Performance report

Regression models for estimating semi-trailer truck loading duration

Application of machine learning models to estimate the duration of semi-trailer truck loading maneuvers for a tissue paper distribution plant, for both bulk and pallet operations, given product quantity, physical dimensions, and number of active operators.

J. Edgar Hernandez Cancino Estrada
ITESM
Candidate for B. S. in Robotics and Digital Systems
(Expected June, 2024)

Monterrey, Nuevo Leon, Mexico
September, 2023

Abstract

This report describes the application of machine learning regression models for estimating the loading times of trailers carrying tissue paper products. The models are based on various factors, including product quantity, weight and spatial dimensions of the products, and the number of operators involved in the loading process. A manually collected dataset was utilized, gathered from a specific distribution plant where these operations usually take place.

Three machine learning methods, Ridge and Support Vector Regression, and a Multi-Layer Perceptron Neural Network, were applied and compared. Furthermore, individual models were developed for both bulk and pallet loading, taking into account the unique characteristics and requirements of each loading scenario.

This work serves as a precursor to an ongoing project, aimed at achieving a substantial improvement in the optimization of the operation of distribution plants. The ultimate goal of this research is to leverage thd dataset to develop a precise and reliable tool for estimating the required times for these operations. The results obtained in this study lay the foundation for the further development of this project, demonstrating the potential and significance of the employed regression machine learning methods for this end.

Keywords: Regression, Ridge, Support Vector Regression, Machine Learning, Neural Network, Multi-Layer Perceptron

Contents

1. Background

In the quest to streamline trailer loading operations and the overall performance of distribution plants, it is crucial to ascertain the amount of a plant's maximum capacity being utilized. This information allows us to assess the current level of efficiency and identify opportunities for improvement and optimization in the operations.

To achieve this goal, to have a reliable time-estimating tool for a scheduled operation, can provide valuable insights into both the capacity and performance of the distribution plant. This study aims to leverage this dataset to gain a deeper understanding of the behavior of truck loading maneuvers, to produce a standard duration estimation model, allowing for informed decision-making and improved logistical planning.

2. Data

The dataset employed in this project comprises the historical records of truck loading maneuvers, manually documented daily over a period of more than 2 years. These maneuvers include bulk and pallet loading cases, for which two independent models were developed. Crucial attributes such as the quantity of loaded product, number of working stevedores per operation, the product's SKU, and the recorded duration of each maneuver are considered, and the dataset was enriched with additional dimensional data, specifically, product weight and physical dimensions associated to each individual SKU. The bulk data contained 29 unique product SKUs, whereas the pallet data 15.

Distinct and separate analyses were conducted for both bulk and pallet data, aiming to produce two distinct models, one for each type of maneuver. The objective behind this approach is to develop specialized models tailored to the specific characteristics and requirements of each type of loading operation.

Originally, the dataset consists of 14080 recorded operations for bulk maneuvers and 13264 operations for pallet maneuvers. Due to the high presence of noise in the data, after the strict cleaning procedure, the final useful data comprises around 6100 and 4100 operations for each maneuver, respectively.

2.1 Initial analysis

Since the dataset was collected manually, high noise and variability in the recorded target variable is expected.

Figures 2.1 and 2.2 display the distribution of recorded durations for the 99^{th} precentile of bulk and pallet maneuvers data. Noticeable peaks are observed at durations that are multiples of 5 or 10, followed by a significant decrease in occurrences of other, more precise durations. This pattern strongly suggests that operations have been rounded to these convenient numbers, indicating the presence of a slight bias in the precision of recorded durations.

Given these considerations, the primary goal of the data cleaning process is to keep

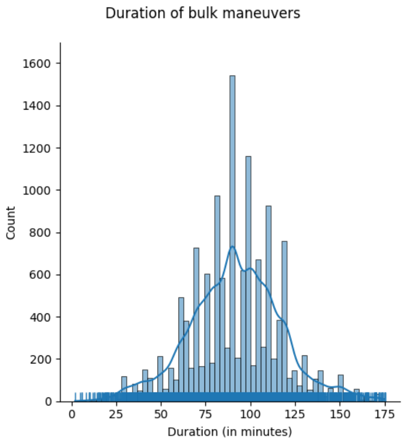

Figure 2.1: Distribution of the duration of bulk maneuvers.

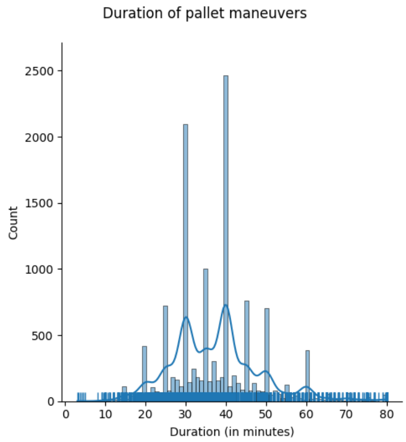

Figure 2.2: Distribution of the duration of pallet maneuvers.

only consistent, relevant operations, and identify and eliminate any potential behavioral outliers present in the data. By doing so, the aim is to enhance the overall data quality and ensure that subsequent analyses are based on our preferred 'standard' maneuvers.

The correlation matrix presented below in Figure 2.3 reveals that for bulk maneuvers, the primary factors affecting the final duration are the product quantity and the number of stevedores involved, while the physical dimensions and weight show little to no correlation. In contrast, Figure 2.4 shows no relevant relationships between the target value and the features involved for pallet maneuvers. This will change after the cleaning process, when the correlation between physical features and the target variable will become more evident for both types of maneuvers.

Although, the correlation matrix for pallet data suggests rather noisy, unrelated data, the same cleaning methodology will be used for both types of maneuvers, as initially the pallet data does not suggest any other evident approach. This, will eventually prove to be a successful approach for both types maneuvers and increase both the correlations and overall reliability of the data.

2.2 Data cleaning and the SPE metric

To mitigate recording bias and replicate a natural variability of the data, Gaussian Noise with a standard deviation $\sigma = 0.5$ and mean $\mu = 0$ was introduced in the target variable for both bulk and pallet data. The resulting distributions of these datasets are shown in figures 2.5 and 2.6.

Now, considering the anticipated high variation among similar maneuvers due to any unaccounted factors and the nature of a manual data collection process, because the interest resides in producing 'standardized' duration estimations, the objective is to focus solely on those recorded operations that exhibit a 'standard' behavior, i.e. those that exhibited a consistent relationship between the most relevant features and the target variable.

Figure 2.3: Correlation matrix for bulk maneuvers.

For this purpose, the metric **SPE** (Spanish acronym for *Segundos-Pieza-Estibador*, which translates to Seconds-Piece-Stevedore) was defined:

$$SPE = (\frac{duration}{n_{pcs}}) * n_{stv} \tag{1}$$

$duration$ = duration in seconds

n_{pcs} = product quantity in the maneuver

n_{stv} = number of stevedores.

SPE Refers to a number of seconds in average it took each stevedore to move a single piece of product. This metric holds the relationship between the most important variables describing each operation, which should remain more or less stable in 'standardized' maneuvers, as we assume a linear, direct relationship between the number of workers and the speed with which the maneuver is completed.

Having computed the SPE for all data, strong restrictions were applied. Specifically, data was limited to the 96^{th} percentile of duration, and then, only data contained in the 60^{th} percentile of SPE was accepted. This number was found empirically, as resulted in a satisfactory performance of the regression models. Other basic cleaning filters were applied, such as removing NaN values or duplicated data.

The first cleaning approach resulted in a significant loss of training data, but it brings us closer to a model for estimating standardized maneuver durations. Through this methodology, we aimed to isolate maneuvers that exhibit a more reliable pattern amidst

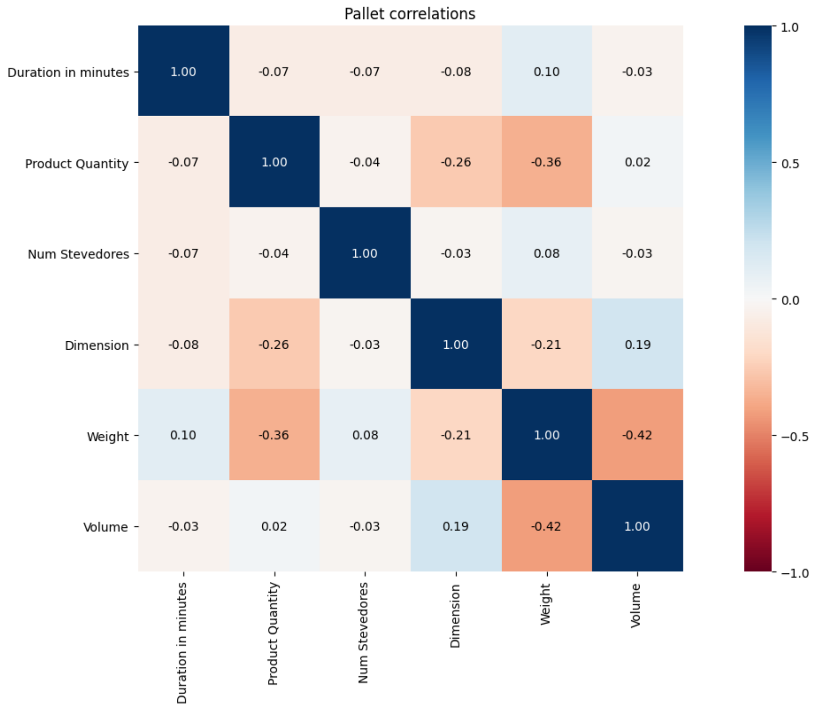

Figure 2.4: Correlation matrix for pallet maneuvers.

the overall variability in the dataset and improve the performance of the resulting predictive models.

It is important to mention that another fundamental part of the cleaning process, was addressing *multi-maneuver* problem, which will be further explained and discussed in subsection 2.3. This issue rendered 42% of the bulk data and 32% of the pallet as invalid.

Figures 2.7 and 2.8 present the coefficient of variation for SPE within various time intervals and the number of occurrences for each group are presented. It is important to mention that for these graphs, the data affected by the *multi-maneuver problem* was not included.

It is evident that for both, bulk and pallet maneuvers, there exists an inverse relationship between the variation and the duration of the maneuvers. Higher variation is encountered in shorter maneuvers, while lower variation is observed for longer maneuvers. This does not appear to be strictly related to the number of operations registered in the duration group.

2.3 The multi-maneuver problem

One of the most concerning issues encountered in the dataset, is the one refered as the *multi-maneuver* problem, which affected approximately 40% of bulk maneuvers data and 32% of pallet maneuvers data.

The problem arose when multiple types of products were loaded into a single truck, which is a typical occurrence, thus resulting in a significant number of affected ma-

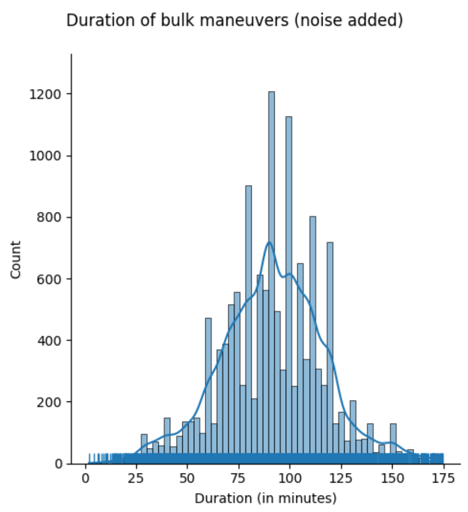

Figure 2.5: Distribution of the duration of bulk maneuvers (Gaussian noise added).

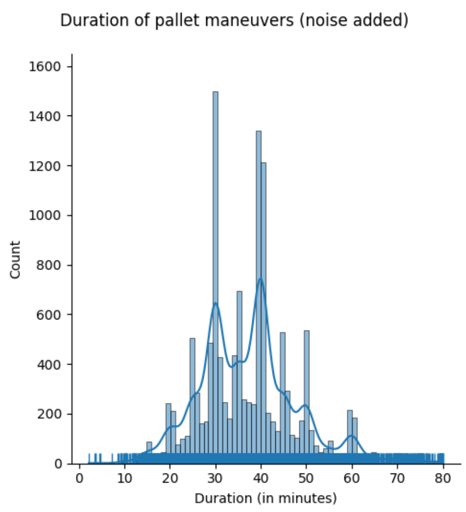

Figure 2.6: Distribution of the duration of pallet maneuvers (Gaussian noise added).

neuvers. In such cases, the operators recorded these maneuvers as one, using the same maneuver unique identifier but recording different product SKUs while maintaining the same total registered duration (see figure 2.9). Consequently, for these maneuvers involving multiple products, only the total duration was recorded, while the individual duration for each sub-operation remained unknown.

Initially, a possible approach to address this problem was to disregard the distinctions between different products, as the correlation matrices indicated that the physical attributes of the products did not exhibit a significant relationship with the total duration. However, during the modeling stage of this study, it became evident that including these attributes actually enhanced the models' performances. Therefore, it became necessary to device an imputation method to recover this data, which had been previously declared invalid.

To address this issue, the following strategy was followed:

1. Extract the average *SPE* for each individual product (based on their SKU).
2. Group together all registers that belong t the same maneuver.
3. For every individual sub-operation, estimate each product's individual duration as a fraction of the whole operation as follows:

   First define the useful vectors:

$$\vec{v}, \quad \text{where } v_i = i^{th} \text{ product's average } SPE$$
$$\vec{q}, \quad \text{where } q_i = i^{th} \text{ product's quantity}$$

   Loading speed ratio between each product and one reference product (the one with the highest *SPE* value) is extracted in $\vec{r}$:

$$\vec{r} = \frac{\max(\vec{v})}{v} \tag{2}$$

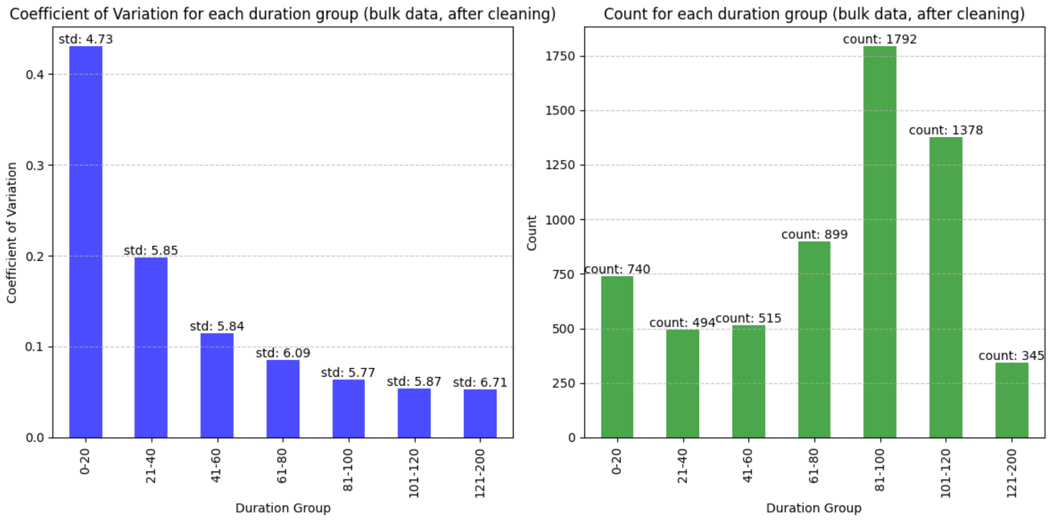

Figure 2.7: Coefficient of variation and standard deviation of SPE per duration group in bulk maneuvers.

To use the ratios to estimate individual durations that sum to the total duration registered, a constant $k = \frac{t}{\vec{q} \cdot \vec{r}}$ is introduced, where t is the total duration registered.

Finally, the individual durations are calculated:

$$\vec{d} = k \, (\vec{q} \cdot \vec{r}) \tag{3}$$

At this point, $\vec{d}$ would hold every i^{th} product's individual duration, which elements will sum to t.

This strategy allows to recover $> 90\%$ of the data lost due to the *multi-maneuver problem* for both bulk and pallet operations. However the same strict data cleaning restrictions are applied to this data, before entering the training and modeling stages.

2.4 Final data analysis

Once the data is cleaned and the *multi-maneuver problem* has been addressed. A descriptive analysis can be performed on the data that will ultimately be used to train the regression models.

2.4.1 Target variable distribution

The duration of the collected bulk maneuvers shows a bimodal distribution (see figure 2.10). This occurs as a consequence to a high number of short duration maneuvers that were recovered from longer, multi-composed operations. This can be seen in figure 2.11, where distribution plots for recovered and original maneuvers are compared.

The majority of bulk operations are concentrated within the 80-120 minutes duration range.

Figures 2.12 and 2.13 show plots for pallet maneuvers, where we find similar behavior.

