# **EXPERIMENT REPORT 1**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Engineering Students’ Salary Prediction |
| **Date** | March 9th 2024 |
| **Deliverables** | 36106-AT1-25100660-experiment-1.ipynb  Multivariate Regression Model |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | Explain clearly what is the goal of this project for the business? How will the results be used? What will be the impact of accurate or incorrect results?  We are seeking insights from the information related to Engineering students across multiple colleges and their salaries in order to tailor curriculums to the market requirements and assist students in evaluating the Return on Investment of their educational expertise. On the other hands, the system also helps employers in negotiating fair compensations in the recruiting process. Accurate assessments can enhance job satisfaction and economic efficiency and in contrast, unreliable predictions might fool stakeholders and lead to wage inequities. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it.  The hypothesis that I’m bringing above is some attributes such as academic performance, college tier and certain personalities immensely affect the salary of engineering student. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.  The objective of the project is to evaluate multiple models to see which model is the most efficient when predicting the students’ salaries, staring from comparing with a naïve baseline model matrix and examine the RMSE score, aiming for the lowest one. Throughout the experiments, we seek to identify the most significant attributes in predicting the output and explore the feasible application of the model. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. Also list the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments  - 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)  - No encoded process needs to apply on the categorical data because of the lack of correlation between these features with the target out. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. Also list the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments  - 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.  - The numerical features were normalized using MinMaxScaler to make sure that they were on the same scale for the models. No encoded process needs to apply on the categorical data because of the lack of correlation between these features with the target out.  The project provided 3 different datasets:   * Training Dataset: 2998 records * Validation Dataset: 599 records * Testing Dataset: 599 records |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. Also list the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments  The baseline model metrics were created by the mean of the output salaries and calculating the root mean squared error of the difference between sum of the mean and real target values.  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  For the first experiment, a multivariate linear regression model was chosen as requirement and to the best of my knowledge, it’s because of its simplicity and interpretability. The model was first trained and validated using K-fold cross-validation to ensure its generalization through the data and then apply the model to the data for the performances which in this case is the RMSE score. No hyperparameter tuning was needed for this model. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyze in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.  The performance of the model after training is an RMSE of 125320.506 on the test set, which is noticeably improvement over the baseline model. Nevertheless, certain features may not have been as impactful as anticipated and will require further analysis. |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)  The improvement over the baseline model was drastically, however the baseline model was construct out of naïve metrics which show that the model still needs further refinement to become a practical tool for stakeholders. On the other hand, the predictive model has demonstrated potential for evaluating students’ salaries and show promising results if we are continuing on performing experiments on other models. Thus, we can bring valuable information for students, career counsellor and recruitment strategies. |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Also highlight the issues that may have to be dealt with in future experiments.  The amount of data is about 2998 records which is incredibly low and can be extremely difficult to enhance the complexity of this and further models without encountering overfitting. The college tier is negatively correlate to the target output, this requires human touch on modifying the dataset to set up the right algorithms in predicting salaries. Further investigation into feature interactions or non-linear models may necessary. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it, is a dead end.  Insights gained include the potential of the predictive model for evaluating students' salaries and the necessity for further refinement and investigation into feature interactions or non-linear models.  RMSE value of baseline model performance on test set: 169575.84318922673  RMSE value of my multivariate linear regression model performance on testing set: 159874.1581536953 |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.  - The given dataset is still too small for the model and experimenting with larger datasets might enhance model complexity without encountering overfitting. Nevertheless, this suggestion is in contrast in with the requirement of the project itself.  - One of the steps to improve the performance of the model is further refining the dataset, particularly with college tier attribute which I have apply above and received some slightly enhancement. Follow in the next experiments, we should explore the feature interactions or non-linear models. |