Delivery Time Prediction Using Support Vector Machine Combined with Look-back Approach

Oğuz Erdi Erkmen[1], Ece Nigiz[1], Z. Sude Sarı[2], H. Şebnem Arlı[2], M. Fatih Akay[2]

[1]*Trendyol, Istanbul, Turkey, erdi.erkmen@trendyol.com, ORCID: 0000-0002-3936-1087*
[1]*Trendyol, Istanbul, Turkey, ece.nigiz@trendyol.com, ORCID: 0000-0002-6563-9345*
[2]*Çukurova University, Adana, Turkey, zsudesarii@gmail.com, ORCID: 0000-0002-0341-6488*
[2]*Çukurova University, Adana, Turkey, ssebnemarlii@gmail.com, ORCID: 0000-0001-8236-7487*
[2]*Çukurova University, Adana, Turkey, mfakay@cu.edu.tr, ORCID: 0000-0003-0780-0679*

Delivery time refers to the time elapsed from the moment the order is created to the delivery of the product to the final consumer. It is extremely important to predict the delivery time in order to ensure customer satisfaction and smooth logistics processes. The aim of this study is to predict the delivery time using Support Vector Machine (SVM) with and without consecutive look-back and periodic look-back approaches. A sample dataset obtained from Kaggle was used. Mean Absolute Error (MAE) was utilized to evaluate the performance of the prediction models. According to the results, the average MAE obtained with the look-back approach (3.81) was 59.12% lower than that obtained without the look-back approach (9.32).

Keywords: Delivery time prediction, Look-back approach, Machine learning, Support Vector Machine

1. Introduction

The concept of logistics is an improved expression of the concept of transportation and has become one of the most important tools in economic development today. The increasing order quantity with the development of the logistics sector causes irregularities and disruptions in the transportation processes of the supplier companies. With the increasing demand, it is of great importance to predict the delivery time in order to maintain order within the supplier company and to prioritize customer satisfaction. With this prediction, the supplier will be able to inform its customers about the arrival time of their packages and create customer satisfaction by removing uncertainty.

In the last few years, numerous methods have been used for delivery time prediction. [1] applied tree-based models to generate distributional forecasts by exploiting the complicated relationship between delivery time and relevant operational predictors. The simulation results showed that the proposed lead-time promise policy could improve sales volume by 6.1% compared to the current policy. [2] described three novel approaches for delivery time predictions, combining a machine learning model with human input. Also, they found that the pure machine learning model delivered better results than a combination of humans and machines. Although the pure machine learning model delivered superior prediction accuracy than the human-machine combination, the systematic qualitative analysis of the results presented insights for future development in this area. [3] studied the use of deep learning for solving a real-world case of last-mile parcel delivery time prediction. They presented the solution under the Internet-of-Things paradigm and discussed its feasibility in a cloud-based architecture as a smart city application. The authors thought that the system not only had the potential to improve the user experience by better modeling user expectations but also to help last-mile mail logistics as a whole. [4] showed the applicability and usefulness of travel time i.e. delivery time prediction for postal services. They investigated several methods such as Linear Regression models and tree-based ensembles such as Random Forest, Bagging, and Boosting, that allowed them to predict delivery time by conducting extensive experiments and considering many usability scenarios. Results revealed that travel time prediction can help mitigate high delays in postal services. [5] proposed a machine learning approach to predict ship speed over the ground using the Automatic Identification System and noon-report maritime weather data. The results showed that the proposed data-driven model has a satisfactory capability to predict the shipping speed based on the chosen features. [6] conducted a comprehensive investigation of the state-of-the-art developing machine learning models for applications to different aspects of International Freight Transportation Management (IFTM). They gave an overview of various fundamental machine learning methods. Then, how different machine learning methods were employed, adapted, and applied to a multitude of subject areas in IFTM were discussed, including demand forecast, operation and asset maintenance, vehicle trajectory, and on-time performance prediction. To increase transparency on delay, [7] developed a prediction model based on 315 explanatory variables, 10 regression models, and 7 classification models. Using machine learning algorithms, they achieved the best results for the neural network and support vector machine, with a prediction accuracy of 77% compared to only 59% of a basic model. Although machine learning algorithms are well suited for solving nonlinear and complex relationships in data generated by mobile sensors, only a small minority of recent publications have used machine learning for travel time estimation in multimodal transport. [8] applied the Extremely Randomized Trees, Adaptive Boosting and Support Vector Regression algorithms to this problem because of their ability to deal with low data volumes and their low processing times. They used different combinations of features derived from the data and built several models for travel time prediction. Results showed that their model performed better than average-based methods. [9] tried to solve the cargo unit discharge time estimation problem of short sea freight transport in collaboration with a European short-sea freight transport company. They proposed and tested a module-based framework using statistical analysis for predicting the discharge time. The results showed the potential for improving performance and accuracy. [10] proposed a novel spatial-temporal sequential neural network model (DeepETA) to take full advantage of the above factors. DeepETA is an end-to-end network that mainly consists of three parts. Experiments on the real logistic dataset showed that the proposed approach performed better. [11] aimed to determine whether machine learning and predictive analytics could improve the estimated time of arrival for a shipment. Using machine learning computing, they developed a model capable of predicting shipping times by training the algorithms on historical shipment data and incorporating external sources of data related to the most impactful factors regarding schedule reliability (e.g. holiday seasons and port congestion levels). Also, they found that machine learning in this instance might be a partial answer to this problem, as it performed better on long lead time than on short lead time when compared to more classical approaches.