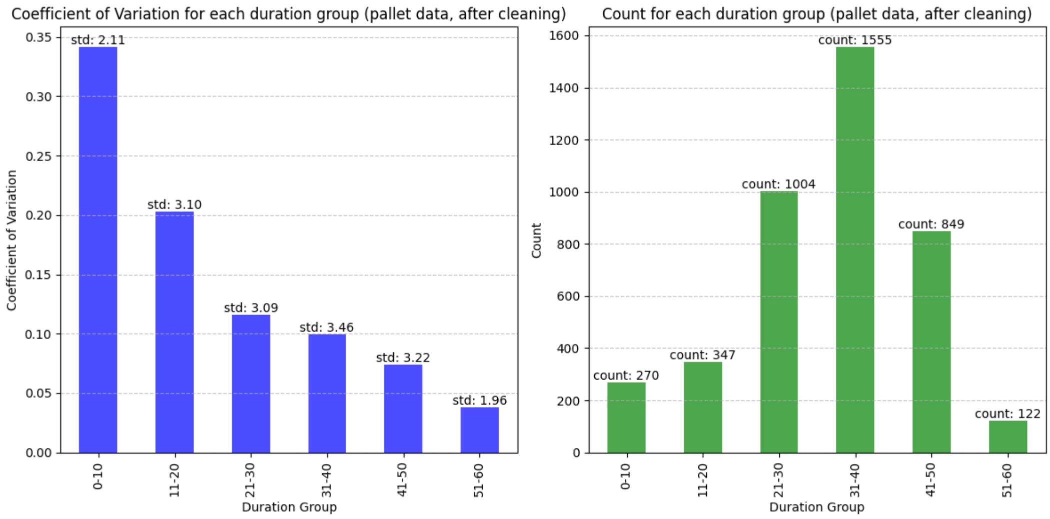

Figure 2.8: Coefficient of variation and standard deviation of SPE per duration group in pallet maneuvers.

index	id	date	sku	product qty	maneuver type	start time	end time	duration	num stevedores
5805	134830	2021-03-29	90217	568	CARGA GRANEL	2021-03-29 07:45:00	2021-03-29 09:00:00	01:15:00	2
5806	134830	2021-03-29	837	83	CARGA GRANEL	2021-03-29 07:45:00	2021-03-29 09:00:00	01:15:00	2

Figure 2.9: Two data registers, showing the multi-maneuver problem. Both have matching id number, date, and start and end times, while product SKU and quantities are different, but still the same duration is recorded.

The duration distribution appears left-skewed, concentrating most of the operations in the range between 30 and 40 minutes.

2.4.2 Feature correlations

Finally, to get an understanding of the relationship existing between the operation's features, the clean data's correlation matrices and scatter plots are analyzed.

Figure 2.14 depicts improved correlations between the target variable (duration) and the most significant features in bulk operations clean data. Notably, there are substantial increases in correlation values observed for the following key features:

- Product quantity, which has risen from 0.14 to 0.81.
- Number of stevedores, which has improved from -0.16 to -0.24.
- Dimensional data, which has transitioned from having insignificant correlation to coefficients ranging from around 0.10 to 0.15.

Figure 2.15 exhibits a similar behavior for pallet data's correlations, with significant increases for the following features:

- Product quantity, which has risen from -0.07 to 0.58.
- Number of stevedores, which has improved from -0.07 to -0.24.

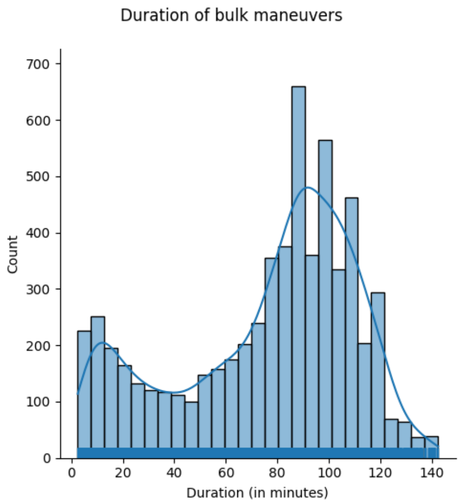

Figure 2.10: Distribution of the duration for clean bulk maneuvers (in minutes).

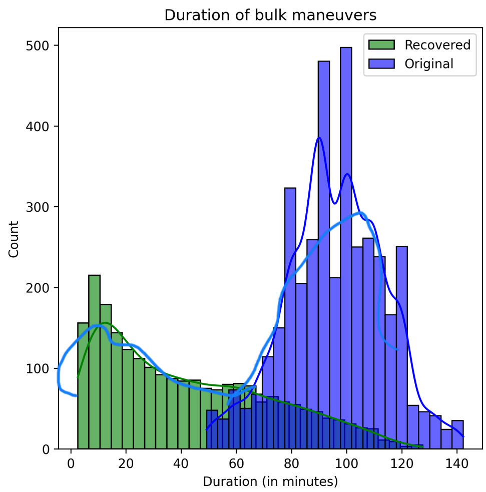

Figure 2.11: Comparison of duration distribution for original and recovered bulk maneuvers.

All these enhanced correlation values indicate a stronger relationship between the total duration and the specified features in the operations data.

The correlation matrices also show certain degree of collinearity present among the explanatory variables for both bulk and pallet operations, specifically among the physical dimension features.

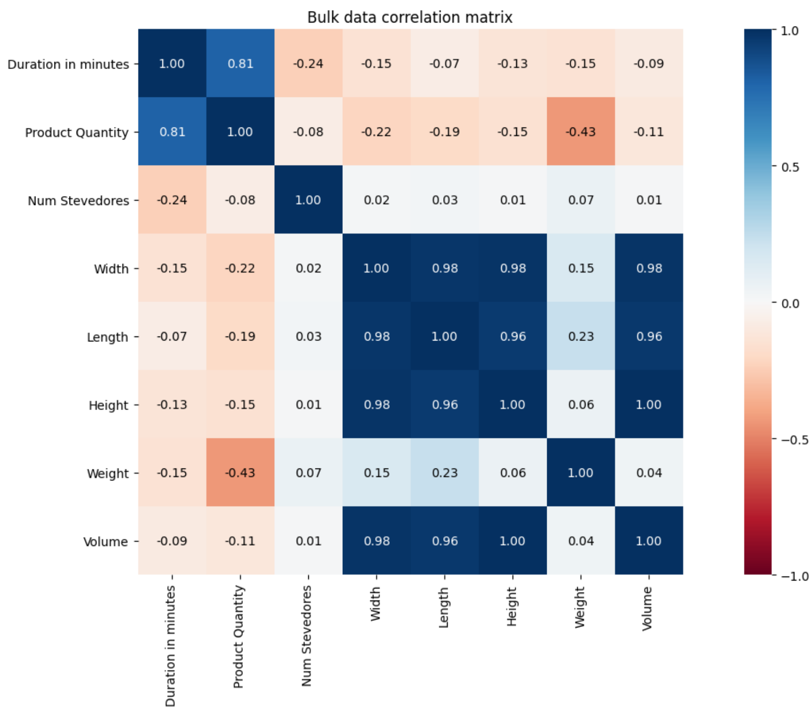

Figure 2.14: Correlation matrix for clean bulk maneuvers.

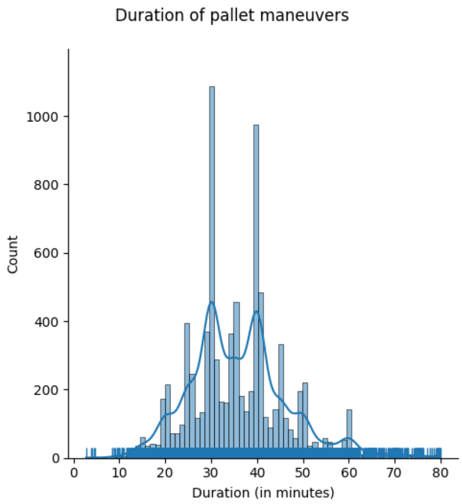

Figure 2.12: Distribution of the duration for clean pallet maneuvers (in minutes).

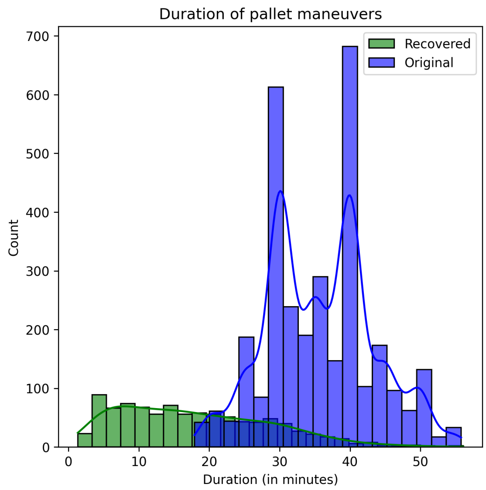

Figure 2.13: Comparison of duration distribution for original and recovered pallet maneuvers.

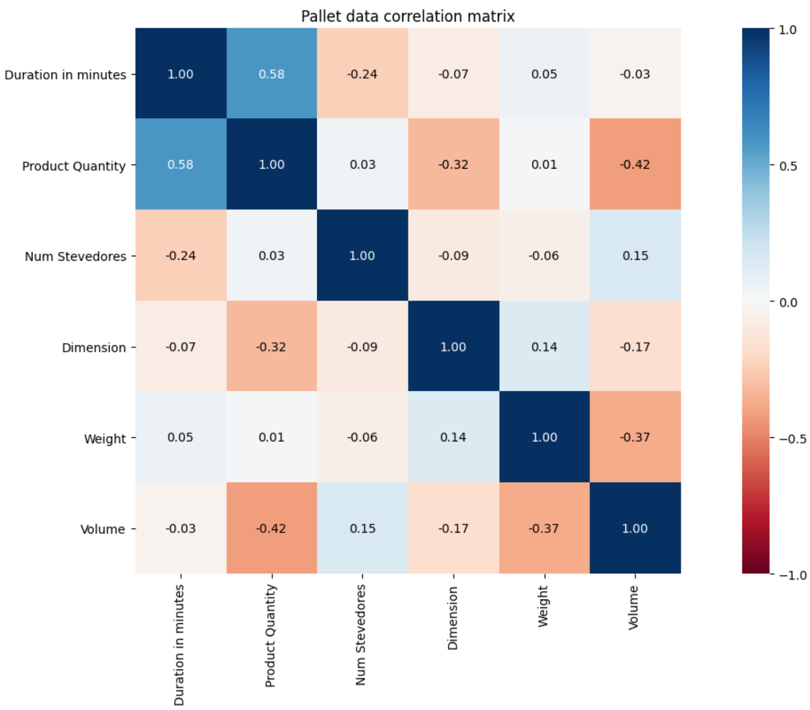

Figure 2.15: Correlation matrix for clean pallet maneuvers.

2.4.3 Feature scatter plots (bulk)

Displayed below are plots illustrating the relationship between the most significant features and the target variables. In these plots, the feature with the highest correlation, namely '*product quantity*', serves as the x-axis, while the target variable is represented on the y-axis. Additionally, the dots on the plots are color-coded to unveil latent relationships between less correlated features and the aforementioned ones. The graphs include a randomly selected sample, constituting 30% of the dataset's total size.

In Figure 2.16, it becomes evident that a discernible divergence in the duration patterns emerges for operations with 3 stevedores while compared to those with only two: the duration time tends to be reduced as the number of stevedores increase for operations of similar quantities. These patterns appear to exhibit a significant degree of linear separability for an important subset of cases. However, a few operations display a marginal deviation, indicating a notable upsurge in duration for the same product quantity within the range of approximately 600 to 800 products. This phenomenon can be attributed to the dimensional attributes inherent to these operations. Figures 2.17 and 2.18 associate duration, product quantity, and dimensional characteristics, providing compelling evidence that these features are responsible for the observed variance in behavior. It is important to notice that this behavior does not appear to be linear.

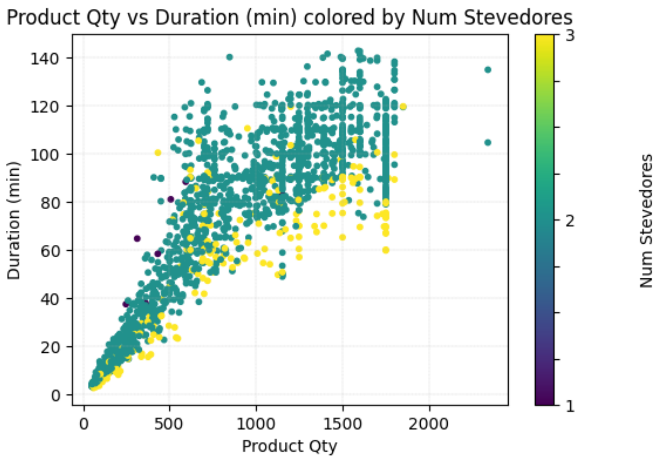

Figure 2.16: Product quantity vs. duration, colored by number of stevedores. Notably, 98.5% of the operations recorded two to three stevedores (87.3% and 11.2% respectively).

These plots provide a comprehensive visualization of evident interrelationships within the dataset features. However, a consistent observation across all these plots is clear: as the product quantity decreases, the variations in the behavior of operation duration due to its dimensional features become less discernible.

Figures 2.19 and 2.20 present similar plots where the quantity values have been constrained to a maximum of 650 products. As previously mentioned, the discernible patterns that were evident for higher quantities now exhibit a less pronounced behavior as the product quantity decreases.

2.4.4 Feature scatter plots (pallet)

In this section, analogous plots for pallet data depict relationships between duration, quantity, and additional attributes. The data illustrates a similar distribution behavior to bulk data in scatter plots when factoring in product quantity, duration and number of stevedores (figure 2.21). The dimension feature mildly shapes the colored distribution. having a similar effect as the dimensional features for bulk operations (see figure 2.22). Nonetheless, weight and volume features exhibit little to indistinct effects, rendering the corresponding plots noisy (figures 2.23 and 2.24). The homogenization effect of low quantities is also present in pallet data.

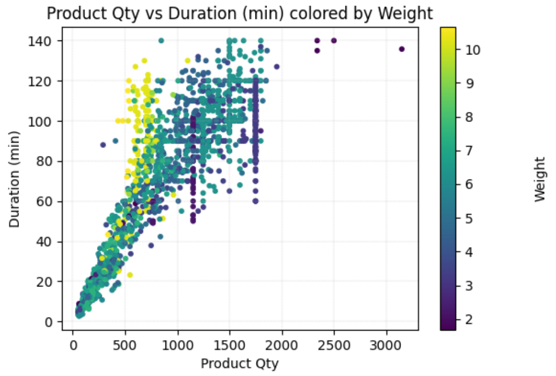

Figure 2.17: Product quantity vs. duration, colored by weight. Weight data has been limited to its 95th percentile.

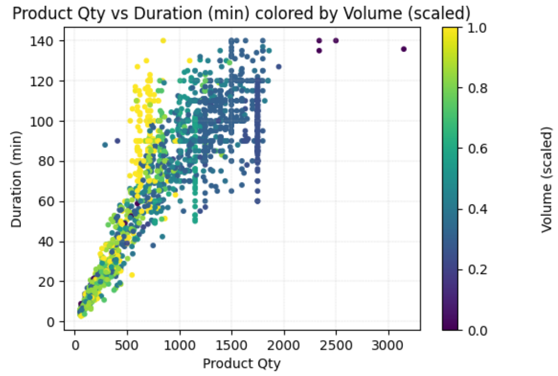

Figure 2.18: Product quantity vs. duration, colored by volume. Volume data has been scaled and limited to its 95th percentile.

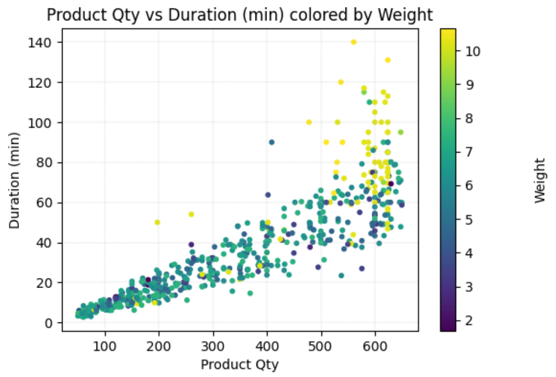

Figure 2.19: Product quantity vs. duration, colored by weight. Weight data has been limited to its 95th percentile.

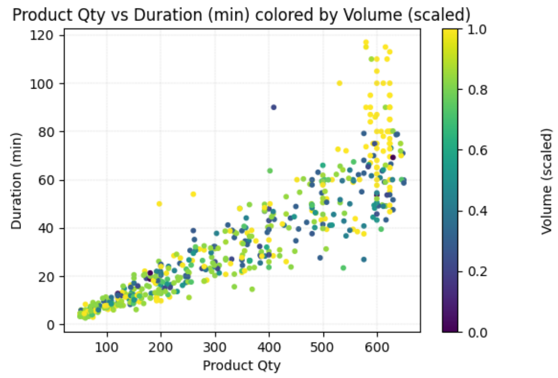

Figure 2.20: Product quantity vs. duration, colored by volume. Volume data has been scaled and limited to its 95th percentile.

