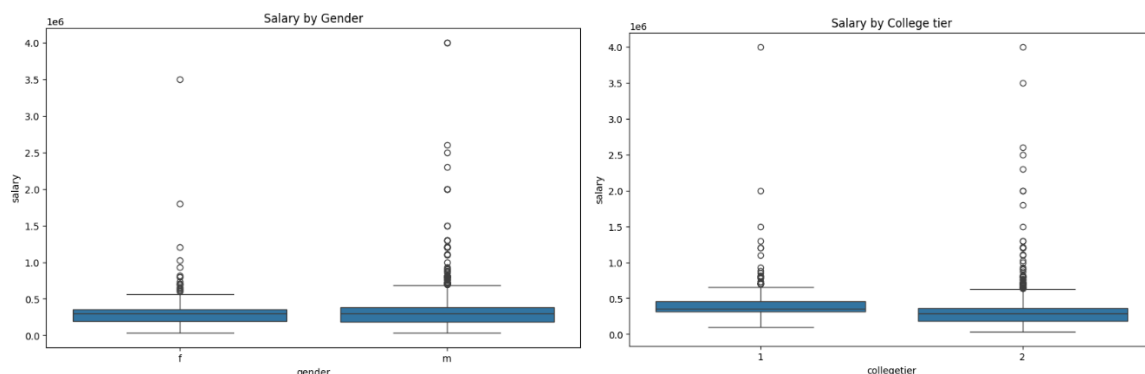
Fig 2.3 indicates a normal distribution for scores in English, logical, and quantitative abilities. The domain module score is not normally distributed, which is within expectation as different cohorts and specialization tracks can have inherently different scoring distributions.



**Fig 2.4. Conscientiousness, Agreeableness, Extraversion, Neuroticism distributions**

From Fig 2.4, the personality traits conscientiousness, agreeableness, and extraversion are left skewed while neuroticism is right skewed.

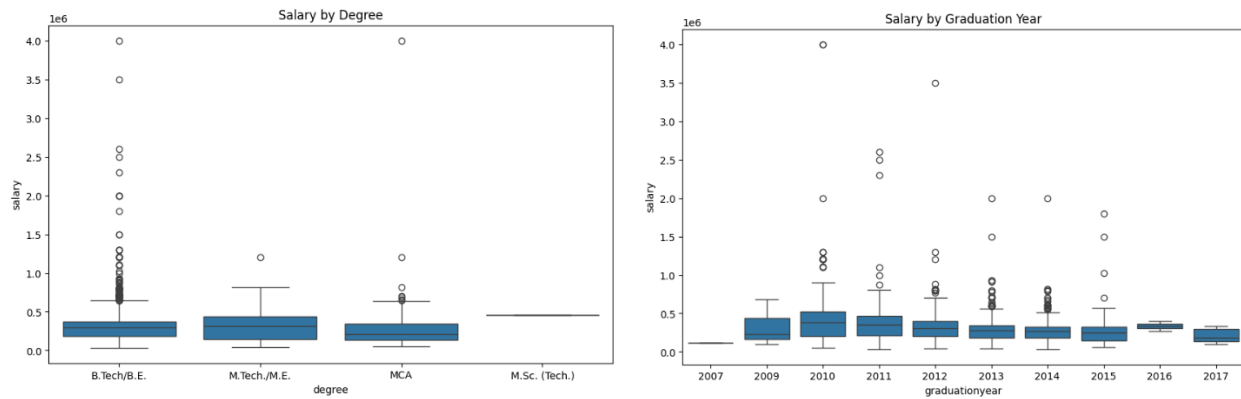
### Bivariate analysis:



