LUMEN
ET VIRTUS

# Prediction and control of delivery times in service platforms with Artificial Intelligence

https://doi.org/10.56238/levv15n39-005

Guilherme Barrado Pereira[1] and Rogério de Oliveira[2]

**ABSTRACT**
Delivery service platforms, or delivery platforms, play an increasingly important role in the consumption of different products and services. This work uses data from a platform for the sale and distribution of cooking gas to end consumers to build a machine learning model to predict delivery times for use in service management. Different machine learning models are applied and compared, and the analysis allows the identification of the main offenders for better control of delivery times and business decision-making in the operation of the platform. The final model uses the Extremely Randomized Trees algorithm and is more accurate than the method currently employed by the company.

**Keywords:** Deliveries, Service platforms, Artificial Intelligence.

[1] Faculty of Computing and Informatics (FCI)
Mackenzie Presbyterian University (UPM) – São Paulo – SP – Brazil
Email: gbp.guilherme@gmail.com
[2] Faculty of Computing and Informatics (FCI)
Mackenzie Presbyterian University (UPM) – São Paulo – SP – Brazil
Email: rogerio.oliveira@mackenzie.br

## INTRODUCTION

The development of e-commerce has contributed to changing consumption patterns and, increasingly, consumers prefer to use online platforms to order products driven by the expansion of broadband services and electronic payment methods. This is particularly present in food delivery, characterizing the so-called O2O operations, *online to offline*, in which the transaction is initiated online and concluded offline with the delivery of the meal (Shankar, et al., 2022; Li, et al. 2020).

This model emerges as a popular trend and a tool to reach a greater number of consumers profitably (Shankar, et al., 2022). This same business model has also been applied to other types of products, such as grocery shopping and medicines. There are companies that opt for their own app and services, and others that choose to employ a multi-vendor aggregator platform and help them reach a wider market. Much is discussed and there is a lot of work about how this *uberization* of services impacts markets, changes social behaviors and generates precarious work. But there seems to be little work on ways to manage these complex platforms.

Currently, more than 90% of the Brazilian population depends on the distribution of LPG gas (cylinders) every day, whether in homes or in activities related to industry (Consigaz, 2023), and one of the important options available today for sale is through online applications, bringing practicality to consumers, but also challenges for management in a more competitive environment.

This work presents and implements a solution based on machine learning models for the prediction and control of delivery times in a platform for the sale and delivery of cooking gas. It thus seeks to provide data-based mechanisms for better management of services and suppliers.

The platform consists of a mobile application where the user can select, from an address, the nearest gas resellers, the desired products, and search for the fastest delivery time or the lowest price. Resellers have access to the platform and can accept or reject orders placed by customers.

One of the challenges of the business is better informing the estimated delivery time of each reseller when a customer makes an inquiry. Currently, the delivery time is the average of the difference between the moment the reseller delivered the order to the customer and the moment of confirmation of receipt of the order, considering the orders of the last ninety (90) days, and the values are presented rounded in a range of 10 minutes. This method does not take into account any factors that may cause variability in delivery times, such as the time or day of the week, and provides fairly inaccurate weather forecasts.

The general objective of this work is to employ machine learning models to provide a better estimate of the delivery times of orders placed on the platform, increasing the quality of service by providing more reliable information to customers.

## THEORETICAL FRAMEWORK

In this project, several concepts and techniques of artificial intelligence and machine learning are employed. Concepts used in the corporate and business segment are also used. Techniques and metrics used in other scientific articles with problems related to this project are analyzed. This section provides a brief review of each of these themes.

### DATA-DRIVEN DECISIONS

Several companies around the world have been racing against time to modernize by digitizing their operations to become data-driven, as the companies that came out ahead in this race became more competitive than the others (Henke et al., 2016; Bughin et al., 2018).

