# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ronik Jayakumar |
| **Project Name** | Salary Predictor – KNN |
| **Date** | 29/03/2024 |
| **Deliverables** | Machine Learning Assessment 1  Experiment 5  KNN |

|  |  |
| --- | --- |
| 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?  The business's objective with this project is to create a prediction model that can reliably forecast earnings based on characteristics such as education, talents, and experience. The findings of this project will be utilized by the company, namely the HR department or recruiting managers, to make educated judgements on wage offers for job hopefuls or current workers. Accurate pay estimating findings can have a number of positive consequences on the business:   * Improved recruiting Process: Accurate pay estimations can assist to speed the recruiting process by making fair and competitive compensation offers to candidates. This can attract top people and increase the overall quality of hiring. * Employee satisfaction can be improved by offering competitive salaries based on reliable predictions. Employees that feel appropriately compensated are more likely to stay with an organisation in the long run. * Compliance and Fairness: Fair and transparent compensation policies are critical for ensuring regulatory compliance and a great employer brand. Accurate pay estimates guarantee that the organisation meets regulatory obligations and encourages equitable remuneration practices.   On the other hand, incorrect results from the salary estimation model can have negative consequences:   * Hiring Challenges: Inaccurate wage estimates might make it difficult to recruit and retain top workers. Candidates may reject offers that they believe are unfair or below market value, making it harder to fill important positions. * Employee Dissatisfaction: Offering salaries that are significantly lower than market rates can result in employee dissatisfaction and low morale * Budget Overruns: Overestimating salaries can lead to budget overruns and financial strain on the company.   All this will be implemented by using the Cross Industry Standard Process for Data Mining (CRISP-DM) process model for efficiently setting up the business. The 6 phases of CRISP-DM are:   * Business Understanding – What does the business need? * Data Understanding – What data do we have/need? Is it clean? * Data preparation – How do we organize our data for modelling? * Modelling – What modelling techniques should be apply? * Evaluation – Which model best meets the business objectives? * Deployment – How do stakeholders access the results? [4]   The CRISP-DM framework would give structure to our business objective to drive it in a systematic way based on customer needs. |
| **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 being tested is to check whether a KNN model can increase efficiency in predicting salary using a combination of specific features, including quantitative aptitude, logical reasoning, academic performance (10th and 12th percentages), English proficiency, college tier, programming skills, domain knowledge, age, and specialization.  The investigation into this hypothesis holds significance due to its potential on streamlining wage determination. The investigation holds great significance for the placement cell as it directly impacts the career trajectory and earning capacity of a graduate. With the implementation of such a model, the placement cell can guide and support students as they navigate the job market.  Furthermore, uncovering the drivers of salary can help a particular cell make more informed recommendations which would improve their placement rates. This in turn would help ensure positive relations between employers and students. |
| **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 expected outcome of the experiment is to develop a predictive model that can estimate salaries based on background characteristics of an individual.  Potential outcomes of this experiment include:   * A well performing model with good predictive abilities and generalization capabilities towards unseen data. * Feature identification that drives salary discrepancies. * Hyperparameter tuning steps * Optimal machine learning model identification * Insights on the relationship between the independent variables with the target variable.   An estimate of the expected goals would be as follows:   * A model which gives decent predictions for the company and the client to use as an estimate. * Increased hiring rates due to data driven decision making * Client and employee satisfaction   The possible scenarios resulting from this experiment are:   * The existence of a model that gives good accuracy scored that could be used as a base model to build on for future and more comprehensive projects.   A model that predicts salaries at an accuracy level lower than what we are expecting but can prove to be a model that can be used as a baseline guide on problems you may face. |

|  |  |
| --- | --- |
| 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 data has been provided in 3 csv files, namely – salary training, salary validation, salary testing with each of these files containing 31 columns or categories. The following are the steps taken to ensure the data is ready for analysis.   * The first process that was carried out was the conversion of the dob column which represents date of birth into an age column using the datetime package in python. * The second was exploring into duplication that may exist in the dataset. The provided validation and testing set had all their values available in the training set. Due to this they were disregarded, and a new train test split will be generated before modelling. For the rest of the data preparation, we will focus on the training set only. * A box plot and a histogram of the target variable was generated to visualize the skew and range of salaries that exist in the dataset. * The next order of data preparation was to explore numerical columns in the dataset. These columns are checked for correlation and the top 9 values except 12th graduation are made into a new pandas DataFrame. 