Machine Learning for Demand Forecasting of Steel Products in the Automotive Industry

Freek Merijn Hobbenschot (s3428397)

A thesis for the MSc Econometrics, Operations Research and Actuarial Science, track Econometrics

university of groningen

Faculty of Economics and Business (FEB)
University of Groningen
The Netherlands
January 1, 2024

In collaboration with:

TATA STEEL

Master's Thesis
Econometrics, Operations Research, and Actuarial Studies

Supervisor: Dr. B.J.P. Achou
Second Assessor: Dr. M. Kesina

Machine Learning for Demand Forecasting of Steel Products in the Automotive Industry

A Tata Steel Case Study: Improving Demand Forecasts in a Dynamic Market Environment

Abstract

This study explores demand forecasting in the steel industry and focuses on the use of machine learning techniques. A Long Short-Term Memory model, Random Forest Regression, Support Vector Regression and an Ensemble model based on simple average are trained and tested on data from April 2008 till January 2023. These models predict European steel demand for products in the automotive industry, using historical demand data as well as economic and industrial indicators. The model's performance are tested using Residual Mean Squared Error, Mean Absolute Error and Mean Percentage Error. Results show that Long Short-Term Memory outperforms other models for two products across all metrics, while Random Forest Regression and Support Vector Regression excel for one product. For the other two products, the results vary between performance metrics. Evaluating the model's performance on data that excludes the COVID-19 period, shows that Long Short-Term Memory and Random Forest Regression are the best performing models. The model's performance appears to be highly sensitive to changes in the data. This underscores the need to keep looking for improvements and to not rely on the use of one single model. It is suggested that further research should be done to improve modeling performance and practical usage.

Contents

1 Introduction **4**

2 Theoretical Background **7**
2.1 Sales Contracts 8
2.2 Forecasting Considerations 9

3 Literature Review **10**
3.1 Econometric Demand Models and VAR 10
3.2 Intensity of Use Models 11
3.3 SWIP Models 12
3.4 Hybrid Demand Models and Machine Learning 13

4 Data **15**
4.1 Available Features 17
4.1.1 Economic Performance Indicators 17
4.1.2 Purchasing Managers' Indices 19
4.1.3 Confidence and Sentiment Indicators 20
4.1.4 Price and Market Indicators 21
4.2 Data Processing 22
4.2.1 Stationarity and Transformation Techniques 23
4.3 Feature Selection 25
4.3.1 Dimensionality Reduction 25
4.3.2 Target Based Feature Selection 27

5 Methodology **28**

6 Results and Analysis **31**
6.1 Forecasting Without the COVID-19 Pandemic 39

7 Conclusion **46**

8 Research Limitations and Further Recommendations **48**

9 Acknowledgements **50**

References **51**

Appendix **55**
A Definitions of Steel Demand 55
B The Bullwhip Effect 56
C Basics of Machine Learning 58
C.1 Learning Scenarios 59
C.2 Deep Learning, Transfer Learning and Ensemble Methods 59

1 Introduction

Accurate demand forecasts help steel-making companies to optimize sales allocation in profitable market segments and regions. Accurate forecasts reduce uncertainties and contribute to more profitable contract negotiations and more efficient use of production capacity. When forecasts do not provide reliable results, it will be difficult to adapt to new market conditions. This leads to risks such as selling excess volume at non-profitable prices or incurring high fixed costs due to underutilized production capacity.

Timely adjustments to shifts in demand help to improve strategic and tactical business plans. On a strategic level it is important to focus on high-growth market segments and regions to drive profitability. On a tactical level it is important to reallocate sales and implement price adjustments to obtain immediate and substantial value gains. This is only possible when forecasting models provide accurate results.

The steel industry's outlook for the next ten years remains fragile, mainly due to negative market expectations and high energy prices (Baroyan et al., 2023). Therefore, it is highly important for steel-making companies to obtain accurate forecasts and prevent unnecessary risks. Otherwise, it will become more and more difficult to survive.