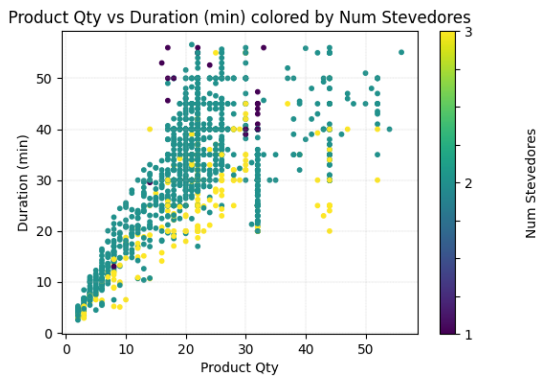

Figure 2.21: Product quantity vs. duration, colored by number of stevedores.

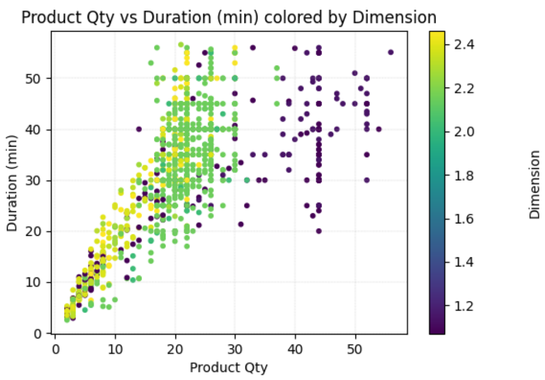

Figure 2.22: Product quantity vs. duration, colored by Dimension.

For pallet plots, 50% of the pallet data was utilized. Due to the comparatively weaker impact of the available attributes on duration compared to bulk data, one might antic-

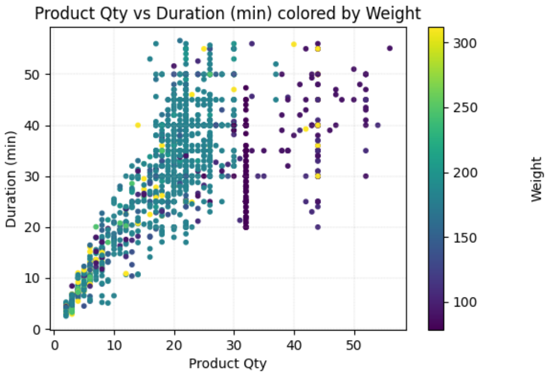

Figure 2.23: Product quantity vs. duration, colored by weight. Weight data has been limited to its 95$^{\text{th}}$ percentile.

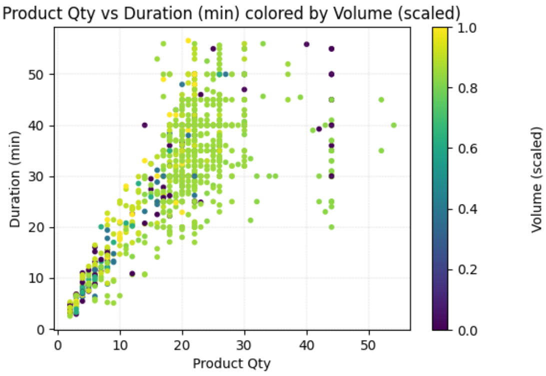

Figure 2.24: Product quantity vs. duration, colored by volume. Volume data has been scaled and limited to its 95$^{\text{th}}$ percentile.

ipate reduced influence of these features to shape the pallet estimation models. This phenomenon could be attributed to factors such as the lesser amount of available data, increased operational noise, or potentially weaker inherent relationships among the available predictor variables.

2.5 Data variability

An important finding regarding the available data's behavior, is the high presence of variability on recorded maneuvers, while the available features remain unvarying.

2.5.1 Variability on bulk products maneuvers

Let's recall the most significant predictor features available for bulk operations and their correlation to the target variable:

Product quantity	*Num. of stevedores*	*Width*	*Length*	*Height*	*Weight*	*Volume*
0.81	-0.24	-0.15	-0.07	-0.13	-0.15	-0.09

If products are grouped by their SKU, all the dimensional features remain constant, and a visual analysis of operations' duration behavior can be carried out, when considering only the remaining variables: *product quantity* and *number of stevedores*. Figures 2.25 to 2.28 show these scatter plots for the four SKUs with the most number of occurrences.

It is evident that for constant dimensional attributes, there still exists a high amount of variability in the target variable, which could be caused either for recording noise, or either undiscovered or random factors. This also explains the relatively low correlation shown for the dimensional variables, as even when they remain unvarying, there is substantial variation in the recorded duration for the operations.

At last, it is relevant to notice how SKU 222 originally presented a high accumulation of points at exactly *product quantity* $= 1750$. This is explained because the product

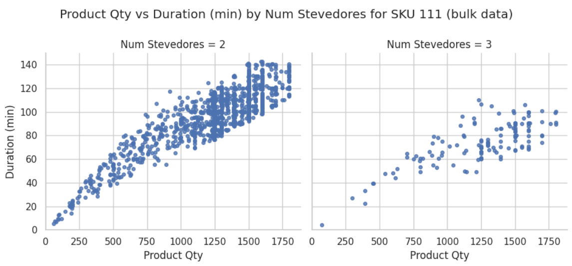

Figure 2.25: SKU 111

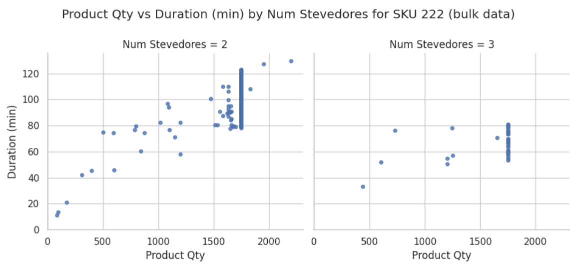

Figure 2.26: SKU 222

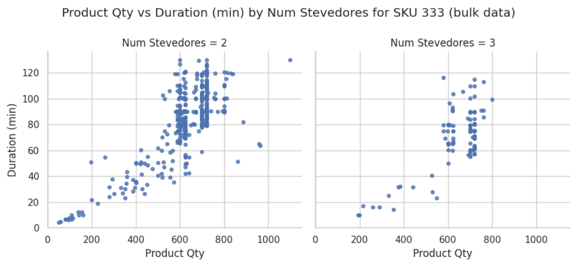

Figure 2.27: SKU 333

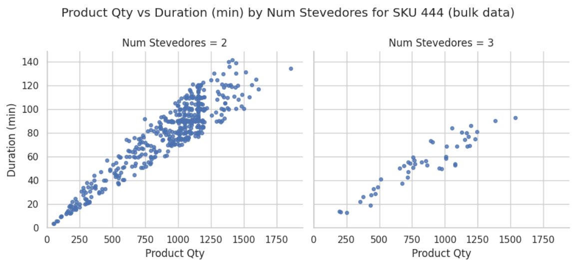

Figure 2.28: SKU 444

in question is usually dispatched in this singular quantity. Here, is evident how there can exist high variation on the recorded durations, even while all the remaining known variables stay unvarying. Because of this, several duplicate values with these varying targets might exist in the training dataset. To mitigate the potential adverse effects of these duplicated data, Gaussian noise with $std = 1.5$ was added to the quantity values, and only 25% of data with a high number of duplicate values was kept while the rest was discarded this (see SKUs at appendix C).

2.5.2 Variability on pallet products maneuvers

To perform a similar analysis on pallet operations, the most significant features are considered:

Product quantity	*Num. of stevedores*	*Dimension*
0.58	-0.24	-0.07

Figures 2.29 to 2.32 display scatter plots of durations for the four most frequently occurring SKUs in the pallet data. Several insights emerge:

- The product quantities in pallet data exhibit an important level of discretization, with frequently recurring numbers of ordered and loaded products.
- This highlights that the target variable (duration) also demonstrates significant variability, even when quantity and dimension remain constant.
- These plots encompass all the available relevant features, and given the continuing variability, it becomes apparent that any model will only generalize the data's real-life behavior to a certain extent.

To try to minimize the effect of the duplicate values in training, Gaussian noise of $std = 0.2$ was added to the product quantity attribute of pallet data, and only 25% of the occurences of those SKU's with a high number of duplicate values was kept (take a special look to SKU 999 in pallet operations in appendix C).

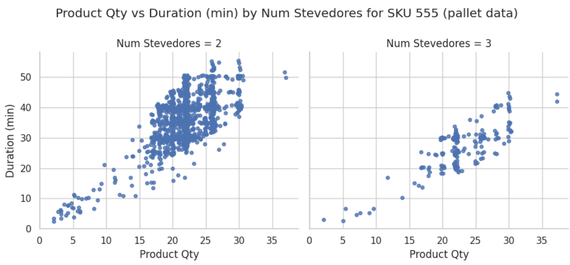

Figure 2.29: SKU 555

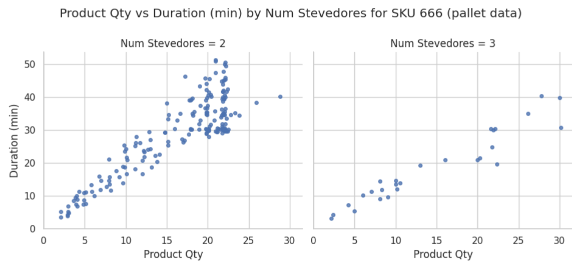

Figure 2.30: SKU 666

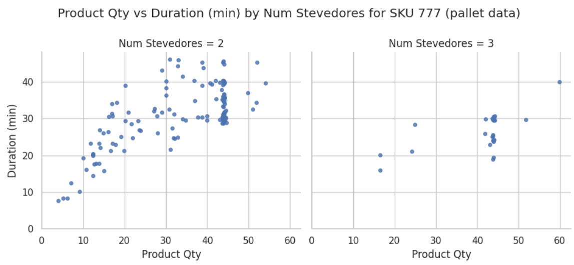

Figure 2.31: SKU 777

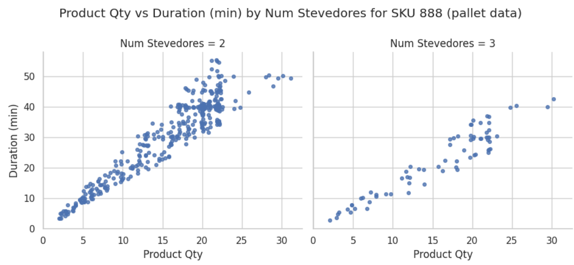

Figure 2.32: SKU 888

3. Modelling

Three regression techniques were employed and assessed: Ridge Regression, Support Vector Regression (SVR), and a Multi-Layer Perceptron Neural Network (MLP).

- **Ridge Regression**:
  Chosen due to its robustness to outlier and collinearity in multiple linear regression [4]. It was expected to face a strong presence of outliers, which would likely undermine typical linear regression methods.

- **Support Vector Regression**:
  This model is ideal to work with a non-linear dataset as the non-linear kernel 'rbf' is utilized. It also has excelent generalization capability coupled with high prediciton accuracy. SVR demonstrates robustness to outliers, as it imposes equal penalties for both high and low misestimates [1]. The inherent use of an ϵ-insensitive loss function allows the model to 'accept' and ignore errors up to a certain degree (the ϵ-insensitive region) [2]. This makes the model robust to minor noise, which aligns with the anticipation of potential biases in duration recordings. This model uses the ϵ-sensitive loss function.

- **Multi-layer Perceptron**:
  Neural networks have the capability to unveil complex and even non-linear relationships for regression problems [5]. Gaudart, Giusiano, and Huiart [3] show that while MLPs cannot replace linear regression models and may not outperform them under optimal usage, they highlight the utility of MLPs when the

real underlying regression function of a phenomena cannot be approached by classical methods, or when colinearity or non-linear regression functions become present. A Multi-Layer-Perceptron Neural Network is a popular, flexible architecture choice that is frequently used for regression tasks and given its capacity to work under data with complex patterns, it is worth evaluating for the goal of this research.

An empirical, constructive approach was employed to determine a suitable Multi-Layer Perceptron (MLP) architecture. The code implementation incorporates the early stopping technique and L2 regularizer to every perceptron layer to mitigate overfitting concerns (see apendix B.3).

All models were provided with the most pertinent attributes for each maneuver type, and polynomial features of second order were generated to potentially capture higher-order interactions among them.

The procedure for training and evaluating each of these three models for both bulk and pallet maneuvers followed a similar course, encompassing the following steps:

1. Data normalization using *sklearn*'s *Standard Scaler*.
2. Train-test split leaving 85% of data for training and 15% for testing.
3. Generate quadratic features for the data (both degrees 1 and 2 were tried).
4. Conduct 8-fold Cross-Validation to derive performance metrics, including R^2, MAPE, and MAE.
5. Train the model using the training data, using the most relevant features for each type of operation.

   For bulk maneuvers: *Product quantity*, *Number of stevedores*, *Width*, *Length*, *Height*, *Weight*, and *Volume*.

   For pallet maneuvers: *Product Quantity*, *Number of stevedores*, and *Dimension*
6. Generate predictions

Various other train-test splits and normalization techniques were experimented with, but the ones ultimately chosen are those that demonstrated the best performance.

4. Results

In this section each model's results and scores will be analyzed. Ridge Regression of degree 2 will be referred as Ridge2, and the Multi-Layer Perceptron and Support Vector Regression models will be MLP1, MLP2, SVR1, and SVR2, depending on the degree of their features.

Model	R^2	MAPE	MAE
Ridge (deg=2)	0.868	18.781	9.848
MLP (deg=1)	0.859	16.301	9.491
MLP (deg=2)	0.864	15.951	9.431
SVR (deg=1)	0.879	15.375	9.338
SVR (deg=2)	0.874	16.633	9.549

Table 4.1: Test performance metrics for bulk maneuvers models.

4.1 Test performance analysis for bulk data

Table 4.1 presents each model's performance metrics, where most models exhibit similar performances. The following insights are drawn from these evaluation scores:

- SVR1 outperforms other models in all R^2, MAPE, and MAE scores.
- SVR2 maintains a similar R^2 but slightly decreases MAPE and MAE performance in comparison to SVR1.
- MLP1 and MLP2 display similar performances to the other models, with a competitive MAE.
- There are no important differences in the performances of MLP and SVR models, although the small scoring differences remained present in several trials.
- Also, when adding complexity to the available features, i.e., their quadratic relationships, there is no significant improvement in the prediction accuracy. However, the non-linear models have an improved performance in comparison to the Ridge Regression linear model.

Table 4.2 shows an extract of 10 random samples of the test data used for MLP2 model, where the predicted and target values can be compared.

Prediction	Target	APE	AE
120	141	0.15	21
86	91	0.05	5
102	84	0.21	18
100	92	0.09	8
100	114	0.13	14
90	95	0.05	5
33	39	0.17	6
89	64	0.38	25
51	52	0.02	1
51	50	0.03	1

Table 4.2: Comparison between Predictions and Targets (MLP2 model)

Model	R^2	MAPE	MAE
Ridge (deg=2)	0.659	19.147	5.188
MLP (deg=1)	0.615	15.931	4.661
MLP (deg=2)	0.625	15.976	4.68
SVR (deg=1)	0.735	15.875	4.556
SVR (deg=2)	0.733	16.198	4.573

Table 4.3: Test performance metrics for pallet maneuvers models.

4.2 Test performance analysis for pallet data

Table 4.3 shows the performance metrics corresponding to pallet data. Here, a similar pattern in performance metrics is observed, when compared to bulk operations.

- R^2 is considerably lower for all of the pallet models, when compared to bulk models' performances. Based on this score, SVR1 and SVR2 appear to be the best descriptive models.
- MLP and Ridge models have a significantly lower R^2 value than SVR, but MLP still has competitive MAPE and MAE performances.
- MAPE remains relatively consistent in the majority of the models, with the exception of Ridge Regression where the performance is decreased considerably.
- SVR1 and SVR2 have the best MAE performances, both scoring around 4.6 minutes, while the next best models, MLP1 and MLP2 score about 4.7 minutes of absolute error. This difference, although of small value and probably significance, remained consistent as several trials were carried out.
- The increasing of data features' order did not result in an improvement of the predictions accuracy, but rather resulted in a slight increase of the errors.

Table 4.4 shows 10 random samples of the test predictions of the SVR1 model for pallet data.

Prediction	Target	APE	AE
31	31	0.03	0
3	3	0.31	0
38	35	0.11	3
30	31	0.01	1
36	30	0.21	6
40	41	0.04	1
13	11	0.10	2
36	44	0.17	8
37	25	0.47	12
18	10	0.70	8

Table 4.4: Comparison between pallet Predictions and Targets (SVR1 model).

4.3 Visual analysis for bulk maneuver models

Figure 4.1 shows the target versus prediction plots for the bulk maneuver models. Consistently, the dots follow linear-like distributions; however, as the duration increases, the points start to spread more widely around the $y = x$ line.