12th graduation year was excluded as it does not align with our business needs. * The new DataFrame of numerical values are checked for outliers using box plots and dealt with using the Inter Quartile Range. Since the overall number of rows is small at 2998, we have decided to adjust the outliers rather than drop them. This keeps our overall training data at the same number while ensuring no outliers exist. The same has been verified using the box plots for a second time. * An ANOVA test is carried out on the categorical columns to find the p-values of each of them. We see that the field pertaining to specialization had a value very close to 0 which shows there is a strong correlation between the two. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments.   * Imputation has been carried out with the column ‘computerprogramming’. The existence of a total of 650 values corresponding to -1 meant these rows could not be dropped. Hence the mean value of the overall column has been replaced in every row with -1. * The chosen categorical column i.e. Specialization has been encoded using the Label Encoder in python. This was the choice of encoding methodology due to the large number of variables within the column which would make other encoding methods like one hot encoding a bit more complex to handle. * As mentioned earlier, the numerical columns were separated, and outliers were removed. The encoded specialization column is integrated into this DataFrame containing the rest of our independent variables. * Finally, scaling has been performed over the entire dataset to ensure all values exist under a single scale. This has been carried out using the Standard Scalar library in python.   Note: Scaling has also been done over the target column as high and difficult to interpret RMSE scores were being observed when not scaled.  The features that have been removed are listed below:   * ID * Gender * 12graduation * College ID * Degree * College GPA * College city ID * College city tier * Graduation year * Electronics and semicon * Computer science * Mechanical engineering * Electrical engineering * Telecom engineering * Civil engineering * Conscientiousness * Agreeableness * Extraversion * Openness to experience   Some of these features were dropped as they had low correlation with the target variable. Other features such as ID were dropped as it just an identification column.  Features that would be of future importance would be the per subject expertise that a student can portray. This would need further analysis and more data for models to learn In an efficient manner but it could be a definite next step in the quest for better and more accurate models. |
| **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. List also 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 K-Nearest Neighbors (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values.  Reason for choosing KNN:   * KNN is a basic, non-parametric technique that may be used for classification and regression. * It is based on the similarity of data points in the feature space, making it appropriate for instances where the underlying data distribution is complicated or poorly defined. * KNN is versatile and resilient since it makes no significant assumptions about the distribution of the underlying data. Given the nature of the salary prediction problem and the possibility of non-linear correlations between characteristics and income, KNN is an appropriate method for capturing complicated patterns in data.   Hyperparameters tuned:  The N Neighbors parameter has been tuned. The ideal number was found using the KFold methodology. The ideal K number in our case being 47.  Models Not Trained:   * Decision Trees or Random Forests: These models were not trained due to the possible complexity of the decision boundary and the risk of overfitting, particularly with insufficient data and a high-dimensional feature space. However, they might be used in future trials to compare the performance of KNN. * Support Vector Machines (SVM): SVM was not picked because of its computational complexity and dependence on kernel function selection. Given the simplicity and interpretability of KNN for this assignment, SVM was judged less appropriate.   Potentially Important Hyperparameter:   * Metric: The distance metric used (e.g., Euclidean, Manhattan) can have a major influence on KNN performance. While the default Euclidean distance is widely used, experimenting with various distance metrics may be beneficial in future studies, particularly if the feature space has unexpected properties. * Overall, KNN was chosen due to its simplicity, versatility, and ability to detect non-linear correlations in data. We want to identify the best configuration for properly forecasting wages by adjusting the number of neighbours and the weight function. |

|  |  |
| --- | --- |
| 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.  Baseline KNN RMSE scores (K=5)  Root Mean Squared Error (RMSE) on Training Set: 0.773  Root Mean Squared Error (RMSE) on Validation Set: 0.927  Main model (K=47)  Root Mean Squared Error (RMSE) on Training Set: 0.867  Root Mean Squared Error (RMSE) on Validation Set: 0.921  Root Mean Squared Error (RMSE) for Testing Set: 0.816   * The models training RMSE was calculated at 0.867. The model showed overfitting with the Validation set with an RMSE of 0.921 but it fared well with the testing set portraying an overall RMSE score of 0.816. * Underperforming situations/Observations: It is critical to investigate particular situations or observations in which the model's predictions differ considerably from real pay levels. These situations may indicate trends or outliers that the model misses, offering insights into the underlying reasons of model underperformance. * Overfitting: The primary model may be overfitted, as shown by a larger validation RMSE than training. This might be because the model memorises noise in the training data rather than learning the underlying patterns. * Insufficient Feature Engineering: The training features may fail to capture the underlying associations with salary, resulting in inferior model performance.   Next Steps:   * On exploration, it was found that **dropping the age** column greatly reduced overfitting and increased model performance. This indicates KNN requiring different parameters for this experiment. |