Steel is an intermediate good with characteristics of derived demand. High global uncertainty and volatility have decoupled steel demand from economic and industrial indicators, leading to disruptions in traditional market cycles. Furthermore, high volatility has impacted traditional stock-cycle behaviour of parties in the supply chain. This makes it difficult for steel-making companies to predict how shifts in market demand will affect stock levels of customers and therefore incoming orders.

The third-quarter market outlook from the European Steel Association EUROFER (2023) reveals an extent of wishful thinking in its current demand forecasts. This leads to multiple downward revisions and therefore less reliable results. Because of this, steel-companies rely more and more on other in-house forecasting models. Examples are SWIP Models, which rely on the Steel Weighted Industrial Production index. While these models are often used, their reliance on traditional stock cycle behaviour imposes challenges. To obtain reliable results, new methods are required. Current studies on the use of machine learning models for macro-economic demand forecasting show promising results. This makes it interesting to asses their usage for the steel

industry.

In this paper, I discuss the limitations of traditional and current forecasting models that are used in the steel industry and I will identify machine learning techniques that are suitable as alternatives. The literature review shows that the Long Short-Term Memory model stands out for macroeconomic forecasting. To construct a Long Short-Term Memory model, the lstm_nowcast library is used. This library is provided by Hopp is his paper on Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (Hopp, 2021). To determine the effectiveness of this model, I will compare it with Support Vector Regression and Random Forest Regression. These models show promising results in the study of Raju et al. (2022) on demand forecasting in the steel industry. Additionally, I will compare the Long Short-Term Memory model to a third benchmark, Ensemble 1. This model combines Support Vector Regression and Random Forest Regression through a simple average of the predictions. This might help to mitigate weaknesses from both individual models and therefore yield more accurate predictions.

The Long Short-Term model and benchmark models are tested for their ability to produce reliable forecasts of steel demand. In this paper, I specifically focus on the demand for steel products in the European automotive market. The decision to concentrate on the European automotive market is driven by its interesting characteristics. The market shows a consumer-driven nature and has significant influence in the global automotive industry. The demand for steel products in this market is influenced by environmental regulations, the growing demand for more sustainable production materials and the shift towards electric vehicles. Exploring the dynamics of steel demand in the European automotive market provides valuable insights into current trends and in factors that shape the industry's future direction. This knowledge is crucial for steel-making companies when developing strategic business plans. The six steel products that are include in this study, are Hot Rolled Narrow Strip, Lengths Cut From Hot Rolled Wide, Cold Rolled Sheets, Hot Dipped Metal Coated Sheets, Electrolytically Metal Coated Sheets and Grain Non-Oriented. These products form the basis of most manufactured components, both in the interior and exterior of vehicles. Demand data from these products come from consolidated data sets, provided by steel-making company Tata Steel Netherlands.

To improve the demand forecasts for these six products, I analyse 27 economic and industrial features. These 27 features are selected based on

consultations with domain experts from Tata Steel Netherlands. These features fall into four categories: Economic Performance Indicators, Purchasing Managers' Indices, Confidence and Sentiment Indicators, and Price & Market Indicators. For each steel product I select the most informative subset of features. The feature selection is based on principle component analysis and recursive feature elimination. Data for the features is monthly available and obtained via Refinitiv, from the sources Eurostat, business information company Wards, the German Association of the Automotive Industry (VDA), S&P Global, the Directorate-General for Economic and Financial Affairs (DG ECFIN), the European Central Bank and the European Banking Federation. The data is seasonally adjusted and covers the period from April 2008 to January 2023. I exclude more recent data from the model evaluation due to publication lags and data revisions in macro-economic indicators.

I use three different performance metrics to evaluate the performance of the models. Similar to Raju et al. (2022), I use the Root Mean Squared Error, the Mean Absolute Error and the Mean Absolute Percentage Error. Since each of these performance metrics captures different aspects of a model's behaviour, this approach helps to obtain a more comprehensive understanding of the performance.

