Machine Learning Capstone Project: Investigating Automated Salary Prediction

Project Overview and Business Problem

This project had the goal of investigating a difficult business issue: how to possibly provide greater fairness and less subjective bias in new hire salary decisions at Company X. Apparently, the **Human Resources group** was seeking a more standardized approach to recommend salary ranges for comparable candidates. **Subjective judgments** in this process could, maybe, cause deviations or perceived inconsistencies in offers.

The idea was to attempt to construct a **machine learning model** that would, in principle, be able to automatically recommend an appropriate salary offer for a candidate, derived from **past data** of the profiles of other applicants. The hope was that this procedure could somewhat **standardize the selection procedure** and possibly **equalize it**, potentially lessening salary discrepancies among employees that appear to have comparable qualifications.

Data Used

For this project, we had a dataset with historical data about applicants to Company X. This information included several details such as **overall work experience**, **experience in the field** that they had applied for, **department**, **particular roles**, **industry background**, **qualifications**, **previous appraisal ratings**, **present compensation**, and other professional attributes. The most important piece of information that we were attempting to forecast was the `Expected_CTC` (**Cost to Company**), which appeared to be the salary that could possibly be offered.

Data Preprocessing and Preparation

The raw data required some preparation before we could construct the model. We did a few things to clean and convert the data into something that machine learning algorithms can typically function on:

- **Removing Identifiers:** `IDX` and `Applicant_ID` columns were removed since they're just individual identifiers and probably wouldn't be useful in predicting salary.
- **Missing Information Handling:** It seemed that numerous columns, particularly those for department, roles, education information, and university names, had some missing values. Where columns were text-based, these were completed with **'Unknown'**, treating the lack of data as a separate category, which could be useful. For numeric

fields such as `Passing_Year_Of_Graduation`, where a missing year may indicate the degree was not applicable or was not received, these were usually replaced with **'0'**.

- **Converting Categorical Variables:** Some of the fields, like 'Department', 'Role', 'Education', and 'Location', were initially in text form. As most machine learning algorithms like numerical input, we employed a method known as **one-hot encoding**. This method actually creates new binary (**0 or 1**) columns for every distinct category, which might assist the model in understanding these differences.
- **Binary Conversion:** The column **'Inhand_Offer'**, which showed **'Y' or 'N'**, was replaced with simple numerical **'1' or '0'** values for model compatibility.
- **Splitting Data:** Lastly, the prepared data was split into two primary parts: a **training set** (**80%** data) that the model would train on, and a **testing set** (**20%** data) that was used to test how well the model may do on unseen data during training. This division is considered to mostly aid in determining whether the model can generalize.

Model Building and Training

For the task of predicting `Expected_CTC`, a **Random Forest Regressor** model was chosen. This type of model is often considered suitable for prediction tasks because it can handle many input features (especially after one-hot encoding) and might be quite robust to unusual data points. It also seems capable of finding potentially complex relationships within the data. The model is really learning from patterns that have been seen within the training data, trying to see how various applicant attributes appeared to have lined up with previous salary offers.

Model Performance and Evaluation

After training the model, we tested its probable accuracy against the test data. The output looked very encouraging:

- **Mean Absolute Error (MAE):** Approximately
  11609.15 $$. This indicates that, on average, the model's predictions of salary could be approximately$$11,609.15$ off. Given that salaries can sometimes differ significantly, this appears to be a relatively small mistake.
- **Mean Squared Error (MSE):** Approximately
  . This measure places greater emphasis on larger errors, and its figure here appears to reflect overall good performance.
- **Root Mean Squared Error (RMSE):** Approximately
  22999.39
  . It is like MAE but maybe even more sensitive to individual prediction errors that are large, and it's in the same units as salary.
- **R-squared** (R2) **Score:** Around
  0.9996
  on an R-squared scale. An R-squared value close to 1 (here, 0.9996) generally indicates that the model explains a significant percentage of the variation in `Expected_CTC`. This

could mean that the model is highly efficient in predicting salary using the given applicant information.

Conclusion and Potential Impact

This machine learning project appears to have been able to build a model for predicting employee wages with seemingly high accuracy. Automating this, the model could help meet some of the project's overall objectives:

- **Potentially Reducing Human Judgment:** It provides a more objective, data-driven means of making salary recommendations, potentially reducing dependence upon subjective judgments.
- **Potentially Promoting Fairness and Reducing Discrimination:** By making suggestions based on measurable professional traits identified in the past record, the model hopes to promote more uniform and possibly fairer offers to similarly profiled candidates. This reproducible mechanism hopes to curtail any unintentional biases potentially introduced by strictly manual review.

The strong performance metrics might suggest that this model could be a useful tool for Company X, potentially helping to streamline their HR processes and contributing to more consistent compensation practices.