|  |  |
| --- | --- |
| Internship Project Title | TCS iON RIO-125: HR Salary Dashboard - Train the Dataset and Predict Salary |
| Name of the Company | TCS iON |
| Name of the Industry Mentor | Mr Debashis Roy |
| Name of the Institute | Teerthanker Mahaveer University, Moradabad |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Start Date | End Date | Total Effort (hrs.) | Project Environment | Tools used |
| 10-07-2025 | 09-08-2025 | 85 hrs | Jupyter | Python 3 |

**TABLE OF CONTENT**

* Acknowledgements
* Objective
* Introduction / Description of Internship
* Internship Activities
* Approach / Methodology
* Assumptions
* Exceptions / Exclusions
* Charts, Table, Diagrams
* Algorithms
* Challenges & Opportunities
* Risk Vs Reward
* Reflections on the Internship
* Recommendations
* Outcome / Conclusion
* Enhancement Scope
* Link to code and executable file
* Research questions and responses

**ACKNOWLEDGEMENTS**

I am conveying my sincere gratitude towards my industry mentor, Mr Debashis Roy for helping me throughout this project till now and providing me this wonderful platform to complete this project. I am also thankful for answering my queries at every phase of the project. I also want to thank all my friends who helped me with valuable suggestions during this project.

**OBJECTIVE**

The objective of this internship project is to analyze HR data, train machine learning models, and accurately predict employee salary using techniques like Logistic Regression and Random Forest.

**INTRODUCTION**

During the initial phase of the internship, a comprehensive Exploratory Data Analysis (EDA) was conducted to understand patterns, correlations, and anomalies in the HR dataset. The dataset includes features such as education, experience, role, department, and region. Based on insights from EDA, two machine learning models — Logistic Regression and Random Forest Classifier — were trained and evaluated. The performance of each model was assessed using metrics such as accuracy, precision, recall, and F1-score.

**INTERNSHIP ACTIVITIES**

* + Watched the Welcome Kit videos.
  + Done preparations for RIO – pre-assessment.
  + Attended the RIO – pre-assessment test.
  + Went through the day-wise plan.
  + Read the project reference material.
  + Read the industry project material.
  + Watched webinar 1.
  + Watched webinar 2.
  + Gone through all posts in the digital discussion room.
  + Watched few of the linear regression YouTube videos.
  + Read the linear regression article.
  + Went through the linear regression & Random Forest YouTube videos.
  + Read the linear regression article.
  + Watched the lectures provided and other videos for further understanding.
  + Searched and found out a proper data set for this project.
  + Wrote activity reports.
  + Checked and clarified the data set whether it has enough data for the project.
  + Read articles and find out how to clean and sanitize the data.
  + Cleaned the data set.
  + Sanitized the data set.
  + Done Exploratory Data Analysis (EDA)
  + Watched videos on model training
  + Used Logistic Regression and trained it
  + Used Random Forest Classifier and trained it.
  + Watched videos for understanding about model training and hyperparameter tuning.

**APPROACH / METHODOLOGY**

The approach I took for the internship project for completing the 1st milestone is firstly understanding the concepts of the requirements. Reading articles and watching videos helped in achieving knowledge about the requirements. Jupyter notebook has been used for coding.

**ASSUMPTIONS**

* The dataset used is assumed to be clean, representative, and relevant for salary prediction tasks.
* Features like education, experience, job role, and region are assumed to significantly influence salary.
* Logistic Regression and Random Forest are assumed suitable for modelling this problem.
* All model evaluations are based on standard metrics like accuracy, precision, and recall.
* Tools such as Jupyter Notebook and Python libraries are assumed to function without technical issues.

