Cukurova University
Journal of
Natural & Applied Sciences

Original Paper

Development of Cargo Delivery Time Prediction Models

Selim Hanedar[1], Ceren Ulus[2], M. Fatih Akay[2]

[1]Trendyol, Department of Technology, İstanbul, Turkey
[2]Çukurova University, Department of Computer Engineering, Adana, Turkey

ORCID IDs of the authors: S.H. 0009-0004-7037-1840; C.U. 0000-0003-2086-6381; M.F.A. 0000-0003-0780-0679.

Cite this article as: Hanedar, S., Ulus, C., Akay, F.A. (2024). Development of Cargo Delivery Time Prediction Models, Cukurova University Journal of Natural & Applied Sciences 3(1): 31-35.

Abstract

E-commerce stands out as the sales form with the fastest growth momentum with high sales volumes. Managing sales volumes efficiently is of great importance in maximizing customer satisfaction. By accurately predicting delivery times, effec-tive logistics optimization is achieved and customers are informed about how long it will take for their cargo to be delivered. In this study, it is aimed to develop cargo delivery time prediction models with machine learning-based Categorical Boosting (CatBoost), Decision Tree (DT), Extreme Learning Machine (ELM), Light Gradient Boosting Machine (LightGBM) and Support Vector Machine (SVM). The 5113-row dataset contains delivery history information for the 16-month period between February 14, 2019, and June 13, 2020. The performance of the developed models has been evaluated using Mean Absolute Percentage Error (MAPE) by utilizing 5-fold cross-validation on the dataset. The results show that the models developed using SVM exhibited the most successful prediction performance.

Keywords: Delivery Time Prediction, Machine Learning, E-Commerce

1. Introduction

E-commerce is attracting attention as a fast-growing sector and is causing radical changes in the world of commerce. In the first half of 2023, the share of trade in total trade increased to 19.1%, which illustrates the growth potential of the sector [1]. In this context, the development of e-commerce offers small and medium-sized enterprises the opportunity to compete and grow in global markets. Establishing themselves in this highly competitive market allows companies to expand their cus-tomer base, increase their market share, strengthen brand awareness and improve the customer experience through continu-ous innovation. However, to thrive in the highly competitive e-commerce environment, customer satisfaction must be max-imized. Providing a user-friendly website, solution-oriented customer service and fast and reliable delivery are strategic measures that increase customer satisfaction.

In recent years, fast delivery has become increasingly important in the modern e-commerce world as customer expectations shift towards instant gratification, offering competitive advantages and increasing customer satisfaction. The proliferation of services such as Amazon Prime, Hepsiburada and Trendyol has enabled consumers to view fast delivery as the norm. Busi-nesses that offer fast delivery can stand out in a competitive market, expand their customer base and build a loyal customer base. The combination of these factors has led to fast delivery, which is becoming an essential element in the e-commerce industry [2].

Research shows that customer satisfaction is 20-30 higher for companies that offer fast delivery than others. In addition, it has been observed that the likelihood of customers making repeat purchases increases by up to 70% for companies that offer fast delivery. It has been determined that companies with customer profiles that share their fast delivery experiences on social media have lower complaint and return rates than other companies [3].

Cargo companies focus on basic components such as shortening delivery times, reducing costs, and delivering products to customers without damage. These components are evaluated using data received from cargo companies and are reflected in agreements made with parameters such as regional deadlines and the number of packages for regional and non-regional de-liveries. Although these standards are shaped within a certain framework, differences in delivery times may occur due to the product diversity and customer profile of each brand. These differences are of great importance as they can cause the loss of potential

Address for Correspondence:
Ceren Ulus, e-mail: f.cerenulus@gmail.com

Received: May 24, 2024
Accepted: June 3, 2024

customers. To overcome these differences, brands must constantly optimize their logistics processes, increase operational efficiency, and provide flexibility to quickly respond to customer demands.

In this context, delivery time estimation stands out as a strategic tool that enables the optimization of logistics processes. This study aims to predict delivery time using machine learning-based methods. For this purpose, prediction models have been developed using CatBoost, DT, ELM, LightGBM and SVM.

This study is organized as follows: Section 2 includes relevant literature. Methodology is presented in Section 3. Section 4 presents delivery time prediction models. Results and discussion are given in Section 5. Section 6 concludes the paper.