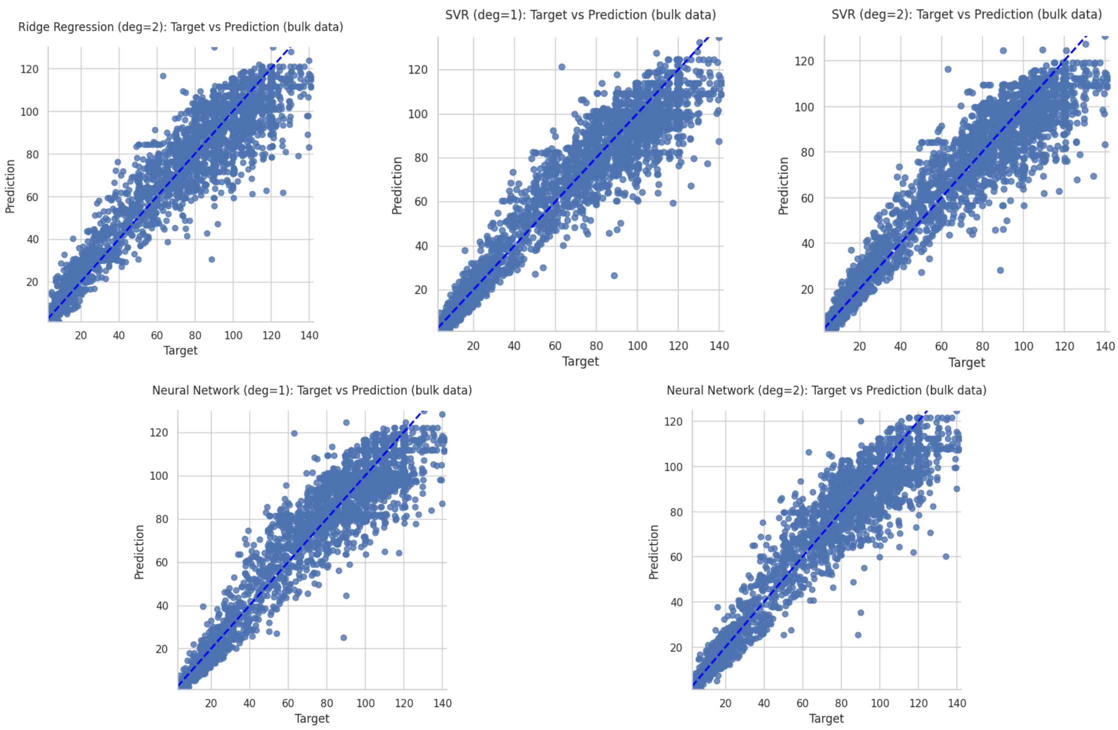

Figure 4.1: Target versus Prediction plots for all models, bulk data.

Figures 4.2 and 4.3 depict new *Target vs. Prediction* graphs, where the colors distinguish between underestimated and overestimated durations for MLP2 and SVR2 models. To categorize a prediction as normal, over, or underestimated, a threshold of a general, rounded MAPE of 16% was employed. MLP2 had 64.5% of its predictions classified as "Normal", whereas SVR2 had 63.8%. These percentages increased to 69.3% for MLP2 and 69.9% for SVR2 when considering only durations longer than 25 minutes.

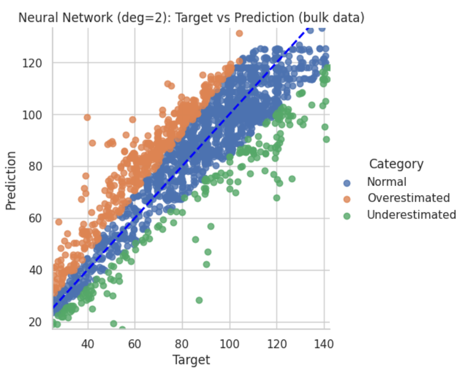

Figure 4.2: Target vs. prediction (MLP2).

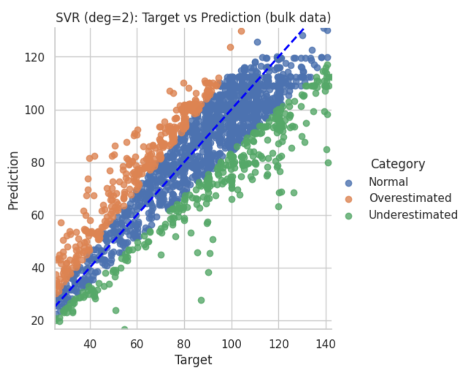

Figure 4.3: Target vs. prediction (SVR2).

While these plots illustrate the models' estimations of data's duration, they do not distinctly distinguish any model's clear superior performance from another.

Next, a multi-variate analysis of the model's behavior is presented. Figures 4.4 and 4.5 illustrate the data's response to variations in the relevant variables: *product quantity*, *number of stevedores*, and two new categorical variables, *weight category* and *volume category*.

These categories label data points based on their weight and volume values. After sorting the data by the respective value (weight or volume), the category label adopts three distinct values: 'Light' or 'Small' for the initial 30% of operations, 'Heavy' or 'Big' for the top 20% and 'Normal' for the rest of the data.

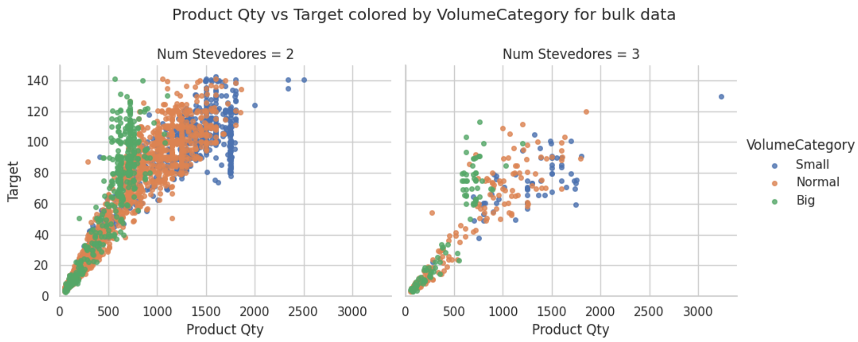

Figure 4.4: Distribution of target variable considering product quantity, volume, and duration for 2 and 3 stevedores (bulk data).

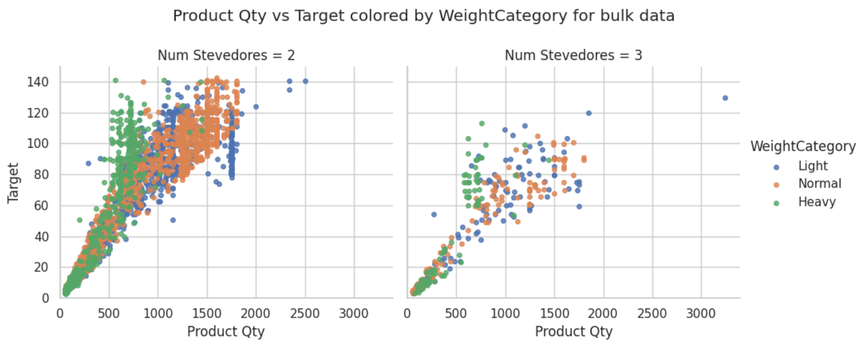

Figure 4.5: Distribution of target variable considering product quantity, weight, and duration for 2 and 3 stevedores (bulk data).

Upon analyzing figure 4.4, a clear correlation emerges between larger volumes and extended durations, while holding the item count constant. This correlation is particularly evident once the product quantity surpasses 500 units. Notably, for the largest products, only a handful of operations exceed 1000 units, aligning with the logistical constraints posed by limited truck loading capacity.

Additionally, it's discernible that as the product quantity increases, the data points corresponding to larger products exhibit a distinct behavior compared to those for normal or small products.

Then, figure 4.5 shows that as product weight gets heavier, the scatter plot's behavior mimics that of the volume plot, as weight often corresponds to volume. Yet, the distinctions in behavior between normal and light products are less pronounced. It is desirable for models to unveil the patterns resulting in this behavior.

Figures 4.6 and 4.7 feature similar plots than the previous ones, showing MLP2's predictions. In appendix A, every model's behavior can be found.

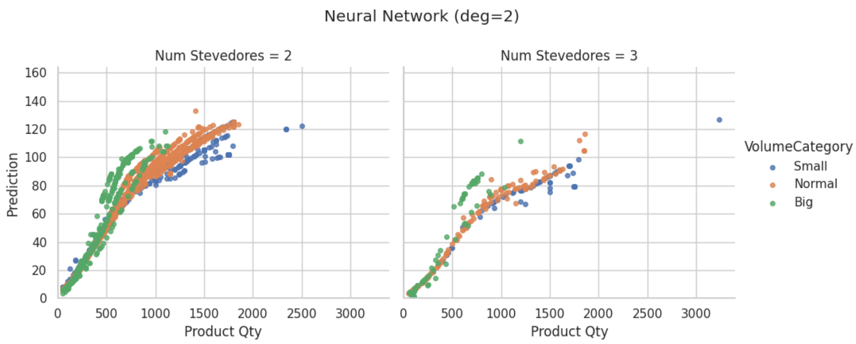

Figure 4.6: Bulk MLP2 predictions colored by volume.

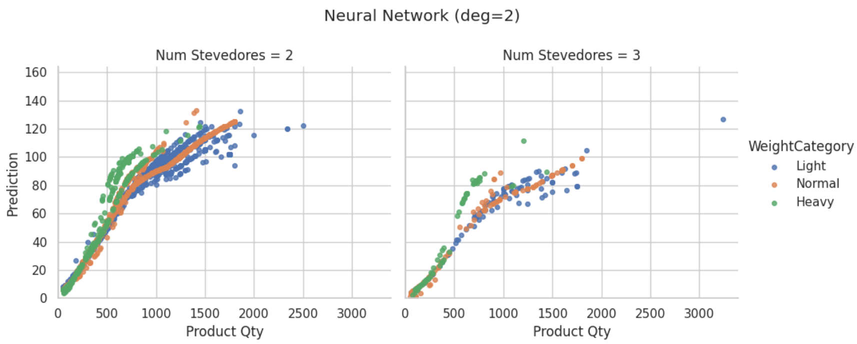

Figure 4.7: Bulk MLP2 predictions colored by weight.

Hereunder are the key findings of the visual analysis of the bulk models behaviors:

- All models effectively reproduce the separation observed by changing the dimensional categories in 4.4 and 4.5. Quadratic MLP and both SVR models perform almost identically, although MLP1 and MLP2's heavier/bigger product predictions seem to have a more dynamic distribution, similar to the real data distribution for this category (see figures A.5 and A.10), while he rest of the models show a more constrained pattern on the predictions.

- It is observed that the performance of the Ridge Regression model deteriorates as the product quantity increases. This is attributed to the natural parabolic behavior of quadratic models, which makes it unfit to estimate the duration of maneuvers with high numbers of product quantity (see figures A.1, and A.6).

- MLP and SVR models have similar behavior, and the predictions mimic the variations on duration distribution for distinct physical dimension's categories (see figures A.2, A.3, A.4, A.5 for volume, and A.7, A.8, A.9, A.10 for weight).

The models also adapt to the number of active stevedores, establishing an inverse relation to the final duration, as product quantity remains unchanged. Figures 4.8 and

4.9 denote this behavior for MLP2. Similar plots for the rest of the models can be found at appendix A.4.

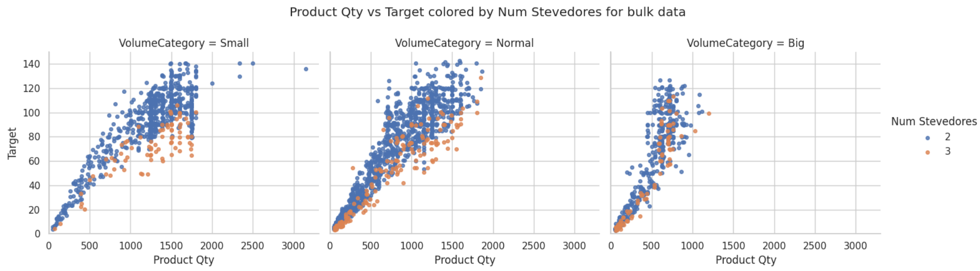

Figure 4.8: Target variable's behavior by volume category and number of stevedores.

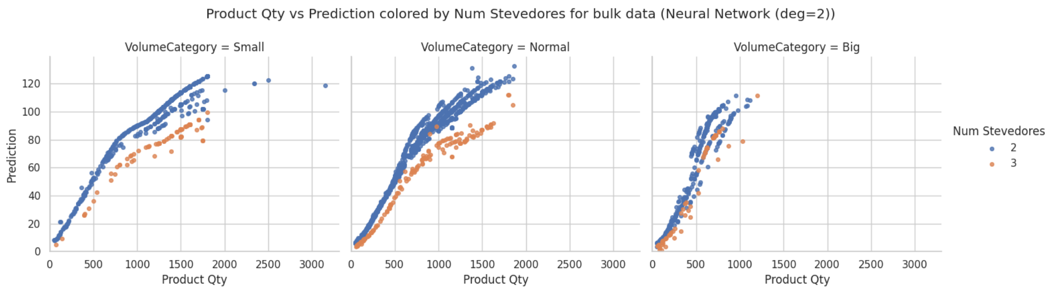

Figure 4.9: MLP2's response to changes in number of stevedores, by volume category.

4.4 Visual analysis for pallet models

Similarly, 4.14 shows the target vs. prediction plots for each model estimating pallet data durations.

These plots are considerably more dispersed than those for bulk data, specially for durations larger than 30 minutes. In this case, however, it is clear that MLP2 and SVR2's prediction distributions follow the most linear-like behaviors, as other models start to increasingly underestimate the target for larger durations.

This behavior, however is expected due to the high variability of the training data, which is specially true but maneuvers of longer durations (see again figure 4.10).

This scatter diagram demonstrates that, in the case of pallet data, the correlation between the available features (dimension, quantity, and number of stevedores) and the target variable is not as pronounced as observed in bulk data. As previously mentioned, operations involving lighter and smaller pallet loads exhibit notably more dispersed durations for constant product quantities and the number of stevedores, which also corresponds to longer durations. Conversely, for larger and heavier loads, the behavior clusters within a similar range, although lacking a distinct separation between these two categories. This circumstance holds the potential to substantially reduce the models' performance due to the considerable variability in duration recordings of the training data.

Figure 4.11 depicts SVR2's predictions for the same data than figures 4.10. The rest of the models' predictions can be found in appendix A.

Figure 4.10: Distribution of target variable considering product quantity, dimension, and duration for 2 and 3 stevedores (pallet data).

Presented here are the principal discoveries:

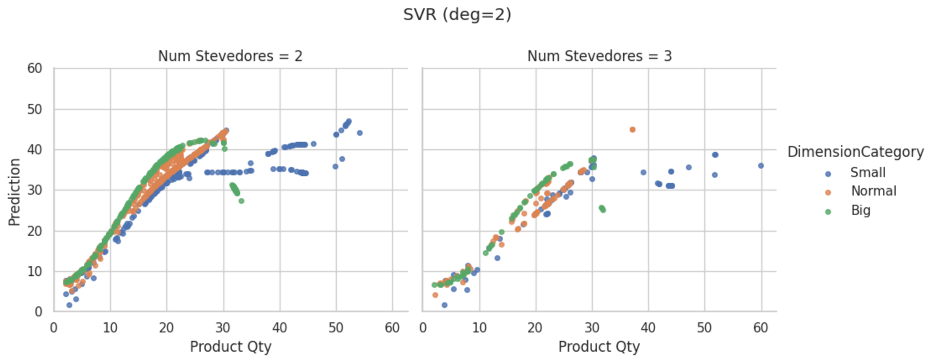

Figure 4.11: Pallet SVR2) predictions colored by dimension.

- Similarly than in bulk data, the Ridge model shows its natural parabolic behavior in the prediction of durations for large product quantities. This results in a poor performance for an important subset of the operations (figure A.11). This model also fails to effectively represent the effect of changing physical dimensions in the target predictions.
- The predictions of MLP1 and SVR1 exhibit analogous behavior, much like MLP2 and SVR2 do. All of these models predict similar values for loads in their mid and top categories, but the behavioral distinction is clear for smaller loads, precisely where high variability was present in the training dataset (see figures A.12 to A.15).
- When categorized by dimension, the MLP1 and MLP2 models' predictions exhibit a semi-linear behavior for lighter and smaller loads operations (see figures A.14 and A.15)). The remaining models have more disperse predictions that show different, non-linear distributions, that still present well-defined continual trajectories (A.12 and A.13).
- The presence of these distinct trajectories cold be considered as evidence that the models effectively generalize the authentic behavior of the data, rather than displaying scattered and seemingly random points that could indicate overfitting.

Now, similarly to pallet data, the models also effectively react to the number of stevedores. Figures 4.12 and 4.13 depict the real data and models' response to these changes. The rest of the models' plots can be found in the Appendix A.5.

