HR Salary Dashboard – Train the Dataset and Predict Salary PROJECT REPORT

submitted by manav b v

2022

|  |  |
| --- | --- |
| Name Of the Student | Manav BV |
| Internship Project Topic | HR Salary Dashboard – Train the Dataset and Predict Salary |
| Name of the Organization | TCS iON |
| Name of the Industry Mentor | Debashis Roy |
| Name of the Institute | ICT academy Kerala |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Start Date | End Date | | Total Effort (hrs.) | | Project Environment | Tools used |
| 07/06/2022 | 23/06/2022 | | 56 | | Jupytor notebook | Microsoft word, Microsoft excel,  Python libraries, Git hub |
| Milestone # | 5 | Milestone: | | Create a dataset, clean the dataset and sanitize it. | | |

**TABLE OF CONTENT**

1. **Acknowledgements**
2. **Objective**
3. **Introduction**
4. **Internship Activities**
5. **Methodology**
6. **Assumptions**
7. **Charts, Table, Diagrams**
8. **Algorithms**
9. **Challenges & Opportunities**
10. **Reflections on the Internship**
11. **Conclusion**
12. **Enhancement Scope**
13. **Link to code and executable file**
14. **ACKNOWLEDGEMENT**

At the completion of my project, I express our sincere gratitude to all have supported and provided with the necessary guidance to do so. I would use it as wonderful opportunity to thank almightily God for all kind of blessings to bring out this project work successfully.

During the period of my internship work, I have received generous help from many quarters, which I like to put on record here with deep gratitude and great pleasure.

First and foremost, I am grateful to my guide Debashis Roy, the Industry Mentor. He allowed me to encroach upon his precious time freely right from the very beginning of this research work until the completion of my internship. His guidance, encouragement, and suggestions provided me the necessary insight into the research problem and paved the way for the meaningful ending of the work in a short duration.

Last but not the least I expressed my sincere thanks to TCS iON for giving me the opportunity to the internship. It was an excellent learning experience and it allowed me to confirm my interest in pursuing a career as data scientist.

1. **OBJECTIVE**

In this project, dealing with the HR Dataset and analysis about the trends in the data and predict the salary

1. **INTRODUCTION**

We know how employees are important for a company. And also know that how salary important to an employee. The purpose of this project is to use data transformation to create a model that will predict a salary when given years of experience, job type and education qualification of employ etc. A prediction is an assumption about a future event. A prediction is sometimes, though not always, is based upon knowledge or experience. Future events are not necessarily certain, thus confirmed exact data about the future is in many cases are impossible, a prediction may be useful to help in preparing plans about probable developments. In this paper salary of an employee of an organization is to be predicted on basis of Special project count.

The Human Resource Dataset is collected from Kaggle. DR. Carla Patalano and DR. Rich set out to create their own HR-related dataset, which is used in one of their graduate MSHRM courses called HR Metrics and Analytics, at New England College of Business.

There are different stages of processing the data set where first we load the data set into the python environment, basic understanding of the data set and exploratory data analysis. When we loaded the data set into the python environment, we were in the stage of understanding about the data which we have. There we check the number of rows and columns data types, memory, index and number of cells in each column. And also, we deal with missing values in order to fill the datasets and also did some statistical function to get the mean median mode and percentile of the data sets. When come to the explorative data analysis, we done with the salary distribution of the employees, gender distribution, and the salary distribution of employees according to their Marital status, Gender, Employment status, Position, State, Department, Recruitment source, Performance score etc. Then we go to predictive analysis, Clean the data by removing outliers and treating missing data. Identify a parametric or nonparametric predictive modeling approach to use. Preprocess the data into a form suitable for the chosen modeling algorithm. Specify a subset of the data to be used for training the model. Train, or estimate, model parameters from the training data set. Conduct model performance or goodness-of-fit tests to check model adequacy. Validate predictive modeling accuracy on data not used for calibrating the model. Use the model for prediction if satisfied with its performance.

1. **Internship Activities**

* Preparing Daily report
* Group discussion on Digital discussion room

1. **Methodology**

The data under study contains 311 rows and 36 columns. In order to gain useful insights about the data, we compare and contrast different strategies and machine learning models. The methodology follows the best practices in the literature and in the industry, including different phases.

1. *Data collection*: The Human Resource Dataset is collected from Kaggle. DR. Carla Patalano and DR. Rich set out to create their own HR-related dataset, which is used in one of their graduate MSHRM courses called HR Metrics and Analytics, at New England College of Business.
2. *Data preprocessing*: posts with missing values are removed and possible conflicts in the data format are fixed. Outlier detection and treatments are done.
3. *Visualization:* To get useful insights about the HR Dataset some meaningful visualizations are done.
4. *Predictive analysis:*
   * Clean the data by removing outliers and treating missing data
   * Identify a parametric or nonparametric predictive modeling approach to use
   * Preprocess the data into a form suitable for the chosen modeling algorithm
   * Specify a subset of the data to be used for training the model
   * Train, or estimate, model parameters from the training data set
   * Conduct model performance or goodness-of-fit tests to check model adequacy
   * Validate predictive modeling accuracy on data not used for calibrating the model
   * Use the model for prediction if satisfied with its performance
5. **Assumptions**

