Assignment 1: Predictive Models for Estimating Delivery Times

QBUS2820 Predictive Analysis

Anabel Geraldine 520360707

Page Count: 11
Date: 20 April 2025

Table of Contents

1. Introduction

Accurate prediction of delivery is very important for optimizing logistics in modern e-commerce and supply chain management. This ability to predict delivery duration ***allows businesses to improve resource allocation, enhance customer satisfaction, and reduce operational inefficiencies.*** In this report, we will discuss and analyze a dataset containing order specific attributes to construct a predictive regression model for delivery time. We will employ feature engineering and machine learning models, including regularized regressions and nonparametric methods, to identify the most effective strategy. The significance of precise delivery time forecasting has been emphasized in recent literature, which highlights its impact on both consumer trust and last-mile logistics performance *(Nguyen et al., 2023)*. Our goal is to develop a robust, generalizable model with strong predictive accuracy across diverse delivery scenarios.

2. Data Introduction

The dataset consists of 10,000 delivery records with numerical and categorical features relevant to predicting delivery time. The data will be processed and cleaned before usage. A summary of the dataset is presented below.

Variable	Description
Delivery Time	Total Delivery time (minutes)
Distance	Distance to customer (km)
OrderSize	Number of items in the order
WeatherImpact	Weather Severity (1 = low, 10 = high)
CourierExperience	Year of experience of the courier
ProcessingTime	Time spent at warehouse before dispatch (minutes)
WeekendDelivery	1 if delivered on a weekend, 0 otherwise
TrafficCondition	Traffic Level (1 = low, 2 = medium, 3 = high)

Figure 1. Table of Data Variables

3. Data Preprocessing

To ensure data quality and reliability, initial preprocessing steps focused on identifying and handling missing or error values. A ***check for missing values*** was conducted, along with ***inspection for outliers*** in the DeliveryTime column, where negative values were found to be invalid. Since a ***negative delivery time is not feasible*** in a real-world context,we replace it with the median of the Delivery Time distribution. This preserves the dataset's integrity without introducing bias that could result from removing records entirely. We then check the data description to make sure that the data is ready to be used.

4. Exploratory Data Analysis

4.1 Descriptive Statistics

To gain a foundational understanding of the dataset, descriptive statistics were computed for all numerical variables. The average delivery time was approximately 65 minutes, with a wide range from just 0.22 minutes up to nearly 377 minutes, indicating potential outliers or variability due to external conditions. The standard deviations across features such as Distance, WeatherImpact, and ProcessingTime also highlight substantial variance, which justifies the use of scaling and normalization in subsequent modeling.

4.2 Visualization

The exploratory data analysis reveals key patterns and distributions in the dataset. Scatter plots suggest a positive, ***nonlinear relationship between delivery time and variables like Distance, Distance_Traffic, and WeatherImpact,*** supporting the case for **polynomial regression**. Histograms show that most numeric variables such as CourierExperience, WeatherImpact, and ProcessingTime are right-skewed, while OrderSize is discretely distributed. Boxplots highlight the presence of several outliers across variables, especially in DeliveryTime and Distance, though most values remain within a reasonable range. Overall, the data shows moderate variability and meaningful structure, supporting the potential for accurate predictive modeling.