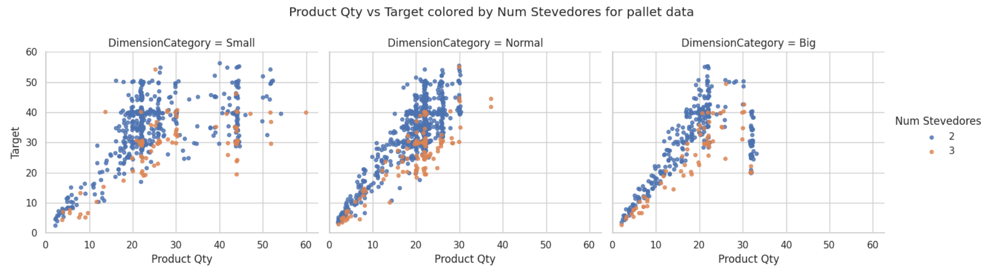

Figure 4.12: Target variable's behavior by volume category and number of stevedores.

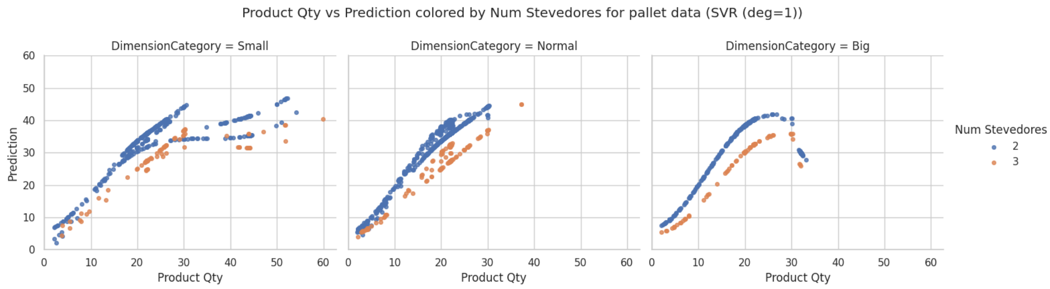

Figure 4.13: SVR1's response to changes in number of stevedores, by volume category.

4.5 Specific error analysis for bulk data

Analyzing smaller groups with selected ranges of duration can be beneficial to get a more in-depth understanding of the models' performance. This approach allows to understand the performance of the models in the most and least important duration categories, based on their number of occurrences, which is essential for understanding the potential of the models for practical use.

In A.6 section of the appendix A, MAPE, MAE, and an overestimation and underestimation count of the predictions can be found for MLP and SVR models.

Upon analyzing the graphs, although the exact metrics vary, all models perform in a similar way. Hereunder are the key findings:

- Mean Absolute Percentage Error values increase from groups of lower to higher duration. This behavior is expected as small absolute errors have a much more significant impact on this metric for groups of smaller durations, i.e, a 5 minutes error would represent a MAPE of 0.5 for an operation of 10 minutes, while the same absolute error would only mean a MAPE of 0.056 for an operation with an actual duration of 90 minutes. This makes the Mean Absolute Error a more appropriate metric to analyze the models' performances.
- All models have similar performances in the duration groups with the biggest number of occurrences in the data [81-100] minutes and [101-120] minutes, with a MAE of around 10 minutes.

- The third most occuring duration group is that with durations ranging from [61-80] minutes. In this case, the SVR models outperform the rest of them, with a MAE of approximately 10 minutes, almost 2 minutes less than the rest of the models. However, these models significantly decrease their performance for durations longer than 120 minutes, scoring a MAE of 20 minutes, which is 4-5 minutes over the MLP models. These longest durations, however, hold a significantly lesser presence in the data than the [61-80] minutes group.
- The models tend to underestimate durations, as these become longer. This is specially true for durations longer than 120 minutes, where practically all the predictions underestimated their target duration. In these durations there is also an importantly higher MAE score of about 15 minutes.

Figures 4.15 and 4.16 show the overestimated and underestimated values for MLP2 and SVR2 pallet predictions. A threshold of a rounded averaged MAPE of the models (16%) was used to categorize the predictions. 57.5% of MLP2's predictions fall into the 'Normal' category, while 59.2% of SVR2's do for all predictions. When considering only those with durations longer than 10 or 15 minutes, these numbers are not affected meaningfully.

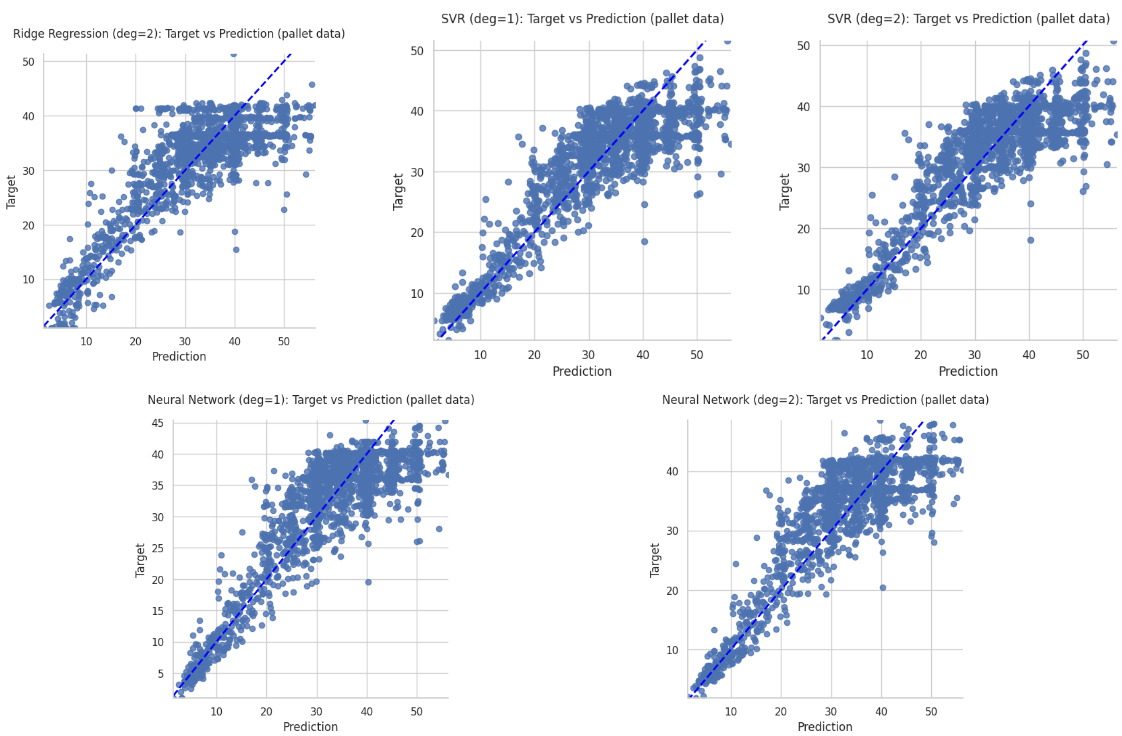

Figure 4.14: Target versus Prediction plots for all models, pallet data.

4.6 Specific error analysis for pallet data

In A.7, the same MAPE, MAE, and over/underestimation analysis plots for pallet predictions can be found. These are the analysis results:

- Similarly than to pallet data, the MAPE is not precisely an appropiate error metric to compare the performance of the models.

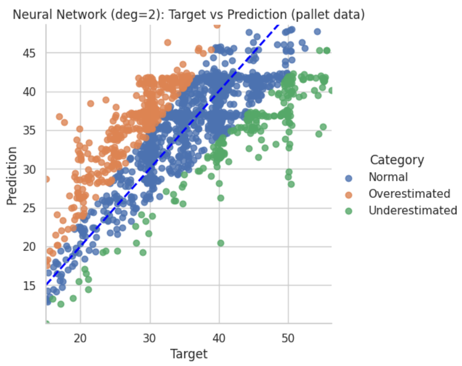

Figure 4.15: Target vs. prediction (MLP2).

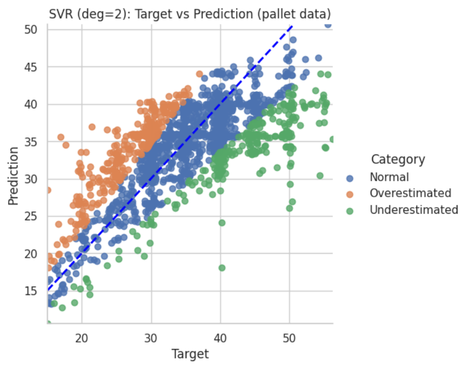

Figure 4.16: Target vs. prediction (SVR2).

- For those operations in the most occuring range of durations ([31-45] minutes), the MAE stays practiaclly unvarying in the models (only Ridge Regression has a slightly worse performance).
- SVR1 and SVR2 MAE scores and over/underestimation counts remain almost identical to each other, meaning than the more complex SVR2 model does not appear to have a significant advantage in predicting more accurate durations than its simpler counterpart. Only for operations < 15 min exists in SVR1 a slightly increased MAE by less than 1 minute, which corresponds to a MAPE increase of around 5%.
- MLP1 performance is comparable to that of SVR1, with small variations in the duration ranges of [0-15] minutes and [46-60], which perform slightly better in MLP1 and SVR1, respectively. However, these variations are minimal and hold little to no importance in defining a models' superior performance over the other.
- MLP2 seems to slightly reduce the error in groups [0-15] and [46-60] minutes when compared to SVR1 and SVR2 (a reduction of about 1 and 2 minutes), while maintaining similar performance for the most represented [31-45] minutes group; this, at the cost of a slight decrease in the performance of the [16-30] minutes group, which is increased between 1 and 2 minutes in average.

5. Conclusion

The preliminary analysis of the models' descriptions regarding the target variable reveals their potential to serve as effective representations of the target feature's behavior. They adeptly emulate the behaviors and trends characteristic of a defined "standard".

Most models exhibit similar performance, yielding a Mean Absolute Percentage Error (MAPE) of approximately 15-16%. While an improvement in performance would be desirable, it is crucial to acknowledge the substantial variability within the target variable and the substantial presence of noise within the recorded durations of the training data. This is further compounded by the relatively low correlation observed between most available features and the target variable.

The main relationships identified with respect to the target variable pertain to the quantity of product being loaded, followed by the number of working operators, with the dimensional characteristics of the products holding a comparatively lower but still present level of importance. This is applicable to both bulk and pallet maneuvers.

Although various features were extracted from the data in an attempt to uncover potentially latent relationships, such as the temporal proximity to the end of a work shift (i.e., whether the operation occurred at the beginning or end of a shift) or the specific shift itself (e.g., day, morning, afternoon), these did not result in any improvement at all, as they had near-zero correlation with the maneuver durations.

Future research aimed at enhancing the performance of regression models could entail the development of tailored and specialized models designed to accommodate specific subsets of data. For instance, the creation of models tailored to handle smaller or larger product quantities, or even the segmentation of products based on similar dimensional characteristics, could be explored as potential avenues for improvement.

Lastly, it is pertinent to note that as durations increase, the models tend to underestimate the duration values. This phenomenon is particularly pronounced for bulk maneuvers exceeding 100 minutes in duration or pallet maneuvers extending beyond 40 minutes. This bias may be attributed to the relatively limited representation of long-duration maneuvers in the training data, although even when a custom loss function was implemented to penalize errors associated with longer durations more severely, there was no discernible improvement in mitigating this effect. Nonetheless, it is important to emphasize that this does not pose a significant impediment to the overarching goal of this project, which is the development of a standardized tool capable of approximating desirable, standardized durations to be pursued by the working stevedores. This pursuit aims to enhance productivity, and the underestimation of these relatively unusual long operations could prove beneficial to enhance the plants performance, as the envisioned goals are shorter than the typical work duration. However, the practical application of these models must validate these assertions.

5.1 Limitations and Challenges

Since the inception of this research endeavor, it has been recognized that the manual collection of operation recordings exhibit inherent noise and a lack of precision. Furthermore, a substantial amount of data has been lost due to errors in the data capture process. To enhance the reliability of the data and thereby improve the performance of regression models, it would be beneficial to fully or partially automate the data acquisition process. One illustrative approach would be the utilization of specialized applications designed to streamline the recording of start and end times which could be as simple as pressing a button.

A. Appendix: Models Visualization

A.1 Models predictions by volume category (bulk data)

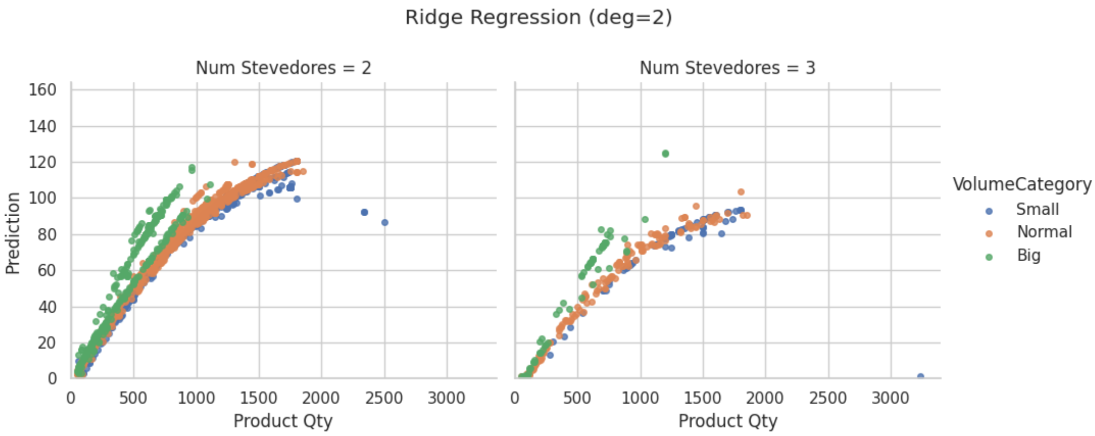

Figure A.1: Bulk Ridge2 predictions colored by volume.

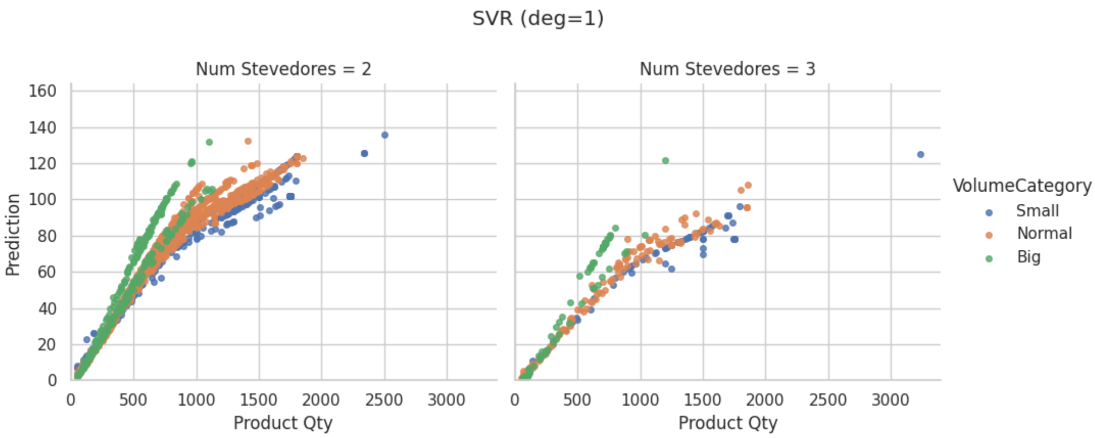

Figure A.2: Bulk SVR1 predictions colored by volume.

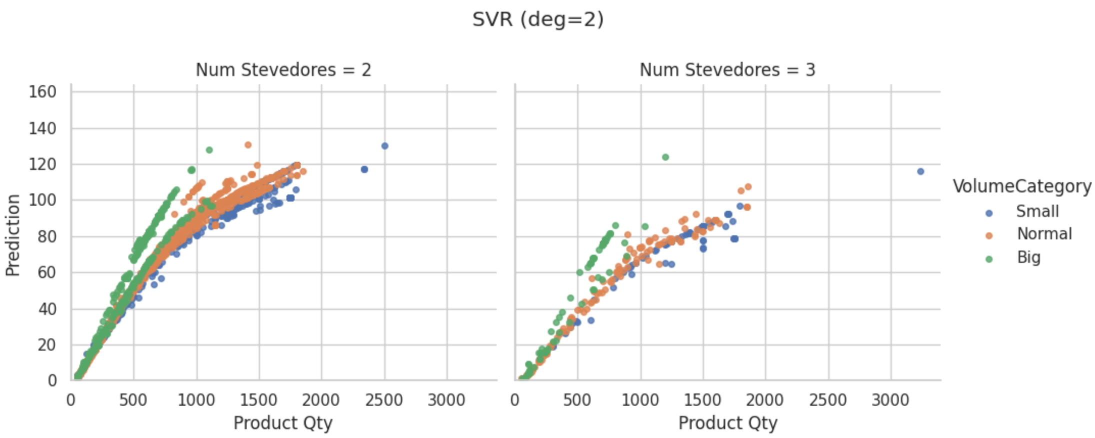

Figure A.3: Bulk SVR2 predictions colored by volume.

