# 

# Salary

# Prediction by Aman Verma

# Business Problem

To ensure there is no discrimination between employees, it is imperative for the Human Resources department of Delta Ltd. to maintain a salary range for each employee with similar profiles

Apart from the existing salary, there is a considerable number of factors regarding an employee’s experience and other abilities to which they get evaluated in interviews. Given the data related to individuals who applied in Delta Ltd, models can be built that can automatically determine salary which should be offered if the prospective candidate is selected in the company. This model seeks to minimize human judgment with regard to salary to be offered.

**What did you wish to achieve while doing the project?**

**Goals and achievements:**

From the given business problem, we wish to achieve is to build a model that automatically determines the salary to be offered to prospective candidates applying to Delta Ltd. The goal is to minimize human judgment in the salary determination process and eliminate any discrimination in salary among similar employee profiles.

To achieve this objective, we plan to use historical data related to individuals who have applied to Delta Ltd. The model will take into account various factors such as an employee's experience, qualifications, and other abilities that are evaluated during the interview process. By using this data, the model will predict the appropriate salary to be offered to a prospective candidate if they are selected to join the company.

**The benefits of this approach include:**

* **Objective Salary Determination**: The model will provide an objective and data-driven approach to salary determination, reducing the chances of bias or discrimination.
* **Consistency**: The model will ensure consistency in salary offers for employees with similar profiles, creating a fair and transparent process.
* **Efficiency**: By automating the salary determination process, the model will save time and effort for the Human Resources department.
* **Reduced Human Judgment:** Minimizing human judgment in salary decisions can lead to more accurate and equitable salary offers.
* **Overall, the goal is to create a robust and unbiased model that will help Delta Ltd. offer salaries based on an individual's qualifications and experience, ensuring fairness and transparency in the hiring process.**

**2.**

**EDA - Uni-variate / Bi-variate / multi-variate analysis to understand relationship b/w variables. - Both visual and non-visual understanding of the data.**

**EDA plays a critical role in the data analysis process by providing valuable insights into the data, aiding in decision-making, and ensuring that the subsequent analysis and modelling steps are based on a solid understanding of the data's characteristics.**

**Uni-variate**

As we can see from the pic chard there are most of the count comes under other department, apart from this, most of the count belongs from Marketing and analytics Bi.

From the location graph, we can see that most people belong from Bangalore, and Surat have the least once.

From the Education level we can clearly conclude the people with higher education have higher count, which suggested that people with higher education have higher chance of getting placed as Post graduate people have higher count and higher change of being placed.

IT software and account have the least count along with Top Managements.

**Number of people with no of company worked with zero have lesser number of counts, also suggesting lower number of getting placed.**

**People with number of companies with 2 & 3 have higher count and indicating higher change of getting selected from the people having a greater number of companies.**

**From Graduation specializations, chemistry specialization has higher count and statistics have the least once.**

**From the give distribution plot, we can see the distribution of all the integer variables.**

**Bi-variate**

**s**

From the above plot, we can see the distribution of appraisals in the given department. Most of the appraisal ratings belong to the B category.

Key performers have the least count in all the departments, as they are highly valued for their skills, expertise, and contributions, and they often have a considerable impact on the entire organization.

People with key performance appraisals can be more suitable for the company, and we should give them some more preference, especially during hiring.

From the box plot on the left, it is evident that people with a Doctorate degree have higher average salaries amount 2,122,986k compared to under graduate 1,389,347k.

* Total Experience is the 3rd highest significant predictor.
* From the given scatter plot above, we can observe the distribution of salary vs. Experience.
* The above graph shows that salary increases with the number of years of experience in both the Current and Expected CTC.
* Individuals with no experience have a very low count, indicating fewer chances of getting hired.
* We should pay attention to candidates without experience as well, as it can help us find new talents, expertise, and promote diversity within the organization.
* From the right plot, we can clearly see a strong linear relationship between Current and Expected CTC, suggesting that the feature 'CTC' is highly important in predicting the target variable.