Although delivery time prediction models have been developed with various machine learning methods in many different sectors, the lookback approach has not been utilized in the studies so far. For this reason, this study aimed to develop delivery time prediction models using SVM with look-back approach. A sample data set obtained from Kaggle was used for developing the predictions models. Firstly, models were generated on the dataset without look-back approach, then consecutive look-back and periodic look-back approach were used to improve the performance of the models. The MAE metric was utilized to evaluate the performance of the models.

This paper is structured as follows: Section 2 provides a description of the dataset. Section 3 provides information on the methodology. Section 4 presents results and discussion. Section 5 concludes the paper.

2. Dataset Generation

The dataset contains 5115 rows of data from February 14th, 2019 to June 13th, 2020 and includes delivery times and some other attributes given in Table 1. The dataset was obtained from Kaggle [12]. Categorical variables in the dataset were converted to numerical values using One-Hot Encoding method. The attributes in the dataset and their explanations are given in Table 1.

3. Methodology

A. Support Vector Machine

The formulation of the standard SVM is defined as a maximum margin classifier, that is, a classifier whose decision function is a hyperplane that maximally separates samples from different classes. SVM can also be used as a regression method, maintaining all the main features that characterize the algorithm (maximal margin). The SVM for regression uses the same principles as the SVM for classification, with only a few minor differences. First, because the output is a real number it becomes very difficult to predict the information at hand, which has infinite possibilities. In the case of regression, a margin of tolerance (epsilon/nu) is set in approximation to the SVM which would have already been requested from the problem. However, the main idea is always the same: to minimize error, individualizing the hyperplane which maximizes the margin [13].

Table 1. *Attributes in the dataset*

Attribute 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

B. Consecutive Look-back and Periodic Look-back:

These two approaches are algorithmic improvements developed in order to use the historical values of the target variable to be predicted as new attributes in the training dataset. The consecutive look-back is how many past values of the target variable will be used as an attribute in the training process of the model. This process takes this name because it is done consecutively. Periodic look-back is how many of the past values of the target variable will be added as an attribute to the training data in the specified period. Periodic look-back has two parameters. The first parameter is the measure of how many times the addition will be made. The second parameter is a measure of how far back to go from the end of the training data. This process is done periodically, not consecutively. Although the two approaches work similarly in logic, the general difference is in how they do the backward operation.

Table 2 and Table 3 show the hyperparameters of the models that were developed with and without consecutive look-back and periodic look-back approach.