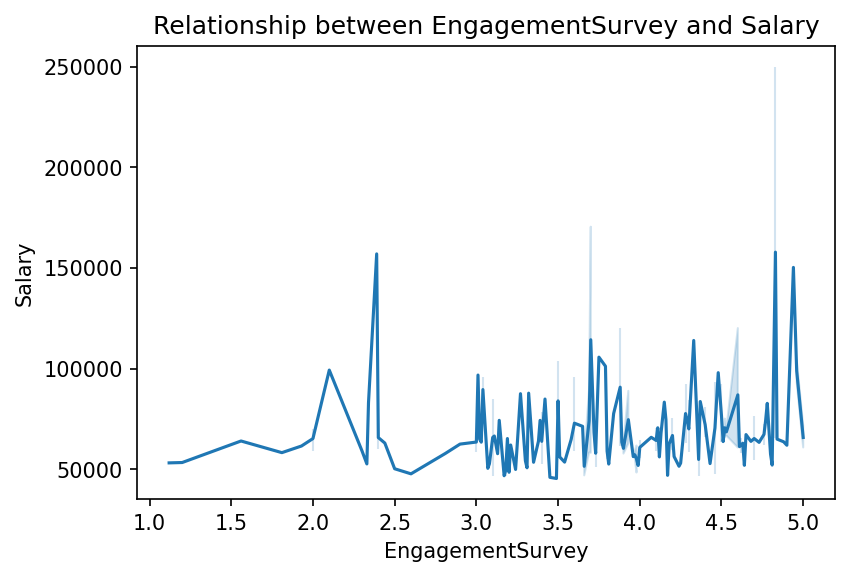
* Salary of employee is corelated with position of employee.
* Missing values, present in data is very less in number.
* The data set is suitable for salary prediction

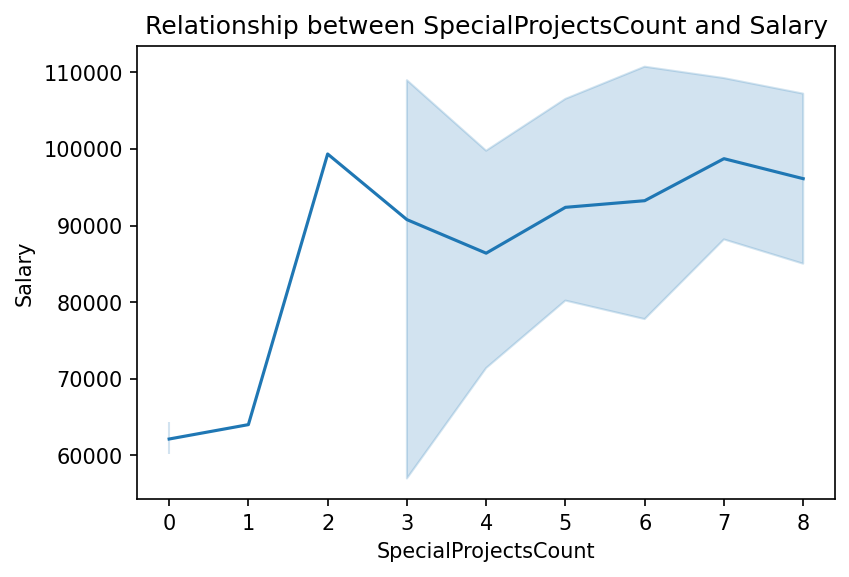
1. **Charts, Table, Diagrams**

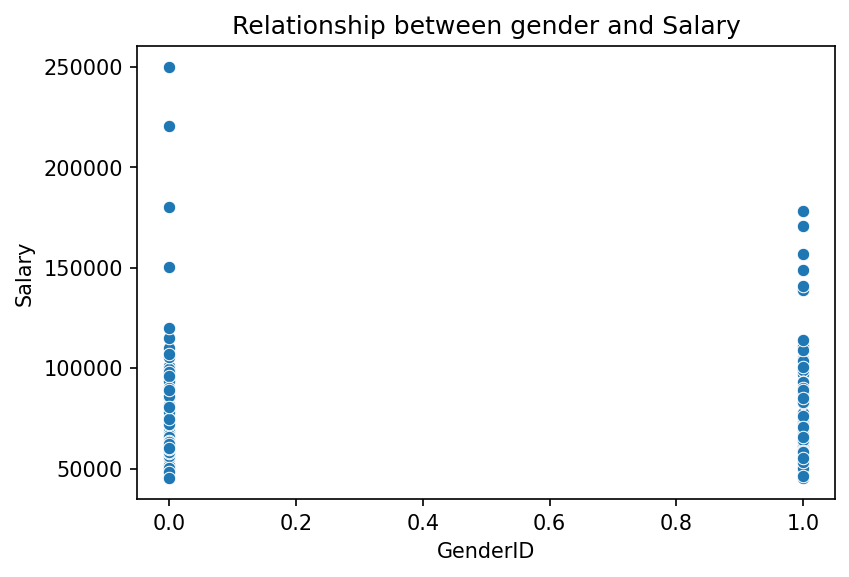
* **Correlation coefficients between variables**



