**3. Model Development**

To estimate pickup and delivery ETAs accurately in a last-mile logistics setting, separate machine learning pipelines were created and optimized for each stage. The approach focused on model robustness, interpretability, and scalability.

**3.1 Models Explored**

* **Random Forest Regressor**  
  Initially used to evaluate feature importance and eliminate non-contributing features such as pickup\_weekend, city, and aoi\_type.
* **LightGBM Regressor**  
  Final model of choice due to its speed and superior performance on tabular datasets. It was trained separately for pickup and delivery datasets with 100 estimators.
* **Linear SVR**  
  Applied on standardized features to serve as a fast linear baseline. Though performant on some tasks, it showed signs of overfitting in certain configurations.

**3.2 Feature Engineering**

Domain-specific features were created to enrich the model inputs:

* **Pickup Data Features**  
  location\_sum = lng + lat  
  time\_difference = pickup\_hour - ETA\_pickup  
  region\_aoi\_product = region\_id × aoi\_id  
  ETA\_ratio = ETA\_pickup / (pickup\_hour + 1)
* **Delivery Data Features**  
  Included: location\_sum\_delivery, time\_difference\_delivery, and region\_aoi\_product\_delivery

**3.3 Dimensionality Reduction**

PCA was applied to reduce feature dimensionality while preserving more than 90% variance using 5 components. This enhanced stability while reducing model complexity.

**3.4 Preprocessing Summary**

* Unnecessary columns removed based on feature importance
* **Standardization** applied via StandardScaler (especially for SVR)
* **80/20 split** used for train-test partitioning
* Separate pipelines maintained for pickup and delivery models

**4. Model Evaluation and Results**

The trained models were evaluated using **Mean Squared Error (MSE)** and **R² Score**, providing insight into predictive performance.

**4.1 Pickup Results**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error** | **R² Score** |
| LightGBM | 0.0186 | 0.9812 |
| LinearSVR | 1.0397e-23 | 1.0000 |
|  |  |  |

***Note:* LinearSVR reported near-perfect accuracy, which may suggest overfitting or numerical instability due to feature scaling.**

**4.2 Delivery Results**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean Squared Error** | **R² Score** |
| LightGBM | 0.3506 | 0.6121 |
| LinearSVR | 1.8668e-23 | 1.0000 |

***Insight:* LightGBM provided more generalizable performance, especially on the delivery dataset, which had greater variance in its features.**