**Multi-variate-** **Education is one the important among the top five highly significant predictors and ranked at 4th.**

* From the above multivariate analysis, we can examine the relationship between certifications and education sectors.
* People with a Doctorate degree have a lower number of certifications but are still receiving higher packages, suggesting the importance of a Doctorate degree over certifications in determining salary.
* Individuals with below undergraduate degrees tend to rely more on certifications, as indicated by the plot, with a target across all certification categories.
* People with Postgraduate and Graduate degrees have fewer certifications compared to those with Undergraduate degrees.

**From the above correlation matrix, we can see some variable are gihhly correlated like Current CTC and expected CTC,**

**Curratnt CTC and Total experience in applied field is also highly correlated.**

**Certification have no impact on current and expected CTC. This can an importance factor during hiring.**

* **From the above grid plot, we can observe the distribution of Current CTC across multiple features such as Department and Industries with Experience.**
* **Most of the data points are represented in red and pink colors, suggesting that Analytics, IT, BFSI, and Training are the major industries based on the volume of data points.**
* **Top management positions are held by experienced individuals, as indicated by the absence of data points for individuals with no experience in the above graph.**
* **The Healthcare and Banking sectors have a higher number of freshers (individuals with no experience) compared to other departments**

**Data Cleaning and Pre-processing - Approach used for identifying and treating missing values and outlier treatment (and why) - Need for variable transformation (if any) - Variables removed or added and why (if any)**

**Data Description**

* **Visual inspection of data (rows, columns, descriptive details)**

The data is consisting of one Target variable name Expected CTC and other independent variables. The dataset contains 25,000 rows and 29 columns.

Each row represents an applicant, and the columns represent different attributes of the applicants, such as their ID, experience, department, role, education, current salary, etc. The data is consisting of 3 float & 10 integers & 16 object variables.

* **Understanding of attributes (variable info, renaming if required)**
* Based on the column names and data types, we can gain some understanding of the attributes present in the dataset.
* It can be observed that some columns have missing values (non-null count is less than 25,000), indicating that not all applicants have provided information for those attributes.
* The dataset has total 86853 missing values both NA and blank values.
* There is not duplicate value found in the dataset.
* However, there are some other anomalies as some feature has Duplicate variable like other & others which need to be treated before and EDA and it is replaced with their actual feature name other only.
* Rest other missing values can be treated after EDA.
* Max expected CTC is 5599570.0 & max current CTC is 3999693.0 & max experience is 25 years.
* There are total 86853 data point missing which are total of 16.12 % of total 538900 data point.
* These are some highly are some highly missing columns with their respective missing percentage in the data.

**Graduation Specialization 24.720**

**University Grad 24.720**

**Passing\_Year\_Of\_Graduation 24.720**

**PG\_Specialization 30.768**

**University\_PG 30.768**

**Passing\_Year\_Of\_PG 30.768**

**PHD Specialization 47.524**

**University PHD 47.524**

**Passing\_Year\_Of\_PHD 47.524**

* After running the ANOVA test and checking their significance level using a certain criterion of 0.05 p-values, we have decided to drop the following variables:
* Due to the high percentage of missing values in these columns (ranging from 24% to 47%), imputing them with their median or mode would alter a significant portion of the data, potentially introducing bias into the modelling process.
* Additionally, from a modelling perspective, we are creating a regressor model, and these variables do not fall under the integer type. Therefore, dropping these variables will simplify the modelling process.
* While these variables may be important from a domain perspective, they did not make it to the top five features with the highest significance for the target variable. Hence, we have decided to drop them from the model to improve simplicity and focus on the most significant predictors