The analysis show that the performance metrics are relatively close among the models. In general, the Long Short-Term Memory model shows competitive results. It outperforms the benchmark models on all performance metrics in the predictions for Cold Rolled Sheets and Electrolytically Metal Coated Sheets. Long Short-Term Memory does not outperform the benchmark models of the other steel products. For both the Long Short-Term Memory model and the benchmark models, the predictive accuracy was not consistent during the whole test period. Testing the model performance on the data excluding COVID-19, shows better model performance. However, even-though this data contains less fluctuations, the performance metrics show errors that are relatively high compared to the average apparent steel demand of the products each month. While Long Short-Term Memory and Random Forest Regression in general outperform Support Vector Regression and Ensemble 1, the predictive performance of the models is highly sensitive to changes in the data. This underscores the need to keep looking for improvements and to not rely on the use of one single model.

In the next section I discuss the theoretical background, which includes elementary knowledge of steel demand, sales contracts and forecasting considerations. Subsequently, section 3 contains the literature review including

traditional and conventional forecasting models for steel demand. It also includes modern demand models and the use of machine learning techniques. In section 4, I present an overview of the data for both the steel products and the features. This section also includes information on data processing and the feature selection process. Section 5 consists of the methodology, which includes the models and performance metrics that are used in the analysis. Section 6 consists of the analysis and includes the presentation and discussion of the findings. Finally, section 7 offers the conclusion and section 8 provides recommendations for further research.

2 Theoretical Background

The demand for steel-intensive goods is linked to economic cycles and directly influenced by economic fluctuations. Demand in the automotive market is consumer driven and responds early in the economic cycle. Similar, steel demand in the packaging market is also consumer driven. The differences is that the packaging market focuses on the food industry. Since food is a necessary good, steel demand in this market shows less intense reactions to economic changes.

Goods such as trucks and trailers are categorized under the engineering market, which is distribution-driven. Demand in this market shows delayed responses to economic upturns and is affected in the mid-stages of the economic cycle. Another steel intensive market is the construction market. This market is economically-driven and responds in the mid to late stages due to the long lead-time for capital-intensive projects.

Understanding the influence of economic cycles is crucial for steel companies as it serves as a guide for the predicting of market trends. This influences the tactical plans with a three-month horizon and the strategic plans with a 12 to 60 month focus. These plans consist of sales strategies and product allocations.

Recent developments, such as increases in inflation, interest rates and volatility levels, have disrupted the traditional market cycles. In its 2022 demand outlook, the European Steel Association Eurofer predicts a recession until the second quarter of 2023 (Eurofer, 2023). Increases in volatility levels is linked to Russia's war in Ukraine, while the economic uncertainty is driven by high inflation. Eurofer expects that positive developments will become visible from the third quarter of 2023 onward, but the overall trajectory of

steel demand remains uncertain.

This negative demand outlook directly impacts the profitability of steel-making companies. Accurate demand forecasts help to improve sales strategies and product allocations, which is crucial to survive.

2.1 Sales Contracts

Sales departments in steel-making companies negotiate contracts with customers. These customers are manufacturers in all types of industries. Contracts typically span 3, 6, or 12 months, with exceptions for longer periods such as 24 or 36 months.

There are two types of contracts, namely volume contracts and order book contracts. Volume contracts specify a fixed amount of steel volume for delivery. Such contracts are often used in engineering markets. Order book contracts focuses on the steel volume that is needed during the contract term. This steel volume depends on the customer's own production needs. This is common in the automotive industry, among others. Order book contracts lack a fixed volume and make it difficult to predict incoming orders. These contracts are highly affected by market disruptions which impact the manufacturers' production levels and steel demand.

Effective management of incoming orders is crucial to avoid extended delivery times and to avoid failure to comply with contractual obligations. It also helps to optimize production capacity and minimize order shortages. Steel-making companies may consider attracting last-minute customers by lowering prices in times of order shortages, but this strategy can lead to longer periods of reduced prices and a decrease in profitability.

An alternative strategy involves selling steel at a reduced price in general (foreign) markets, commonly referred to as 'dumping'. While not necessarily profitable, it safeguards prices in the primary market. Constructing smarter contracts with customers helps to reduce general sales, while it increases the market share and overall profits when done timely.