**Fig 2.5 Salary by Gender & College Tier**

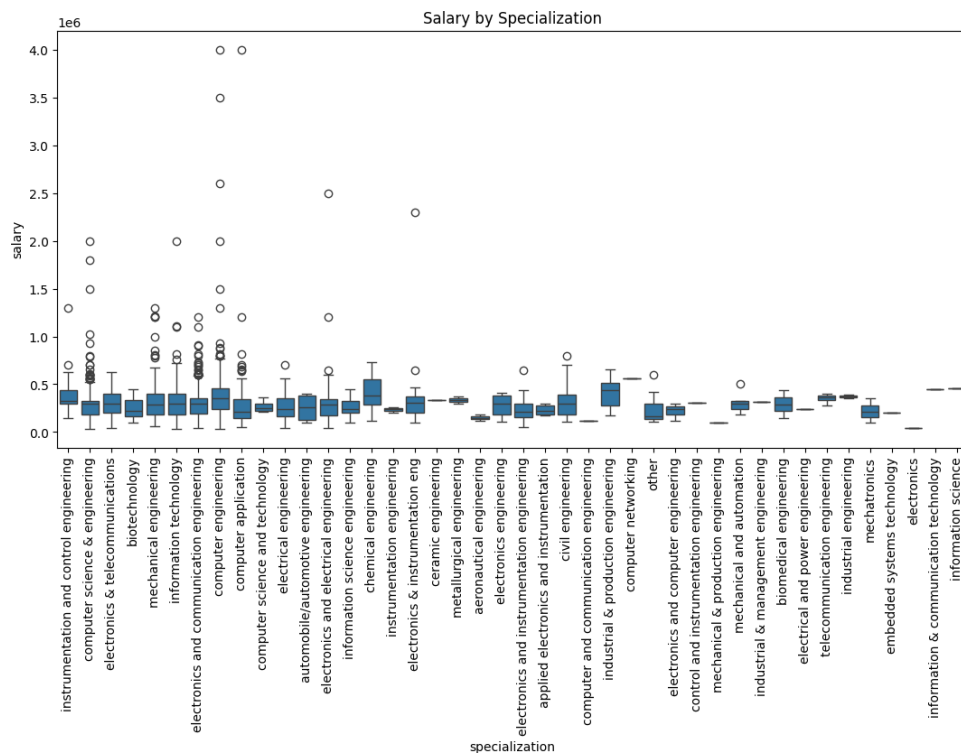


Fig 2.5 indicates similar variance and distribution between gender and college tier categories. This means that these categorical variables may not be useful in predicting salary.




**Fig 2.6 Salary by degree & graduation year**

Similarly, Fig 2.6, shows a similar variance between salary as separated by degree and graduation year, which indicates that they are not useful features to incorporate into the model.



**Fig 2.7. Salary by specialization**

Fig 2.7 showcases a clear variation between specialization and salary outcome. This feature will help with the model's prediction.



The chosen features are:

**Numerical:** 10percentage, 12percentage, collegegpa, English, logical, quant, domain, conscientiousness, agreeableness, extraversion, neuroticism, openness\_to\_experience

**Categorical:** specialization

10percentage, 12percentage, and College GPA are measures of academic performance and are well linked to higher wages (Zou, Zhang & Zhou 2022). Similarly, English, Logical, and Quantitative scores are all in some sense related to academic performance. Personal traits such as Conscientiousness, Agreeableness, Extraversion, Neuroticism, and Openness to experience have also been shown to impact salary (Vella 2024).

Specialization is chosen to discern fields within the engineering discipline that pay more as different branches of engineering command different salary ranges. For instance, aerospace engineers might have different salary ceilings compared to civil engineers or software engineers.

Features not included:

- Individual subject score e.g *civilengg*
- Gender
- Degree
- Year 12 and College Graduation year
- ID and College ID
- College tier
- College city ID & College city tier (see ethical consideration section below)



## 4. Data Preparation

The data preparation stage slightly differs for Linear Models and KNN Models.

For Linear Models:

- Outliers (salaries above 400k) are removed.
- Specializations with a count below 50 are grouped into 'others' to prevent overfitting.
- Specialization is one hot encoded into dummy variables and concatenated into the dataset.
- All features are standardized and scaled before modeling

For the KNN Model:

- **Hyperparameters tuned:** distance = 'Manhattan' and n\_neighbor = 43

For Random Forest Models:

- Outliers are removed the same as above
- **Hyperparameters tuned:** Using GridCV the optimal hyperparameters are: *max\_depth = 5, max\_features = 'sqrt', min\_samples\_leaf = 2, min\_sample\_split = 10, n\_estimators = 50.*

■ ■ ■

## 5. Modeling

Three models were trained:

- **Linear Regression:** is fundamental and interpretable and serves as a good starting model to capture linear relations between salaries and chosen features. No hyperparameter tuning is required.
- **KNN:** The KNN model utilizes feature space by calculating the distance between the feature vectors of the students and finding its 'nearest' neighbor for prediction, allowing it to capture non-linear relationships and handle complex interactions between predictors. *(Hyperparameter:  $n\_neighbors=43$ ,  $metric='manhattan'$ ).*
- **Random Forest:** utilize the ensemble method and introduce more complex decision boundaries.

