IEOR 142 Project Report: Predicting Package Delivery Delays
Team Members: Christina Kim, Matthew Lim, Ryan Stevenson, Brandon Concepcion, Mark Barranda

---

Introduction: Motivation and Impact

On-time package delivery is a cornerstone of success in logistics and supply chain operations. Delays can disrupt operations, drive up costs, and lead to customer dissatisfaction, ultimately threatening business relationships and revenue. Having a reliable system to predict delivery delays offers significant potential for improvement in operational efficiency and customer satisfaction.

Predictive insights can help logistics companies proactively allocate resources, prioritize at-risk shipments, and enhance overall communication with customers. For instance, providing early warnings about potential delays allows for better schedule management and more transparent communication, which can reduce frustration and maintain trust. Predictive analytics also enables companies to deploy additional personnel or vehicles where necessary, ensuring smooth operations and better service reliability.

Data Source and Preprocessing[1][2]

The dataset initially included 6,880 rows and 32 columns related to package deliveries, including booking details, locations, transportation distances, timestamps, and delivery outcomes (on-time or delayed) in India. It was extracted specifically from Kaggle, but after a deep dive into the dataset description (and even a little bit of an investigation into the dataset owner's LinkedIn), we were unable to exactly pinpoint how and where this data was collected.[3]

After looking at all of the columns and looking at the number of null values in each of them, we decided to remove unnecessary categorical columns, such as driver information, booking ID, and customer details, and keep necessary columns, such as dates, location, and estimated times of arrival (ETA). These columns were converted to datetime format, and additional features like booking year, month, and day were extracted. Latitude and longitude pairs were split into separate columns for better usability. A binary column, is_delayed, was created to consolidate the on-time and delay columns, marking packages as delayed (1) or on-time (0). Categorical features, such as vehicleType, Market/Regular, and GpsProvider, were encoded numerically using one-hot encoding or label encoding since we believed these were

---

[1] Link to Entire Google Folder:
https://drive.google.com/drive/folders/1cmhnxX8Eno8Dvkw7NIYNdmtjn_KruUKv?usp=sharing
[2] Link to Code:
https://drive.google.com/drive/folders/1lH-JPpPGXjy4WD8Jwrw_El8qe6GZcUvG?usp=drive_link
[3] Link to Original Data:
https://www.kaggle.com/datasets/ramakrishnanthiyagu/delivery-truck-trips-data?resource=download

crucial factors that could help determine delivery status. Finally, the dataset was filtered to remove rows with inconsistencies or missing values. After preprocessing, the dataset was reduced to 6,141 rows and 80 columns, ensuring it was clean, consistent, and ready for analysis and modeling. Finally, the dataset was split into an 80/20 training and testing split to assess model performance on unseen data. These preprocessing steps helped prepare a comprehensive and clean dataset for our analysis.

Modeling Procedures

We implemented three machine learning models to predict delivery delays: Linear Discriminant Analysis (LDA), Decision Tree Classifier (CART), and Gradient Boosting Classifier (GBC). Each model was selected for its unique approach and strengths in classification tasks.

Model 1: Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a classification method that works by finding the best linear combination of features to separate our two classes of data. After initializing and training the LDA model, it predicts the labels for the test set and is evaluated against the true test labels. It achieved an accuracy of 88.69%, with a True Positive Rate (TPR) of 97.28% and a False Positive Rate (FPR) of 30.36%.

Model 2: Cross-Validated Decision Tree Classifier (CART)

Classification and Regression Trees (CART), a tree-based algorithm that builds decision trees by recursively splitting the data, was selected for its interpretability and ability to capture non-linear relationships. We used 5-fold cross-validation with GridSearchCV to perform cross-validation and select the optimal ccp_alpha value. We first created a grid of 201 equally spaced ccp_alpha values between 0 and 0.10 and a DecisionTreeClasifier set to a random state of 2024 for reproducibility. Next, we used GridSearchCV with a 5-fold cross-validation strategy to find the value of ccp_alpha that maximizes accuracy. This approach splits the training data into 5 subsets, training the DecisionTreeClassifier on 4 subsets and validating on the 5th subset, repeating the process for all folds. After training the model, it predicts the labels for the test set and is evaluated against the true test labels. The best ccp_alpha value from the cross-validation process is 0.0010, which gave us an accuracy of 93.00%, a TPR of 97.16%, and an FPR of 16.23%.

Model 3: Gradient Boosting Classifier (GBC)

The Gradient Boosting classifier (GBC) builds a series of trees sequentially, where each tree corrects the errors made by the previous one and minimizes loss. In the provided code, a gradient boosting classifier is initiated with specific hyperparameters, including 200 estimators, a maximum of 10 leaf nodes per tree, and a fixed random state of 2024 for reproducibility. After training the model, it predicts the labels for the test set and is evaluated against the true test labels. It achieved the highest accuracy of 93.25%, a TPR of 95.99%, and an FPR of 12.83%.