Since contracts with manufactures lack a fixed volume, production is highly depended on demand in the market. Accurate demand forecasting helps to account for developments in the market. Therefore, it plays a crucial role in the management of incoming orders. Furthermore, it helps to prioritize the production capacity and to decide on the profitability of orders. A positive demand outlook makes it for example possible to be more selective when entering new contracts and accepting new orders. On the other hand,

a negative demand outlook might be reason to accept less profitable orders.

The profitability of orders depends on volume and price agreements, but also on product type. Steel products with value-added features such as coating or painting yield higher profit margins. Contracts for these types have therefore a higher priority.

Accurate demand predictions also help to improve product differentiation. If demand forecasts show positive outlooks for the automotive market, there is reason for steel-making companies to increase their resources on the development of new products. These products helps them to distinguish themselves from competitors who want to jump in the same market. Furthermore, accurate demand forecasts help to decide which geographical regions to target. This determines where to strengthen the relationships with customers and where to spend budget on marketing.

2.2 Forecasting Considerations

In this paper, I specifically focus on demand of steel products in the automotive industry and EU27 region. The demand for steel products in this market is consumer-driven. Furthermore, it is impacted by environmental regulations and increasingly affected by the demand for electric vehicles. This makes it interesting for steel-making companies to predict further developments in steel demand for this market. Given the various ways in which steel demand can be defined (see Appendix A), it is important to clarify what exactly will be predicted.

Steel demand can be represented as real demand, true demand an apparent demand. Real steel demand is the demand at the end of the production chain. It is affected by inventory changes of consumers and therefore subject to the Bullwhip effect (see Appendix B). Currently. it has limited relevance for steel companies in terms of tactical and strategic planning. It does not accurately reflect customer needs, because of its reliance of traditional stock-cycle behaviour.

True steel demand reflects the total steel consumption as seen from the consumers or end-users in a specific market, not the direct customers of steel-making companies. True steel demand accounts for exports and imports of steel through steel intensive products. For example, true steel demand in the EU27 includes the import of steel as component of an electrical vehicle that is imported from outside this region. In other words, true steel demand does not reflect the needs of the direct customers or manufacturers. Therefore it

is not suitable to determine the profitability of that region.

Apparent steel demand does reflect steel demand from direct customers in the market. It accurately reflects the needs of manufacturing companies and is not affected by changes in inventory levels. As a result, apparent demand is the primary interest when forecasting steel demand. Apparent steel demand can be examined at both an aggregated level and on a product level. To determine the best sales strategies and production allocations, apparent demand will be examined on a product level.

Current methods such as SWIP rely on a 'top-down' approach, whereby apparent steel demand is derived from the predictions of real steel demand. The derivation of apparent steel demand from real steel demand is based on traditional market cycles and stock-cycle behaviour. Due to disruptions in traditional market cycles and changes in stock-cycle behaviour, new forecasting methods should predict apparent steel demand directly.

3 Literature Review

This literature review aligns with Chu and Kong's (2018) survey, titled 'Comprehensive Survey of Steel Demand Forecasting Methodologies and their Practical Application for the Steel Industry,'. This survey provides a broad overview of techniques that are used in the steel industry.

In this literature review, I discuss traditional and modern modeling techniques that are used to forecast steel demand. I discuss the limitations and applications of these techniques. Furthermore, I discuss the applicability of machine learning techniques and their interest for the steel industry.

It is worth noting that the terms 'steel demand' and 'steel consumption' are often used interchangeably in the literature. Sometimes steel demand is examined from the perspective of a specific industry and sometimes from the perspective of a specific region or country. Whether real steel demand, true steel demand or apparent steel demand is examined, is also not often clear. Sometimes it follows from the context of an article.

3.1 Econometric Demand Models and VAR

Following Chu and Kong (2018), traditional demand models can be categorized into two main categories: econometric demand models and vector

auto-regression (VAR) models. Each category has its own principles and assumptions.

Econometric demand models can be sub-categorized into single equation models and simultaneous equations models. Single equation models use a single equation to predict the behavior of a dependent variable given one or more independent variables. They are suitable when clear causal relationships are present between the variables. Simultaneous equations models are similar, but these use multiple equations to predict various dependent variables simultaneously. This makes it possible to capture more complex relationships. Despite their simplicity and interpretability, these models often rely on external forecasts for explanatory variables. This introduces uncertainty and potential inaccuracies into the prediction results.