2. Literature Review

[4] presented the Marine Predators Algorithm (MPA) for Shipment Status Time Prediction (STP). Initially validated on nu-merical benchmark problems, STPMPA surpassed all algorithms, demonstrating superior performance. Employing Extreme GB optimizers, STPMPA outperformed compared optimization algorithms for the STP problem, illustrating its efficacy in producing efficient forecasts for real-time systems. [5] proposed the Heterogeneous HyperGraph Neural Network (H2GNN) model for estimating package arrival time. This model introduced an order heterogeneous hypergraph, where hyperedges represent orders and nodes represent order attributes. Leveraging Hyper-Graphsage, H2GNN extended hypergraph learning to large-scale e-commerce data, enabling informative representations of packages by preserving both structure-based infor-mation learned by hypergraph and feature-based information captured by Transformer, ultimately outperforming baseline methods. [6] introduced the Knowledge Distillation Graph (KDG) NN based package Estimated Time Arrival (ETA) pre-diction model, termed KDG-ETA. This model utilized information densification to condense past trip knowledge into Origin-Destination pair embeddings, thereby combining context embeddings from the feature extraction module with com-prehensive trip information. An adaptived attention module was incorporated for delivery time estimation, resulting in a 3.0% to 39.1% reduction in Mean Absolute Error. [7] proposed Heterogeneous Tasks Aware Package Pick-Up Time (HTAPT), a package pickup time prediction framework comprising a pre-trained arrival time prediction module and a pickup time and route prediction module. HTAPT demonstrated improved forecast accuracy by up to 10% compared to state-of-the-art methods. [8] aimed to develop a novel machine learning forecasting method by integrating time series data features and the Adaptive Neuro-Fuzzy Inference System. Their proposed framework established a four-stage operations model, facilitating systematic forecasting of real-time e-order arrivals at distribution centers to enhance third-party logistics providers' efficiency in handling hourly e-order arrival rates. [9] introduced the Graph Structure Learning-based Quantile Regression (GSL-QR) model for e-commerce ETA prediction. GSL-QR dynamically updated spatial and temporal order relationships using graph structure learning optimized for ETA forecasting tasks. This multi-objective quantile regression model guaran-teed both forecasting accuracy and order fulfillment rate, demonstrating superiority over baseline models. [10] presented the Inductive Graph Transformer (IGT), which leverages raw feature information and structural graph data to estimate package delivery time. IGT's discrete pipeline-trained transformer captured various information from raw features and dense embed-ding data as a regression function, thus outperformed state-of-the-art methods in delivery time prediction. [11] aimed to assess the effectiveness of e-scooter sharing services in the delivery of postal services in Turkey by estimating the delivery time and energy cost of e-scooter vehicles for distributing mail or goods. Random Forest (RF), GB, K-Nearest Neighbor, and NN were among the machine learning algorithms used. They discovered that the GB algorithm had better prediction per-formance for energy cost and delivery time, demonstrating the financial and environmental benefits of e-scooter delivery vehicles over conventional vehicles. [12] developed a data-driven framework for managing customer expectations and im-proving satisfaction. Utilizing tree-based models, their approach generated distribution estimated and employed a new quan-tile regression forest partitioning rule with a cost-sensitive decision-making structure, achieving superior prediction perfor-mance compared to traditional commitment times. [13] proposed the Graph2Route model, a dynamic spatial-temporal graph-based model leveraging underlying graph structure and features for encoding and decoding. Demonstrating superiority over existing models, Graph2Route captured evolving relationships between different problem instances, offering promising prospects for route optimization.

3. Methodology

3.1. Support Vector Machines

SVM is a supervised learning technique utilized in high-dimensional space to discern a hyperplane effectively segregating data points, enabling their analysis for both regression and classification purposes. In the pursuit of segregating distinct catego-ries of data points, SVM may utilize various hyperplanes. The principal objective of the algorithm is to ascertain the hyper-plane maximizing the margin, defined as the largest separation distance between data points belonging to different classes [14].

3.2. Light Gradient Boosting Machine

The LightGBM, a gradient-boosting framework, leverages tree-based learning methods. Distinguished from its predecessors, it offers several advantages, including accelerated processing, minimal RAM usage, support for parallel GPU learning, and heightened computational velocity. Employing a histogram-based approach, LightGBM efficiently segregates discrete-value continuous variables to mitigate processing costs. The granularity of divisions required for computation significantly impacts the training duration of decision trees. LightGBM optimizes resource consumption and expedites training by iteratively partitioning the tree per leaf and selecting the leaf with the maximal reduction in loss, thereby averting overfitting despite potential complexities inherent in smaller datasets [15].

3.3. Categorical Boosting

CatBoost, an open-source machine learning methodology rooted in GB, stands out within the field. Its notable attributes include rapid learning convergence, adept handling of diverse data types such as text, categorical, and numeric variables, compatibility with GPU acceleration, and comprehensive visualization capabilities, distinguishing it from conventional methods. Furthermore, CatBoost excels in the classification of categorical data and adeptly manages missing values without necessitating supplementary coding interventions during the data preprocessing phase [16].