The explanations of the hyperparameters in Table 2 and Table 3 are given below:

- In regression problems, a margin of tolerance is set in approximation to the SVM. There are two different options for this; 'Epsilon' and 'Nu'.
- Kernel function specifies the kernel type to be used in the algorithm. There are 4 kernel functions namely ‘linear’, ‘poly’, ‘rbf’ and ‘sigmoid’.
- C is the penalty parameter of the error term.
- There are three gamma types which are scale, auto and value. If the gamma type is selected as 'scale', the gamma value is going to be “1 / (number of attributes* standard variation of input)”. If the gamma type is selected as ‘auto’, the gamma value is going to be “1 / number of attributes”. If the gamma type is selected as ‘value’, the gamma value is going to be accepted as the entered value.
- Degree variable is used only when the kernel function 'poly' is selected. Ignored when other kernel functions are used.
- Coef0 is an independent term in kernel function. It is only significant for ‘poly’ and ‘sigmoid’.

Table 2. *The hyperparameters of the models without look-back approach.*

Model No	Amount of Training Data	Consecutive Lookback	Periodic Lookback	Type of SVM	Kernel Function	Epsilon/Nu	C	Gamma	Degree	Coef()
1	100%	-	-	Epsilon	Polynomial	$9*10^{-1}$	2^{-5}	Auto	3	
2	95%	-	-	Epsilon	RBF	$9*10^{-1}$	2^{1}	Auto	-	-
3	65%	-	-	Epsilon	Polynomial	$1*10^{0}$	2^{-5}	Auto	1	0
4	65%	-	-	Epsilon	Sigmoid	$3*10^{-1}$	2^{-2}	Scale	-	0
5	70%	-	-	Epsilon	RBF	$9*10^{-1}$	2^{-3}	Auto	-	-
6	70%	-	-	Nu	RBF	$1*10^{-1}$	2^{-2}	Scale	-	-
7	100%	-	-	Epsilon	Polynomial	$7*10^{-1}$	2^{-4}	Scale	2	0
8	75%	-	-	Epsilon	Polynomial	$1*10^{0}$	2^{-1}	Scale	3	0
9	85%	-	-	Nu	Polynomial	$9*10^{-2}$	2^{-1}	Value (2^{-4})	3	0
10	90%	-	-	Nu	Sigmoid	$5*10^{-2}$	2^{5}	Value (2^{-5})	-	1

Table 3. *The hyperparameters of the models with look-back approach.*

Fold No	Amount of Training Data	Consecutive Lookback	Periodic Lookback	Type of SVM	Kernel Function	Epsilon/Nu	C	Gamma	Degree	Coef()
1	100%	5	1-5	Epsilon	Polynomial	$1*10^{-2}$	2^{-2}	Auto	2	1
2	95%	5	2-5	Nu	Polynomial	$5*10^{-1}$	2^{1}	Auto	2	1
3	85%	5	2-5	Epsilon	RBF	$1*10^{-2}$	2^{-3}	Auto	-	-
4	100%	5	2-5	Epsilon	RBF	$3*10^{-2}$	2^{-2}	Auto	-	-
5	70%	5	3-5	Epsilon	Polynomial	$3*10^{-2}$	2^{-3}	Value (2^{1})	2	0
6	85%	5	2-5	Epsilon	Linear	$5*10^{-2}$	2^{-2}	-	-	-
7	70%	5	2-5	Nu	RBF	$8*10^{-1}$	2^{1}	Value (2^{-5})	-	-
8	95%	5	2-10	Epsilon	Polynomial	$7*10^{-2}$	2^{-1}	Auto	3	0
9	70%	5	2-5	Epsilon	RBF	$3*10^{-2}$	2^{-1}	Value (2^{-4})	-	-
10	90%	5	5-5	Nu	Linear	$6*10^{-2}$	2^{-5}	-	-	-

4. Results and Discussion

For delivery time prediction, models with and without consecutive look-back and periodic look-back approach were developed with SVM. 10 different models were created. (i.e. For creating the 1st model, the first 500 rows of the data set were used as the test set and the remaining data was used as the training set. For the 2nd model, rows 500 through 999 were used as the test set and the remaining data was used as the training set. This process was repeated 10 times and models were created). MAE was utilized to evaluate the performance of the models.

Table 2 shows the MAE's of the models with and without look-back approach.

Table 2. MAE's of the models with and without look-back approach.