**Treating missing values.**

* It is essential to handle outliers in certain features like certification and experience thoughtfully during data preprocessing and model development. Removing outliers helps maintain data integrity and can improve the accuracy and generalization of the model.
* Furthermore, there are some variables with missing values, such as certification, international degree, and total experience in the applied field. We have applied appropriate techniques to handle these missing values, ensuring that the data is properly imputed or processed to avoid bias and ensure a robust model.

|  |  |
| --- | --- |
|  | **Top contributing factors** |
| **1** | **Current salary of the job applicants.** |
| **2** | **Number Of Companies worked** |
| **3** | **Total Experience:** |
| **4** | **Education** |
| **5** | **In-hand Offer** |

* **Treating Skewness-**

By performing Z-score normalization on the data, we 'll now have a standardized distribution with a mean of 0 and a standard deviation of 1. This process will reduce the skewness and bring the data closer to a more symmetric distribution.

**Model building - Clear on why was a particular model(s) chosen. - Effort to improve model performance.**

**1.Linear regression**

**Model Metrics:**

**R2 Score (Train): 0.9701086069492951**

**Adjusted R2 Score (Train): 0.970079539698302**

**R2 Score (Test): 0.9689252457487594**

**Adjusted R2 Score (Test): 0.968854640185772**

**RMSE (Train): 0.00769421969536993**

**RMSE (Test): 0.007732960011962491**

**MSE (Train): 5.920101672061853e-05**

**MSE (Test): 5.979867054661094e-05**

**MAE (Train): 0.006304941983356228**

**MAE (Test): 0.006352030183357052**

**MAPE (Train): 0.7992817763847834**

**MAPE (Test): 0.805875326541253**

**Train Score: 0.9701086069492951**

**Test Score: 0.9689252457487594**

**Linear regression assumes a linear relationship between the independent variable(s) and the dependent variable. It seeks to model the relationship as a straight line that best fits the data points.**

**Analysis & Approach-**

Linear Regression model appears to perform well on both the training and test sets, with high R-squared values indicating a good fit to the data. The RMSE and MSE values are relatively low, suggesting that the model's predictions are close to the actual salary values. Additionally, the MAPE values indicate a reasonable level of accuracy in the model's predictions

## **2(i) Ridge regression or L2 regularization with different value of alpha**

**Alpha=1.0**

* The Ridge regression model with alpha=1.0 achieves a good fit to the data, but not as alpha (0.001), indicated by high R2 scores of 0.96875 (train) and 0.96734 (test).
* The MAPE values are relatively low, indicating a small average percentage difference between predicted and actual values.

Model Metrics:

R2 Score (Train): 0.9687591694343084

R2 Score (Test): 0.96734168955898

RMSE (Train): 0.007865979147907709

RMSE (Test): 0.007927545977735078

MSE (Train): 6.187362795531888e-05

MSE (Test): 6.284598522910361e-05

MAE (Train): 0.00641740057762926

MAE (Test): 0.006476917062034105

MAPE (Train): 0.008158396003493513

MAPE (Test): 0.008240425472763959

Train Score: 0.9687591694343084

Test Score: 0.96734168955898

**Alpha=0.5**

Model Metrics:

R2 Score (Train): 0.9687717684610588

R2 Score (Test): 0.9673657277584911

RMSE (Train): 0.007864392863648702

RMSE (Test): 0.00792462789948177

MSE (Train): 6.184867511380864e-05

MSE (Test): 6.279972734524486e-05

MAE (Train): 0.006421640787560747

MAE (Test): 0.006480447893758761

MAPE (Train): 0.008159035408168633

MAPE (Test): 0.008240319609266605

Train Score: 0.9687717684610588

Test Score: 0.9673657277584911

* The Ridge regression model with alpha=0.5 demonstrates a strong fit to the data, as indicated by high R2 scores of 0.968771 (train) and 0.96736 (test).
* The model's predictions are accurate, with low MSE and RMSE values, and the MAPE values suggest a relatively small average percentage difference between predicted and actual values.
* Overall, the Ridge regression model with alpha=0.5 performs well in terms of fit, accuracy, and minimizing prediction errors.

