# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Engineering Students’ Salary Prediction |
| **Date** | March 9th 2024 |
| **Deliverables** | 36106-AT1-25100660-experiment-2.ipynb  Ridge Linear 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 the goal of this project for the business is. 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. And furthermore, the Ridge Model will apply the regularization terms to the cost function to control the trade-off between 2 goals: fitting the data well to increase the model and keeping the parameters small to avoid overfitting so we can see some fundamental improvement in the performance. |
| **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.  In this experiment, we will evaluate the Ridge Regression model on the set of attributes that have been thoroughly picked up from the previous experiments. As I mention in the last report the objective of the project is to compare among multiple models to evaluate the model that best fit for predicting the students’ salaries. Thus, for this experiment, we should focus on tuning the best possible hyperparameters for the Ridge Regression Model and figure out any refinement for the dataset that we are having now. |

|  |  |
| --- | --- |
| 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. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments  The criteria of the project instruct to running multiple models on the same chosen set of features, so the data preparation step is same as the previous experiment. The data shows up with no null value and no duplicate which can help us in the cleaning process and moving toward the numerical normalization and categorical encoded process. From the date of birth data (‘dob’), I created a column ‘Age’ to see the diversity and differences of salaries 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. |
| **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  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.  Attributes with less impact on salary predictions were excluded from the process to simplify the model and avoid overfitting.  Then, by using train\_test\_split from sklearn library, I have split the dataset into 3 pair of set with the following proportion of population:  Training set (X\_train, y\_train): 75%  Validation set (X\_valid, y\_valid): 15%  Testing set (X\_test, y\_test): 15% |
| **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: 179329.24728108424  RMSE value of baseline model performance on validation set: 159782.64825701702  RMSE value of baseline model performance on test set: 127763.42679157657  For this experiment with the Ridge Regression model, we are obligated to generate a regression model with the regularization terms in the cost function to reduce overfitting and enhance the model performance at the same time. The penalty component of the Ridge Regression model is L2 Regularization which is the squared value of the weights. Thus, I have done GridsearchCV with the model to see which ‘alpha’ is the best for the Ridge Model (‘alpha’ or so-called lambda is the regularization parameter)  Best parameters: {'alpha': 5} |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse 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:  RMSE value of Ridge linear regression model performance on testing set: 120792.53792538412  The result shows drastically improvement in the performance of the model compared to the baseline model or to the multivariate logistic regression model in experiment 1. Other than the above attributes that have high correlation with the target, I have added 3 more features that are moderately high correlation score. |
| **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)  In my opinion, it’s still too soon to assess the application of the model for the business since we haven’t evaluated other regression models. However, this predictive model noticeably shows incredible potential for evaluating universities’ graduate’s salaries because of the improvement when we apply the regularization term to the regression model, even on the poor dataset with only 600 data points. Imagine what impact can it bring to the recruitment teams and the career counselling services when we can finalize a complex and refined model for prediction salaries. |
| **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 most concerned problem is still the data limitations, especially in this and the following experiments where we tend to apply significantly fine-tuning applications to increase the complexity of the models which can lead to overfitting with such a small quantity of information. |

|  |  |
| --- | --- |
| 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.  The experiment highlights the effectiveness of regularization in enhancing Ridge Model performance for salary prediction. The improvement in the root mean squared error has shown the importance of feature selection process and the potential impact of regularization techniques on the score. |
| **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 results show incredible improvement in the model performance which indicated that the application of L2 regularization is working, and the overfitting problem is being dealt with. The recommendation for the next experiment is to apply L1 Lasso Regression model to see how absolute value of the weights in the cost function affect the model and is it better than the squared values of the weights. This L1 regularization can help in feature selection by reducing the coefficients of less important features to zero. |