Model No	Without look-back approach	With look-back approach
1	7.89	2.55
2	7.26	2.68
3	7.12	2.62
4	11	2.64
5	7.64	2.82
6	7.86	3.81
7	7.24	2.92
8	7.93	3.75
9	7.36	4.55
10	21.88	9.77

- When the results of delivery time prediction models with the look-back approach are examined, the lowest MAE (2.55) has been observed in the 1st model and the highest MAE (9.77) has been observed in the 10th model.

- When the results of delivery time predictions models without the look-back approach are examined the lowest MAE (7.12) has been observed in 3rd model and the highest MAE (21.88) has been observed in the 10th model.

- When all results are compared it was observed that using the look-back approach significantly reduced the MAE's of the models.

- The MAE's of the models created without the look-back approach were decreased by 67.68%, 63.09%, 63.2%, 76%, 63.09%, 51.53%, 59.67%, 52.71%, 38.18%, and 55.35%, respectively, after applying the look-back approach.

- According to the results, the average MAE obtained with the look-back approach (3.81) was 59.12% lower than that obtained without the look-back approach (9.32).

5. Conclusion

In this study, delivery time prediction models were developed by using with and without look-back approach. A sample data set obtained from Kaggle was used for the predictions generated using SVM. Firstly, models were generated on the dataset without look-back approach, then consecutive look-back and periodic look-back approach were used to improve the performance of the models. MAE was used to evaluate the performance of the folds. According to the results, the MAE obtained with the look-back approach (3.81) was 59.12% lower than that obtained without the look-back approach (9.32). The results showed that using the look-back approach significantly reduced the MAE's of the models.

REFERENCES

[1] N. Salari, S. Liu, Z. J. M. Shen., "Real-Time Delivery Time Forecasting and Promising in Online Retailing: When Will Your Package Arrive?", Manufacturing & Service Operations Management Sep., 2022.
[2] M. Lochbrunner, H. F. Witschel. "Combining Machine Learning with Human Knowledge for Delivery Time Estimations.", FHNW - University of Applied Sciences Northwestern Switzerland, Olten, 2022.
[3] A. C. de Araujo, A. Etemad, "End-to-End Prediction of Parcel Delivery Time with Deep Learning for Smart-City Applications", IEEE Internet of Things Journal, vol. 8, no. 23, pp. 17043-17056, Dec., 2021.
[4] J. Khiari, C. Olaverri-Monreal, "Boosting Algorithms for Delivery Time Prediction in Transportation Logistics," in 2020 International Conference on Data Mining Workshops ICDMW, pp. 251-258, Sorrento 2020.

[5] M. Abebe, Y. Shin, Y. Noh, S. Lee, I. Lee, “Machine Learning Approaches for Ship Speed Prediction towards Energy Efficient Shipping”. Applied Sciences vol. 10, no.7, pp. 2325, March, 2020.
[7] L. Barua, B. Zou, Y. Zhou, “Machine learning for international freight transportation management: A comprehensive review”, Research in Transportation Business & Management, vol. 34, pp. 100453, 2020.
[8] A. Viellechner, S. Spinler, “Novel Data Analytics Meets Conventional Container Shipping: Predicting Delays by Comparing Various Machine Learning Algorithms”, Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020
[9] N. Servos, X. Liu, M. Teucke, M. Freitag, “Travel time prediction in a multimodal freight transport relation using machine learning algorithms”, Logistics vol. 4, no.1, pp. 1, Dec., 2019.
[10] B. Jia, N. G. M. Rytter, L. B. Reinhardt, G. Haulot, M. B Billesø, “Estimating Discharge Time of Cargo Units– A Case of Ro-Ro Shipping” in International Conference on Computational Logistics pp. 122-135, Barranquilla, 2019.
[11] F. Wu, L. Wu. "DeepETA: a spatial-temporal sequential neural network model for estimating time of arrival in package delivery system." in Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii 2019.
[12] https://www.kaggle.com/datasets/salil007/1-shipping-optimization-challenge?select=test_2.csv
[13] A. C. Jonquais, F. Krempl. "Predicting Shipping Time with Machine Learning." 2019.