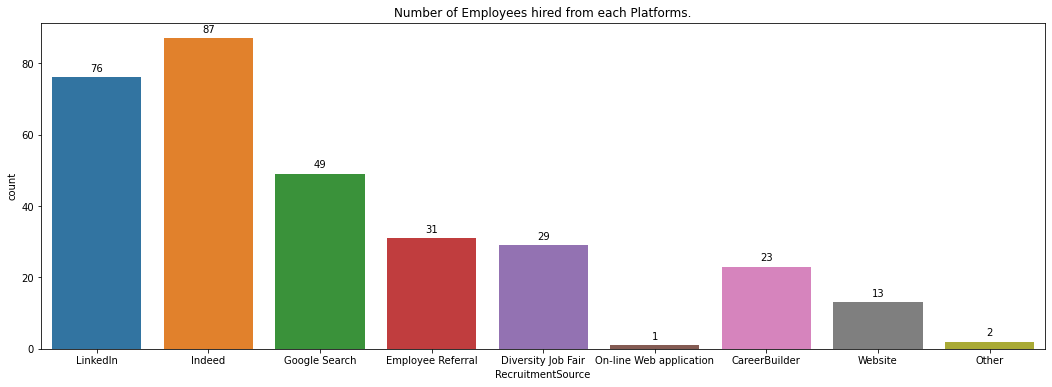
* There is stronger and weaker correlation present and regular and inverse correlation present.
* **Relationship between Engagement Survey and Salary**



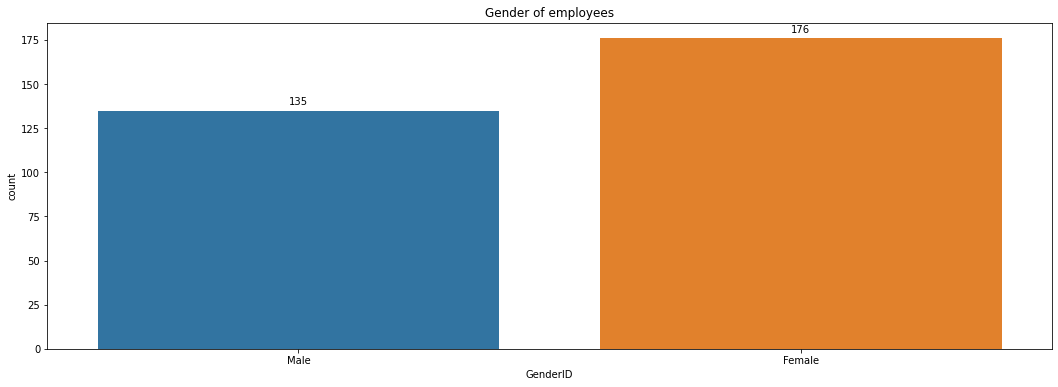
* We can see, there is a Salary hike with respect to Engagement Survey.
* **Relationship between Special Projects Count and Salary**
* We can see an increase in Salary with respect to increase in Special projects count.
* **Relationship between gender and Salary**



* Employee King, Janet has the highest salary that is 250000, as President & CEO.
* Employee Zima, Colleen has the lowest salary that is 45046, as Production Technician.
* **Number of Employees hired from each Platforms**



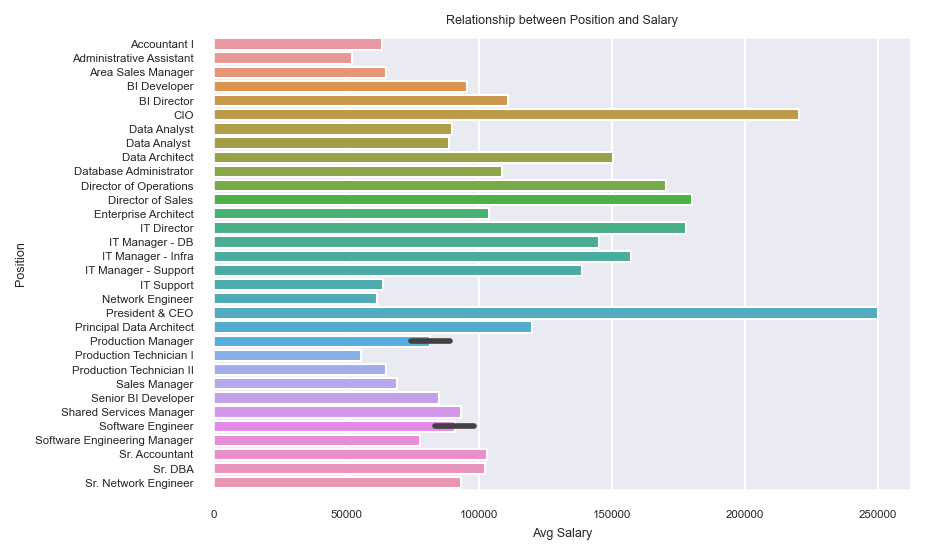
* Highest recruitment is from Indeed (87) platform and least is from On-line web application (1) platform
* **Gender of employees**