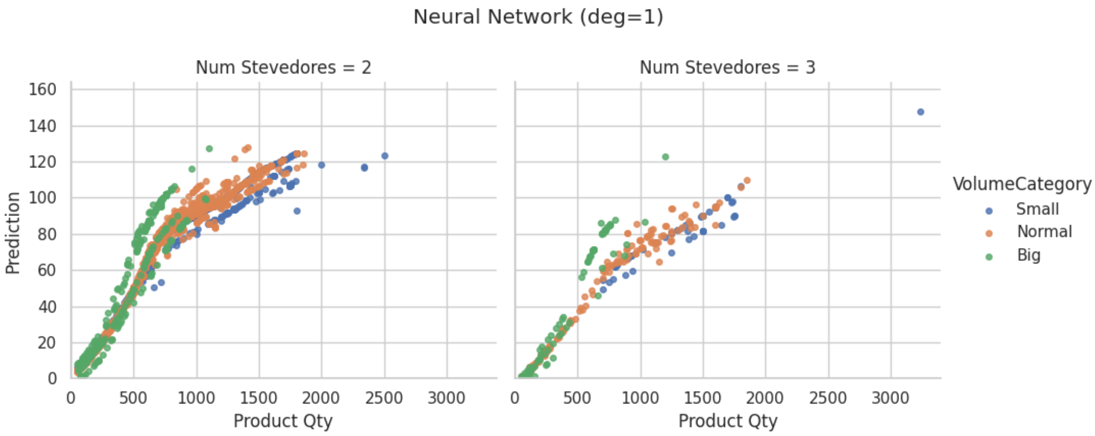

Figure A.4: Bulk MLP1 predictions colored by volume.

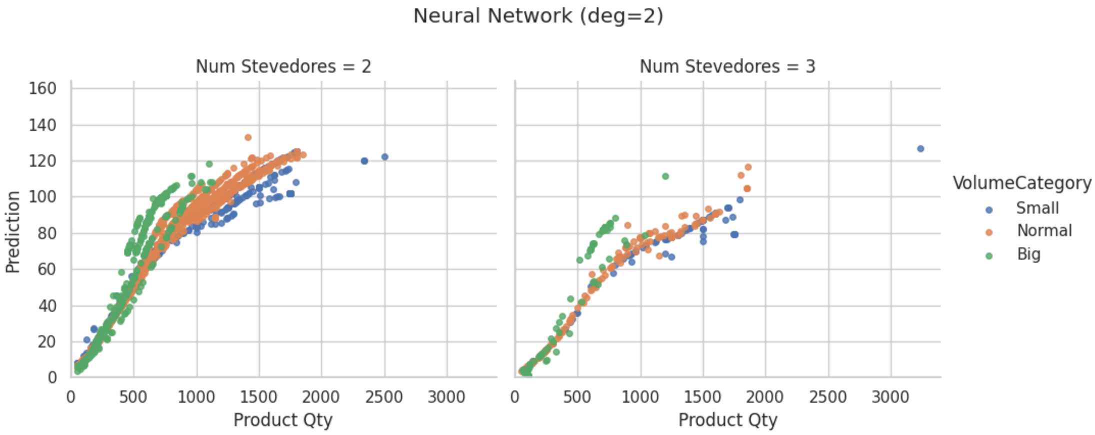

Figure A.5: Bulk MLP2 predictions colored by volume.

A.2 Models predictions by weight category (bulk data)

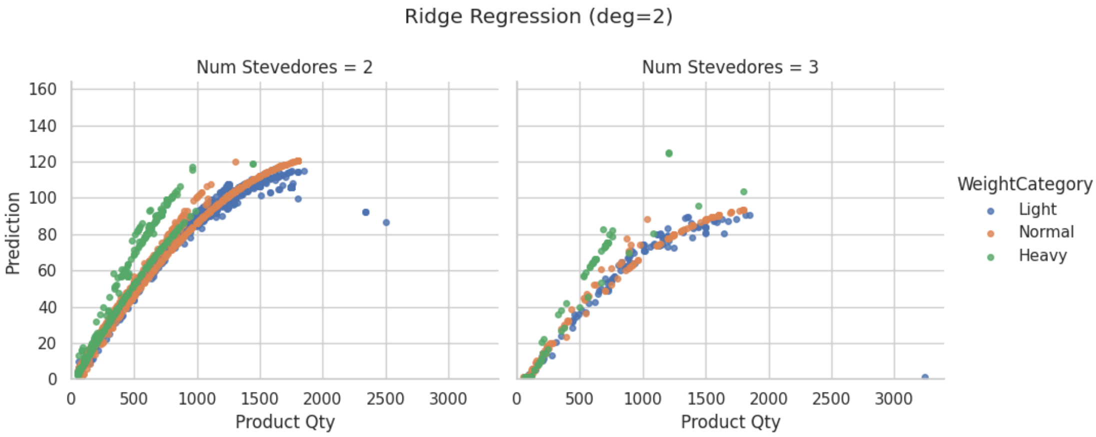

Figure A.6: Bulk Ridge2 predictions colored by weight.

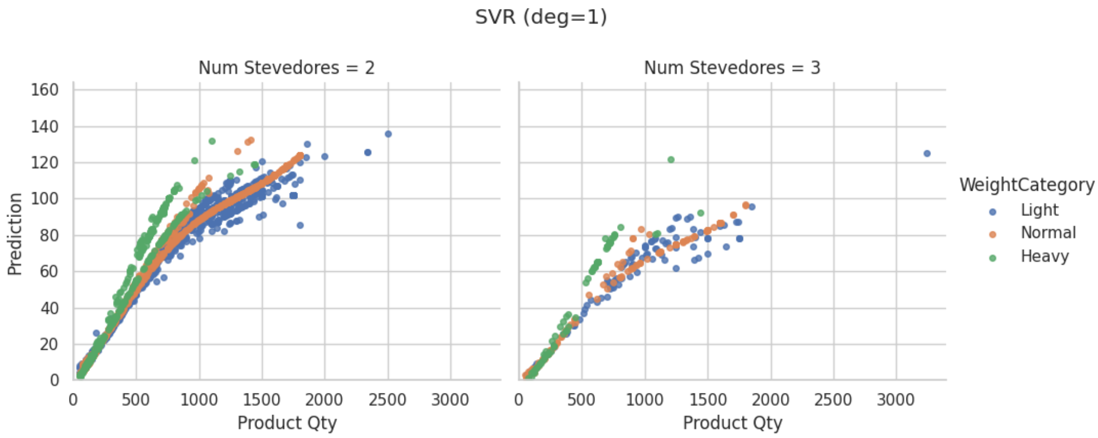

Figure A.7: Bulk SVR1 predictions colored by weight.

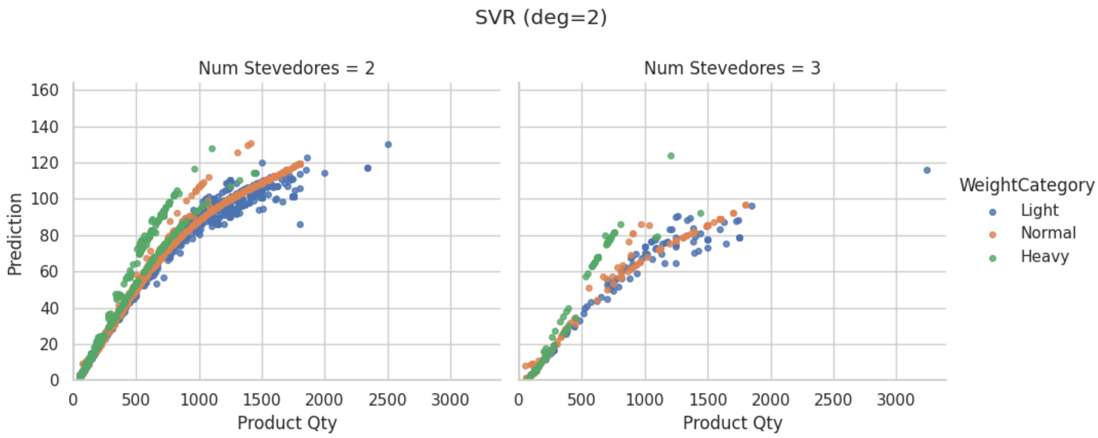

Figure A.8: Bulk SVR2 predictions colored by weight.

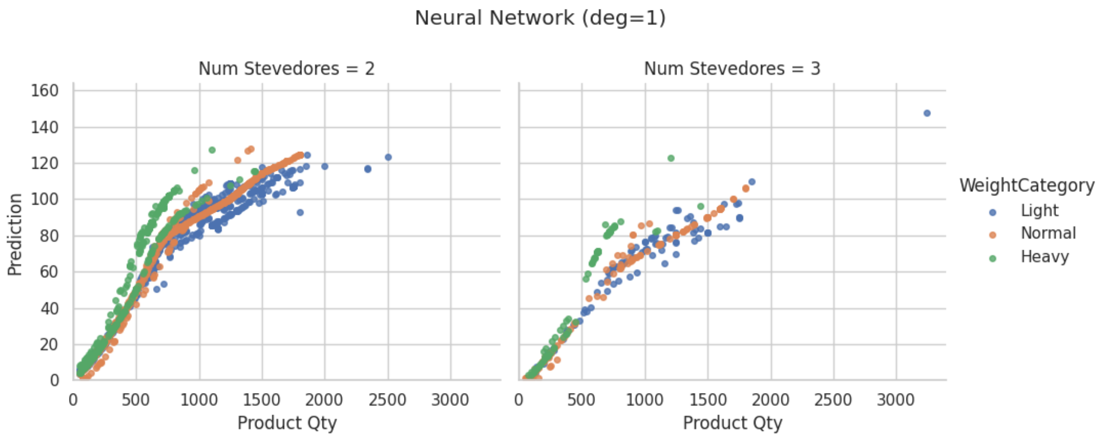

Figure A.9: Bulk MLP1 predictions colored by weight.

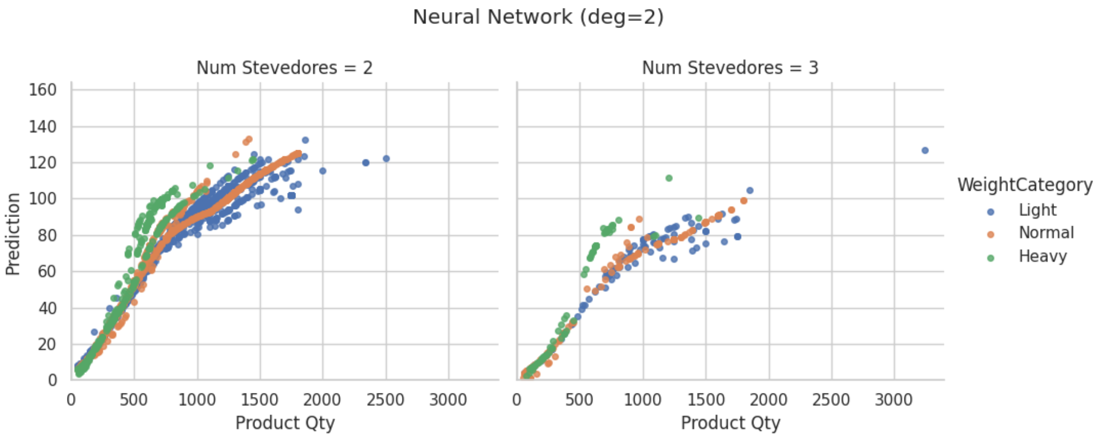

Figure A.10: Bulk MLP2 predictions colored by weight.

A.3 Models predictions by dimension category (pallet data)

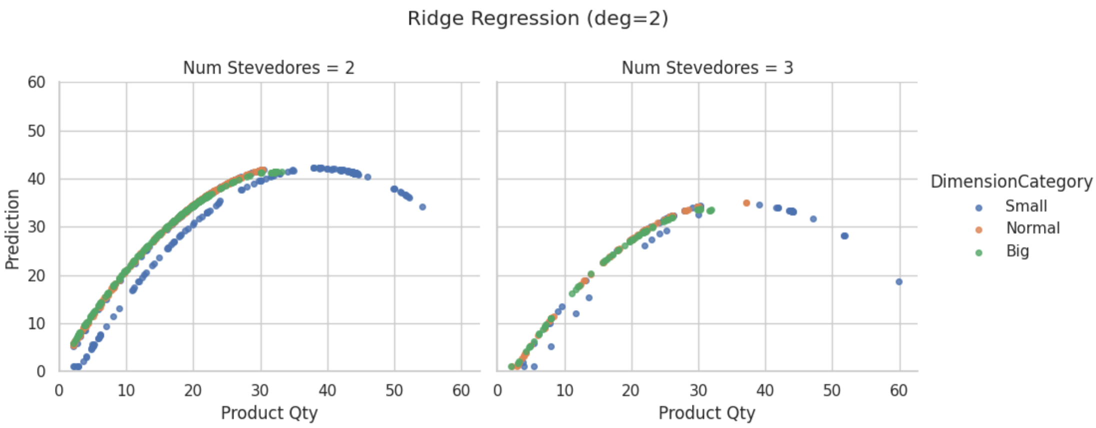

Figure A.11: Pallet Ridge2 predictions colored by dimension.

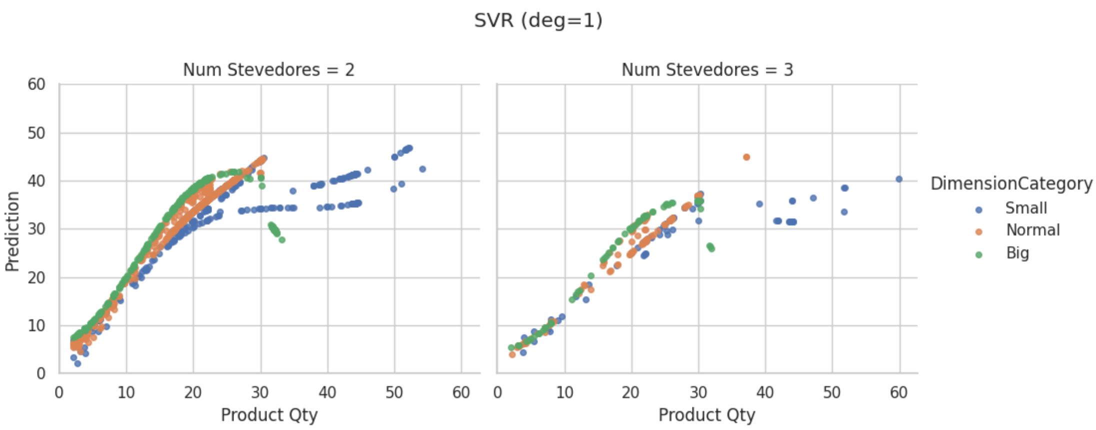

Figure A.12: Pallet SVR1 predictions colored by dimension.

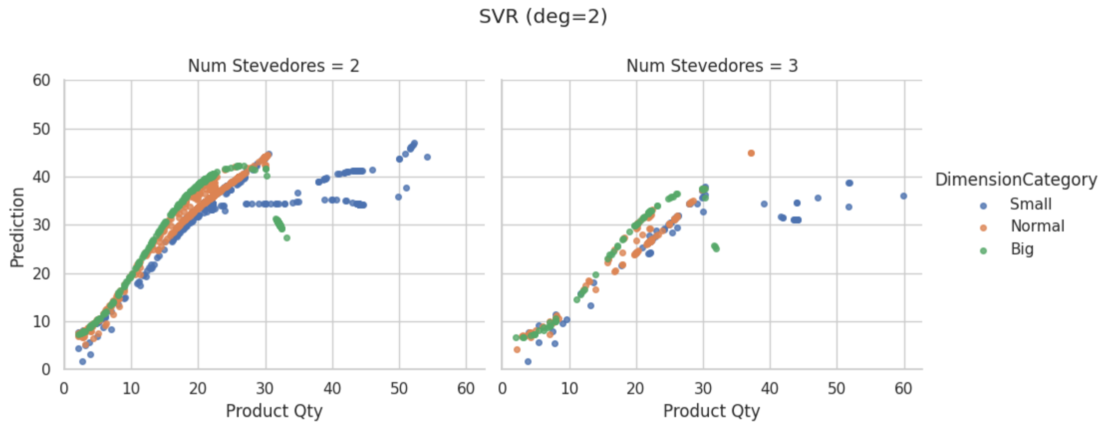

Figure A.13: Pallet SVR2 predictions colored by dimension.