3.4. Extreme Learning Machine

ELM stands as a training algorithm exhibiting promising performance particularly within the realm of single-hidden-layer Feed Forward Neural Networks (FFNN). It manifests notably expedited convergence in contrast to traditional methodolo-gies [17]. Diverging from conventional FFNN learning paradigms, such as the Back Propagation Algorithm (BPA), ELM eschews reliance on gradient-based techniques. Notably, it obviates the need for iterative training, as all model parameters are deterministically established in a single pass during the learning process.

3.5. Decision Tree

DT approach represents a supervised machine learning technique applicable to problems encompassing both regression (for continuous output values) and classification (for discrete output values). Its nomenclature derives from the tree-like structure, wherein features (or conditions) serve as branches, while class labels constitute the terminal nodes or leaves. The principal strength of DT lies in its inherent simplicity, interpretability, and visualizability. Moreover, it accommodates the integration of decision-making processes within the tree framework. Notably, this approach is adept at modeling datasets characterized by intricate nonlinear relationships between input and output variables. However, it is pertinent to acknowledge its suscepti-bility to overfitting and its limited efficacy in handling classification tasks involving multiple output classes [18].

4. Dataset Overview

The dataset obtained from Kaggle [19] consists of 5113 rows, containing historical delivery data for the 16-month period between February 14, 2019, and June 13, 2020. Categorical variables in the dataset have been converted to numerical values using One-Hot Coding. The features and their descriptions are presented in Table 1.

Table 1. Features in the Dataset

Feature Name	Definition
Year	Year
Month	Month of the year
Day	Day of the month
Hour	Hour of the day
Minute	Minute of the hour
Second	Second of the minute
PuP	Pick up point
DoP	Drop off point
Source_country	Country from where the product needs to be delivered
Destination_country	Country to where the product needs to be delivered
Freight_cost	Cost of transportation / kg
Gross_weight	Gross weight in kg which needs to be delivered
Delivery_charges	Fixed cost per delivery
Delivery_mode	Method of delivery
Delivering_company	Candidate delivering company
Shipping_time	The time that it takes for a product to reach their destination

5. Delivery Time Prediction Models

Models have been developed using CatBoost, DT, ELM, LightGBM and SVM to predict delivery times. 5-fold cross validation has been applied for each method. To increase the model's success, the optimal values of the specified hyperparameters have been found using grid search. The developed models have been evaluated using the MAPE metric since delivery time prediction is a regression problem. The hyperparameter ranges used as a basis for developing prediction models are given in Table 2.

Table 2. Hyperparameter Ranges

Methods	Hyperparameter Range
CatBoost	"Max_Depth": [5, 16] "Iterations": [3, 4] "Learning_Rate": [0.6, 1.0]
DT	"Max_Depth": [5, 12] "Min_Samples_Split": [2, 4] "Min_Samples_Leaf": [1, 4]
ELM	"Alpha": [0.1, 0.4] "Neuron_Size": [25, 75]
LightGBM	"N_Estimators": [75, 130] "Max_Depth": [5, 10] "Learning_Rate": [0.1, 0.2]
SVM	"Nu": [0.5, 0.9] "C": [1, 7] "Degree": [3]

6. Results and Discussion

The MAPE's obtained with the developed prediction models are given in Table 3.

Table 3. MAPE's of Delivery Time Prediction Models

	CatBoost	DT	ELM	LightGBM	SVM
1. Fold	21.23%	19.55%	21.94%	23.21%	19.61%
2. Fold	21.18%	19.29%	23.23%	23.95%	19.02%
3. Fold	21.57%	18.43%	22.74%	24.98%	18.04%
4. Fold	19.93%	19.36%	23.7%	19.5%	18.95%
5. Fold	18.98%	20.75%	20.78%	19.98%	18.06%
Average	20.58%	19.48%	22.48%	22.32%	18.74%

- The performance of SVM-based prediction model is 1.84% higher than that of CatBoost-based prediction model.
- SVM-based prediction model provided 3.74% lower MAPE than ELM-based prediction model.
- LightGBM-based prediction model has 3.58% higher MAPE than SVM-based prediction model.
- The DT-based prediction model shows a performance close to the SVM-based prediction model with a difference of 0.74%.
- The performance of DT-based prediction model is 1.1% higher than that of CatBoost-based prediction model.
- ELM and LightGBM-based prediction models have 3% and 2.84% higher MAPE, respectively, than DT-based prediction model.

Upon examination of the obtained results, it has been observed that, in general, the models developed using SVM exhibited the most successful prediction performance. ELM has been the model with the lowest performance and the highest MAPE among the developed forecasting models. As a result of the analysis, it has been observed that SVM and DT models provide satisfactory results in delivery time prediction.

