Evaluating Socio-Economic Drivers of Service Completion in Last-Mile and First-Mile Reverse Logistics

Antonio Lorenzo-Espejo(✉), Kenny-Jesús Flores-Huamán, Luis Onieva, and Ana Pegado-Bardayo

Departamento de Organización Industrial y Gestión de Empresas II, Escuela Técnica Superior de Ingeniería, Universidad de Sevilla, Cm. de los Descubrimientos, s/n, 41092 Seville, Spain
{alorenzo,kflores1,onieva,apegado}@us.es

Abstract. The COVID-19 pandemic significantly accelerated the growth of e-commerce, placing unprecedented strain on last-mile and first-mile reverse logistics (LM&FMRL) operations. This surge led to increased service demand, requiring companies to expand their workforce rapidly. However, the long-term impact of this scaling on service effectiveness remains uncertain. This study examines the role of socio-economic factors in LM&FMRL service completion rates using a machine learning regression approach. Data was collected from operational records of carriers across more than 400 postal codes in Spain, each with at least 200 attempted services in February 2022. Socio-economic indicators, such as average age, population density, income levels, and land use, were sourced from a proprietary database. The analysis includes multiple regression models, namely Linear Regression, Decision Tree, Random Forest, Support Vector Regression (SVR), and Gradient Boosting. Results indicate that the Random Forest model demonstrated the best performance, though overall predictive accuracy remained moderate. SHAP analysis provided insights into the socio-economic drivers of service disruptions. Understanding these factors could support strategic improvements in LM&FMRL effectiveness.

Keywords: last-mile logistics · first-mile reverse logistics · socio-economic factors · regression · service disruptions

1 Introduction

After the surge of e-commerce following the democratization of internet access and the lockdown due to the COVID pandemic, the last-mile logistics industry had to endure demands, in volume and in service and cost expectations, that far exceeded the capacity installed in some of the existing companies [1]. Similarly, the demand for first-mile reverse logistics services experienced a noteworthy increase, with most e-commerce retailers offering their clients cheap or even free product returns. For instance, in the U.S., online returns jumped from 10% in January 2020 to 18% in January 2021 [2].

E. Barbadilla-Martín et al. (Eds.): CIO 2025, LNDECT 276, pp. 121–127, 2025.
https://doi.org/10.1007/978-3-032-10126-6_20

The industry, characterized by its high competitiveness and low access barrier, responded by absorbing the excess demand and hiring novel workers. For instance, Amazon announced plans to hire an additional 100,000 warehouse and delivery workers to handle the increased workload during the pandemic [3]. Similarly, Instacart expanded its workforce, hiring 300,000 more workers in response to the surge in grocery delivery demand [4]. Due to the lack of longitudinal studies evaluating the long term effects of this significant scaling, there is still logical concern over a potential decrease in service level in today's operations.

There are many different types of service disruptions that can occur in a typical B2C last-mile and first-mile reverse logistics (LM&FMRL) setting. In their framework, Lorenzo-Espejo et al. [5] identify 8 potential service disruptions in LM&FMRL operations. As per the authors, LM&FMRL service disruptions can be grouped into eight types: Missing Information, Wrong Address, Customer Absence, Service Rejection, Service Modification, Lack of Time, Incomplete Freight, and Dubious Disruptions.

By implementing a machine learning regression model, we attempt to predict the incompletion rate due to the previously described service disruptions based on the socio-economic characteristics of the service area. In order to do so, we have collated some of the main available socio-economic indicators available in a proprietary database for each postal code in many Spanish provinces. We also have extracted data from the operational databases of carriers performing delivery and collection services in these provinces. The posited methodology is explained in Sect. 2, and the main results and conclusions are presented in Sects. 3 and 4, respectively.

