# Predictive Modeling Documentation: Salary and Delivery Time Prediction Models

Manasi Odassery

B94

This document outlines the methodologies, evaluations, and business impacts of two predictive models developed for strategic HR planning and logistics optimization: the Salary Prediction Model and the Delivery Time Prediction Model. It provides insights into the data-driven approaches taken and recommendations for future improvements.

1. Project Objectives:

- Salary Prediction Model: To develop a predictive model that accurately forecasts salary increments based on the years of experience, aiding HR in strategic compensation planning and equitable salary distribution.
- Delivery Time Prediction Model: To construct a predictive model to estimate delivery times from sorting times, enhancing operational efficiency and customer service in logistics.

2. Data Collection and Preparation:

- Salary Model:
    - Data Source: HR payroll data over the last ten years from company records.
    - Data Volume: 30 employee records involving years of experience and corresponding salaries.
    - Data Cleaning: Removed anomalies and standardized data for consistency.
    - Normalization: Applied where necessary to align scales of salary figures across different divisions.
- Delivery Model:
    - Data Source: Logistics records from the previous year, detailing package sorting and delivery timings.
    - Data Volume: 21 data points covering various sorting conditions.
    - Data Cleaning: Focused on removing outliers that could skew model predictions.
    - Feature Engineering: Created new variables where needed to better capture the dynamics of delivery operations.

3. Exploratory Data Analysis (EDA):

- Salary Model:
    - Techniques Used: Correlation matrices, scatter plots, and line plots to determine the strength and nature of the relationship between years of experience and salary.
    - Findings: Identified a robust positive correlation indicating that salary increases with years of experience. Variance analysis suggested consistent salary growth patterns across different departments.
- Delivery Model:
    - Techniques Used: Scatter plots and regression line analysis to assess the relationship between sorting time and delivery time.
    - Findings: Confirmed a positive correlation, with increased variability in delivery times associated with longer sorting times. Indications of potential operational inefficiencies were noted for higher sorting times.

4. Model Building:

- Methodology:
    - Both models employed simple linear regression due to the linear characteristics observed during EDA.
- Implementation:
    - Python's Scikit-learn library was utilized for its robust linear regression capabilities.
    - Both models were trained on 80% of the data, with 20% held back for testing to prevent overfitting.

5. Model Evaluation:

- Salary Prediction:
    - MSE: 49838096.86, indicating the model's average squared prediction error.
    - $R^2$: 0.9024, demonstrating that the model explains approximately 90.24% of the variance in salaries from years of experience.
- Delivery Time Prediction:
    - MSE: 14.0467, reflecting the model's precision in predicting delivery times.
    - $R^2$: -1.0208, suggesting model inadequacies possibly due to omitted variable bias or external disturbances not accounted for.

6. Business Impact:

- Strategic Implications:
    - Salary Model: Facilitates more accurate budget forecasts for salary increments and supports competitive and fair compensation strategies.
    - Delivery Model: Enables more precise planning and resource allocation in logistics, potentially reducing costs and improving service quality.

7. Challenges and Recommendations:

- Enhancements Needed:
    - Salary Model: Incorporate additional predictors such as educational qualifications, performance metrics, and possibly departmental distinctions to refine accuracy.
    - Delivery Model: Explore multivariate regression models that include external factors like traffic conditions, weather, and perhaps package dimensions.

8. Conclusion: The models serve as critical tools in their respective domains, driving data-informed decisions that can significantly impact the company's strategic planning and operational effectiveness. Ongoing model updates, data enrichment, and methodological advancements are recommended to keep pace with changing business needs and environmental variables.