**Alpha=0.0001**

* **Ridge regression model with alpha=0.001 demonstrates good performance.**
* **It achieves high R2 scores on both the train and test sets, indicating that it can explain a significant portion of the variance in the target variable.**
* **The RMSE values are relatively low, suggesting accurate predictions with small errors. The MAPE values are also low, indicating good accuracy in predicting the target variable as a percentage of the true values.**

**Model Metrics:**

**R2 Score (Train): 0.9687778139992179**

**R2 Score (Test): 0.9673832726198388**

**RMSE (Train): 0.007863631584733856**

**RMSE (Test): 0.007922497390963752**

**MSE (Train): 6.18367017004239e-05**

**MSE (Test): 6.276596490982745e-05**

**MAE (Train): 0.006425909305997616**

**MAE (Test): 0.006484003398766132**

**MAPE (Train): 0.008159691288508483**

**MAPE (Test): 0.008240223878100845**

**Train Score: 0.9687778139992179**

**Test Score: 0.9673832726198388**

## **Creating a Decision tree model for regression-**

The Decision Tree algorithm offers a powerful and interpretable approach to building the salary prediction model. It complements the project's goals of fairness, transparency, and eliminating discrimination, making it a suitable choice for this specific application in HR data analysis for Delta Ltd

* The decision tree model demonstrates an exceptional fit to the data, as reflected by the near-perfect R2 scores of 0.9999 (train) and 0.9999 (test).
* The model achieves extremely low MSE and RMSE values, indicating minimal prediction errors.
* The MAPE values are negligible, suggesting highly accurate predictions with minimal average percentage difference.
* Overall, the decision tree model performs exceptionally well in terms of fit, accuracy, and minimizing prediction errors.
* Both the train and test scores are very close to perfection, indicating a robust and reliable model.

**Analysis & Approach**-The Decision Tree model appears to perform exceptionally well on both the training and test sets. The extremely high R-squared values near 1 indicate an almost perfect fit of the model to the data. However, the model's performance raises concerns about potential overfitting, especially considering the significantly lower RMSE, MSE, and MAPE values on the training set compared to the test set.

Model Metrics:

R2 Score (Train): 0.9999999999985139

Adjusted R2 Score (Train): 0.9999999999985124

R2 Score (Test): 0.9999989611293149

Adjusted R2 Score (Test): 0.9999989587688762

MSE (Train): 2.9433673080600167e-15

MSE (Test): 1.999149706526371e-09

RMSE (Train): 5.4252809218141105e-08

RMSE (Test): 4.4711851969319847e-05

MAE (Train): 3.1732223293744e-08

MAE (Test): 5.705779468759757e-07

MAPE (Train): 4.061431475911196e-08

MAPE (Test): 7.012044488414191e-07

Train Score: 0.9999999999985139

Test Score: 0.9999989611293149

## **Creating Random Forest model-** Random Forest, an ensemble of Decision Trees, yields a powerful and robust salary prediction model. It reduces overfitting and enhances accuracy by combining multiple trees' predictions, ensuring stable estimates for job applicants' salaries. While maintaining interpretability & providing reliable and data-driven salary estimations from the HR data analysis.

* The random forest model demonstrates an outstanding fit to the data, with near-perfect R2 scores of 0.9999 (train) and 0.9999 (test).
* The model achieves exceptionally low MSE and RMSE values, indicating minimal prediction errors.
* The MAPE values are negligible, suggesting highly accurate predictions with minimal average percentage difference.
* Overall, the random forest model performs exceptionally well in terms of fit, accuracy, and minimizing prediction errors. Both the train and test scores are very close to perfection, indicating a robust and reliable model.