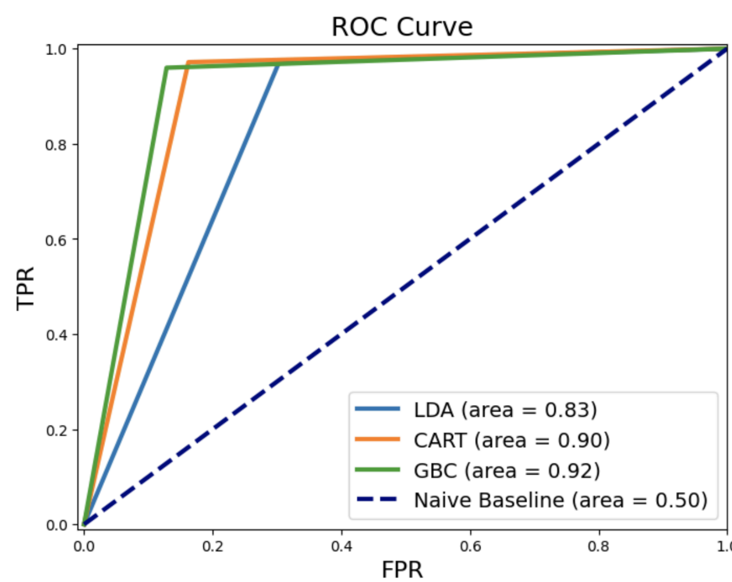
Discussion and Results

Performance Metrics

Accuracy measures the overall correctness of a model's predictions, providing a general sense of how well it performs. The True Positive Rate (TPR) highlights the model's ability to correctly identify actual delays, reflecting its sensitivity to true outcomes. Meanwhile, the False Positive Rate (FPR) assesses the proportion of non-delays that are incorrectly classified as delays, offering insight into the model's tendency for false alarms.

Comparison of Models[4] [5]

	Model	Accuracy	TPR	FPR
0	LDA	0.886900	0.972845	0.303665
1	CART	0.930024	0.971665	0.162304
2	Gradient Boosting Classifier	0.932465	0.959858	0.128272



Based on the table and ROC curve above, the three models - LDA, CART, and GBC- demonstrate varying levels of performance. LDA, while achieving a TPR of 97.28%, has the lowest accuracy (88.69%) and the highest FPR (30.37%), along with the lowest ROC AUC of 0.83. CART has a higher accuracy (93.00%) and ROC AUC (0.90) with a slightly lower TPR (97.17%) compared to LDA but exhibits a much lower FPR (16.23%). The GBC model has the highest accuracy (93.25%) and the lowest FPR of 12.83% while maintaining a strong TPR of

---

[4] ROC curves for LDA, CART, and GBC

[5] Performance comparison of machine learning models based on Accuracy, TPR, and FPR.

95.99%. It also achieves the highest ROC AUC of 0.92, indicating superior overall performance in distinguishing between delayed and on-time packages.

Synthesis of Results within the Application Context

Linear Discriminant Analysis (LDA) offers the advantage of computational efficiency; however, it was our worst performing model as it had the lowest accuracy and ROC AUC and the highest FPR. Our CART and GBC models had very similar accuracy, ROC AUC, and TPR, but the GBC model produced the lowest FPR. In the context of the application, we decided that minimizing false negatives (i.e., predicting a package is on time when it is actually delayed) is more critical to minimize than false positives (i.e., predicting a package is delayed when it is actually on time). Therefore, while GBC achieves slightly better accuracy and ROC AUC and a lower FPR, CART's higher TPR indicates that it is less likely to miss delays (i.e., has fewer false negatives). Since false negatives are the primary concern, CART is our best model as it offers a better trade-off for this specific priority.

Conclusion and Future Work

This project demonstrates the value of using machine learning to predict package delivery delays. Out of the models we tested, the Cross-Validated Decision Tree Classifier model stood out as the best performer, offering a high True Positive Rate while maintaining a high accuracy and low False Positive Rate. Its strong performance makes it a promising option for use in logistics, especially in situations where timely and accurate predictions for on-time or delayed packages are essential.

Some limitations to consider include our models could be biased as we dropped null values and potentially important columns. For instance, Dropping columns related to drivers eliminates the possibility of specific drivers' impact on delivery delays. In addition, Encoding a large number of one-hot encoded variables (e.g., vehicleType, GPsProvider) could increase the dimensionality of the dataset, risking overfitting in complex models like Gradient Boosting.

Looking ahead, there are several ways to build on this work. Adding external data, like weather conditions or traffic patterns, could make the predictions even more accurate by accounting for additional factors that impact delivery times. Testing the models in real-world logistics settings would also provide a better sense of how well they perform in practice and help identify areas for improvement. Finally, combining multiple models into an ensemble could capitalize on the strengths of each, making the system more robust and reliable. These steps would make the predictive system even more effective at optimizing and predicting package delivery.