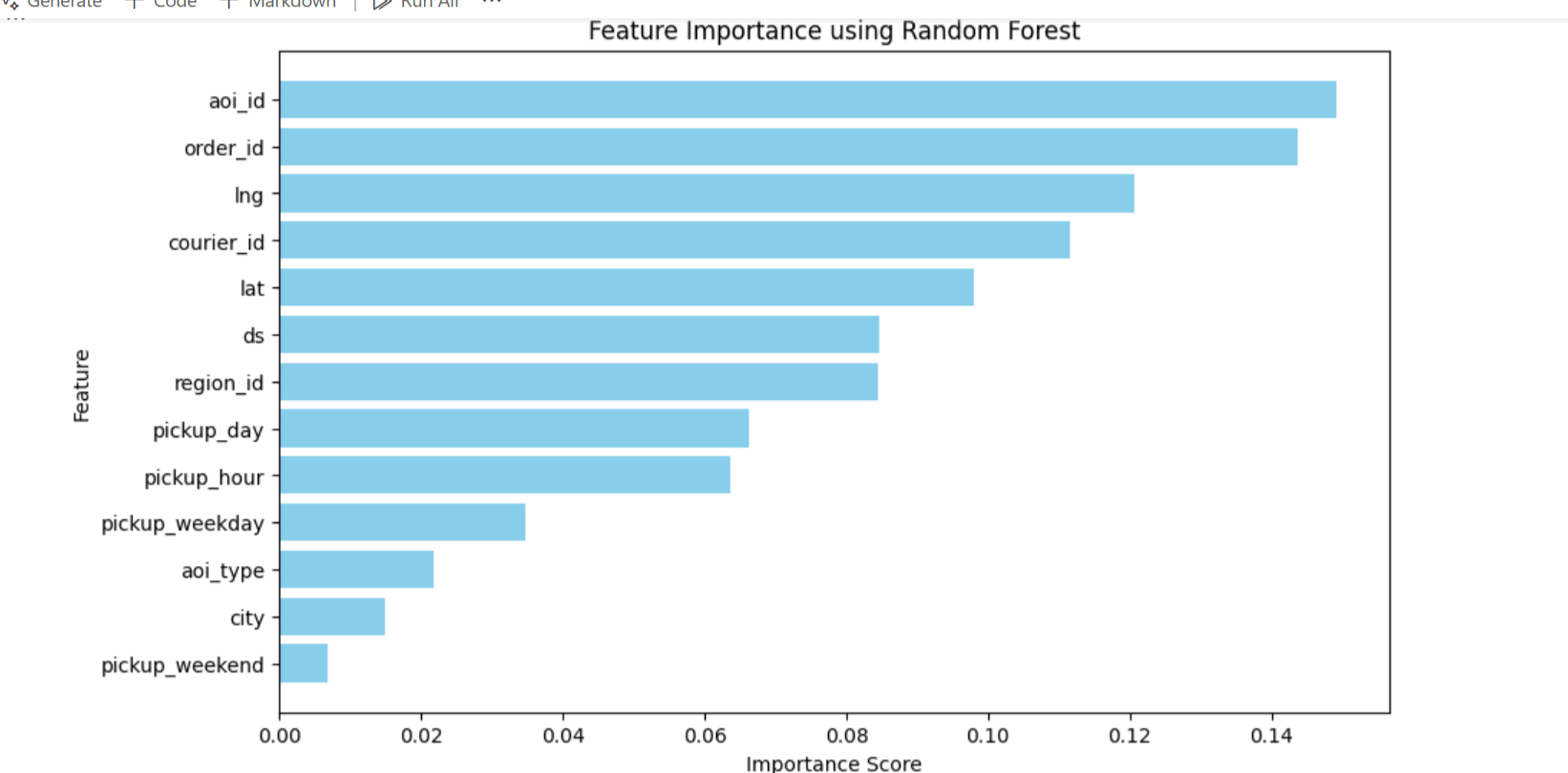
**4.3 Summary Observations**

* LightGBM outperformed across both tasks with consistent and explainable results
* Pickup predictions were more accurate than delivery, possibly due to less variability in pickup features
* LinearSVR’s performance was unusually high, suggesting a need to further validate data preprocessing steps