Model Metrics:

R2 Score (Train): 0.9999999999990216

Adjusted R2 Score (Train): 0.9999999999990206

R2 Score (Test): 0.999999164403017

Adjusted R2 Score (Test): 0.9999991625044405

MSE (Train): 1.9378432234114236e-15

RMSE (Train): 4.402094073746521e-08

MSE (Test): 1.6079801724836307e-09

RMSE (Test): 4.00996280841061e-05

MAPE (Train): 4.045425197906077e-06

MAPE (Test): 6.242895890862345e-05

Train Score: 0.9999999999990216

Test Score: 0.999999164403017

**Analysis & Approach**

**The Random Forest Model Worked Well On Both The Training And Test Sets, With Very High R-squared Values Near 1, Indicating An Almost Perfect Fit Of The Model To The Data. The Rmse, Mse, And Mape Values Are Relatively Low, Suggesting That The Model's Predictions Are Close To The Actual Salary Values And Are Accurate.**

**This Model Seems To Be A Highly Effective Choice For Salary Prediction In The Given Hr Data. It Offers Robust Performance, Minimizing Bias And Providing Accurate Salary Estimates, Aligning Well With The Project's Objectives Of Fairness, Transparency, And Eliminating Discrimination**

## **Support Vector Machine**

* The SVM regressor model demonstrates relatively low R2 scores of 0.176686 (train) and 0.143824 (test), indicating a moderate fit to the data.
* The model exhibits higher MSE and RMSE values, suggesting larger prediction errors compared to other models.
* The MAPE values are also relatively higher, indicating a noticeable average percentage difference between predicted and actual values.
* Overall, the SVM regressor model performs less favourably compared to other models in terms of fit and accuracy.
* The train and test scores are consistent with the R2 scores, highlighting the model's limitations in capturing the underlying patterns in the data.

SVM Regressor Metrics:

R2 Score (Train): 0.17668695019336478

R2 Score (Test): 0.14382452327553585

MSE (Train): 0.0016306021450799076

RMSE (Train): 0.040380715014470804

MSE (Test): 0.0016475803750142196

RMSE (Test): 0.040590397571522004

MAPE (Train): 3.9857411721958176

MAPE (Test): 4.023291812301816

Train Score: 0.17668695019336478

Test Score: 0.14382452327553585

**KNN Regressor**

KNN Regressor Metrics:

R2 Score (Train): 0.845380970757328

R2 Score (Test): 0.7401109195829985

MSE (Train): 0.00030622874350465815

RMSE (Train): 0.017499392661022788

MSE (Test): 0.0005001172776096042

RMSE (Test): 0.02236330202831425

MAPE (Train): 1.4985338804967459

MAPE (Test): 1.8867092578856115

Train Score: 0.845380970757328

Test Score: 0.7401109195829985

* The KNN regressor model demonstrates relatively high R2 scores of 0.845380 (train) and 0.740110 (test), indicating a good fit to the data.
* The model exhibits lower MSE and RMSE values, suggesting smaller prediction errors compared to other models.
* The MAPE values are relatively lower, indicating a smaller average percentage difference between predicted and actual values.
* Overall, the KNN regressor model performs well in capturing the underlying patterns in the data and provides accurate predictions. The train and test scores are consistent with the R2 scores, indicating the model's ability to generalize well to unseen data.

# Model validation - How was the model validated ? Just accuracy, or anything else too ?

# Model Tuning Methods:

# Random Forest regressor with the tuned parameters Using grid search CV.

**Model Score After Tuning –**

**Analysis & Approach-RF model with tuning has proven to be a powerful and reliable approach for salary prediction in the HR data analysis for Delta Ltd. It effectively addresses overfitting concerns and provides trustworthy salary estimations for job applicants. The model can be deployed with confidence in the real-world hiring process, supporting the company's goal of fair and data-driven salary decisions while promoting diversity and inclusivity in the workplace**.