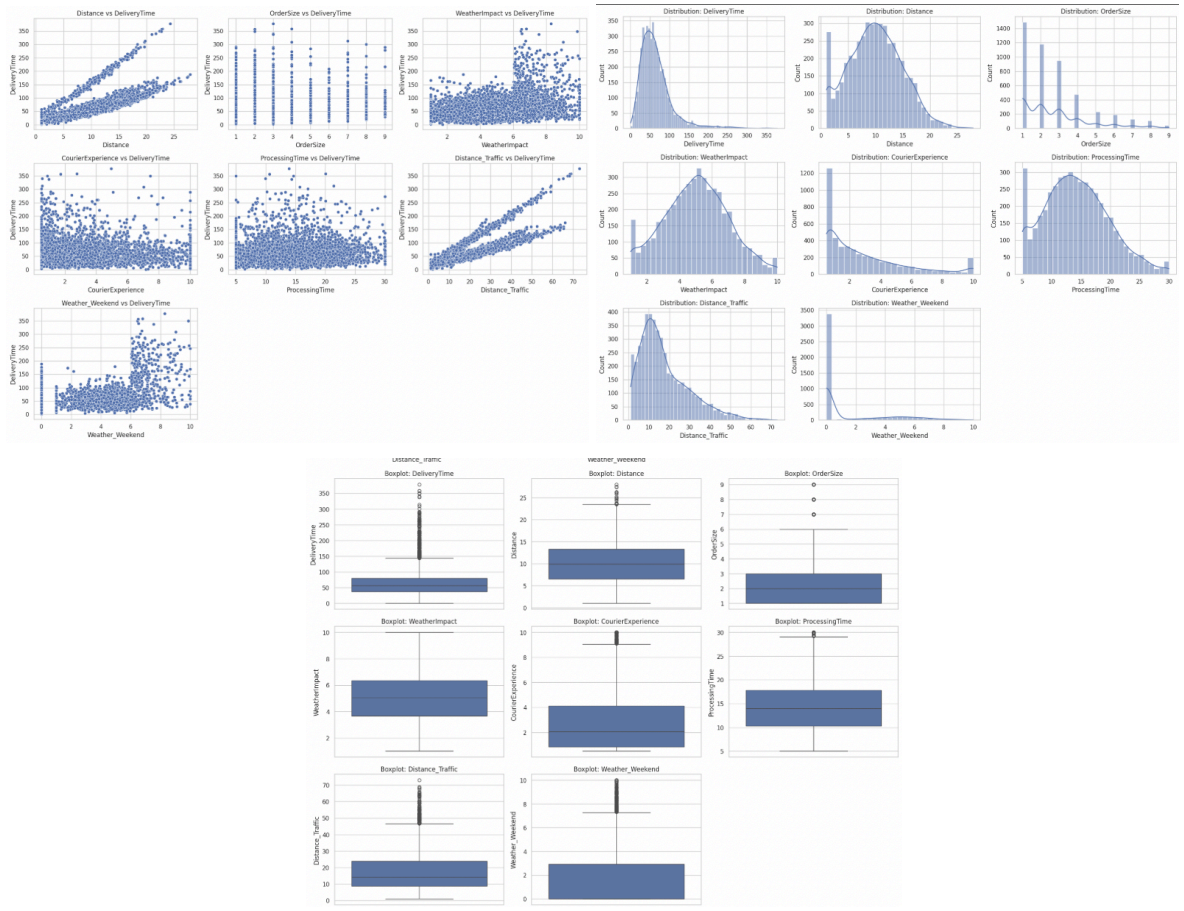


Figure 2. EDA Visualizations: Variable Relationships, Distributions, and Outliers

4.3 Correlation

The ***correlation matrix*** reveals key relationships between features and delivery time. The strongest positive correlations are observed with Distance_Traffic (0.81) and Distance (0.70), suggesting that longer and more congested routes significantly impact delivery duration. Moderate correlations are seen with TrafficCondition, Weather_Weekend, and WeatherImpact, indicating external conditions also play a role. Meanwhile, CourierExperience shows a negligible negative correlation, implying limited direct influence. These analyses guided the selection and engineering of predictive features.