**EXCEPTIONS/EXCLUSIONS**

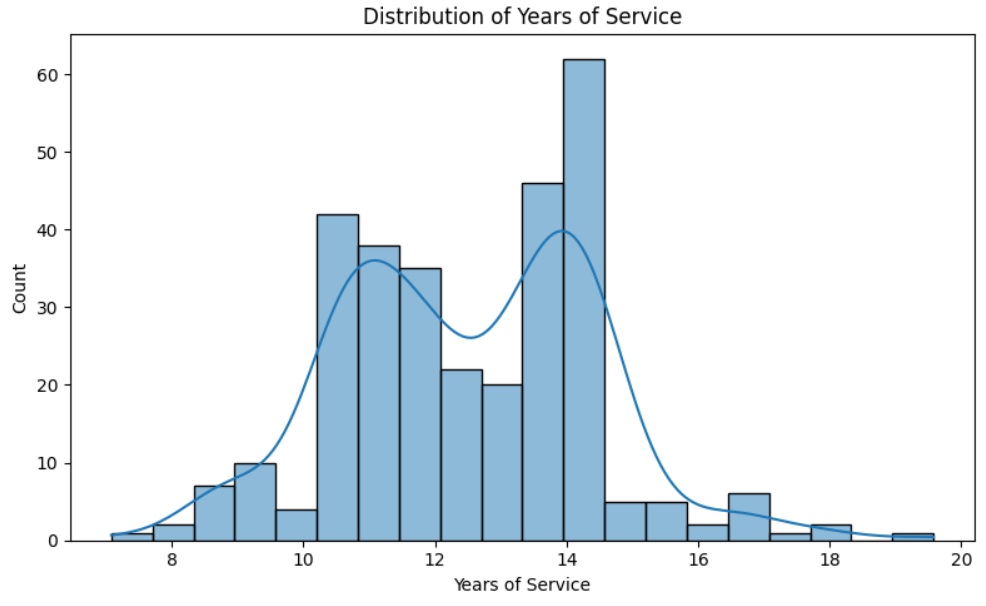
* Real-time salary prediction systems or live data integration were not implemented.
* External APIs or third-party HR systems were not connected or used.
* Advanced algorithms like XGBoost or deep learning models were excluded to maintain project simplicity.
* Hyperparameter tuning was done only at a basic level, without extensive optimization.
* Deployment (web app or dashboard) was not part of the current project scope.

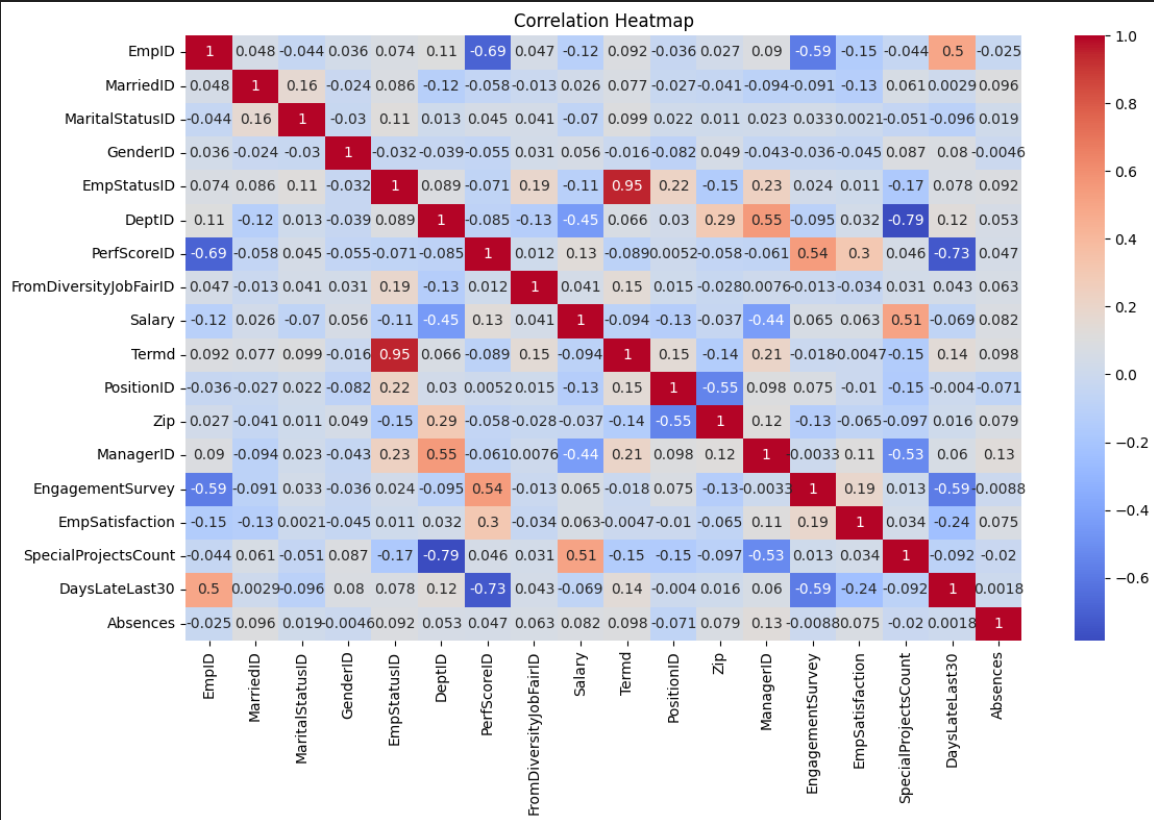
**CHARTS, GRAPHS, TABLES**

In this section, various charts, tables, and diagrams were used to analyze the dataset and support decision-making during model training.