Comparison-**The Tuned Random Forest surpasses the regular Random Forest with better predictive accuracy, reduced overfitting, and improved generalization. Hyperparameter tuning fine-tuned the model, resulting in more precise and reliable salary predictions for Delta Ltd**

Random Forest Model Metrics:

R2 Score (Train): 0.9997673837942151

R2 Score (Test): 0.9997578970778151

MSE (Train): 4.6070505529118447e-07

RMSE (Train): 0.0006787525729536385

MSE (Test): 4.658905027874024e-07

RMSE (Test): 0.0006825617208629578

MAPE (Train): 0.05888759640625341

MAPE (Test): 0.05926151317787093

Train Score: 0.9997673837942151

Test Score: 0.9997578970778151

**Ensemble modelling, wherever applicable**

**Applying Grid Search CV on Random Forest and tuning the model.**

**'max\_depth': [7],**

**'max\_features': [8],**

**'min\_samples\_leaf': [5, 10],**

**'min\_samples\_split': [50, 100],**

**'n\_estimators': [100,50]**

**Best Parameters: {'max\_depth': 7, 'max\_features': 8, 'min\_samples\_leaf': 10, 'min\_samples\_split': 50, 'n\_estimators': 100}**

**Best Score: 0.9997771636850469**

**K-fold cross-validation**-is a valuable technique for accurately evaluating a model's performance and making more informed decisions during model selection and hyperparameter tuning.

## **K FOLD VALIDATION IN LINEAR REGRESSION**

**Average R2 score: 0.9695444564344722**

# K FOLD VALIDATION IN RANDOM FOREST

# Average R2 score: 0.9997676289661455

## **K FOLD VALIDATION IN DECESION TREE**

# Average R2 score: 0.9999996734362503

# Based on the provided Average R2 scores, let's compare the performance of K-Fold Validation in Linear Regression, Random Forest, and Decision Tree:

# 1.Linear Regression:

# - Average R2 score: 0.9695444564344722

# 2.Random Forest:

# - Average R2 score: 0.9997676289661455

# 3.Decision Tree:

# - Average R2 score: 0.9999996734362503

# Comparison

# - The R2 scores indicate the goodness of fit of the models. The closer the R2 score is to 1, the better the model is at explaining the variance in the target variable.

# - Linear Regression has an R2 score of 0.9695, which means it explains approximately 96.95% of the variance in the target variable. While this is a good performance, it may not capture complex relationships between variables.

# - Random Forest and Decision Tree models show significantly higher R2 scores, both approaching 1. Random Forest has an R2 score of 0.9998, and Decision Tree has an even higher R2 score of 0.9999996734362503. These scores suggest that both models can explain nearly 100% of the variance in the target variable.

# - The high R2 scores of Random Forest and Decision Tree indicate that they can capture complex relationships and patterns in the data, making them more suitable for modeling datasets with intricate structures.

# Overall, the choice of model depends on the specific requirements of the problem, interpretability, and generalization performance on unseen data. It's essential to use appropriate evaluation techniques, like K-Fold Cross-Validation, to obtain reliable estimates of model performance and make well-informed decisions.

# Final interpretation / recommendation - Very clear and crisp on what recommendations do you want to give to the management / client.

**Any other model tuning measures (if applicable)**

We can also use some other technique for model tuning like **Random Search**, which is similar to Grid Search which used all the possible combination randomly and selects a defined number of combinations.

**Regularization** – L2 or Ridge regularization is used in Linnear regression to control model complexity and overfitting.