2 Methodology

The methodology followed in this study follows a six-step process: data gathering; data preprocessing; feature selection; regression modeling; hyperparameter tuning; and model evaluation and explainability analysis. The pipeline has been implemented in the following software: data management is performed using SQL Server Management Studio 20 and Python's pandas library [6]; the machine learning pipeline has been implemented in Sci-Kit learn [7] and the SHAP analysis is carried out with the SHAP library [8]. The coding is performed in the PyCharm IDE, and ran in a 11th Gen Intel(R) Core(TM) i7-11800H @ 2.30GHz system.

2.1 Data Gathering

The analysis pipeline starts by gathering two main sources of data: operational data from carriers in Spain; and socio-economic data at a postal code geographical level from a proprietary *Correos* database. Regarding the operational data, 437 postal codes with at least two hundred attempted services in the month of February of 2022 have been identified. Based on that data, the incompletion rate for every postal code is computed, which serve as the target variable of the models. As per the socio-economic data, the next variables are obtained for each postal code: Average Age, Average Age Female, Average Age Male, Businesses, Commercial Area, Density, Estimated Expenses, Estimated Income,

Estimated Population, Foreign Population, Large Businesses, Medium Businesses, Percentage of Single-Family Homes, Population Over 65, Population Over 65 Female, Population Over 65 Male, Population Under 18, Population Under 18 Female, Population Under 18 Male, Residential Area, Residential Weight, Self-Employed Weight, Small Businesses, Spanish Population, and Total Services.

2.2 Data Preprocessing

The process begins by splitting the raw dataset -which now includes a variable with the province in which the postal code is located- into training and test sets. The split is performed in a stratified manner to ensure that the provincial distribution is maintained in both sets, with 80% of the data used for training and 20% held out for the final evaluation. This stratification is necessary for reliably assessing the model's performance.

Data preprocessing, begins with the application of StandardScaler. The goal here is to standardize the features so that each one has a mean of zero and a standard deviation of one. This transformation is essential because many machine learning algorithms, particularly those relying on distance metrics or gradient descent such as Support Vector Regression, perform better when all features are on a comparable scale. By ensuring that the scale of each feature is normalized, the model avoids biasing its training process toward variables with larger numerical ranges, which can lead to faster convergence and more stable performance during training. Finally, the missing values have been imputed using the mean of the corresponding variables, and outliers in the target variable are discarded using the interquartile range (IQR) method.

2.3 Feature Selection

Feature selection follows preprocessing and uses the SelectKBest method in conjunction with the f_regression score function. This step evaluates each feature based on its ability to explain the variance in the target variable by calculating the F-statistic. Only the top k features, where k is treated as a hyperparameter, are selected to be passed to the regression model. This process not only reduces dimensionality but also helps in removing noisy or irrelevant features, thus simplifying the model and potentially improving its generalizability and interpretability. The value of k is tuned as part of the overall hyperparameter optimization, ensuring that only the most predictive features are used in the final model.

2.4 Regression Modeling

Regression modeling is where the selected features are used to train one of several regression models. The pipeline is designed to work with a variety of models, each offering distinct advantages. Linear Regression, for instance, models the relationship between predictors and the target variable using a straight line, making it fast and highly interpretable; however, it may struggle with capturing complex non-linear patterns. In contrast, the Decision Tree Regressor splits the data into segments based on feature values, forming a tree-like structure that is intuitive and can capture non-linear relationships but is prone to overfitting if not carefully controlled. The Random Forest Regressor

builds on this by constructing an ensemble of decision trees, each trained on random subsets of the data, and then averaging their predictions to reduce variance and improve performance. Support Vector Regression (SVR) extends the concepts of support vector machines to regression tasks by fitting a function within a defined error margin, with kernel functions enabling it to handle non-linear relationships in high-dimensional spaces. Lastly, the Gradient Boosting Regressor employs a sequential approach in which each new model attempts to correct the errors of its predecessors, offering a powerful method for capturing intricate patterns but requiring careful regularization to avoid overfitting. Each of these models is integrated into the pipeline, allowing for a comprehensive comparison based on their tuned hyperparameters.