Labson et al. (1994) apply a single equation model for forecasting steel demand in Japan, China and North America. They include variables such as steel prices, industrial production and technological change. Pei and Tilton (1999) use a simultaneous equation model to illustrate the impact of technology and consumer preferences on the income elasticity of metal demand in high- and low-income countries.

The VAR model, a type of simultaneous equation model, treats all system variables as endogenous and models their joint behavior over time. This approach allows simultaneous forecasting of multiple variables and accounts for correlations between them. However, the curse of dimensionality is a drawback. Including more and more variables can complicate the identification and estimation of the model (Basu & Michailidis, 2013). Additionally, VAR models assume joint stationarity over time. This is challenging for accurate long-term forecasts (Hamilton, 1994).

Chen et al. (1991) use a VAR model to forecast steel demand in China up to the year 2000. They capture joint serial correlation between variables related to production, price level, investments and steel demand. Wu and Crompton (2003) use a VAR model to forecast steel consumption in China up to 2010 and include variables such as real GDP and investment expenditure. To provide more flexibility and to obtain higher accuracy, they make also use of Bayesian priors.

3.2 Intensity of Use Models

The intensity of use (IU) approach goes back to 1972 when it was introduced by the International Iron and Steel Institute (IISI), nowadays known as the

World Steel Association (worldsteel). This approach is based on the hypothesis that the quantity of material consumed per unit of output, referred to as the intensity of use, is linked to a country's level of economic development as indicated by its GDP per capita. This hypothesis became popular in the 1970s when member countries of worldsteel experienced a decline in steel demand, despite continuing growth in macroeconomic indicators such as GDP. The lack of strong causal relationships between steel demand and GDP led to concerns about the usability of traditional econometric approaches and underscored the need for new models.

Studies on raw material demand (Malenbaum, 1973, 1978) further popularized the intensity of use approach. Since then, it has become a widely used prediction tool in the steel industry and a popular topic in the literature. For example, Crompton (2000) uses this tool to forecast Japan's crude steel consumption between 1997 and 2005. This study shows a link between investment spending, consumption and the intensity of steel.

Recent discussions question the validity of the intensity of use hypothesis. Wårell (2014) analyzes steel consumption and GDP trends across 61 countries over 42 years, using ordinary least squares and a panel regression model. Her analysis shows that the hypothesis applies mostly to middle-income countries. Crowson (2017) identifies common weaknesses in data that is used to test the intensity of use hypothesis and underlines the misleading nature of the relationship between consumption and GDP per capita.

3.3 SWIP Models

Many steel-making companies companies currently use Steel Weighted Industrial Production (SWIP) models, using a bottom-up approach to estimate steel demand. SWIP models are based on underlying components that reflect the steel-intensive industries. These models consists of a linear equation whereby demand levels from steel-intensive industries are multiplied by their predetermined weights. These weights reflect the industries' relative importance in the steel market.

SWIP models determine weights in advance using industry-specific data, adjusting them based on industry specific demand levels. This approach heavily relies on industrial performance indicators and offers a more accurate prediction of steel demand, compared to approaches that rely on macroeconomic indicators. The reason is that macro-economic indicators rely on the whole economy, including the performance of industries that are less de-

pendent on steel, such as the leisure and tourism industry. This can lead to distorted outlooks, especially for developed countries whereby these industries are relatively large.

SWIP models allow for the derivation of apparent steel demand from real steel demand, taking into account fluctuations in inventory levels. However, due to recent economic developments such as disrupted market cycles and changes in stock-cycle behaviour, the traditional market dynamic is no longer present. Therefore it has become more challenging to derive apparent steel demand from real steel demand. This makes the use of SWIP models less reliable and results in frequent revisions in current demand forecasts (Eurofer, 2023).