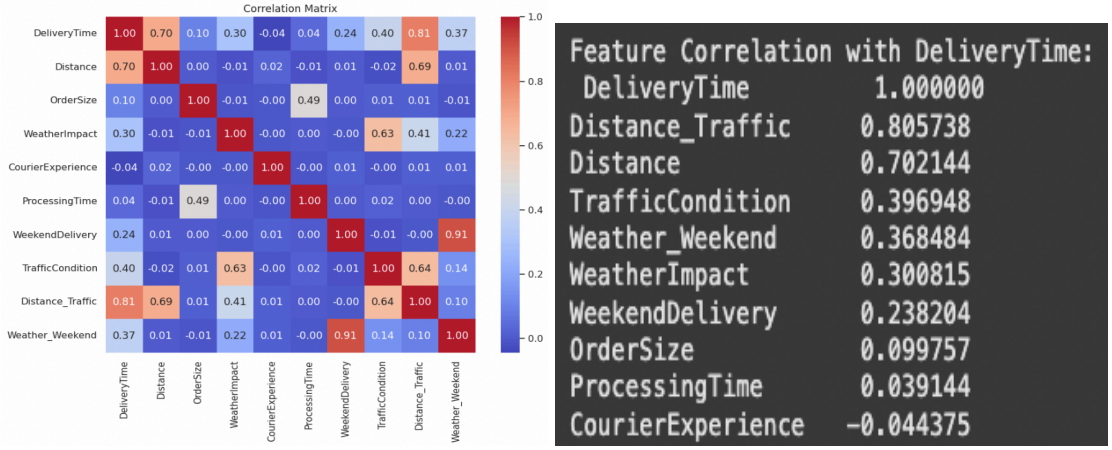


Figure 3. Correlation Matrix and rank with Delivery Time

4.4 Regime-based Visual Inspection

The features selected for modeling were not solely chosen based on their global correlation with DeliveryTime, but also due to their ***interpretability and observed behavior under different conditions***. Distance, TrafficCondition, and WeekendDelivery clearly formed distinct behavioral regimes and were strong predictors across all regimes. While OrderSize and CourierExperience showed weaker overall correlation, they appeared to impact the delivery curve structure within low-traffic contexts, justifying their inclusion. Finally, interaction features like Distance_Traffic and Weather_Weekend were introduced to help linear models capture the non-linear trends observed in the plots.

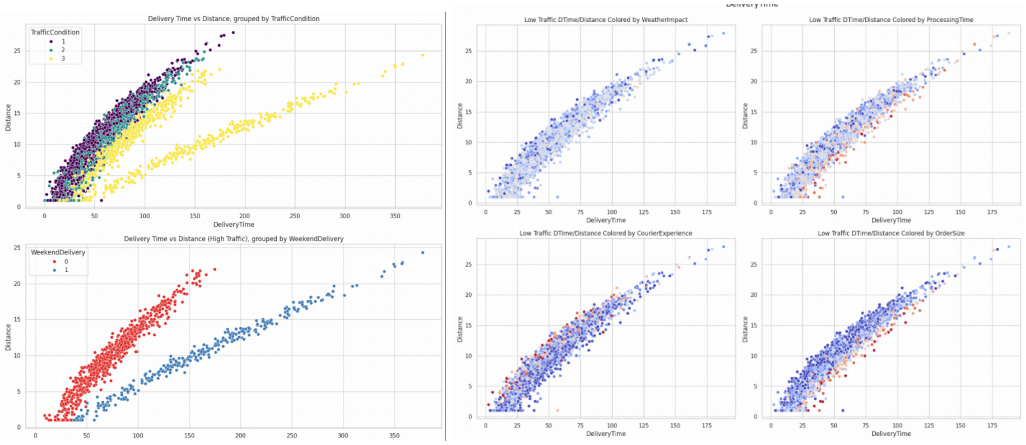


Figure 4. Regime Based visualization

5. Methodology

Following initial data cleaning and exploratory analysis, we proceeded with ***model development and evaluation***. Preprocessing steps included handling outliers in delivery times and engineering interaction terms to better capture combined effects between key variables. ***All features were standardized***, and polynomial transformations were applied to allow linear models to approximate nonlinear patterns.

We will then ***test different regression models*** such as Linear Regression, Ridge Regression, Lasso Regression, and K-Nearest Neighbors. These models were evaluated both ***globally and within traffic condition-based regimes*** to assess robustness under different operating scenarios. ***Hyperparameters were tuned*** using grid search and cross-validation, specifically alpha for Ridge and Lasso, and k for KNN.

Model performance was compared using common ***regression metrics*** (MSE, MAE, RMSE, and R^2). Based on these evaluations, we identified the top-performing models in both global and regime-specific contexts.

6. Feature Engineering

To enhance model flexibility and capture non-linear patterns in the data, several feature engineering techniques were applied. Two ***key interaction terms*** were introduced: Distance_Traffic and Weather_Weekend, designed to model the combined effects of congestion and weather disruptions. Given the convex relationship observed between Distance and DeliveryTime, ***polynomial features up to degree 2*** were included using PolynomialFeatures, enabling the model to capture curvature and variable interactions.

Standardization was applied to all numerical features to ensure compatibility with distance-based algorithms like KNN and to stabilize regularization in Ridge and Lasso regression. These engineered features were incorporated into a flexible sklearn pipeline, ensuring consistency across all models tested.