**Final interpretation / recommendation - Very clear and crisp on what recommendations do you want to give to the management / client.**

**Interpretation of the most optimum model and its implication on the business**

**The most optimum model based on the given testing and model building approach that is the Random Forest model. It has demonstrated exceptional performance with high R2 0.9997578 scores and low 0.00068 RMSE values on both the training and test datasets. This indicates a strong fit to the data and accurate predictions.**

**The implication of using the Random Forest model in a business context, specifically in HR analytics, can be significant. Here are some key interpretations and implications:**

* **Predictive Power: The Random Forest model has proven to be highly accurate in predicting employee outcomes. It can be leveraged to forecast various HR-related metrics, such as Experience, Current\_CTC, No\_Of\_Companies\_worked, Total\_Experience, Education ,Inhand\_Offer. This enables organizations to proactively identify potential issues and make informed decisions.**
* **Feature Importance: Random Forest models provide insights into feature importance. By analysing the importance of different variables, HR professionals can identify the factors that have the most significant impact on employee outcomes. This knowledge can guide decision-making processes, such as prioritizing initiatives to improve employee selection process & satisfaction process with minimal discrepancy.**
* **Robustness and Generalization: Random Forest models are known for their robustness and ability to generalize well to unseen data. This is important in HR analytics as it ensures that the model's predictions can be relied upon in real-world scenarios. It can be used to make data-driven decisions not only based on the existing employee population & salary data but also for new hires or prospective candidates.**
* **Scalability: Random Forest models can handle large datasets and a high number of features efficiently. This makes them suitable for HR analytics applications, where the dataset may contain a wide range of variables related to employee demographics, performance metrics, training records, etc. The model's scalability allows for comprehensive analyses and insights.**
* **Interpretability: Although Random Forest models are considered a "black box" due to their ensemble nature, they still provide some level of interpretability. Feature importance, as mentioned earlier, helps in understanding the drivers of employee outcomes. Additionally, the model can offer insights into decision-making processes by identifying which variables have the most significant impact on the predictions.**

**Overall, the use of the Random Forest model in HR data Salary prediction can enhance decision-making processes, improve workforce planning, Save time in deciding what to offer whom leading to more effective selection procedure & people management strategies and ultimately contributing to the success of the business, and also can help overspending to offer Salary.**

**Insight and Recommendations**

**Insight**

* **Random Forest with hyper-tuning is considered the best model for salary prediction, providing improved accuracy and reduced overfitting.**
* **Current Salary of the job applicant is the most influential contributor in the dataset, indicating its significant impact on salary predictions.**
* **Candidates with a higher number of companies worked tend to receive higher pay, suggesting both advantages and disadvantages for the company.**
* **Job applicants with a Doctorate degree receive higher packages compared to those with PG and Graduation degrees, highlighting the value of higher education.**
* **The provided data was right-skewed, and steps were taken to address biasness and ensure accurate analysis.**
* **The analysis reveals a linear relationship between the dependent variable (Expected Salary) and the independent variable (Current Salary), suggesting a strong correlation.**

**Recommendation**

* **Continuously update and improve the data cleaning and preprocessing techniques to maintain data quality.**
* **Consider the cost of living and regional disparities when determining salary offers for different locations.**
* **Monitor industry trends and salary fluctuations to stay competitive in the job market.**
* **Enhance transparency in the salary prediction model to instill trust among job applicants and employees.**
* **Use Total Experience as a key factor in determining salary offers to match applicants' qualifications and expertise.**
* **The IT industry is the top salary offering sector, indicating its competitive pay scale compared to other industries.**
* **There are very few people belonging from Surat; we should consider hiring more candidates from this city to promote diversity.**
* **Evaluate the pros and cons of hiring candidates with higher numbers of companies worked, considering their potential impact on the company it means also shows that the stability factor**
* **Doctorate degree have a lower number of certifications but are still receiving higher packages, suggesting the importance of a Doctorate degree over certifications in determining salary.**
* **Individuals with no experience have a very low count, indicating fewer chances of getting hired. We should pay attention to candidates without experience as well, as it can help us find new talents, expertise, and promote diversity within the organization.**
* **People with key performance appraisals can be more suitable for the company, and we should give them some more preference, especially during hiring.**
* **People with number of companies with 2 & 3 have higher count and indicating higher change of getting selected from the people having a greater number of companies.**