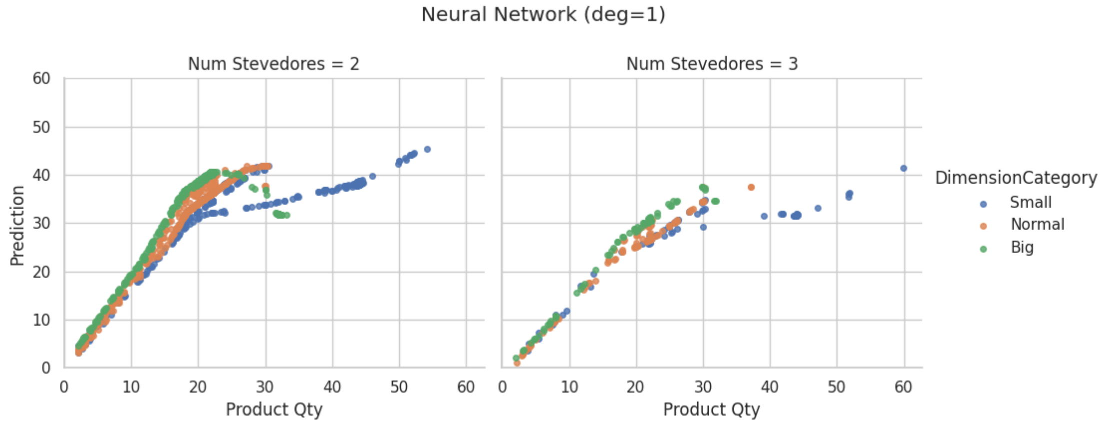

Figure A.14: Pallet MLP1 predictions colored by dimension.

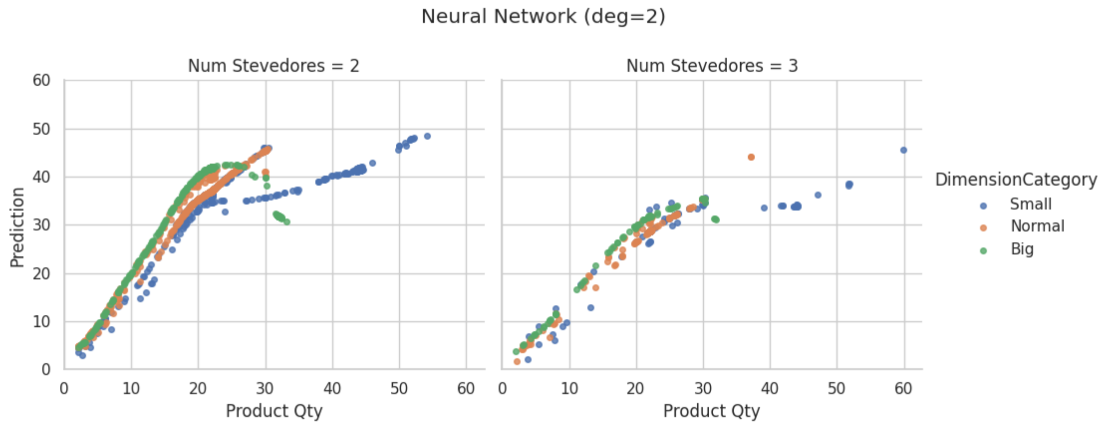

Figure A.15: Pallet MLP2 predictions colored by dimension.

A.4 Model's response to changes in number of stevedores (bulk data)

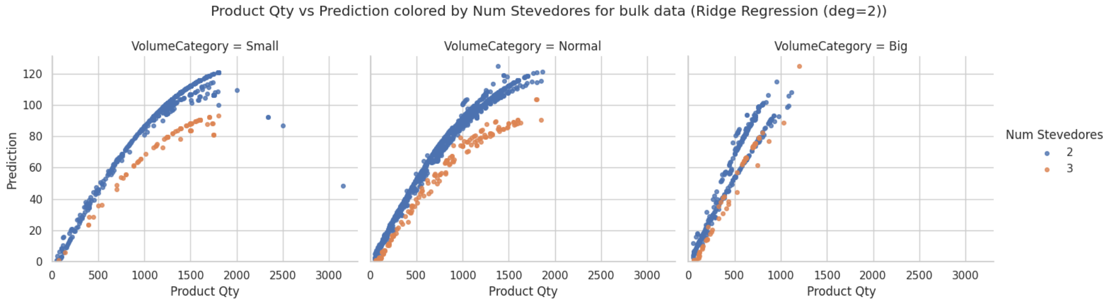

Figure A.16: Ridge2's response to changes in number of stevedores, by volume category.

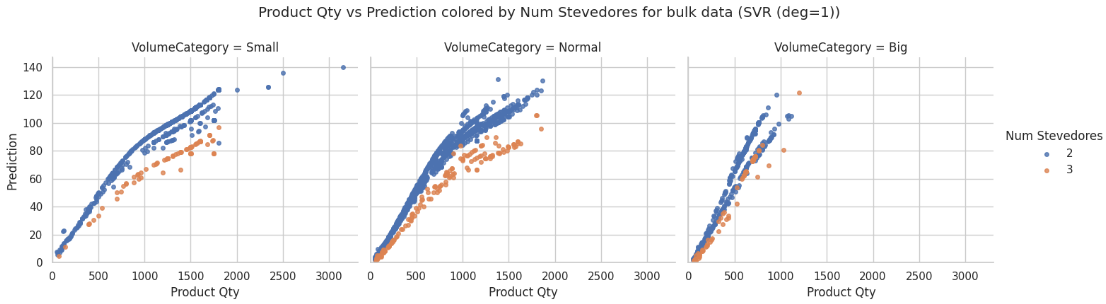

Figure A.17: SVR1's response to changes in number of stevedores, by volume category.

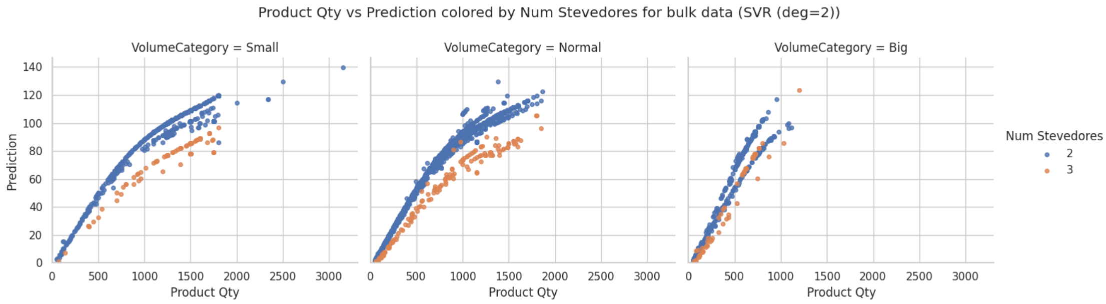

Figure A.18: SVR2's response to changes in number of stevedores, by volume category.

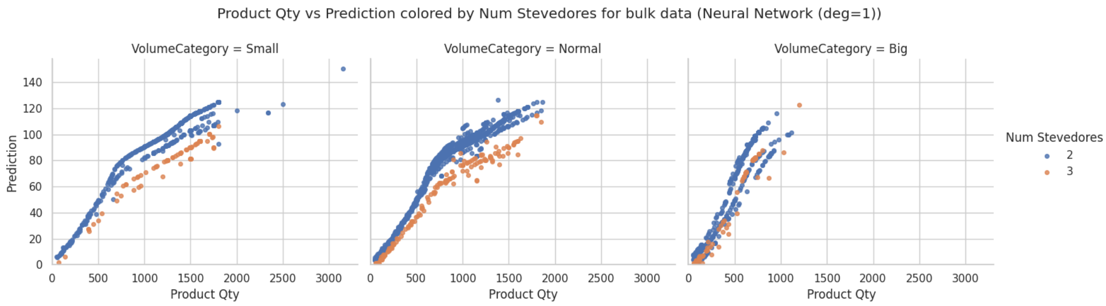

Figure A.19: MLP1's response to changes in number of stevedores, by volume category.

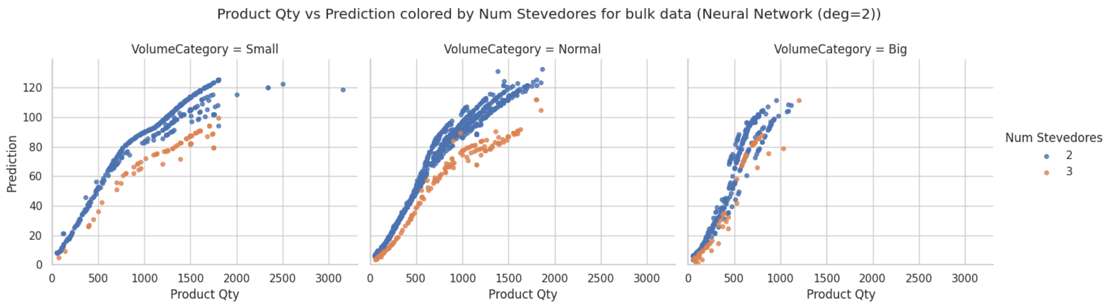

Figure A.20: MLP2's response to changes in number of stevedores, by volume category.

A.5 Model's response to changes in number of stevedores (pallet data)

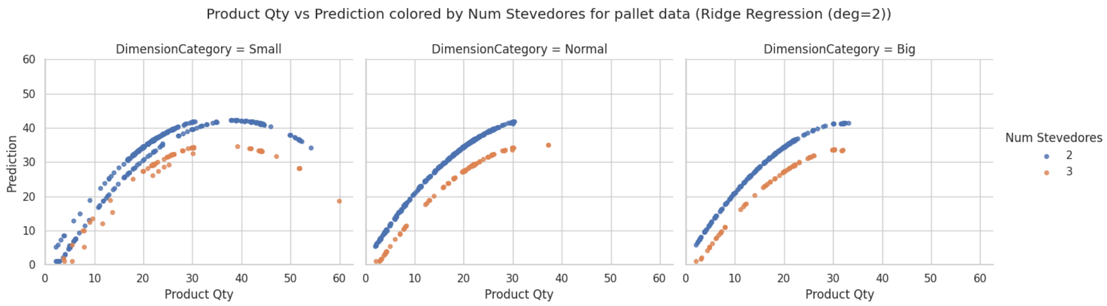

Figure A.21: Ridge2 response to changes in number of stevedores, by dimension category.

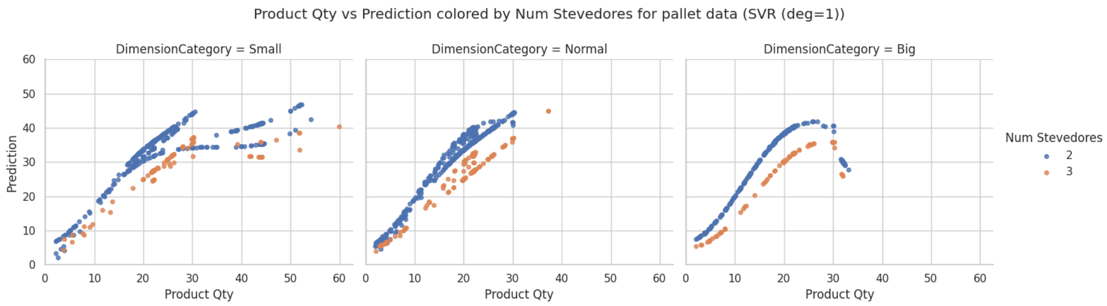

Figure A.22: SVR1's response to changes in number of stevedores, by dimension category.

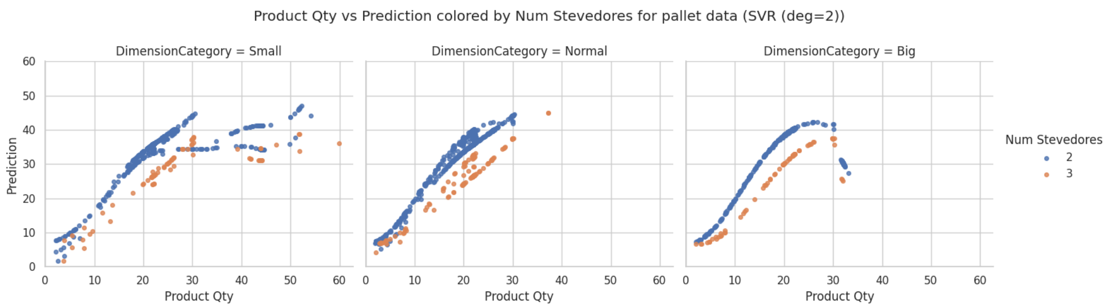

Figure A.23: SVR2's response to changes in number of stevedores, by dimension category.

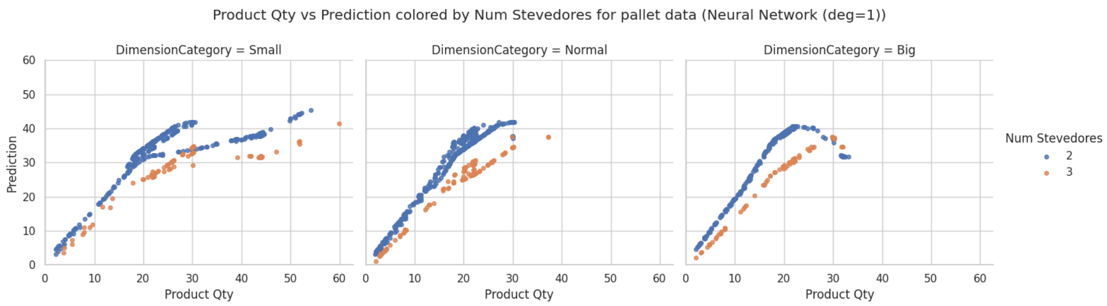

Figure A.24: MLP1's response to changes in number of stevedores, by dimension category.

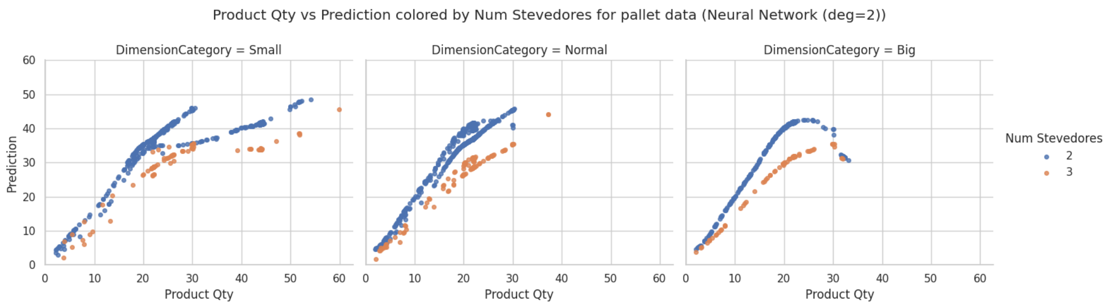

Figure A.25: MLP2's response to changes in number of stevedores, by dimension category.

A.6 Error analysis by duration group, bulk data

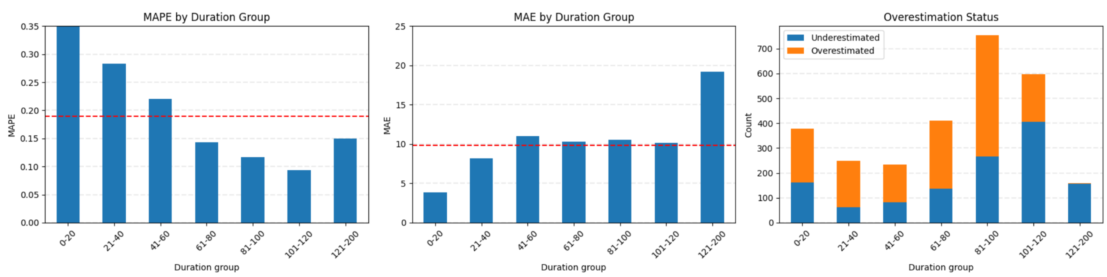

Figure A.26: Ridge2: MAPE, MAE and over/underestimation status by duration group (bulk data.

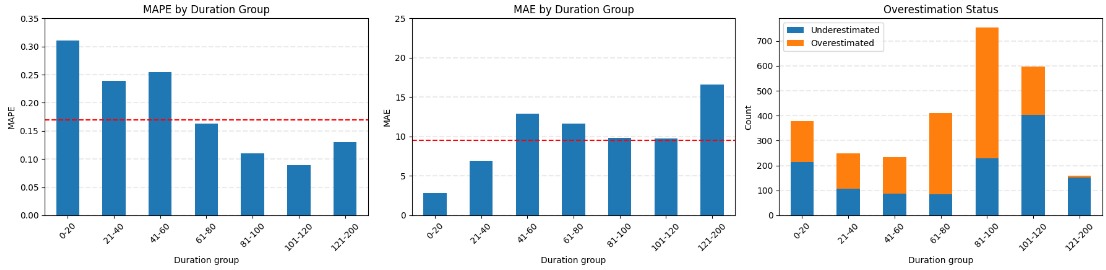

Figure A.27: MLP1: MAPE, MAE and over/underestimation status by duration group (bulk data).