Another limitation of the SWIP model is the high level of aggregation within industries. For the automotive industry, this means a lacks of differentiation in vehicle types. This leads to an inadequate representation of market developments, such as the shift towards electric vehicles and the reduction of vehicles' weights in order to minimize energy consumption. Due to lack of differentiation, these developments are not sufficiently be reflected, which leads to less reliable predictions.

3.4 Hybrid Demand Models and Machine Learning

Other models that have gained popularity are the hybrid models. Hybrid models are a collective name for models that combine different modeling approaches, such as traditional approaches with the more modern ones.

An example is the Grey model. This model is based on Grey System Theory and combines statistical and mathematical techniques to handle limited or uncertain information in data sets. While not directly related to steel demand, Kayacan et al. (2010) examine the performance of Grey-Verhulst models and show that these are effective when it comes to time-series predictions. Evans (2014) propose an alternative approach for parameter estimation in generalized Grey-Verhulst models and show their effectiveness in short-term prediction of steel demand in the United Kingdom.

Hybrid models can also include machine learning techniques to improve traditional forecasting methods. Tseng et al. (2001) introduce a Fuzzy-ARIMA model, combining ARIMA with fuzzy regression and probabilistic neural networks. This technique helps to address uncertainties and inaccuracies in the data. Torbat et al. (2018) apply this Fuzzy-ARIMA model to forecast Iran's crude steel consumption. They find that this hybrid approach

result in higher accuracy, compared to traditional ARIMA models. A short introduction into machine learning, including neural networks, is included to the appendix of this paper (Appendix C).

While hybrid models offer promising forecasting results, they come with challenges due to their mathematical complexity and theoretical nature. Steel-making companies prioritize practical applicability and user-friendliness and are therefore reluctant to use hybrid approaches. This emphasizes the importance of finding a balance between user-friendliness and performance of models, as highlighted in Chu and Kong's (2018) Comprehensive Survey of Steel Demand Forecasting Methodologies.

More and more studies look into the use of machine learning techniques to model steel demand. Kumar et al. (2023) develop an Artificial Neural Network model for steel consumption in India using variables such as GDP, Index of Industrial Production, and population. Using back-propagation they outperform traditional time-series methods. Azadeh et al. (2012) compare Artificial Neural Networks with Fuzzy Linear Regression and Conventional Linear Regression models to predict steel consumption in the United States and Iran. Their study reveals that Artificial Neural Networks have the best predictive performance for steel demand in the USA. Fuzzy Linear Regressions show better predictive performance for steel demand in Iran. The authors discuss that the difference in modeling performance can be attributed to the economic situation, which is stable in the United States and unstable in Iran. This highlights the fact that different situations might ask for different forecasting models.

Raju et al. (2022) introduce a forecasting framework and evaluate several models; Random Forest Regression, Gradient Boosting Regression, Extreme Gradient Boosting Regression, Support Vector Regression, Extreme Learning Machine and Multilayer Perceptron Neural Network. Furthermore, stacking models are included, which combine the predictions of multiple machine learning models. The authors conclude that these stacking models have the best predictive performance. While studies like this offer promising results, it also highlight the complexity of using machine learning models. The diverse range of modeling choices makes it complicated to find the most suitable ones.

During the United Nations Conference on Trade and Development, Hopp (2021) introduced his study on Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks. This study highlights the effectiveness of Long Short-Term Memory models for macro-economic time-series, outper-

forming dynamic factor models in predicting current or near-future scenarios. It is shown that Long Short-Term Memory models are rather easy to apply and work well with small data sets. The use of monthly or quarterly data in macro-economic forecasting results in relatively small data sets, which creates challenges for msot machine learning models. According to Hopp (2021), Long Short-Term Memory models are highly capable of working with data sets that are short in length, but involve multiple additional variables. This makes the Long Short-Term Memory model an interesting candidate for predicting steel demand.

There is extensive literature available, both on single Long Short-Term Memory models as on hybrid models that include Long Short-Term Memory. For example, Zhang et al. (2023) use a hybrid autoregressive model with Long Short-Term Memory to predict COVID-19 cases. Park and Yang (2022) introduce an interpretable Long Short-Term Memory model to predict GDP growth and crises. These studies show promising results, which makes it interesting to examine the applicability of these models for the steel industry.