Figure 5. Quadratic Curve Fitting: DeliveryTime vs Distance

7. Model Selection

7.1 Linear Regression

Linear regression is a foundational statistical method used to model the relationship between a dependent variable and one or more independent variables. The model estimates a linear relationship of the form:

$$y = X\beta + \epsilon$$

In this implementation, we trained a ***linear regression model*** to predict delivery time using a set of engineered and original features. The workflow includes *standardizing all predictors and applying polynomial feature transformation (degree 2)* to capture non-linear relationships. The model was trained on 80% of the data and tested on the remaining 20%, with its performance evaluated using MSE, RMSE, MAE, and R^2 score. A scatter plot was also generated to visually compare the model's predictions against actual delivery times based on distance.

	Metric	Linear Regression
0	Mean Squared Error (MSE)	63.4717
1	Root Mean Squared Error (RMSE)	7.9669
2	Mean Absolute Error (MAE)	6.1172
3	R^2 Score	0.9656

Figure 6. Table of Evaluation Metrics for Linear Regression

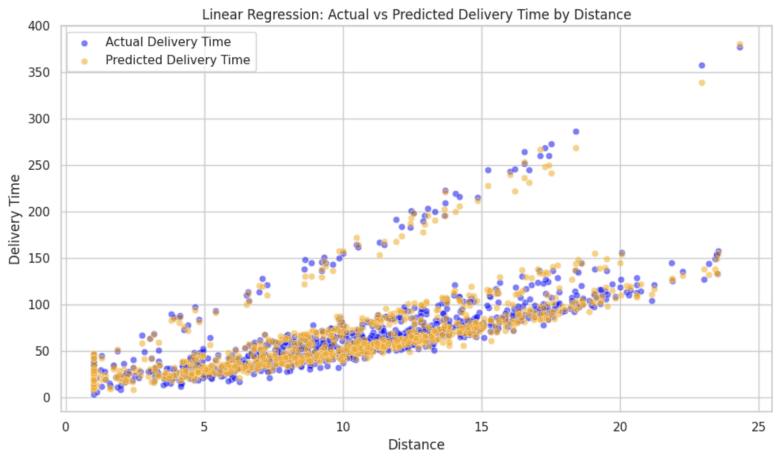


Figure 7. Scatter Plot showing actual and predicted delivery time by distance

The linear regression model showed ***strong performance*** in predicting delivery time. With an ***R^2 score of 0.9656*** and relatively ***low error metrics (MSE: 63.47, MAE: 6.12)***, the model effectively captured the underlying patterns in the data. The scatter plot confirms that predicted values closely follow the actual delivery trends across varying distances. This result gives a solid baseline and demonstrates that even a relatively simple model, when supported by effective feature engineering, can yield accurate and interpretable outcomes.

7.2 K-Nearest Neighbor

KNN is a non-parametric, instance-based algorithm that predicts the target of a query point by averaging the targets of the k closest observations in feature space, typically using Euclidean distance. It is highly flexible and capable of modeling complex, non-linear relationships without requiring assumptions about the underlying data distribution.

We trained and evaluated a KNN regression model using polynomial features and scaling. It tuned the number of neighbors (k) via GridSearchCV on both the full dataset and within traffic-based regimes. Performance is assessed using standard metrics and the optimal k is chosen per regime to better adapt to varying delivery conditions.

Metric	KNN (k=5)
Mean Squared Error (MSE)	74.5067
Root Mean Squared Error (RMSE)	8.6317
Mean Absolute Error (MAE)	6.5317
R² Score	0.9596

Dataset Group	Optimal K	MSE	MAE	R2 Score
TC = 1	6	37.8394	4.7407	0.9459
TC = 2	4	35.9509	4.6917	0.9514
TC = 3, WD = 0	4	39.2954	4.8570	0.9617