7. Conclusion

Today, e-commerce has gained rapid momentum with large sales volumes. Consequently, the number of deliveries has also increased. To accurately plan the increasing number of delivery, the duration of these deliveries must be predicted. Logistics optimization can be achieved by planning the cargo delivery time, and customer satisfaction can be maximized by providing information about when the cargo will be delivered. In this study, the performance of different machine learning methods on cargo delivery time prediction have been compared. The prediction models have been evaluated with the MAPE metric. When the

results obtained from the developed prediction models have been examined, it has been found that SVM had superior performance in predicting delivery times.

References

[1] Nodirovna, M. S., & Sharif o'g'li, A. S. (2024). E-Commerce Trends: Shaping The Future of Retail. Open Herald: Periodical of Methodical Research, 2(3), 46-49.

[2] Muñoz-Villamizar, A., Velázquez-Martínez, J. C., Haro, P., Ferrer, A., & Mariño, R. (2021). The environmental impact of fast shipping ecommerce in inbound logistics operations: A case study in Mexico. Journal of Cleaner Production, 283, 125400.

[3] Cui, R., Lu, Z., Sun, T., & Golden, J. M. (2024). Sooner or later? Promising delivery speed in online retail. Manufacturing & Service Operations Management, 26(1), 233-251.

[4] Özdemir, R., Taşyürek, M., & Aslantaş, V. (2024). Improved Marine Predators Algorithm and Extreme Gradient Boosting (XGBoost) for shipment status time prediction. Knowledge-Based Systems, 111775.

[5] Zhang, L., Liu, Y., Zeng, Z., Cao, Y., Wu, X., Xu, Y., ... & Cui, L. (2024). Package Arrival Time Prediction via Knowledge Distillation Graph Neural Network. ACM Transactions on Knowledge Discovery from Data, 108: 1–19.

[6] Zhang, L., Wu, X., Liu, Y., Zhou, X., Cao, Y., Xu, Y., ... & Miao, C. (2024). Estimating package arrival time via heterogeneous hypergraph neural network. Expert Systems with Applications, 238, 121740.

[7] [Guo, B., Zuo, W., Wang, S., Zhou, X., & He, T. (2023). Attention Enhanced Package Pick-Up Time Prediction via Heterogeneous Behavior Modeling. In International Conference on Algorithms and Architectures for Parallel Pro-cessing Singapore: Springer Nature Singapore,189-208.

[8] Hamdan, I. K., Aziguli, W., Zhang, D., & Sumarliah, E. (2023). Machine learning in supply chain: prediction of real-time e-order arrivals using ANFIS. International Journal of System Assurance Engineering and Management, 14,Suppl 1, 549-568.

[9] Zhang, L., Zhou, X., Zeng, Z., Cao, Y., Xu, Y., Wang, M., ... & Shen, Z. (2023). Delivery time prediction using large-scale graph structure learning based on quantile regression. In 2023 IEEE 39th International Conference on Data Engineering, IEEE, 3403-3416.

[10] Zhou, X., Wang, J., Liu, Y., Wu, X., Shen, Z., & Leung, C. (2023). Inductive graph transformer for delivery time estimation. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 679-687.

[11] İnaç, H., Ayözen, Y. E., Atalan, A., & Dönmez, C. Ç. (2022). Estimation of postal service delivery time and energy cost with e-scooter by machine learning algorithms. Applied Sciences, 12(23), 12266.

[12] Salari, N., Liu, S., & Shen, Z. J. M. (2022). Real-time delivery time forecasting and promising in online retailing: When will your package arrive?. Manufacturing & Service Operations Management, 24(3), 1421-1436.

[13] Wen, H., Lin, Y., Mao, X., Wu, F., Zhao, Y., Wang, H., ... & Wan, H. (2022). Graph2route: A dynamic spatial-temporal graph neural network for pick-up and delivery route prediction. In Proceedings of the 28th ACM SIGKDD Conference On Knowledge Discovery and Data Mining, 4143-4152.

[14] Salcedo-Sanz, S., Rojo-Álvarez, J. L., Martínez-Ramón, M., & Camps-Valls, G. (2014). Support vector machines in engineering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(3), 234-267.

[15] Barros, F. S., Cerqueira, V., & Soares, C. (2021). Empirical study on the impact of different sets of parameters of gradient boosting algorithms for time-series forecasting with LightGBM. In PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part I 18, Springer International Publishing, 454-465.

[16] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[17] Wang J., Lu S., Wang S. H., Zhang Y. D. (2022), A review on extreme learning machine, Multimedia Tools and Applications, 81(29), 41611-41660.

[18] Dabiri, H., Farhangi, V., Moradi, M. J., Zadehmohamad, M., & Karakouzian, M. (2022). Applications of decision tree and random forest as tree-based machine learning techniques for analyzing the ultimate strain of spliced and non-spliced reinforcement bars. Applied Sciences, 12(10), 4851.

[19] URL https://www.kaggle.com/datasets/salil007/1-shipping-optimization-challenge?select=train_2_pr.csv