For the KNN model: Manhattan distance results in better performance than Euclidean. A range of numbers of neighbors was tested {2, 3, 4, 5, 6, 8, 9, 10, 21, 27, 33, 41, 43, 47}. Stopped at 43 due to the conclusion that it cannot outperform Linear, possibly because of the curse of high dimensionality.

■ ■ ■

## 6. Evaluation

### a. Results and Analysis

The metric used to assess model performance is RMSE (root mean squared error). Among all the models, Linear has slightly better technical performance than other models,

The Linear model's technical performance:

Performance Metric	Baseline	Training	Validation	Test
RMSE	87798.91	78154.08	77148.08	81777.44
% Difference to Baseline	0%	10.2%	12.13%	6.8%
% Difference to KNN	0	1%	1.9%	1.12%
% Difference to Random Forest	0	-5.93%	2%	0.05%

In this case, the baseline model is just the average salaries. The Linear model was not able to reach the target objective of 30% over this baseline.

Nonetheless, a linear regression model can extract insightful information about underlying factors that contribute to higher salaries through its coefficients. This will be discussed in detail in the next section.

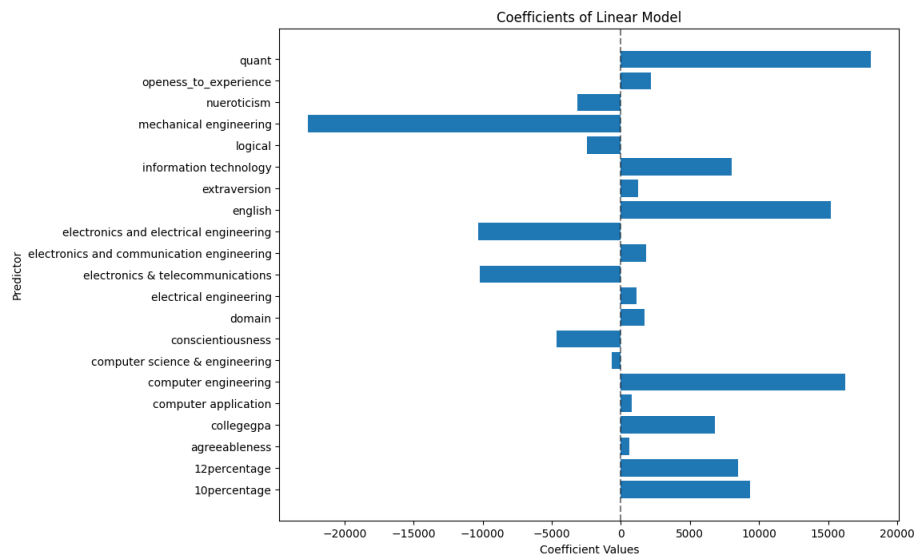
As for the KNN model, the high dimensionality of features may lead to data sparsity and make it challenging to find "nearest neighbors" since points are relatively far apart in high-dimensional spaces. In contrast, linear regression does not inherently suffer from this curse, as it aims to find linear relationships between variables rather than finding the nearest neighbors.

Random Forest achieved similar results to Linear regression. Despite many attempts at hyperparameter tuning using GridSearchCV to find the optimal parameter, overfitting persists.

This is likely due to the relatively small dataset (especially for specialization features when splitting).