The knowledge extracted from this data is used for strategic decision-making and for the automation and improvement of various activities throughout the industrial chain, demonstrating success stories with the application of different Artificial Intelligence techniques in the most varied business segments, whether in the development of autonomous cars, in image recognition applied to health diagnoses,  in text and speech translation, in recommendation systems, etc. (Oluyisola et al., 2020; Marr, 2020; Ludermir, 2020).

### SUPERVISED MACHINE LEARNING

It involves the use of statistics and optimization methods that allow labeled datasets to be analyzed for patterns for the construction of mathematical models that are evaluated based on the predictive ability in relation to the variance measures of the data themselves (Nasteski, 2017). It is composed of classification techniques in which data are classified according to a pre-defined class and regression techniques that predict a numerical value from the analysis (Salian, 2018). Due to the nature of the algorithm that updates autonomously, the error rate is reduced with each execution, that is, the algorithm is able to learn from the analyzed data (IBM Cloud Education; 2020).

The selection of the algorithm to be used is an important step of the project and is done by comparing metrics such as Mean Absolute Percentage Error (MAPE), the Mean Absolute Error (MAE) and Mean Squared Error (MSE). *Gradient Boosting*, *Decision Tree*, *Random Forest* and *Support Vector Machines* (SVM) are among the most used techniques because they have better performance in regression methods (Almaghrebi et al., 2020).

### *DECISION TREE, ENSEMBLE METHODS* E *RECURSIVE FEATURE ELIMINATION*

*Decision Tree* is a recursive algorithm that has nodes related in a hierarchical way in a format that resembles a tree. Each node represents a decision made on top of a variable in the dataset, and the root node represents the most important variable. *Ensemble Methods* combine several techniques for ranking variables, in order to improve selection.

*Random Forest* is a technique that uses sets of decision trees that are generated using a random sample of the data. The cutting variables, for the formation of the nodes, are chosen randomly and the combination of all trees is used to obtain the results (Hartshorn, 2016).

*Recursive Feature Elimination* uses *Random Forest* for variable selection and model training. The use of variable selection techniques or dimensionality reduction brings significant performance gains in model training. The overall goal is to understand which variables have the most impact, eliminating noise and discarding variables that are irrelevant to the prediction or classification of the problem. Models with fewer variables are more interpretable and have less training cost (Kuhn et al., 2018, Seijo-Pardo et al., 2015).

*Extremely Randomized Trees* is a technique that essentially consists of randomizing the choice of variable and cutoff point while splitting a node of the decision tree. The trees are constructed in a totally random manner and their structures are independent of the output values of the learning sample (Geurts et al., 2006).

## RELATED WORKS

For some time now, machine learning algorithms have also been used together with traditional statistical and time series forecasting models in the most diverse segments (Rundo, 2019). Some of these jobs used to predict travel times, travel times, etc., are listed below.

Regression techniques, including *Random Forest* and *SVM*, were applied to predict travel time in passenger transport in the city of Porto, Portugal. The work uses the Mean Squared Error as a metric and concludes that for the problem presented and the data used for training, the *Random Forest* technique was superior to the *SVM technique (*Moreira et al., 2005).

*SVM* and other regression techniques were applied to predict travel time in multimodal cargo transport companies. This work used the Mean Absolute Error as a metric and demonstrates the importance of variable selection to improve the accuracy of models by comparing models created with different data "architectures". The *MVS technique* proved to be superior after being trained and compared with the others (Servos et al., 2020).

*Decision Trees*, *Extremely Randomized Trees*, Artificial Neural Networks and deep learning techniques such as Recurrent Neural Networks were applied to predict the travel time for taxi rides with good performance, comparing and demonstrating the different results obtained from the best selection of variables and different configurations of the models (Joshi et al., 2017; Lam et al., 2015).

This work uses data labeled with the actual delivery times, so the supervised machine learning model is employed for the predictions. Given the nature of the problem, regression techniques such as *Decision Tree*, *Random Forest, Extremely Randomized Trees, SVM* and *Gradient Boosting are used.* In this study, we are interested not only in predicting delivery times, but also in identifying the most predictive variables.

