# **EXPERIMENT REPORT 5**

|  |  |
| --- | --- |
| **Student Name** | Ngoc Quang Vinh Pham |
| **Project Name** | Engineering Students’ Salary Prediction |
| **Date** | March 9th 2024 |
| **Deliverables** | 36106-AT1-25100660-experiment-5.ipynb  k-Nearest Neighbour 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 addition of characteristics like academic standing and college level is thought to have a big impact on pay. The k-Nearest Neighbor (kNN) Regression Model is selected due to its ability to identify intricate, nonlinear correlations that linear models may overlook. |
| **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.  With an emphasis on the usefulness of using the 'k' closest data points in feature space to ascertain wage expectations, this experiment aims to evaluate the kNN Regression Model's capacity to correctly estimate engineering students' incomes. |

|  |  |
| --- | --- |
| 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 recordsa |
| **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 kNN Regression Model Uses a non-parametric method based on the closest 'k' instances, chosen for its ability to capture nonlinear relationships and compare linear model performance by similarity in the feature space. Parameters such as the number of neighbors (k) were optimized through GridsearchCV to find the model configuration that minimized RMSE. The optimal parameters were identified as:  Best parameters: {'leaf\_size': 1, 'metric': 'manhattan', 'n\_neighbors': 90} |

|  |  |
| --- | --- |
| 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 K-Nearest neighbors regression model performance on testing set: 154556.51795681656  This result represents a significant improvement in prediction accuracy, indicating the kNN model's strong performance and suggesting its potential as the preferred model for salary predictions within the scope of this project. |
| **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)  Because the kNN model predicts salaries more accurately, it has great potential for use in recruiting and instructional methods. It provides a data-driven foundation for curriculum creation and wage negotiations. Remarkable forecasts may have a significant impact on graduates' financial prospects and level of work satisfaction. |
| **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 included selecting an optimal 'k' value and determining the most effective metric for distance calculation, critical for the model's ability to generalize from training data to unseen instances. |

|  |  |
| --- | --- |
| 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 highlighted the strength of the kNN Regression Model in handling complex, nonlinear data relationships for salary predictions. This methodology proved superior in capturing the intricate dynamics influencing salary outcomes compared to simpler models. The improvement in RMSE shows that this is the model with the best score out of all 5 experiments and indicates a promising direction for further exploration and model refinement. |
| **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.  Future research need to take into account:  - Extending the Feature Set: Adding more factors might reveal more intricate connections that influence wage projections.  Examining Sophisticated kNN Methods: Model accuracy might be further improved by looking at weighted distance measurements or by changing the number of neighbors depending on local data density.  - Comparative Analysis: To assess the kNN model's performance and determine which advanced regression methodology is the most reliable for salary prediction, this method compares it to other models.  When deploying the kNN model, or a mix of models, into a production setting, more validation would be required to guarantee scalability and stability. Additionally, ongoing monitoring and updating would be necessary to accommodate new data and changing market circumstances. |