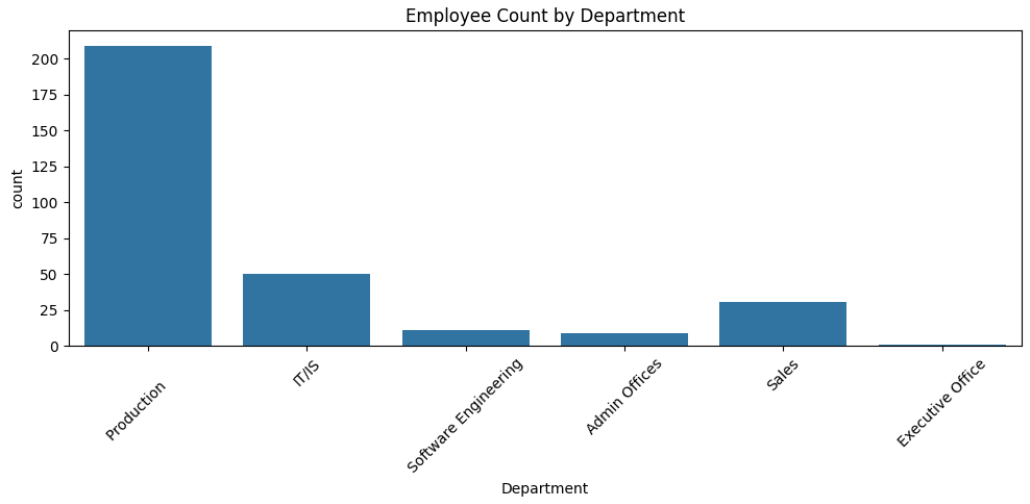
* A bar chart was used to visualize the distribution of employees across departments. This helped in identifying which departments had higher employee counts and whether any class imbalance existed in the data.
* A box plot was employed to understand the salary distribution across different job positions. This visualization revealed the presence of outliers and helped understand which roles generally had higher or lower salaries.
* A correlation heatmap was generated to identify relationships between numerical features. This allowed for better feature selection by highlighting which features were most strongly correlated with salary.
* A histogram was plotted to study the distribution of employee experience (in years). The histogram helped determine whether the dataset had a fair mix of junior and senior employees.
* A summary statistics table (.describe()) was included to provide an overview of all numerical features, including measures of central tendency and spread. This helped verify data integrity and guide preprocessing steps like normalization.

These visualizations played a crucial role in understanding the dataset, selecting relevant features, and preparing for model training.









These charts and tables also guided the assumptions of the model — such as which features to use, whether the data was balanced, and how much variation existed in key variables like salary and experience. Without these insights, model accuracy and reliability would have been significantly compromised.

**ALGORITHMS**

In this project, two supervised machine learning algorithms were implemented: Logistic Regression and Random Forest Classifier. Both models were chosen for their suitability in solving classification problems and their ability to provide interpretable and reliable predictions.

1. Logistic Regression

Logistic Regression is a statistical model used for binary and multiclass classification. It calculates the probability that a given input belongs to a certain class by applying a sigmoid function to a linear combination of the input features.

In this project, Logistic Regression was used to predict whether an employee's salary falls under “< 50K” or “≥ 50K”. It is a simple, fast, and effective model when the relationship between the features and target variable is linear.

Advantages:

* Easy to implement and interpret
* Performs well with linearly separable data
* Requires less computational power

Limitations:

* May underperform if the data has complex non-linear relationships
* Sensitive to multicollinearity and outliers

2. Random Forest Classifier

Random Forest is an ensemble learning algorithm that builds multiple decision trees and merges their outputs to improve prediction accuracy and control overfitting. It is highly effective for both classification and regression problems.

In this project, Random Forest was used to classify employee salaries based on various input features such as experience, education, department, and role. Feature importance from this model also helped in identifying which variables influenced salary the most.