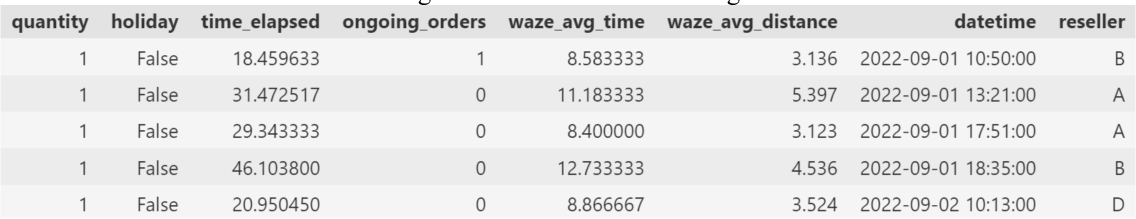
## RESEARCH METHODOLOGY

The following are the tasks and methods used in this work. The initial set of raw data that was used for exploratory analysis is presented. The preparation process is described and details tasks such as deriving variables and removing *outliers*. The models and metrics used are presented as well as the training strategy.

### DATA

The data used in the development of the project were made available by the company in a file in the format . CSV (*comma separated values*). Personal and/or sensitive data have been previously obfuscated for privacy and confidentiality reasons. The data totals 8,421 instances distributed over a 13-month period. Figure 1 presents a sample of the raw data used in the work.

Figure 01: Raw data for training



| quantity | holiday | time_elapsed | ongoing_orders | waze_avg_time | waze_avg_distance | datetime | reseller |
|---|---|---|---|---|---|---|---|
| 1 | False | 18.459633 | 1 | 8.583333 | 3.136 | 2022-09-01 10:50:00 | B |
| 1 | False | 31.472517 | 0 | 11.183333 | 5.397 | 2022-09-01 13:21:00 | A |
| 1 | False | 29.343333 | 0 | 8.400000 | 3.123 | 2022-09-01 17:51:00 | A |
| 1 | False | 46.103800 | 0 | 12.733333 | 4.536 | 2022-09-01 18:35:00 | B |
| 1 | False | 20.950450 | 0 | 8.866667 | 3.524 | 2022-09-02 10:13:00 | D |

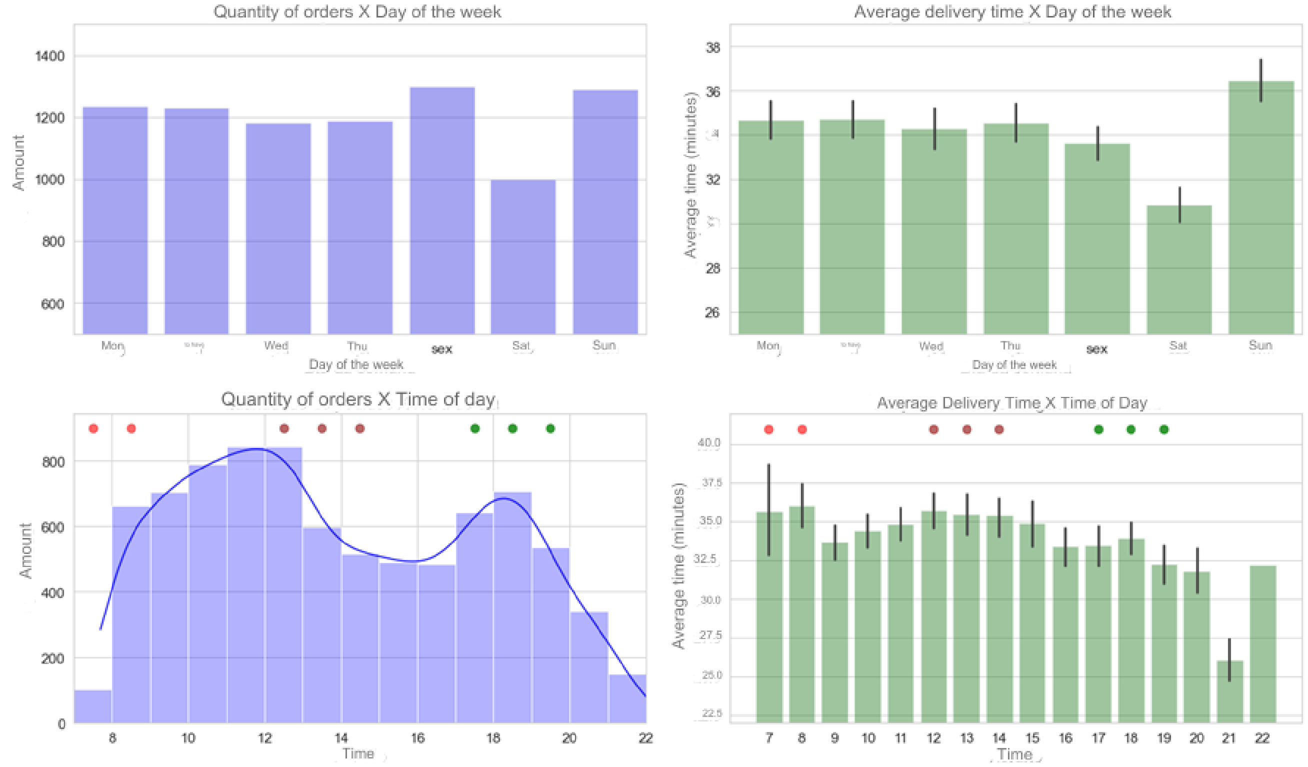
The code repository of this project provides the data dictionary with more details of each variable as well as the raw data itself. The files are available at <https://github.com/gbpereira/time_prediction/tree/main/data>.