Figure 8. Evaluation table for KNN Global and Split Region

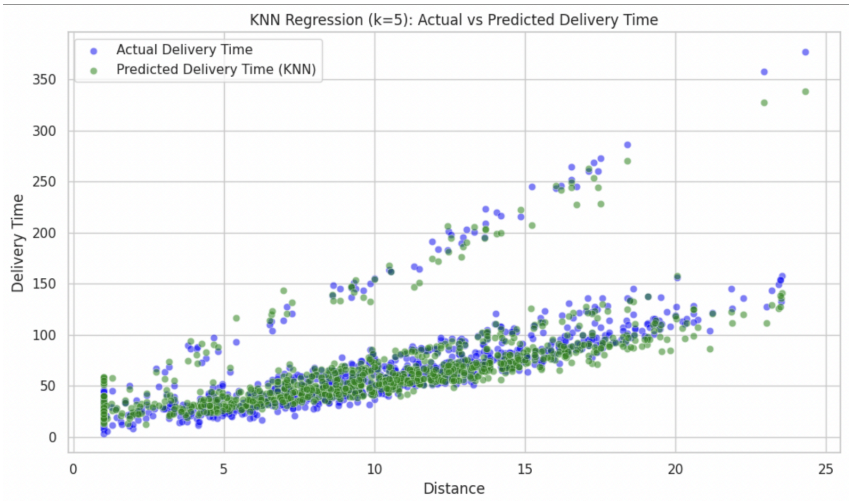


Figure 9. Scatter Plot showing actual and predicted delivery time by distance

The KNN model demonstrated reasonable predictive capability overall, with a global R^2 score of 0.9596 and MAE of 6.53. However, this performance was weaker than that of the regularized linear models, indicating it may not generalize as well. While the split-regime KNN models (based on traffic conditions) achieved improved accuracy—with R^2 scores up to 0.9617 and lower MAE, KNN still struggled with outliers and denser regions in the data. Its non-parametric nature and reliance on distance-based similarity make it sensitive to uneven data distributions and computationally inefficient at scale.

7.3 Ridge Regression

Ridge regression extends linear regression by adding an ℓ2-norm penalty to the loss function, giving

$$\text{Loss} = \sum (y_i - \hat{y}_i)^2 + \alpha \sum \beta_j^2$$

This technique helps reduce model variance and multicollinearity by shrinking coefficients without eliminating them. Ridge is especially appropriate for this task given the high correlation among engineered features and the need to retain all predictors for interpretability. (Hoerl, A. E., & Kennard, R. W., 1970)

In this implementation, we applied Ridge Regression to predict delivery time using both global and regime-specific models. The data was first standardized and expanded with second-degree polynomial features. We used GridSearchCV to identify the optimal regularization strength (alpha) through 3-fold cross-validation. The model was evaluated using standard regression metrics, both on the overall dataset and within distinct traffic regimes, allowing us to assess how Ridge adapts across different delivery contexts.

Metric	Ridge Regression
MSE	63.6332
RMSE	7.9770
MAE	6.1149
R² Score	0.9655,

Dataset Group	Optimal Alpha	MSE	MAE	R² Score
TC = 1	20.4336	26.5870	4.0204	0.9620
TC = 2	9.2367	25.3291	4.0374	0.9657
TC = 3, WD = 0	4.1753	23.0995	3.8923	0.9775