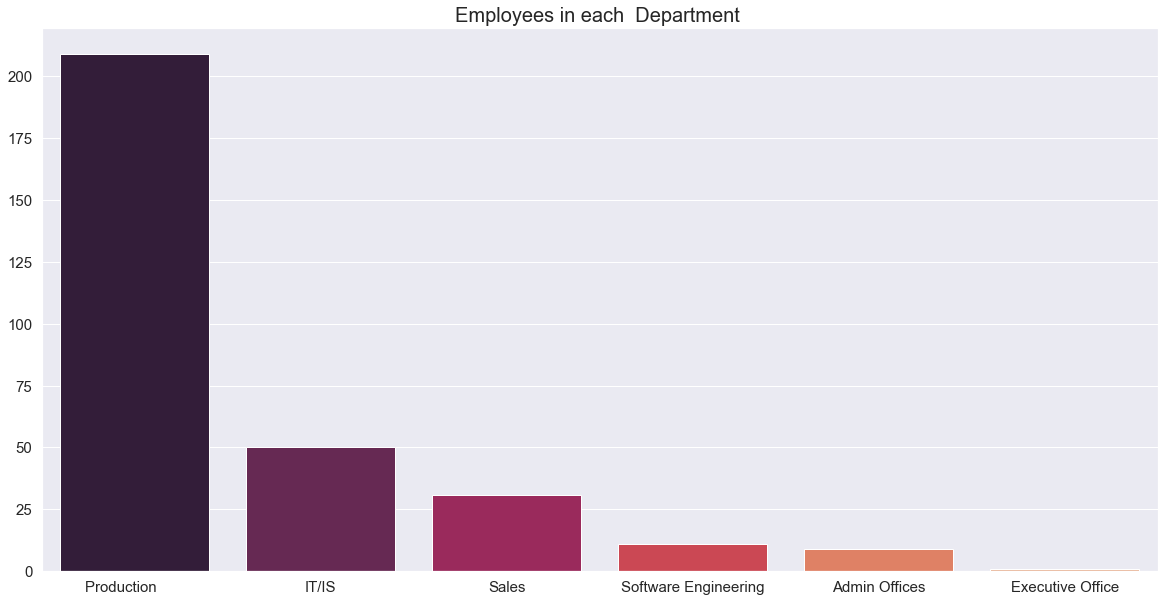


* Number of male employees is 135.
* Number of female employees is 176.
* **Relationship between position and salary**

|  |  |  |
| --- | --- | --- |
| Position | Position ID | Average Salary |
| Accountant I | 1 | 63507.67 |
| Administrative Assistant | 2 | 52280 |
| Area Sales Manager | 3 | 64932.56 |
| BI Developer | 4 | 95465 |
| BI Director | 5 | 110929 |
| CIO | 6 | 220450 |
| Data Analyst | 9 | 89932.57 |
| Data Analyst | 9 | 88527 |
| Data Architect | 7 | 150290 |
| Database Administrator | 8 | 108499.6 |
| Director of Operations | 10 | 170500 |
| Director of Sales | 11 | 180000 |
| Enterprise Architect | 30 | 103613 |
| IT Director | 12 | 178000 |
| IT Manager - DB | 13 | 144959.5 |
| IT Manager - Infra | 13 | 157000 |
| IT Manager - Support | 13 | 138888 |
| IT Support | 14 | 63684.38 |
| Network Engineer | 15 | 61605 |
| President & CEO | 16 | 250000 |
| Principal Data Architect | 29 | 120000 |
| Production Manager | 17 | 88976 |
| Production Manager | 18 | 74242.08 |
| Production Technician I | 19 | 55524.18 |
| Production Technician II | 20 | 64892.21 |
| Sales Manager | 21 | 69240 |
| Senior BI Developer | 22 | 84802.67 |
| Shared Services Manager | 23 | 93046 |
| Software Engineer | 23 | 83363 |
| Software Engineer | 24 | 98203.22 |
| Software Engineering Manager | 25 | 77692 |
| Sr. Accountant | 26 | 102859 |
| Sr. DBA | 27 | 102234 |
| Sr. Network Engineer | 28 | 93070.8 |

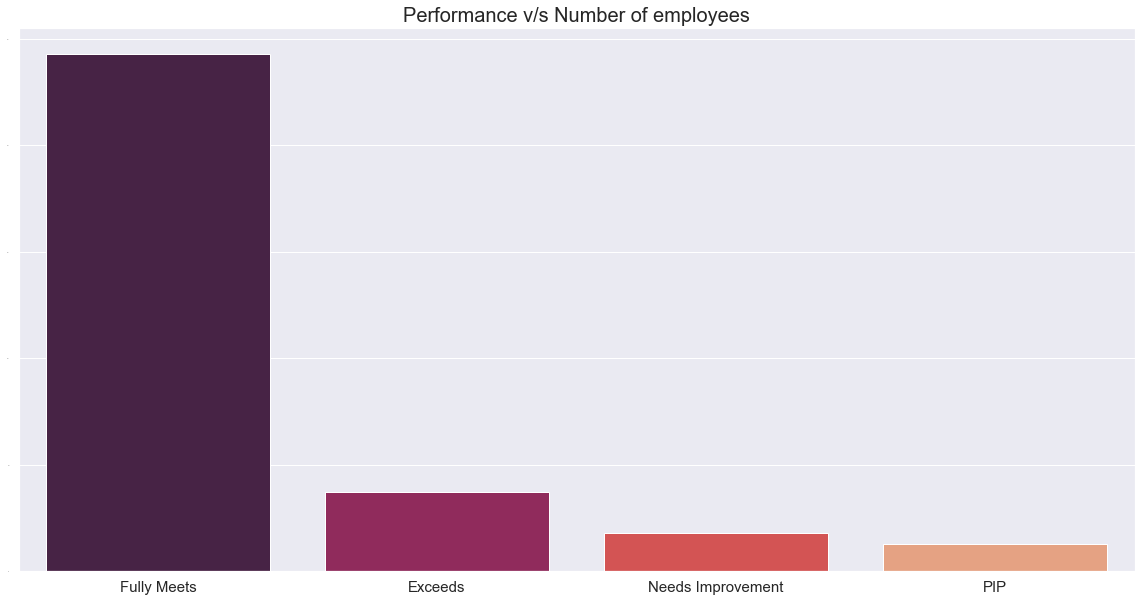
Note: Here we find average salary of every position