### EXPLORATORY ANALYSIS

Exploratory analysis is a crucial step in the work and aims to investigate the dataset to understand the main characteristics in order to reveal patterns, trends, and relationships that can guide more in-depth analyses that can evidence the need to derive existing variables or even the need to add new variables to the initial set.

Figure 2 illustrates the graphs with the distribution of orders and the average delivery times throughout the day and throughout the week. The top graphs demonstrate the relationship between the lowest number of orders that coincides with the lowest average delivery time throughout the week. The distribution of the order quantity and the average delivery times are shown in the lower graphs. The number of orders varies throughout the day with two peaks at meal times, while the peaks of average times occur at times of heavy vehicle traffic.

Figure 02: Average times and number of orders by day of the week and time



Quantity of orders X Day of the week

Average delivery time X Day of the week

Quantity of orders X Time of day

Average Delivery Time X Time of Day

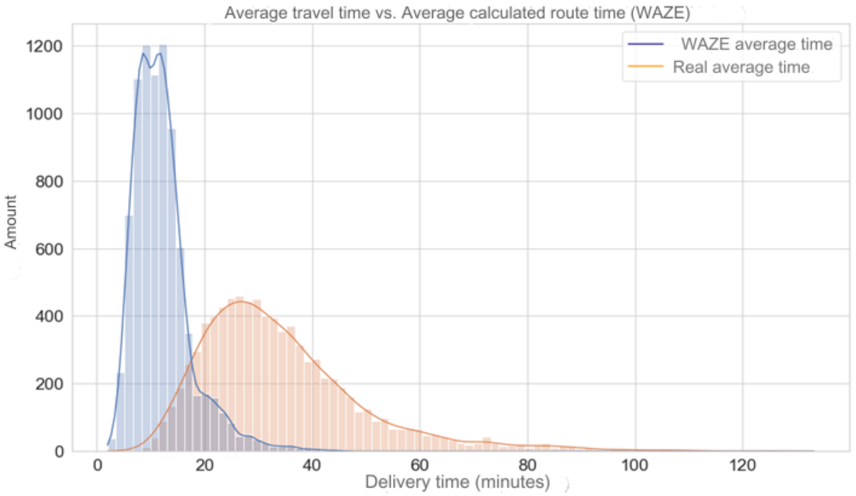
The variation in the number of orders and the average delivery time throughout the days of the week and times of the day were factors that indicated the need to derive the variable "*datetime*" in other variables that were used in the training to increase the explainability and improve the robustness of the models.