| **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 model has shown signs of potential overfitting which means more work needs to be put into its development. Avid testing of various ways on the models betterment would be the next step.  Estimated impacts of incorrect results:   * Underestimation of salary: This could lead to dissatisfaction and demotivation in young employees. This in turn could result in low job acceptance rates which would hamper overall efficiency. * Overestimation of salary: This could create unrealistic expectations in the minds of employees. When these standards are not met, the same dissatisfaction and demotivation sets in which could cause them to remain in our business’ pipeline and jobless for much longer than expected. * Misclassification of employee potential: If the model incorrectly classifies employees' potential based on expected salary, it may result in wasteful resource allocation and talent management. Overestimating or underestimating employees' talents may result in allocating them to tasks that do not effectively use their skills and knowledge, resulting in poor performance and productivity. * Misrepresentation of salaries due to underlying bias: There is a risk that the model may amplify underlying biases that exist in the dataset which needs to be carefully considered. |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.  The issues faced during the experiment were as follows:   * Duplication in the validation and testing datasets. The values existed in the training dataset which gave rise to erratic numbers. This was overcome by ensuring duplicates existed and by creating a new train-validation-test split. * Existence of outliers, specifically the -1 values of the various subjects. * High number of features exist which required a lot of feature scaling * Erratic RMSE scores – The scaling of the target value along with the rest of the dataset fixed this issue. * Finding the ideal value of K to be used over experimentation. * Overfitting of the model   Some issues which may arise in future more comprehensive experiments especially on a larger scale would be as follows:   * Limited data – The limited availability of data could prove to be a challenge with more comprehensive model training. * Privacy of sensitive information – Salary information is considered as sensitive private data, and its unauthorized disclosure could lead to an infringement in an individual’s privacy. Strict authorization standards need to be setup by an organization when dealing with this information.   Data Accuracy and integrity – The accuracy and integrity of the data used to train the model is crucial for its accuracy levels. Verification steps must be undertaken by the organization to ensure inconsistencies and errors are |

|  |  |
| --- | --- |
| 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.  Effectiveness of the Model: The experiment showed that the K Nearest Neighbors model did not work up to expectations due to overfitting of the model. The RMSE score of the testing data proved to be higher than the training data which shows signs of potential overfitting indicating work to be done on its betterment.  Room for Improvement: While the model performed well, there is space for improvement in terms of prediction errors, especially on the testing set. This suggests that there may be more variables impacting compensation levels that have not been captured by the existing collection of characteristics, or that the model's complexity requires modification.  Feature Engineering Opportunities: Exploring new characteristics or improving current ones might boost the model's prediction performance. For example, including industry-specific factors, job titles, years of experience, or performance indicators might offer a more complete knowledge of salary drivers. Model Selection Considerations: While KNN was a good starting point, experimenting with other machine learning algorithms like decision trees, random forests, or gradient boosting could provide alternative modelling approaches that better capture nonlinear relationships and interactions between variables.  Evaluation metrics: In addition to RMSE, investigating alternate assessment measures such as mean absolute error (MAE), R-squared (R2), or quantile loss may offer a more complete knowledge of the model's performance and ability to capture various elements of prediction accuracy.  Data quality and pre-processing: Continuously monitoring data quality, fixing missing values and outliers, and ensuring adequate feature scaling are all critical steps towards improving model resilience and generalization capacity. |
| **4.b. Suggestions / Recommendations** | 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.  Effectiveness of the Model: The experiment showed that the K Nearest Neighbors would be a decent model but portrayed potential overfitting. This was identified due to the RMSE score of testing and validation being higher than the training set.  Room for improvement:  More time and effort to be spent on feature engineering. Causes of overfitting to be identified and worked on to improve model health and accuracy.  Feature Engineering Opportunities: Exploring new characteristics or improving current ones might boost the model's prediction performance. For example, including industry-specific factors, job titles, years of experience, or performance indicators might offer a more complete knowledge of salary drivers. Model Selection Considerations:  K Nearest Neighbors as of now will potentially give us subpar predictions which pushes for the identification of better models. Experimenting with other machine learning algorithms like decision trees, random forests, or gradient boosting could provide alternative modelling approaches that better capture nonlinear relationships and interactions between variables.  Evaluation metrics: In addition to RMSE, investigating alternate assessment measures such as mean absolute error (MAE), R-squared (R2), or quantile loss may offer a more complete knowledge of the model's performance and ability to capture various elements of prediction accuracy.  Data quality and pre-processing: Continuously monitoring data quality, fixing missing values and outliers, and ensuring adequate feature scaling are all critical steps towards improving model resilience and generalization capacity. |