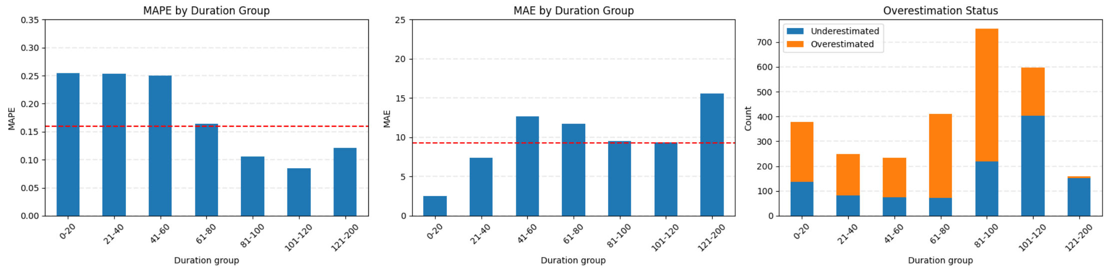

Figure A.28: MLP2: MAPE, MAE and over/underestimation status by duration group (bulk data).

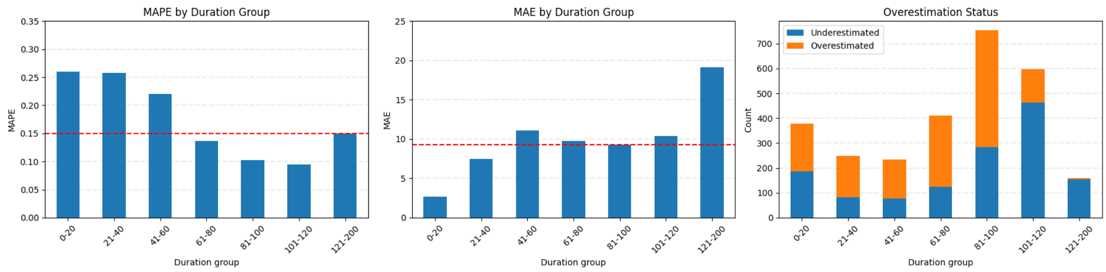

Figure A.29: SVR1: MAPE, MAE and over/underestimation status by duration group (bulk data).

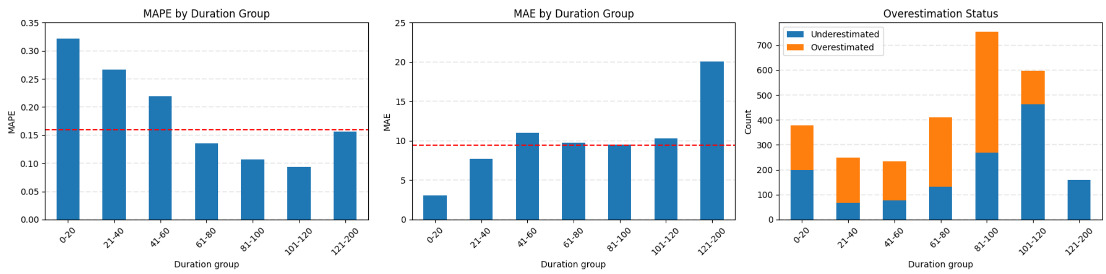

Figure A.30: SVR2: MAPE, MAE and over/underestimation status by duration group (bulk data).

A.7 Error analysis by duration group, pallet data

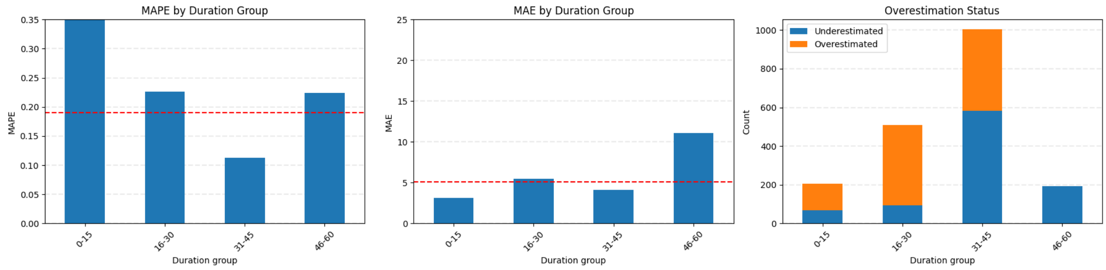

Figure A.31: Ridge2: MAPE, MAE and over/underestimation status by duration group (pallet data).

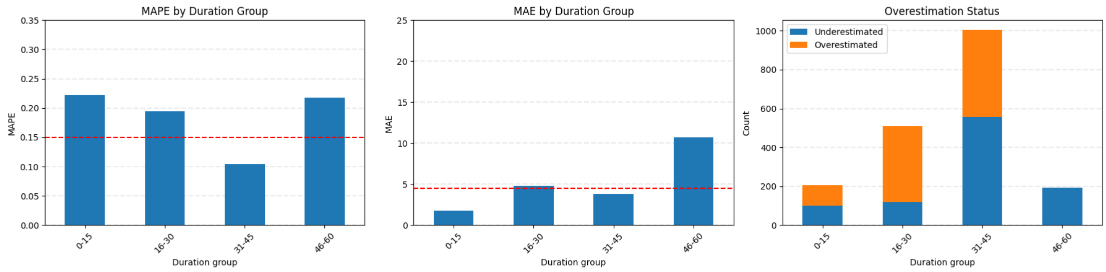

Figure A.32: MLP1: MAPE, MAE and over/underestimation status by duration group (pallet data).

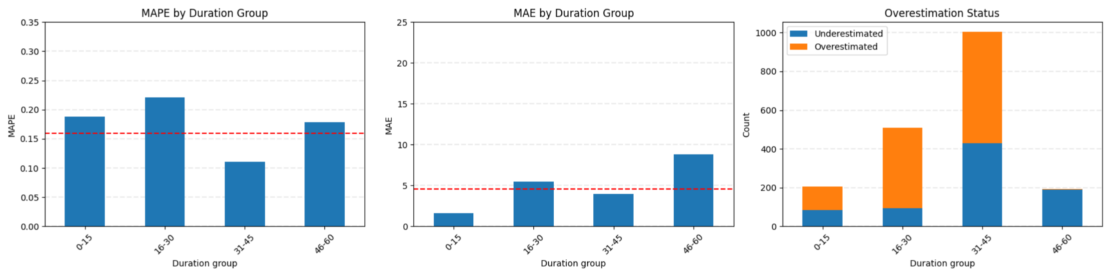

Figure A.33: MLP2: MAPE, MAE and over/underestimation status by duration group (pallet data).

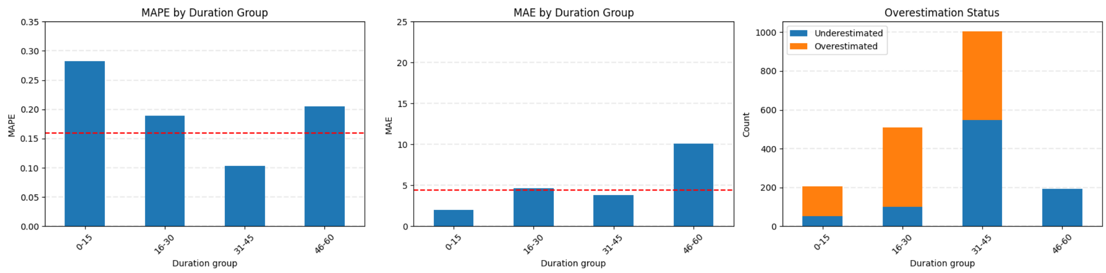

Figure A.34: SVR1: MAPE, MAE and over/underestimation status by duration group (pallet data).

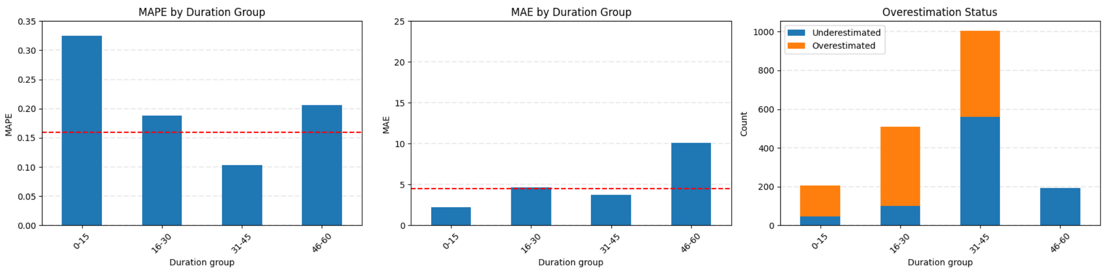

Figure A.35: SVR2: MAPE, MAE and over/underestimation status by duration group (pallet data).

B. Appendix: Code

B.1 Ridge Regression

```
# Make Polynomial
data_train_ = PolynomialFeatures(degree=degree, include_bias=False).fit_transform(data_train)
data_test_ = PolynomialFeatures(degree=degree, include_bias=False).fit_transform(data_test)

# Ridge model using all training data
ridge_model = Ridge().fit(data_train_, target_train)
```

Code extraction 1: Ridge Regression implementation.

B.2 Support Vector Regression

B.2.1 Grid Search

A grid search was carried out to find the most appropriate kernel and C, *gamma*, *epsilon* and C parameters for every individual SVR model for both bulk and pallet maneuvers.

```
# Hyperparameter grid
param_grid = {
    'C': [10, 20, 30, 50, 100],
    'kernel': ['linear', 'rbf'],
    'gamma': ['scale'],
    'epsilon': [2,4,6,10]
}
# Create the GridSearchCV object
grid_search = GridSearchCV(SVR(), param_grid=param_grid, cv=n_cv)

# Fit the GridSearchCV object to the data
grid_search.fit(data_train_, target_train)

parameters = grid_search.best_params_
```

Code extraction 2: Grid search for best SVR parameters.

B.2.2 Model

```
# Create model
svr_model = SVR(epsilon=parameters['epsilon'], C=parameters['C'], gamma=parameters['gamma'], kernel=parameters['kernel'])

# SVR model using all training data
svr_model = svr_model.fit(data_train_, target_train)
```

Code extraction 3: SVR implementation.

B.3 Multi-Layer Perceptron Neural Network

```
import tensorflow as tf
from tensorflow.keras import regularizers

def create_model(input_shape):
    model = tf.keras.models.Sequential([
        tf.keras.layers.Dense(64, activation='relu', input_shape=input_shape, kernel_regularizer=regularizers.l2(0.01)),
        tf.keras.layers.Dense(32, activation='relu', input_shape=input_shape, kernel_regularizer=regularizers.l2(0.01)),
        tf.keras.layers.Dense(16, activation='relu', input_shape=input_shape, kernel_regularizer=regularizers.l2(0.01)),
        tf.keras.layers.Dense(1, activation='linear')
    ])
    return model

    # Make data polynomial
    data_train_ = PolynomialFeatures(degree=degree, include_bias=False).fit_transform(data_train)
    data_test_ = PolynomialFeatures(degree=degree, include_bias=False).fit_transform(data_test)

    # Define the neural network architecture
    l = len(data_train_[0])

    # Create and compile the model
    model = create_model(input_shape=(l,))
    model.compile(optimizer='adam', loss='mse')

    # Train the model on the training data
    early_stopping = EarlyStopping(monitor='loss', patience=10, verbose=1, mode='auto')
    model.fit(data_train_, target_train, epochs=n_epoch, verbose=verbose_, callbacks=[early_stopping])
```

Code extraction 4: Neural Network architecture and implementation.

C. Appendix: Individual SKUs visualization

Bulk SKUs

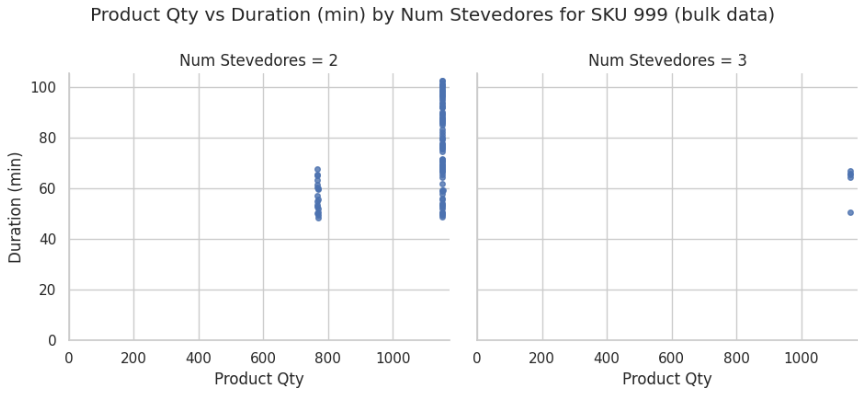

Figure C.1: SKU 999 scatter

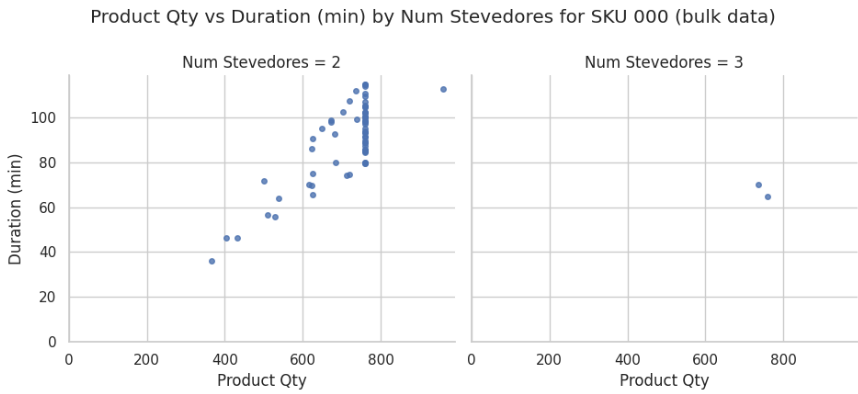

Figure C.2: SKU 000 scatter

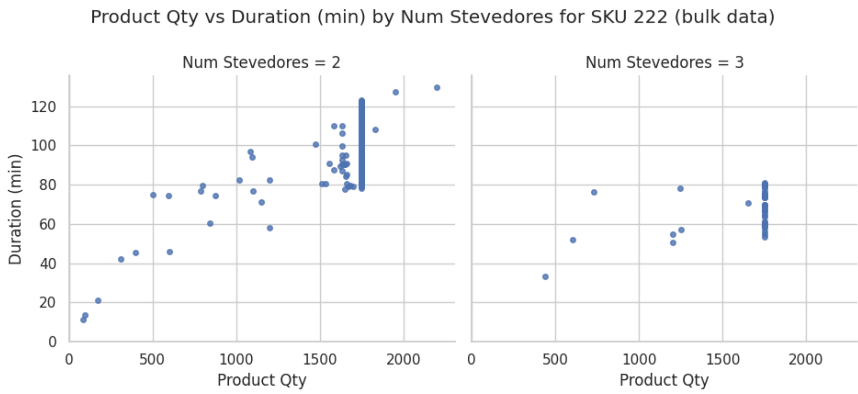

Figure C.3: SKU 222 scatter

Pallet SKUs

Bibliography

[1] Mariette Awad and Rahul Khanna. *Support Vector Regression*, pages 67–80. Apress, Berkeley, CA, 2015. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_4. URL https://doi.org/10.1007/978-1-4302-5990-9_4.

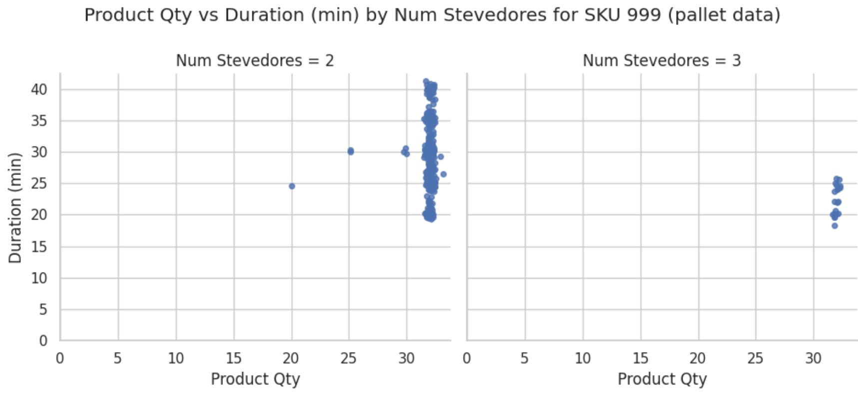

Figure C.4: SKU 999 (pallet) scatter

[2] Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

[3] Jean Gaudart, Bernard Giusiano, and Laetitia Huiart. Comparison of the performance of multi-layer perceptron and linear regression for epidemiological data. *Computational Statistics Data Analysis*, 44(4):547–570, 2004. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(02)00257-8. URL https://www.sciencedirect.com/science/article/pii/S0167947302002578.

[4] Gary C. McDonald. Ridge regression. *WIREs Computational Statistics*, 1(1):93–100, 2009. doi: https://doi.org/10.1002/wics.14. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.14.

[5] Fionn Murtagh. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5):183–197, 1991. ISSN 0925-2312. doi: https://doi.org/10.1016/0925-2312(91)90023-5. URL https://www.sciencedirect.com/science/article/pii/0925231291900235.