## DATA PREPARATION

In the preparation of the data, the time and average distance between the available routes between the reseller address and the order address was surveyed. The two pieces of information were calculated using the *WazeRouteCalculator* library. From this calculation, the information was added to the dataset with the labels "*waze_avg_time*" and "*waze_avg_distance*", containing the time and average distance respectively. For confidentiality and privacy reasons, the information containing the geographic coordinates was previously removed from the dataset.

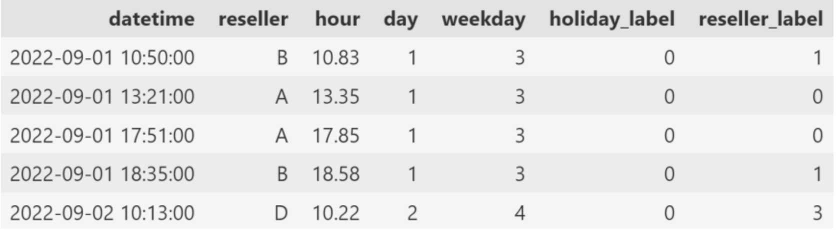
Figure 3 represents a histogram comparing the calculated average route times to the average of the actual delivery time. The difference between the values is due to the operational time of the reseller between accepting the order, separating the product(s) and delivering it.

Figure 03: Comparison between delivery times and average travel time for the delivery route



The "datetime" *variable*, which contains the date and time of creation of the order, has been converted into the "*day*", "*hour*" and "*weekday*" variables, containing respectively the day of the month, the time of the day and the day of the week. The "*holiday*" and "*reseller*" attributes went through the encoding process, which converts categorical values into numerical values that can be used by the models. Figure 4 represents a part of the data with the addition of the new variables after derivation and encoding.

Figure 04: Encoding and derivation of the hour, day, weekday, holiday_label and reseller_label variables



| datetime | reseller | hour | day | weekday | holiday_label | reseller_label |
|---|---|---|---|---|---|---|
| 2022-09-01 10:50:00 | B | 10.83 | 1 | 3 | 0 | 1 |
| 2022-09-01 13:21:00 | A | 13.35 | 1 | 3 | 0 | 0 |
| 2022-09-01 17:51:00 | A | 17.85 | 1 | 3 | 0 | 0 |
| 2022-09-01 18:35:00 | B | 18.58 | 1 | 3 | 0 | 1 |
| 2022-09-02 10:13:00 | D | 10.22 | 2 | 4 | 0 | 3 |

The dataset was also analyzed in order to remove *outliers*, which interfere with the training of the models. For this purpose, the Interquartile Range (IRQ) was calculated and all instances with *time_elapsed* above the upper limit or below the lower limit were removed. The dataset was left with 8,063 instances after this process.

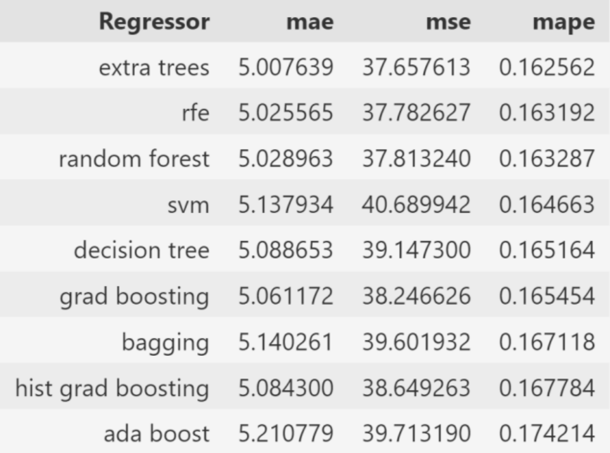
Finally, all variables, now with numerical type, went through the normalization process in order to adjust the scale of the values. This step is necessary to make the models more robust and less sensitive to *outliers*.

