**Predicting Delivery Time Based on Sorting Time**

**Project Overview:**

This project aims to predict "Delivery Time" using "Sorting Time" as the independent variable. The project follows a structured workflow, encompassing data cleaning, exploratory data analysis (EDA), model building, and model evaluation to determine the ideal model for predicting delivery time.

**Project Steps:**

**Data Import and Preprocessing:**

The project begins by importing necessary Python libraries in a Jupyter Notebook environment.

The data was thoroughly cleaned, and it was found that there were no missing or null values in the dataset, ensuring data quality.

**Exploratory Data Analysis (EDA):**

EDA was conducted to gain insights into the dataset and the relationship between variables:

**Regression Plot Analysis:** Used regression plots to assess the linearity of the relationship between Sorting Time and Delivery Time. Here, in this case, it was Positive Linear Regression and the correlation between the variables is 0.8259.

**Histogram Analysis:** Analyzed histograms to understand the distribution of the data and its central tendencies. This aids in assessing the normality or skewness of the data. In this case, it was slightly Right-skewed and the skewness value is 0.33.

**Box Plot Analysis:** Employed box plots to identify potential outliers within the dataset. In this case, there is no Outliers.

**Line Plot Analysis:** Studied line plots to visualize the trends and patterns in the data. In this case, Delivery time is increasing with Sorting time but it’s not perfectly Linear, as if we see in the graph, at a particular point 4 and 6, the delivery time is huge.

**Model Building:**

The core of the project involves building predictive models for "Delivery Time" using "Sorting Time" as the independent variable.

To ensure robustness, models were created using three different data splits:

* 70:30
* 75:25
* 80:20

Model building was implemented using scikit-learn (sklearn), a popular machine learning library.

**Model Evaluation:**

Model performance was assessed using two key metrics:

R2 Score (Coefficient of Determination): Indicates the extent to which the model explains the variance in Delivery Time. An R2 score of 79.8% was obtained for the 75:25 split, making it the ideal model.

Root Mean Squared Error (RMSE): Measures the average prediction error of the model. The RMSE value for the 75:25 split was 2.54 units.

**Project Conclusion:**

The project successfully achieved the objective of predicting "Delivery Time" based on "Sorting Time." The 75:25 split model, with an R2 score of 79.8%, was selected as the ideal model for its accuracy in explaining the variance in Delivery Time. This model has the potential to optimize logistics, enhance customer satisfaction, and provide a competitive edge in the industry.

The project highlights the power of data-driven decision-making and its potential impact on efficiency, cost savings, and customer loyalty. Future steps may include further model refinement and real-world implementation.