# **EXPERIMENT REPORT 3**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Engineering Students’ Salary Prediction |
| **Date** | March 9th 2024 |
| **Deliverables** | 36106-AT1-25100660-experiment-3.ipynb  Lasso 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. The penalty component of L1 Regularization is the absolute value of the weights that will reducing some of the coefficients to zero. In this way, it’ll eliminate less informative variables and simplifying the model. |
| **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 experiment aims to evaluate the performance of Lasso Regression in terms of feature reduction and prediction accuracy enhancement. The goal is to derive a model that not only predicts salaries with lower RMSE but also sheds light on the most influential determinants of salary, providing clear insights for stakeholders. |

|  |  |
| --- | --- |
| 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  The Lasso Regression Model apply L1 Regularization term to the model which as I mentioned, is the absolute value of the weights. Thus, I have done GridsearchCV with the model to see which ‘alpha’ is the best for the Lasso Model (‘alpha’ or so-called lambda is the regularization parameter)  Best parameters: {'alpha': 1} |

|  |  |
| --- | --- |
| 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:  RMSE value of Lasso linear regression model performance on testing set: 159491.14258425796  Lasso Regression shows a promising reduction in RMSE from training to testing, indicating effective feature selection and predictive performance improvement over baseline models. |
| **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 experiment's findings suggest that Lasso Regression can provide a streamlined and interpretable model that is valuable for informing educational programming and fair recruitment practices. Its ability to identify key salary determinants supports targeted interventions and strategic decision-making. However, the performance is kind of similar to the Ridge Regression model which hasn’t proved the improvement of the model. |
| **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.  Challenges encountered during the experiment included addressing data limitations and ensuring the model's robustness across diverse datasets. The fine-tuning process revealed the delicate balance required between model simplicity and retaining predictive power. |

|  |  |
| --- | --- |
| 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.  This experiment illuminated the crucial role of feature selection in enhancing model performance. Lasso Regression's success in simplifying the predictive model without a significant compromise on accuracy underscores its potential utility in salary prediction tasks. Even though the RMSE performance |
| **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.   * Additional Hyperparameter Optimization: To further maximize performance, keep adjusting the Lasso Regression's "alpha" value. * Model Comparison: Comparing Lasso Regression to other models (e.g., ElasticNet) to ascertain which method is optimal for salary prediction in terms of overall model accuracy and feature selection capabilities. * Comprehensive Feature Analysis: Analyzing the characteristics kept by Lasso Regression in further detail in order to determine how they affect pay projections and to provide stakeholders with useful information. |