## TRAINING AND SELECTION OF MODELS

The training algorithms were selected from the analysis of *the MAPE* (Mean Absolute Percentage Error) with the training of the entire database using cross-validation. The following models were trained and compared: *Decision Tree*, *Random Forest*, *Recursive Feature Elimination*, *Extremely Randomized Trees*, *SVM*, *AdaBoost*, *Gradient Boosting*, *Histogram-based Gradient Boosting* and *Bagging*.

Figure 5 shows the comparison between the models ordered by *MAPE*. In addition to *the MAPE,* the Mean Absolute Error (*MAE)* and Mean Squared Error *(MSE)* were also compared. The three (3) highest-ranked models were chosen: *Extremely Randomized Trees, Recursive Feature Elimination* (RFE), and *Random Forest*.

Figure 05: Comparison between metrics of the trained regression algorithms with the entire database



| Regressor | mae | mse | mape |
|---|---|---|---|
| extra trees | 5.007639 | 37.657613 | 0.162562 |
| rfe | 5.025565 | 37.782627 | 0.163192 |
| random forest | 5.028963 | 37.813240 | 0.163287 |
| svm | 5.137934 | 40.689942 | 0.164663 |
| decision tree | 5.088653 | 39.147300 | 0.165164 |
| grad boosting | 5.061172 | 38.246626 | 0.165454 |
| bagging | 5.140261 | 39.601932 | 0.167118 |
| hist grad boosting | 5.084300 | 38.649263 | 0.167784 |
| ada boost | 5.210779 | 39.713190 | 0.174214 |

The average delivery time of each reseller is calculated always based on the last ninety (90) days and fluctuates over time with the company making continuous efforts to make them smaller and smaller. Given this variation, very old data end up generating noise in the results of the models and so the training strategy consisted of training the model with data from ninety (90) days to then test it with data from the next thirty (30) and analyze the results.

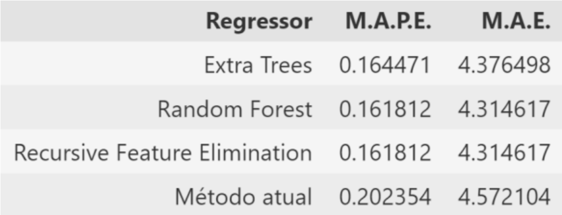
## RESULTS

This section presents the results of the work, where the models were compared looking for the one with the highest accuracy within the established metrics. An analysis of the explainability of the best model was made, providing *insights* for the company.

### DELIVERY TIME FORECAST

For each of the three (3) selected models, the Mean Absolute Percentage Error *(MAPE)* and Mean Squared Error *(MSE)* were calculated using data from 07/01/2023 to 09/31/2023 as training. The data used for the forecast covers the period between 10/01/2023 and 10/30/2023. Figure 6 demonstrates the results of each model, *Extremely Randomized Trees* continued to have the best performance with *MAPE* of 0.164471 and *MAE* of 4.376498 surpassing the current method used by the company.

Figure 06: Comparison between metrics of the models trained at 90 days

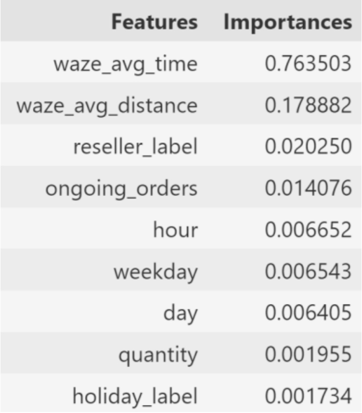


| Regressor | M.A.P.E. | M.A.E. |
|---|---|---|
| Extra Trees | 0.164471 | 4.376498 |
| Random Forest | 0.161812 | 4.314617 |
| Recursive Feature Elimination | 0.161812 | 4.314617 |
| Método atual | 0.202354 | 4.572104 |

## EXPLAINABILITY OF THE MODEL

Each model does its own ranking of variables that have the most impact on delivery time and those variables that are irrelevant during training. The explainability of the models was analyzed in search of *insights* that can help the company in decision making. Figure 7 contains the relationship of variables and their importance determined by the *Extremely Randomized Trees* model, the one that had the lowest mean absolute percentage error (MAPE) in training with the entire data set. Route time and distance between addresses are the main offenders in order delivery time, with 76% and 17% respectively of importance.