Figure 10. Evaluation table for Ridge Regression Global and Split Region

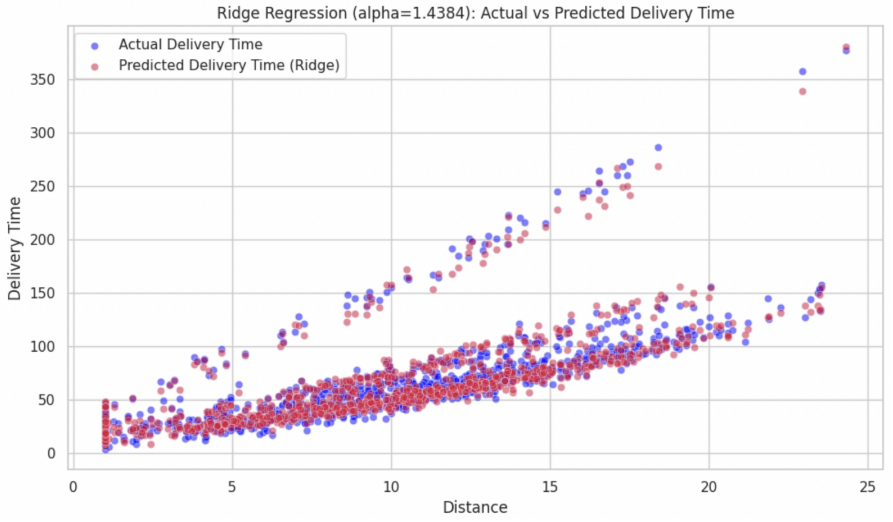


Figure 11. Scatter Plot showing actual and predicted delivery time by distance

The Ridge regression models trained on regime-based subsets of the delivery data demonstrated consistently high performance, with R^2 scores ranging from 0.9620 to 0.9775. This indicates that a significant proportion of the variance in delivery time can be explained by the selected features and engineered interactions. The best performance was observed in the TC = 3, WD = 0 group, high traffic on weekdays, where delivery dynamics are likely more consistent and less impacted by erratic human behavior or unusual scheduling.

7.4 Lasso Regression

Lasso regression introduces an $\ell 1$-norm penalty to the objective function, giving

$$\text{Loss} = \sum (y_i - \hat{y}_i)^2 + \lambda \sum |\beta_j|$$

This promotes sparsity by shrinking some coefficients entirely to zero. This embedded feature selection makes Lasso valuable in high-dimensional datasets or when identifying the most influential predictors is important. (Tibshirani, R., 1996)

We implement Lasso Regression with polynomial features and standardization, using grid search to find the best alpha. The model is evaluated on both the full dataset and traffic-based regimes using common metrics. Lasso's regularization helps reduce overfitting and simplifies the model by shrinking weaker coefficients.

Metric	Lasso (alpha=0.0062)
Mean Squared Error (MSE)	63.4101
Root Mean Squared Error (RMSE)	7.9630
Mean Absolute Error (MAE)	6.1029
R² Score	0.9656

Dataset Group	Optimal Alpha	MSE	MAE	R2 Score
TC = 1	0.1274	26.8639	4.0352	0.9616
TC = 2	0.1274	25.6491	4.0587	0.9653
TC = 3, WD = 0	0.1274	23.3491	3.9202	0.9772

Figure 12. Evaluation table for Lasso Regression Global and Split Region

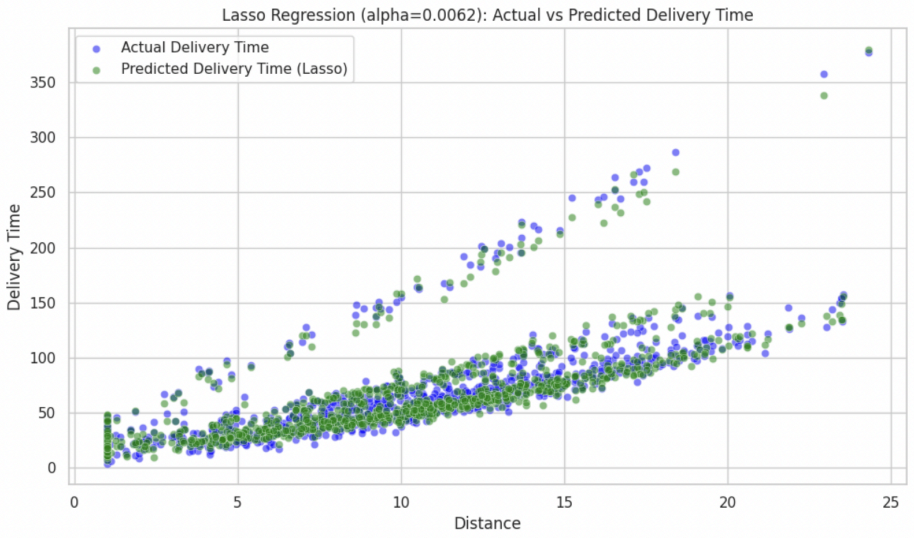


Figure 13. Scatter Plot showing actual and predicted delivery time by distance

The Lasso regression model performs remarkably well on the dataset, achieving a high R^2 score of 0.9656. This implies that the model is able to explain over 96% of the variance in delivery times using the engineered and polynomial features. The low RMSE and MAE indicate that predictions are not only accurate but also consistent. The selected alpha of 0.0062 is relatively low, suggesting only mild regularization, enough to reduce overfitting without excessively shrinking the coefficients of informative features. Compared to Ridge, which distributes penalty more evenly, Lasso can zero out unimportant coefficients entirely,

potentially enhancing interpretability in future steps. However, some convergence warnings hint that certain alpha values struggled to minimize the loss effectively.