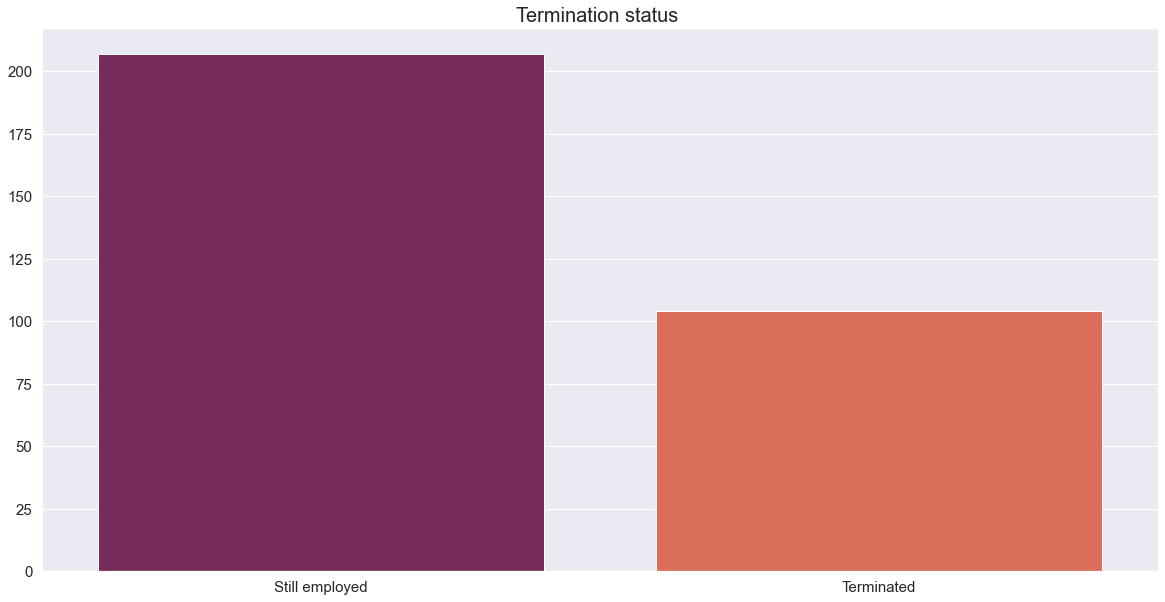
* President & CEO has highest average salary, that is 250000.0
* Administrative Assistant has lowest average salary, that is 52280.0
* **Employees in each Department**
* ****Production Department has highest number of employees, that is 209
* Executive Office has least number of employees, that is 1



* **Performance v/s Number of employees**

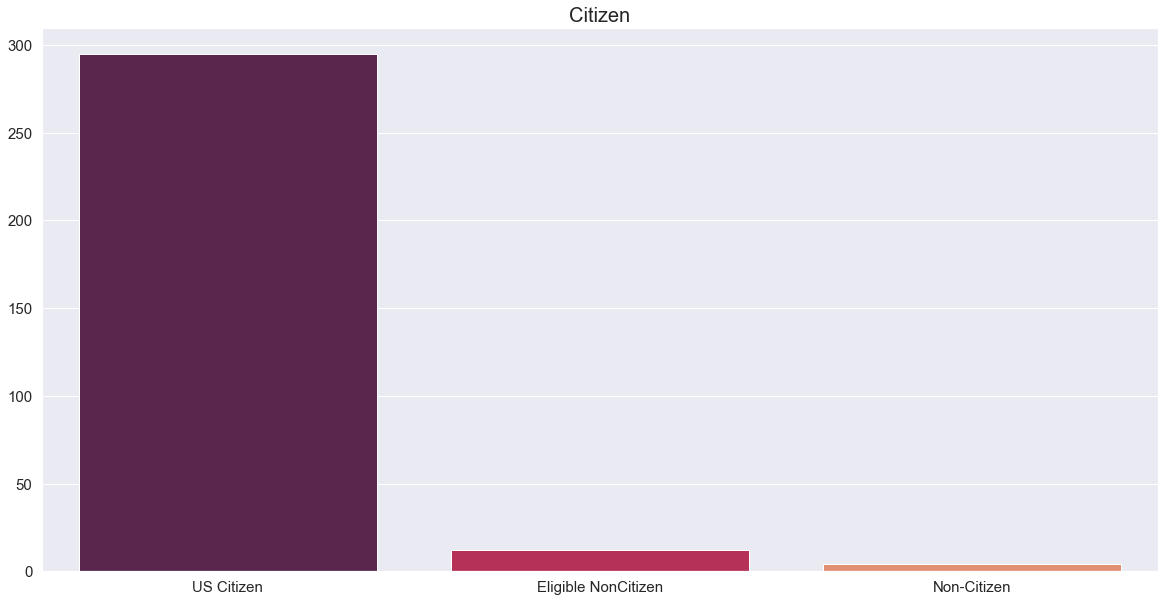
****

* About 243 employee’s performance is fully meets, 37 employee’s performance is Exceeds, 18 employee’s performance is Needs Improvement and 13 employee’s performance is PIP.
* **Termination Status**