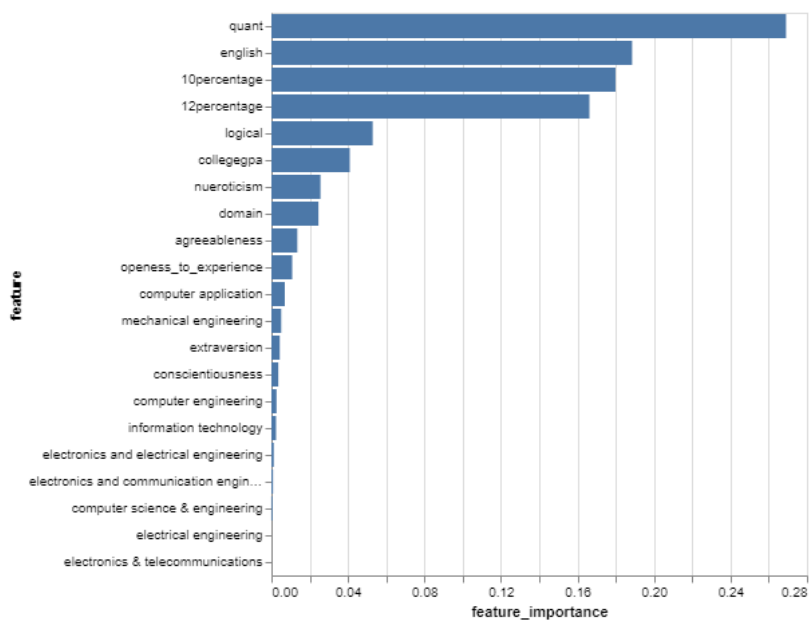
## b. Business Impact and Benefits

From the Linear model coefficient:



The coefficient on each feature provides a crude measure of the importance of each factor contributing to higher or lower expected salaries.

From Random Forest Feature Importance:



From the linear coefficients, computer science engineers are shown to earn more, and mechanical engineering earn less on average compared to other fields. However, Random Forest feature importance analysis reveals that this might not be the case as it does not contribute much to predicting salary outcomes, despite the large coefficient on the linear model. In other words, it may not be statistically significant in the model.

Academic scores such as quants, English, year 10, year 12 grades, and other metrics of academic assessment seem to be the most important factors when predicting salary. Hence, educational institutions such as universities should tailor their courses to boost these skills that are currently in demand. Aspiring engineers can also use this information to consciously build their skills in these areas (Quants and English) with more confidence that it would lead to potential higher salaries.



### c. Data Privacy and Ethical Concerns

As the Data is provided by the stakeholder, i.e. Universities and Educational institutions it is crucial to consider:

- **Data Privacy:** The dataset likely contains personal information about engineering students, including their academic performance, demographic details, and potentially sensitive data such as ethnicity or socioeconomic status. Hence, Data anonymization techniques are employed to protect sensitive information.
- **Ethical considerations:** researchers should avoid using this predictive model to perpetuate existing bias among Indigenous people or for any discriminatory purpose. The features CollegeCity and CollegeTier could pertain to a larger proportion of indigenous minorities purely based on their geographical location. Including these features in the model may inadvertently result in predictions that lead to disadvantages for the indigenous community. Hence, they are excluded from the model
- **Data Securities:** universities should ensure and enforce data-sharing agreements and policies.



## 7. Conclusion

As it is unable to meet the benchmark of 30% over the baseline model, it should not be deployed to predict salaries yet. It should be used in conjunction with more data hypothesized to correlate with income such as economic variables, family size, etc to improve its prediction.

However, the model offers the ability to use real-time data to guide the decision-making process. Educational institutions and career counselors can leverage the model to provide personalized guidance to engineering students regarding potential career paths and expected salary trajectories. By iteratively updating models and incorporating new insights, stakeholders can adapt to changing market dynamics and ensure an optimal allocation of resources.





## 8. References

- Mackay, S 2018, Do female engineering graduates earn more than their male counterparts? - Engineering Institute of Technology: Engineering Institute of Technology, [www.eit.edu.au](http://www.eit.edu.au).
- Vella, M 2024, 'The relationship between the Big Five personality traits and earnings: Evidence from a meta-analysis', Bulletin of Economic Research.
- Zou, T, Zhang, Y & Zhou, B 2022, 'Does GPA matter for university graduates' wages? New evidence revisited', in S Fu (ed.), PLOS ONE, vol. 17, no. 4, p. e0266981.