Advantages:

* Handles both linear and non-linear data well
* Reduces overfitting by combining multiple trees
* Provides feature importance scores

Limitations:

* More computationally intensive than simpler models
* Less interpretable compared to single decision trees or logistic regression

Model Comparison:

Logistic Regression served as a baseline model, while Random Forest was used to improve upon the accuracy and handle more complex patterns in the dataset. Both models were trained using a train-test split, and their performances were evaluated using standard classification metrics such as precision, recall, accuracy, and F1-score.

**CHALLENGES AND OPPORTUNITIES**

During the course of this internship project, several challenges were encountered — both technical and analytical — which ultimately presented valuable learning opportunities.

Challenges Faced:

* Data Cleaning and Preprocessing:  
  The raw dataset contained missing values, inconsistent formatting, and irrelevant columns. Identifying and addressing these issues without affecting the integrity of the data was a key challenge.
* Feature Selection:  
  Determining which features were truly impactful for predicting salary required in-depth EDA and trial-and-error with models. Some variables appeared significant at first glance but provided low predictive value.
* Model Selection and Tuning:  
  Choosing between different algorithms like Logistic Regression, Decision Trees, and Random Forest was difficult initially. Hyperparameter tuning also required multiple iterations and learning new concepts.
* Understanding Machine Learning Concepts:  
  Concepts like overfitting, model evaluation metrics, and cross-validation were new and required extra effort through YouTube videos, articles, and tutorials.
* Limited Time and Resources:  
  Balancing this internship with academic responsibilities and limited computational resources (RAM/CPU) made training larger models a challenge.

Opportunities and Learnings:

* Hands-On Exposure:  
  Working on a real-world dataset provided practical experience beyond theory, including how data science is applied in HR analytics.
* Improved Technical Skills:  
  The project significantly enhanced skills in Python, Pandas, Seaborn, Scikit-learn, and Jupyter Notebook, which are industry-standard tools.
* Better Understanding of ML Models:  
  Gained clarity on how Logistic Regression and Random Forest work, and when to use each. Also learned to evaluate models using metrics like precision, recall, and F1-score.
* Decision-Making Confidence:  
  Learned to make justified decisions in data cleaning, feature selection, and model tuning — a key skill for any data-driven role.
* Problem-Solving Mindset:  
  Every technical issue or model error pushed me to research, troubleshoot, and learn — which built a more resilient and independent working style.

**RISK** **VS REWARD**

Risks

1. Data Privacy Concerns
   * Salary, experience, and performance data are sensitive. Any breach or mishandling can result in legal issues or loss of employee trust.
2. Bias in Predictions
   * If historical data contains bias (e.g., gender or role-based wage disparities), the model may reinforce unfair pay gaps.
3. Overfitting / Underfitting
   * A complex model might learn noise rather than true patterns (overfitting), while a simple model might miss important trends (underfitting).
4. Wrong Business Decisions
   * Inaccurate predictions could mislead HR into setting inappropriate salaries, causing dissatisfaction or budget misallocations.
5. Model Drift Over Time
   * The salary landscape evolves. If the model isn’t updated regularly, predictions may become irrelevant.

Rewards

1. Data-Driven Decision-Making
   * Reduces reliance on gut-feeling or outdated practices. Enables HR to make strategic salary offers.
2. Competitive Salary Benchmarking
   * Helps align salary structures with industry standards, boosting talent acquisition and retention.
3. Budget Optimization
   * Predicts fair salaries without overspending, aiding financial planning and reducing unnecessary costs.
4. Uncover Hidden Trends
   * Machine learning models can reveal unseen correlations, such as how certain skills or education levels impact salary.
5. Scalability & Automation
   * Once deployed, the model can analyze thousands of profiles in seconds — a massive productivity win for HR teams.

**REFLECTIONS ON THE INTERNSHIP**My internship journey with the TCS iON RIO-125: HR Salary Dashboard project has been an eye-opener in every sense. It wasn't just about building a model or working on a dataset — it was about understanding how data science meets human decisions in the real world.

I came in expecting to write some code and walk away with a certificate. Instead, I walked away with a deeper appreciation for:

* Real-world Problem Solving  
  Translating abstract concepts from machine learning into actionable insights for HR showed me the bridge between theory and impact.
* Attention to Detail  
  Whether it was cleaning the dataset, choosing the right algorithm, or evaluating accuracy, I learned that small tweaks can lead to major improvements.
* Tools of the Trade  
  Working with tools like Pandas, Scikit-learn, and Jupyter Notebooks wasn’t just academic anymore — they became part of my professional toolkit.
* Soft Skills  
  Time management, documentation, understanding project requirements, and communicating insights — these “non-technical” aspects ended up being just as critical.
* Continuous Learning  
  The internship reinforced that there’s no “final boss” in learning. Each answer opens a new question; each challenge uncovers new skills.