****

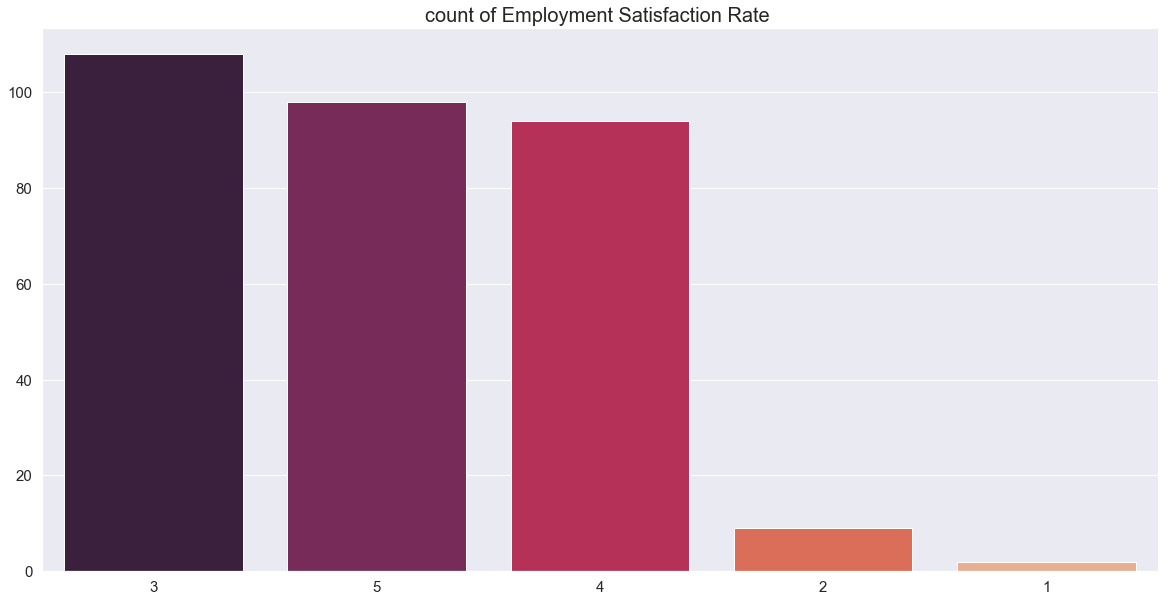
* 207 are still employed and 104 are terminated
* **Citizenship**



****

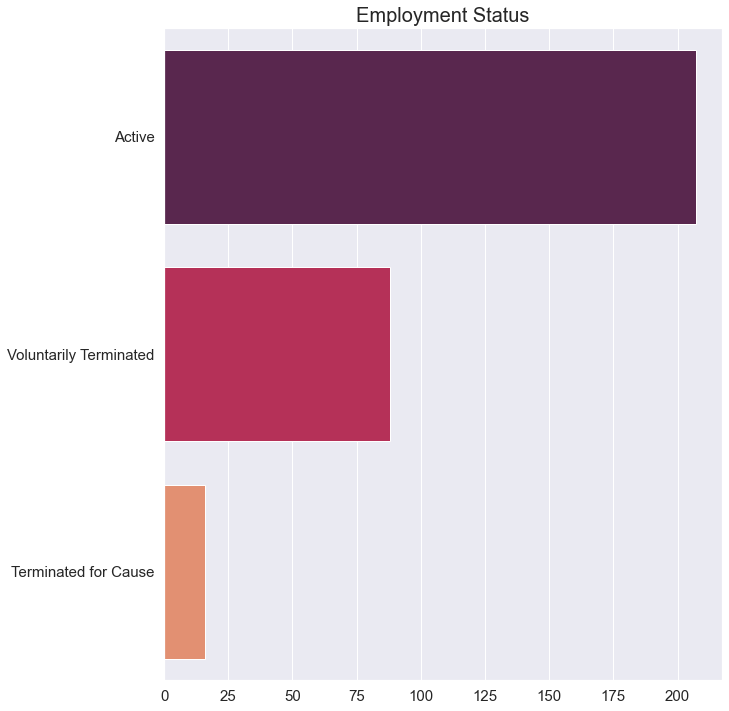
* 295 employees are US citizen's, 12 are Eligible Noncitizen's and 4 are non-citizens.
* **count of Employment Satisfaction Rate**



****

* Most employees are rated Employment Satisfaction Rate as 3, it is 108.
* 98 employees are rated as 5, 94 employees are rated as 4, 9 employees are rated as 2 and only 2 employees are rated as 1.
* **Employment status**





* 207 employees are still employed
* 88 employees are voluntarily terminated
* 16 employees are terminated for cause

**8.Algorithms**

Step 1: Import python libraries like Pandas, NumPy, Matplotlib.pyplot and seaborn.

Step2: Collected data imported into the python environment and done some basic understanding about the data such as shape, datatypes, mean etc.

Step 3: Done some visualization using matplot.lib.pyplot and seaborn and make some meaningful inference about the data.

Step 4: Check and fill the null values present in the data.

Step 5: Check and fix the outliers of data.

Step 6: Converting the labels into a numeric form so as to convert them into the machine-readable form by using Label encoder.

Step 7: To normalize the data we use min-max scaler. And now the data is clean and ready to modelling.

Step 8: Identify a parametric or nonparametric predictive modelling approach to use

Step 9: Pre-process the data into a form suitable for the chosen modelling algorithm

Step 10: Specify a subset of the data to be used for training the model

Step 11: Train, or estimate, model parameters from the training data set

Step 12: Conduct model performance or goodness-of-fit tests to check model adequacy

Step 13: Validate predictive modelling accuracy on data not used for calibrating the model

Step 14: Use the model for prediction if satisfied with its performance

**9.Predictive Analysis**

To split the data, we will be using train\_test\_split from sklearn. train\_test\_split randomly distributes data into training and testing set according to the ratio provided. Here we already remove some columns for getting more accuracy and these data’s have no effect on prediction analysis. Here the prediction variable is continuous so we use regression analysis for prediction,

* **1. Linear regression**

Linear regression is probably one of the most important and widely used regression techniques. It’s among the simplest regression methods. One of its main advantages is the ease of interpreting results. The first step is to import the package NumPy and the class LinearRegression from sklearn.linear\_model. The second step is defining data to work with. The next step is to create a linear regression model and fit it using the existing data. Create an instance of the class LinearRegression, which will represent the regression model. Then we check mean squared error and R-squared values of model.