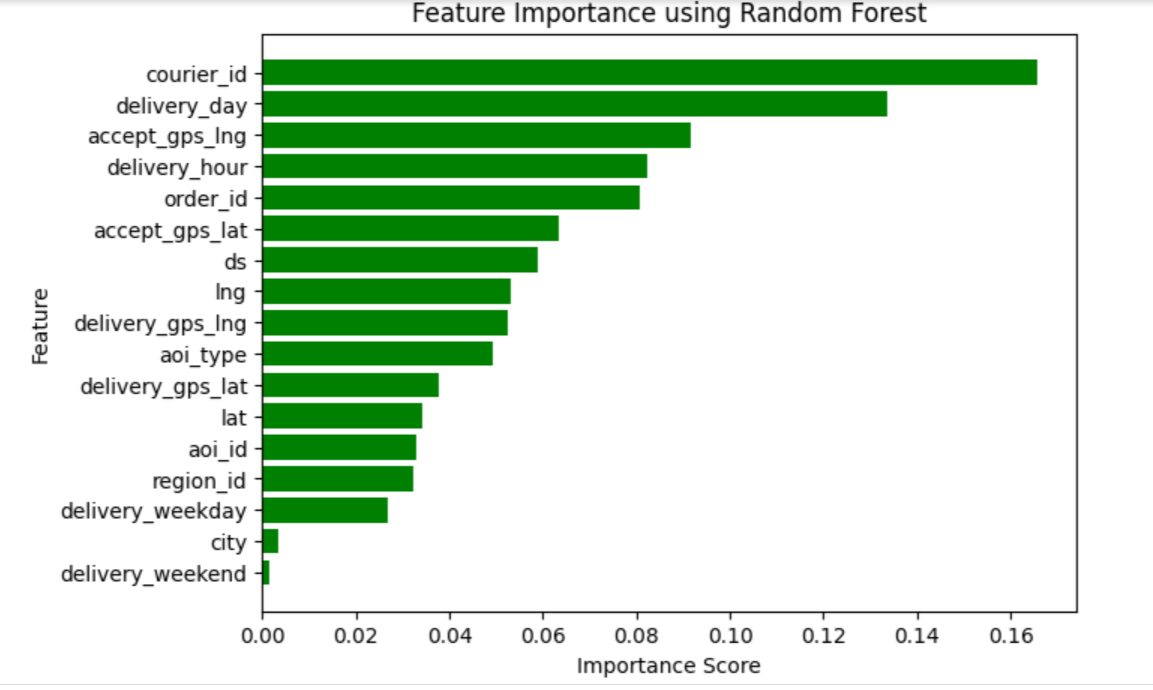
**4.4 Visualizations**

* **Random Forest Feature Importance plots**

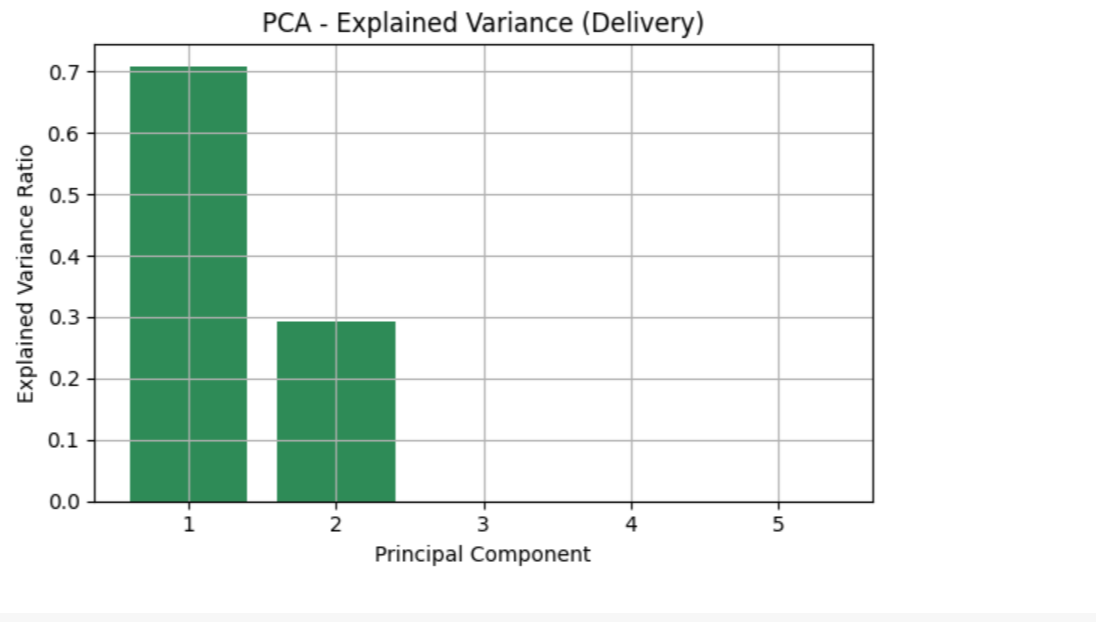
**Pickup**

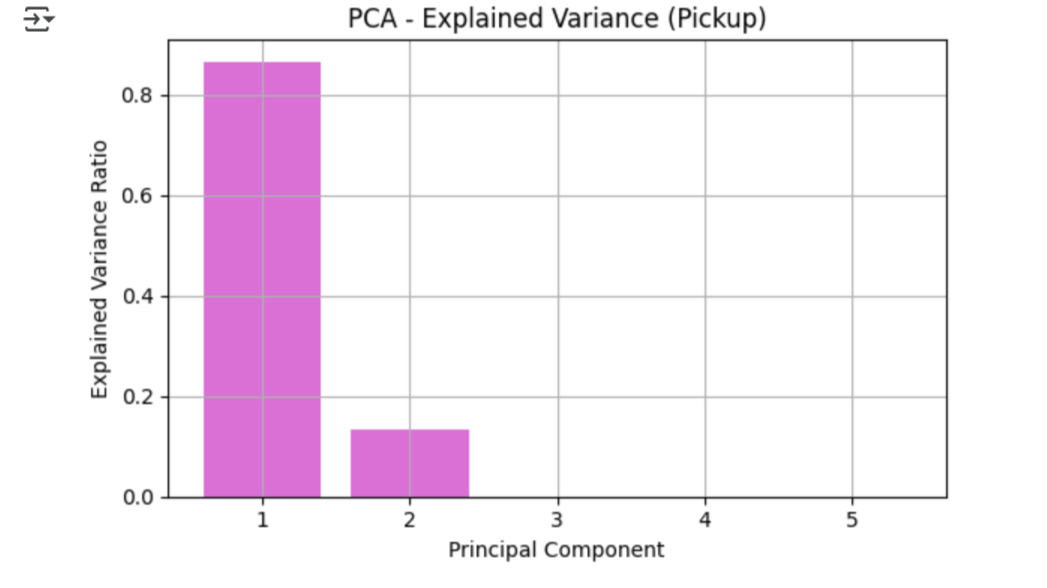


**Delivery**



* **PCA Explained Variance chart**

arts



* **Predicted vs. Actual ETA scatter plots**