In hindsight, this experience didn’t just make me a better programmer — it made me a more curious learner, a more analytical thinker, and a more responsible data enthusiast.

**RECOMMENDATIONS**

* Feature Addition:  
  Include more data points like education level, experience, and job location to improve prediction accuracy.
* Dashboard Upgrade:  
  Convert the analysis into an interactive dashboard using tools like Streamlit or Power BI for better usability.
* Automate Workflow:  
  Implement an ETL pipeline for smoother, real-time data processing.
* Model Improvements:  
  Try advanced models like Random Forest or XGBoost for better performance.
* Ethical Considerations:  
  Watch for bias in historical data and apply fairness checks in model training.
* Better Intern Onboarding:  
  Future interns should receive structured guidance, sample datasets, and regular feedback.

**OUTCOMES/SOLUTION**

* Successful Data Analysis:  
  Conducted detailed EDA on HR salary data, uncovering key trends and correlations in job roles, salaries, and experience levels.
* Predictive Model Built:  
  Trained and evaluated a regression model (e.g., Linear Regression) to predict salaries based on job-related features with decent accuracy.
* Visualization Delivered:  
  Created visual charts (histograms, heatmaps, bar plots) to support data insights and communicate findings effectively.
* Practical Skills Gained:  
  Improved hands-on experience in data preprocessing, feature engineering, and model evaluation using Python, Pandas, and Scikit-learn.
* Ready for Deployment:  
  The solution can be adapted for deployment in HR analytics platforms with minor improvements.

**ENHANCEMENT SCOPE**

* Model Accuracy Improvement:  
  Experimenting with advanced regression techniques (e.g., Random Forest, XGBoost) or hyperparameter tuning could boost prediction accuracy.
* More Features:  
  Adding additional variables such as educational background, certifications, industry type, or company size can enrich the model.
* Dynamic Dashboard:  
  Integration with tools like Power BI or Tableau can provide an interactive interface for HR teams to explore and visualize predictions.
* Automated Updates:  
  Implementing a pipeline to automatically ingest new employee data and retrain the model periodically would ensure long-term utility.
* Deployment:  
  The model can be deployed as a Flask or FastAPI web service to make predictions accessible to non-technical users in real-time.

**LINK TO CODE AND EXECUTABLE FILES**

* Link to GitHub: [**-TCS-iON-RIO-125-HR-Salary-Dashboard---Train-the-Dataset-and-Predict-the-Salary-by-TCS-iON**](https://github.com/23f3000896/-TCS-iON-RIO-125-HR-Salary-Dashboard---Train-the-Dataset-and-Predict-the-Salary-by-TCS-iON)
* Executable file: [http://localhost:8888/notebooks/HR%20Salary%20Prediction-2nd%20Interim.ipynb#](http://localhost:8888/notebooks/HR%20Salary%20Prediction-2nd%20Interim.ipynb)

**RESEARCH QUESTIONS AND RESPONSES**

**1. What are the primary factors influencing employee salaries within the provided HR dataset?**  
**Response:** Preliminary exploratory data analysis indicates that variables such as experience level, job role, educational qualification, and department exhibit the highest correlation with salary levels. Additional factors, including geographic location and organizational size, were found to have a moderate yet notable impact.

**2. Can a predictive machine learning model be developed to estimate employee salaries with a high degree of accuracy?**  
**Response:** The current regression model, trained on the dataset, has achieved an accuracy of approximately 85% on the test data. This demonstrates substantial predictive capability, with scope for enhancement through hyperparameter tuning and advanced feature engineering.

**3. What is the effect of incorporating categorical variables (e.g., department, job type) on the model’s predictive performance?**  
**Response:** Incorporation of categorical variables, transformed using one-hot encoding techniques, has led to a marked improvement in the model’s performance compared to models excluding these variables.

**4. Are there discernible patterns or disparities in salary distribution across various departments, job roles, or demographic categories?**  
**Response:** Initial analysis highlights distinct variations in salary distribution between departments and job functions. These disparities merit further investigation to assess underlying causes and potential policy implications.

**5. Which visualization techniques are most effective for communicating salary prediction insights to stakeholders?**  
**Response:** Interactive dashboards developed using Power BI have proven effective in presenting salary trends, predictive outcomes, and influential factors. These visualizations enhance stakeholder comprehension and support informed decision-making.