2.5 Hyperparameter Tuning

Hyperparameter Tuning, employs RandomizedSearchCV to optimize the entire pipeline, including both the feature selection process and the parameters of the regression models. This approach randomly samples a predefined number of combinations from the hyperparameter space rather than exhaustively testing every possibility. For each sampled combination, the pipeline uses k-fold cross-validation to assess performance—commonly using a five-fold scheme—which ensures that the evaluation is robust and generalizable. The tuning process not only identifies the best model-specific hyperparameters (such as maximum tree depth for Decision Trees or the regularization parameter C in SVR) but also determines the optimal number of features k to retain during feature selection. This comprehensive optimization ensures that each component of the pipeline is calibrated to achieve the best possible predictive performance on the given data.

2.6 Model Evaluation and Explainability Analysis

After tuning, the winning model is selected based on the highest cross-validated R^2 score. This chosen model is then evaluated on the holdout test set, which was set aside earlier, to determine its performance on unseen data. To enhance transparency and provide insights into the model's decision-making process, SHAP (SHapley Additive exPlanations) is applied using a KernelExplainer. SHAP values are computed for a subset of the test data, allowing for a detailed summary plot that illustrates the contribution and importance of each feature to the final predictions.

3 Results

The posited pipeline has been implemented with the available data, and the results are provided in this section. Firstly, the model tuning and selection results are produced, followed by the testing results of the winning model once trained on the full training dataset. These are summarized in Table 1.

The results highlight the superior performance of the Random Forest regressor, especially when compared to the baseline Linear Regression. However, the overall predictive performance can only be considered low to moderate. This is also evidenced if the predicted values are plotted against the real observations, as pictured in Fig. 1.

Figure 1 shows that, while the predictions are relatively accurate for the lower, more typical observations of the incompletion rate, the model is not able to capture the variability of the real values as it increases.

Table 1. Model selection and testing results.

Model	R^2	MAE	RMSE	MAPE	Tuning time (s)	Training time (s)	Testing Time (s)
Linear Regression	0.135	0.108	0.144	0.671	1.069	0.003	0.001
Decision Tree	0.031	0.113	0.152	0.689	26.653	0.003	0.001
Random Forest	0.207	0.103	0.139	0.628	473.228	0.237	0.014
SVR	0.185	0.105	0.140	0.659	23.598	0.005	0.001
Gradient Boosting	0.177	0.105	0.142	0.641	172.759	0.327	0.002
Random Forest (selected)	0.337	0.090	0.118	0.554	–	0.237	0.015

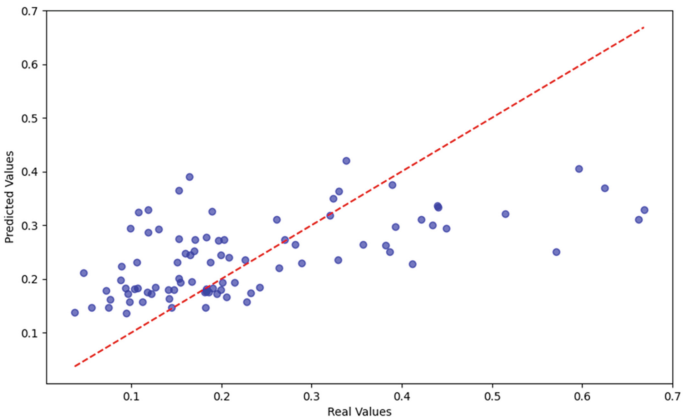

Fig. 1. Predicted values plotted against the real observations of incompletion rate.

Finally, the SHAP analysis results have been summarized in Fig. 2, which shows how features impact the prediction of the percentage of incomplete services.