8. Final Model

Ridge Regression was chosen as the final model due to its consistent and balanced performance across both global and regime-specific evaluations. While Linear Regression performed well, it lacked regularization and risked overfitting with polynomial terms. KNN showed strong performance under regime splits but suffered from higher error globally and was sensitive to local noise. Lasso, although competitive, introduced more variance in predictions and overly shrunk some coefficients, which risked discarding useful information. Ridge offered the best trade-off, capturing complex, non-linear relationships through polynomial features while maintaining coefficient stability via L2 regularization. Its superior regime-level R^2 scores (up to 0.9775) and robustness under multicollinearity further supported its selection.

Model	MSE	RMSE	MAE	R^2 Score
Linear Regression	63.4717	7.9669	6.1172	**0.9656**
KNN (k = 5)	74.5067	8.6317	6.5317	0.9596
Ridge Regression	63.6332	7.9770	6.1149	**0.9655**
Lasso Regression	63.4101	7.9630	6.1029	**0.9656**

Figure 14. Global Evaluation

Dataset Group	Model	Optimal Param	MSE	MAE	R^2 Score
TC = 1	KNN	k = 6	37.8394	4.7407	0.9459
	Ridge Regression	α = 20.4336	26.5870	4.0204	0.9620
	Lasso Regression	α = 0.1274	26.8639	4.0352	0.9616
TC = 2	KNN	k = 4	35.9509	4.6917	0.9514
	Ridge Regression	α = 9.2367	25.3291	4.0374	0.9657
	Lasso Regression	α = 0.1274	25.6491	4.0587	0.9653
TC = 3, WD = 0	KNN	k = 4	39.2954	4.8570	0.9617
	Ridge Regression	α = 4.1753	23.0995	3.8923	**0.9775**
	Lasso Regression	α = 0.1274	23.3491	3.9202	0.9772

Figure 15. Region Evaluation

8.1 Model Tuning

To refine the Ridge Regression model, we conducted hyperparameter tuning not only on the regularization strength (alpha) but also on the polynomial degree of the features.

	Degree	Best Alpha	MSE	RMSE	MAE	R^2 Score
0	1	2.5595	334.9503	18.3016	12.3899	0.8185
1	2	0.8286	63.5667	7.9729	6.1158	0.9656
2	3	11.5140	26.9396	5.1903	4.1614	0.9854