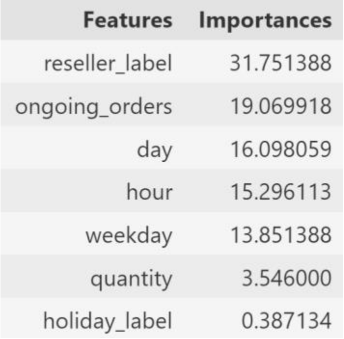
Figure 07: Relevance of each variable for delivery time



| Features | Importances |
|---|---|
| waze_avg_time | 0.763503 |
| waze_avg_distance | 0.178882 |
| reseller_label | 0.020250 |
| ongoing_orders | 0.014076 |
| hour | 0.006652 |
| weekday | 0.006543 |
| day | 0.006405 |
| quantity | 0.001955 |
| holiday_label | 0.001734 |

The variables referring to route time and distance were removed from the set and the remaining variables were normalized for a new analysis of importance. The biggest offender, *"reseller_label",* with approximately 31% of importance, highlights the different "operating times" of each dealer. The comparison between the normalized variables is shown in Figure 8.

Figure 08: Relevance of each variable for delivery time



| Features | Importances |
|---|---|
| reseller_label | 31.751388 |
| ongoing_orders | 19.069918 |
| day | 16.098059 |
| hour | 15.296113 |
| weekday | 13.851388 |
| quantity | 3.546000 |
| holiday_label | 0.387134 |

The second offender, "*ongoing_orders*", shows how the amount of orders in progress impacts delivery times and highlights the need for enough couriers at peak times to ensure that orders are delivered within the allotted time.

## CONCLUSION AND FUTURE WORK

The general objective of this work was to create a model using machine learning techniques, capable of predicting the delivery times of orders in the company's marketplace, providing more reliable information to end customers. During development, it was also possible to identify those variables that have the greatest impact on order delivery time.

The 3 best models were analyzed and trained using the same dataset and, using *MAPE* as a metric, the *Extremely Randomized Trees algorithm was chosen* as the one that makes predictions with greater accuracy. With *a MAPE* of 0.164471, this model is superior to the method currently employed by the company with *a MAPE* of 0.202354 calculated in the same period.

Future work may involve the use of other techniques and algorithms in order to arrive at models with even more accuracy. Given the nature of the company's business (marketplace), the work can also be adapted and used as a basis for forecasting models on other platforms that have deliveries made by third parties. Another possibility is to look for new variables or new derived variables that can be added to the models, increasing their accuracy.

This entire project, data and artifacts can be freely accessed in the repository available at <https://github.com/gbpereira/time_prediction>.

## REFERENCES

1. Almaghrebi, A., Aljuheshi, F., Rafaie, M., James, K., & Alahmad, M. (2020). Data-driven charging demand prediction at public charging stations using supervised machine learning regression methods. *Energies*, 13(16). Disponível em <https://doi.org/10.3390/en13164231>. Acesso em 12/11/2023.

2. Boulic, R., & Renault, O. (1991). 3D hierarchies for animation. In N. Magnenat-Thalmann & D. Thalmann (Eds.), *New trends in animation and visualization* (pp. 1-13). John Wiley & Sons Ltd.

3. Bughin, J., Seong, J., Manyika, J., Chui, M., & Joshi, R. (2018). Notes from the AI frontier: Modeling the impact of AI on the world economy. McKinsey Global Institute.

4. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63, 3-42. Disponível em <http://dx.doi.org/10.1007/s10994-006-6226-1>. Acesso em 07/12/2023.

5. GLP – Gás Liquefeito de Petróleo, Consigaz. (2023). Disponível em: <https://www.consigaz.com.br/gas-glp/>. Acesso em 10/11/2023.

6. Hartshorn, S. (2016). Machine learning with random forests and decision trees. *S.L.*

7. Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). The age of analytics: Competing in a data-driven world. McKinsey Global Institute.

8. Joshi, N., Hotalappa, M., & Gadade, K. (2017). A study on travel time estimation for taxi trips. *International Education & Research Journal (IERJ)*, 3. Disponível em: <https://www.academia.edu/36968291>. Acesso em 09/11/2023.

9. Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling* (2nd ed.). Springer.

10. Lam, H. T., Diaz-Aviles, E., Pascale, A., Gkoufas, Y., & Chen, B. (2015). Taxi destination and trip time prediction from partial trajectories. *arXiv*. Disponível em: <https://arxiv.org/abs/1509.05257>. Acesso em 09/11/2023.

11. Li, C., Mirosa, M., & Bremer, P. (2020). Review of online food delivery platforms and their impacts on sustainability. *Sustainability*, 12(14), 5528.

12. Ludermir, T. (2021). Inteligência artificial e aprendizado de máquina: estado atual e tendências. Disponível em: <https://doi.org/10.1590/s0103-4014.2021.35101.007>. Acesso em 18/11/2023.

13. IBM Cloud Education. (2020). Machine learning. Disponível em: <https://www.ibm.com/cloud/learn/machine-learning>. Acesso em 09/11/2023.

14. Marr, B. (2016). A short history of machine learning every manager should read. *Forbes*. Disponível em <https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read>. Acesso em: 18/11/2023.

15. Moreira, J. M., Jorge, A., Sousa, J. F., & Soares, C. (2005). Trip time prediction in mass transit companies: A machine learning approach. *Advanced OR and AI methods in transportation*. Disponível em: <https://hdl.handle.net/10216/6749>. Acesso em 09/11/2023.

16. Nasteski, V. (2017). An overview of the supervised machine learning methods. Disponível em: <https://www.researchgate.net/publication/328146111_An_overview_of_the_supervised_machine_learning_methods>. Acesso em 10/11/2023.

17. Oluyisola, O. E., Sgarbossa, F., & Strandhagen, J. O. (2020). Smart production planning and control: Concept, use-cases and sustainability implications. *Sustainability*, 12(9), 3791. Disponível em <https://doi.org/10.3390/su12093791>. Acesso em 18/11/2023.

18. Rundo, F., Trenta, F., di Stallo, A. L., & Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24), 5574. Disponível em: <https://doi.org/10.3390/app9245574>. Acesso em 02/11/2023.

19. Salian, I. (2018). SuperVize me: What's the difference between supervised, unsupervised, semi-supervised and reinforcement learning? *NVIDIA Blog*. Disponível em: <https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/>. Acesso em 10/11/2023.

20. Seijo-Pardo, B., Bolón-Canedo, V., Porto-Díaz, I., & Alonso-Betanzos, A. (2015). Ensemble feature selection for rankings of features. Disponível em <https://www.researchgate.net/publication/300786598_Ensemble_Feature_Selection_for_Rankings_of_Features>. Acesso em: 18/11/2023.

21. Servos, N., Liu, X., Teucke, M., & Freitag, M. (2020). Travel time prediction in a multimodal freight transport relation using machine learning algorithms. *Logistics*, 4(1), 1. Disponível em: <https://doi.org/10.3390/logistics4010001>. Acesso em 08/11/2023.

22. Shankar, A., Jebarajakirthy, C., Nayal, P., Maseeh, H. I., Kumar, A., & Sivapalan, A. (2022). Online food delivery: A systematic synthesis of literature and a framework development. *International Journal of Hospitality Management*, 104, 103240.