Each point on a row represents a single postal code's SHAP value for that feature (i.e. that feature's contribution to the model's prediction for that postal code). The features are listed on the y-axis, sorted by overall importance (mean absolute SHAP value), and only

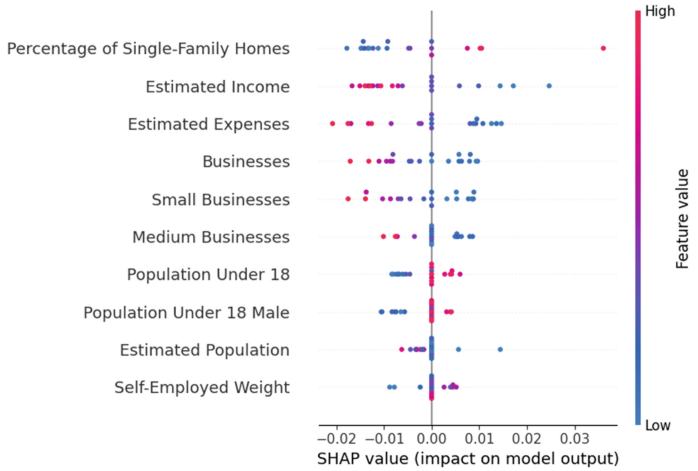

Fig. 2. SHAP plot.

the ten most relevant are shown. The horizontal position of a point shows whether that feature increased the predicted incomplete service percentage (positive SHAP value, to the right) or decreased it (negative SHAP value, to the left) for that specific postal code. The color indicates the actual feature value in that postal code – red means the feature's value is high, and blue means it's low for that area. This way, we can see not only which features are most important, but also how high vs. low values of each feature affect the outcome. For example, according to the model, more single-family homes correspond to more incomplete services while, in postal codes with higher income and expenditure, less incomplete services are expected. Moreover, for areas with lower business density, more incompletions are predicted. Conversely, for postal codes with younger population, a higher service disruption rate is expected.

4 Conclusions

In this work, a study for the analysis of the input of socio-economic factors in LM&FMRL service completion effectiveness is presented. Although the configured regression model is only able to produce low to moderate prediction results, the fact that a R^2 over 33% can be achieved simply with postal code-aggregates of several socio-economic indicators shows potential in this direction. A deeper understanding of these factors may pave the way for targeted strategies to boost LM&FMRL effectiveness.

Acknowledgments. This research has been financially supported by project CAROLUM (PID2021-125125OB-I00), funded by MICIU/AEI/10.13039/501100011033 and the ERDF, UE.

References

1. Li, Z., Gu, W., Meng, Q.: The impact of COVID-19 on logistics and coping strategies: a literature review. Reg. Sci. Policy Pract. **15**(8), 1768–1794 (2023)

2. DHL: The impact of COVID-19 on consumer returns. https://www.dhl.com/discover/en-global/logistics-advice/logistics-insights/what-impact-has-covid-19-had-on-consumer-returns, last accessed 2025/03/19
3. CNBC: Amazon to hire 100,000 warehouse and delivery workers. https://www.cnbc.com/2020/03/16/amazon-to-hire-100000-warehouse-and-delivery-workers.html, last accessed 2025/03/19
4. Business Insider: Instacart's army of shoppers has exploded from 180,000 to 500,000 since the start of the pandemic—and some workers say it's making the job more difficult for everyone. https://www.businessinsider.com/instacart-hiring-spree-coronavirus-working-conditions-worse-for-everyone-report-2020-5, last accessed 2025/03/19
5. Lorenzo-Espejo, A., Muñuzuri, J., Pegado-Bardayo, A., Guadix, J.: A framework for analyzing service disruptions in last-mile and first-mile reverse logistics. Res. Transp. Econ. **108**, 101485 (2024)
6. McKinney, W.: Data structures for statistical computing in python. In: van der Walt, S., Millman, J. (eds.) Proceedings of the 9th Python in Science Conference, pp. 51–56 (2010)
7. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
8. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774. Curran Associates, Inc. (2017)