MSE is: 103857987.54780628

r2 score is: 0.5235465653547282

* **2. Lasso regression**

Lasso regression is also called Penalized regression method. This method is usually used in machine learning for the selection of the subset of variables. It provides greater prediction accuracy as compared to other regression models. Lasso Regularization helps to increase model interpretation. The first step is to import the package NumPy and the class Lasso from sklearn.linear\_model. The second step is defining data to work with. The next step is to create a lasso regression model and fit it using the existing data.

MSE is: 103803017.94587804

r2 score is: 0.5237987410059051

* **3. DecisionTree regressor**

Decision tree is one of the well-known and powerful supervised machine learning algorithms that can be used for classification and regression problems. The model is based on decision rules extracted from the training data. In regression problem, the model uses the value instead of class and mean squared error is used to for a decision accuracy. Decision tree model is not good in generalization and sensitive to the changes in training data. A small change in a training dataset may affect the model predictive accuracy. The first step is to import the package NumPy and the class DecisionTreeRegressor from sklearn.tree. The second step is defining data to work with. Then we check mean squared error and R-squared values of model.

MSE is: 91358974.67952128

r2 score is: 0.5808863786072105

* **4. Random forest regressor**

Random forest is a supervised learning algorithm that uses an ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests run in parallel, meaning is no interaction between these trees while building the trees. The first step is to import the package NumPy and the class RandomForestRegressor from sklearn.ensemble. The second step is defining data to work with. Then we check mean squared error and R-squared values of model.

MSE is: 83325169.20564754

r2 score is: 0.6177418415492129

**fine tuning:** Fine-tuning is a way of applying or utilizing transfer learning. Specifically, fine-tuning is a process that takes a model that has already been trained for one given task and then tunes or tweaks the model to make it perform a second similar task,

After fine tuning:

MSE is: 83325169.20564754

r2 score is: 0.6177418415492129

* **5. SGD regressor**

he class SGDRegressor() implements a plain **stochastic gradient descent learning routine** which supports different loss functions and penalties to fit linear regression models. In stochastic gradient descent, we repeatedly run through the training set one data point at a a time and update the parameters according to the gradient of the error with respect to each individual data point. The first step is to import the package NumPy and the class SGDRegressor from sklearn.linear\_model. The second step is defining data to work with. Then we check mean squared error and R-squared values of model.

MSE is: 8.202436007075973e+29

r2 score is: -3.762906349625243e+21

* **6. GradientBoosting regressor**

The idea of gradient boosting is to improve weak learners and create a final combined prediction model. Decision trees are mainly used as base learners in this algorithm. The weak learner is identified by the gradient in the loss function. The prediction of a weak learner is compared to actual value and error is calculated. Based on this error, the model can determine the gradient and change the parameters to decrease the error rate in the next training. The first step is to import the package NumPy and the class GradientBoostingRegressor from sklearn.ensemble. The second step is defining data to work with. Then we check mean squared error and R-squared values of model.

MSE is: 55117578.05215957

r2 score is: 0.7471455013492125

After tuning:

MSE is: 64061357.29399316

r2 score is: 0.706115490667374

* **7. Support vector regression**

Support Vector Regression (SVR) is quite different than other Regression models. It uses the Support Vector Machine (SVM, a classification algorithm) algorithm to predict a continuous variable. While other linear regression models try to minimize the error between the predicted and the actual value, Support Vector Regression tries to fit the best line within a predefined or threshold error value. The first step is to import the package NumPy and the class SVR From sklearn.svm. The second step is defining data to work with. Then we check mean squared error and R-squared values of model.

MSE is: 234534295.12117416

r2 score is: -0.07593718202131239

After tuning:

MSE is: 234534295.12117416

r2 score is: -0.07593718202131239

**conclusion:**

After tuning some of the models shows the better result. Here Gradient boosting regressor shows the most accuracy, so we choose Gradient boosting regressor as the prediction model.