Figure 16. Performance across 3 different polynomial degrees

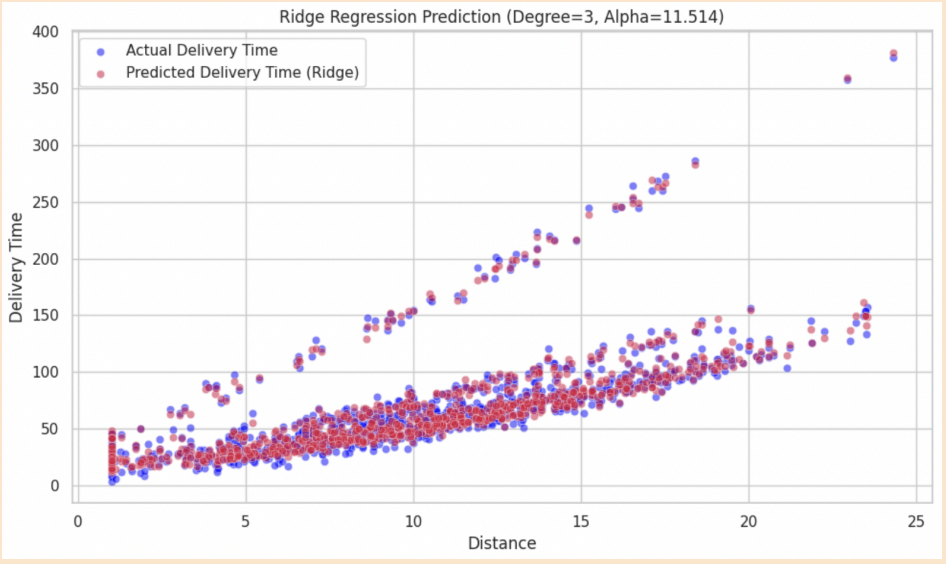
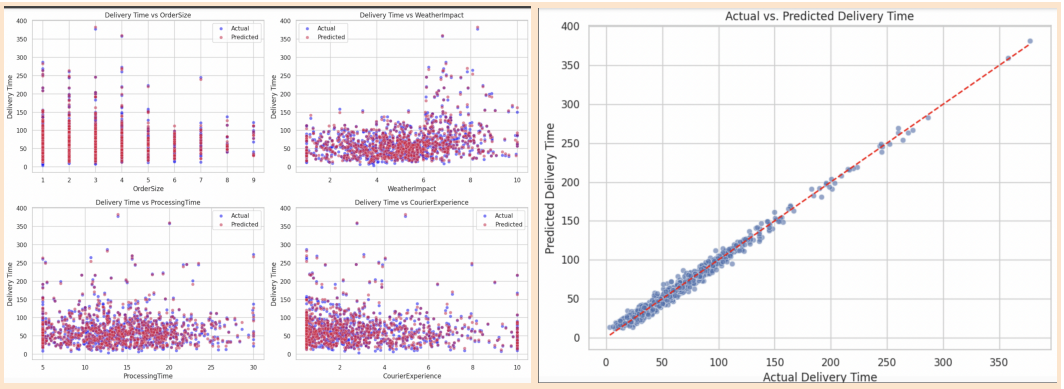


Figure 17. Scatter Plot showing actual and predicted delivery time by distance

The final selected model, which is a Ridge Regression with polynomial features of degree 3 , showed a superior performance in capturing the underlying complexity of delivery time predictions. With an R^2 score of 0.9854, it significantly outperformed simpler models while maintaining robustness against overfitting. The incorporation of higher-order interactions proved essential in modeling nonlinear relationships between features. Ridge's ability to retain all predictors, stabilize coefficients under multicollinearity, and generalize effectively to unseen data justifies its use as the most balanced approach for this task. (Ng, A.,2021)

8.3 Residual Plots



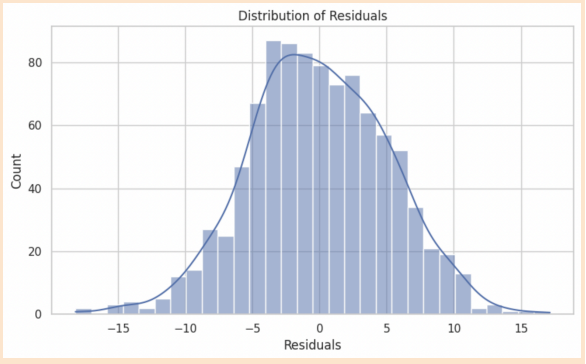


Figure 18. Residual plots for the final model

9. Conclusion

The final model selected for delivery time prediction is a R***idge Regression model with polynomial features of degree 3***. This choice was informed by its strong performance across both global and segmented regime evaluations. Ridge outperformed alternatives like KNN and Lasso by consistently producing lower error rates and higher R^2 values, particularly within high-traffic weekday conditions. Its use of L2 regularization made it particularly effective in handling the multicollinearity introduced by polynomial and interaction terms, allowing it to generalize well without overfitting. The final tuned model achieved a R^2 score of 0.9854, indicating that it explains over 98% of the variance in delivery times, a significant improvement over simpler linear baselines.

Looking ahead, this model serves as a strong foundation for operational deployment in delivery time forecasting. However, future improvements could involve the use of ensemble methods or gradient boosting models, which may further enhance performance on complex patterns or outliers. We can also incorporate external features such as real-time traffic data, courier shift length, or regional weather conditions that could yield even greater predictive power.

10. References

Nguyen, V., et al. (2023). A predictive framework for last-mile delivery routes considering customer satisfaction. *Computers & Industrial Engineering*, 180, 109346. https://www.sciencedirect.com/science/article/pii/S0360835223003704?via%3Dihub

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55–67. https://doi.org/10.2307/1267351

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x