|  |  |  |
| --- | --- | --- |
| index | Predicted Salary | Actual salary |
| 290 | 71296.52097 | 88976 |
| 9 | 71189.88926 | 50178 |
| 57 | 88003.68526 | 83552 |
| 60 | 76191.03736 | 65729 |
| 25 | 88631.53645 | 96837.75 |
| 63 | 64992.69195 | 56294 |
| 92 | 57843.36742 | 58530 |
| 184 | 66969.67855 | 63291 |
| 244 | 96100.35115 | 96837.75 |
| 46 | 53345.04028 | 64786 |
| 75 | 69489.09631 | 70621 |
| 163 | 58193.46769 | 63676 |
| 297 | 54928.05199 | 50274 |
| 308 | 91235.18114 | 96837.75 |
| 286 | 65161.20116 | 74813 |
| 291 | 68974.55234 | 55875 |
| 5 | 56183.68496 | 57568 |
| 155 | 65952.55144 | 71339 |
| 164 | 88133.68424 | 93046 |
| 168 | 56250.68277 | 52624 |
| 73 | 56903.71552 | 61584 |
| 104 | 67026.82876 | 59370 |
| 137 | 68414.57122 | 83082 |
| 206 | 65587.00512 | 71966 |
| 76 | 94883.56524 | 96837.75 |
| 173 | 56682.08013 | 52057 |
| 113 | 66313.00104 | 61242 |
| 33 | 63535.38672 | 63763 |
| 295 | 59450.8618 | 66541 |
| 251 | 56636.04045 | 50923 |
| 108 | 95416.25277 | 96837.75 |
| 3 | 59590.61872 | 64991 |
| 82 | 58088.4445 | 54285 |
| 260 | 58682.99642 | 58939 |
| 93 | 62821.15164 | 72609 |
| 101 | 50433.6783 | 53171 |
| 45 | 74092.57572 | 66808 |
| 17 | 62442.39966 | 59026 |
| 197 | 92264.48379 | 87921 |
| 119 | 54858.22337 | 63813 |
| 42 | 85429.24643 | 96837.75 |
| 24 | 66195.10926 | 57815 |
| 179 | 56731.10592 | 61349 |
| 242 | 55008.62719 | 53180 |
| 299 | 84348.8687 | 96837.75 |
| 196 | 58330.43394 | 57575 |
| 226 | 57019.56961 | 46430 |
| 7 | 54813.83577 | 59365 |
| 288 | 70100.97587 | 57859 |
| 77 | 66696.35763 | 74241 |
| 114 | 66788.49195 | 66825 |
| 239 | 78482.02333 | 96837.75 |
| 145 | 68217.1911 | 63322 |
| 84 | 56974.2359 | 60340 |
| 181 | 57708.22244 | 54132 |
| 230 | 63927.67134 | 61809 |
| 303 | 54379.10762 | 59728 |
| 198 | 55028.05859 | 50470 |
| 195 | 64595.33286 | 74417 |
| 210 | 71195.94248 | 68829 |
| 224 | 61969.65067 | 46799 |
| 158 | 68340.06507 | 66074 |
| 126 | 54226.4191 | 46998 |
| 109 | 66948.67693 | 74679 |
| 247 | 56342.33228 | 46428 |
| 234 | 65130.6416 | 55578 |
| 202 | 65111.11155 | 63695 |
| 78 | 70199.33356 | 75188 |
| 147 | 72145.27697 | 68999 |
| 59 | 90007.10973 | 92329 |
| 118 | 69445.24398 | 62957 |
| 6 | 93244.42614 | 95660 |
| 182 | 65620.76136 | 55315 |
| 30 | 59032.42768 | 65288 |
| 22 | 59336.61703 | 62910 |
| 267 | 64758.77394 | 58273 |
| 56 | 55428.0946 | 63381 |
| 148 | 56325.61927 | 50482 |
| 140 | 53042.9161 | 45069 |
| 208 | 52819.72027 | 47414 |
| 279 | 53868.47385 | 48413 |
| 203 | 58518.1108 | 62061 |
| 144 | 93585.77674 | 96837.75 |
| 167 | 72795.67241 | 77915 |
| 90 | 55683.54906 | 64057 |
| 223 | 63548.93001 | 61844 |
| 66 | 87364.84917 | 96837.75 |
| 116 | 60344.80382 | 66149 |
| 172 | 63217.52989 | 73330 |
| 250 | 51270.30011 | 56147 |
| 19 | 59192.12637 | 53250 |
| 146 | 54321.89019 | 61154 |
| 79 | 57916.73621 | 62514 |
| 300 | 54183.8571 | 58371 |

**10.Challenges & Opportunities**

* There is no proper description about the data.
* The correlation between the variables is weak.
* Some of the values are already encoded and it is very difficult to understand.
* The data is not perfect for salary prediction.
* Some of the information’s are already encoded into numerical form, so it’s very helpful in the time of encoding.
* The missing values present in the data is comparatively less in number.

**11.Reflections on the Internship**

My internship as a Data Science and Analytics for TCS iON has been the most rewarding and motivational experiences. I connected with cooperating mentor and other interns, all of whom supported my growth as a teacher, both directly and indirectly. With such empathetic, compassionate and supportive mentors, this experience has helped me achieve my goal of completing my project on HR salary prediction. I am confident that I will continue to grow and develop professionally and in my personal endeavours, and I look forward to where my career in Data Science and Analytics takes me in the future.

**12.Conclusion**

We were complete the basic data understanding, analysing and some visualization in order to understand the data and its trends. There are many categorical, non-numerical and missing values in the data set and we were converting it in to numeric variables in order to get a good outcome and possible result from the data set. From the data pre-processing steps the data became cleaner and more standardized. After splitting the data into train and test data we conduct implementation of prediction model to check the most accurate model. Here the prediction variable salary is continuous so we use regression analysis for modelling. We did Linear regression, Lasso regressor, DecisionTree regressor, RandomForest regressor, SGD regressor, GradientBoosting regressor and Support vector regressor and also done fine tuning for some models. Fine tuning shows better result for some models. From these models Gradient Boosting regressor shows the highest accuracy that is 74.7%. So we fit Gradient boosting regressor as prediction model.

**13.Enhancement Scope**

The project has a very vast scope in future. Now the data is cleaned and standardized so the data set is ready for modelling.

**14. Link to code and executable file**

* **Data set**

<https://www.kaggle.com/datasets/rhuebner/human-resources-data-set>

* **Code**

<https://github.com/ManavBVijayan/Internship/blob/1696e13ffe7a8c1bafd9cf560c227a9413226c01/internship%20%20(3).ipynb>