4 Data

To forecast steel demand, I use consolidated data obtained from Tata Steel Netherlands. The data set consists of European demand for 39 types of steel products. The data is monthly available from April 2008 to June 2023. There is no missing data, so the time series consists of 183 data points per product. The time span includes significant economic events, such as the financial crisis and the COVID-19 pandemic.

I focus on the prediction of demand for steel products that are used in the automotive industry. From the list of 39 products, I have chosen six products that relate mostly to this industry. These products will serve as target variables in the forecasting models. The products are:

- **Hot Rolled Narrow Strip:** Commonly used in the automotive industry for various structural components in vehicles.
- **Lengths Cut From Hot Rolled Wide:** Commonly used for manufacturing automotive components that require wider and thicker steel sheets, such as vehicle frames.

- **Cold Rolled Sheets:** Characterized by a smooth surface finish and often used for exterior parts such as door panels and hoods.
- **Hot Dipped Metal Coated Sheets:** Commonly used for parts that require heavy resistance to corrosion, such as undercarriage components.
- **Electrolytically Metal Coated Sheets:** Commonly used used for parts that require high-quality surface finish as well as moderate corrosion resistance. This can range from interior to exterior components.
- **Grain Non-Oriented:** Commonly used for electrical components, such as the motors and transformers in vehicles.

Figure 1 displays the demand, from the perspective of steel-making companies seen as 'market supply', of these products over time. Notice that these products show similar behaviour in terms of demand.

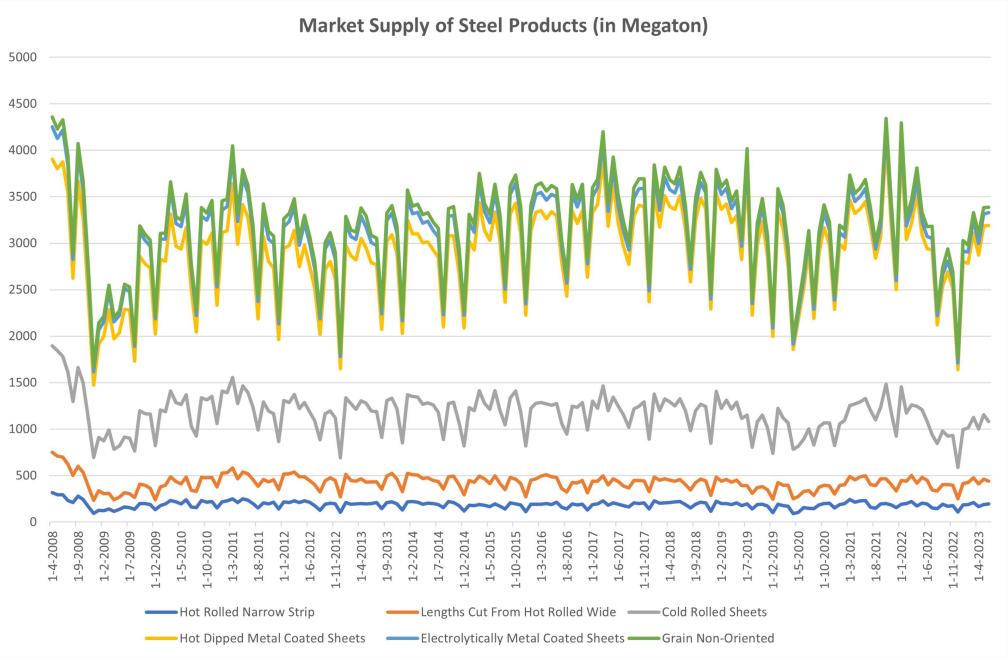

Figure 1: Market Supply of EU27 Automotive Steel Products (2008-2023) - Tata Steel Netherlands

To improve the demand predictions for these products, I include additional variables into the forecasting models. These variables, also knows as features, provide additional information related to steel demand.

4.1 Available Features

I consider 27 macro-economic and industrial indicators, mostly related to the automotive industry, to include as features. The selection of these features is based on domain knowledge from industrial experts at Tata Steel Netherlands. Domain knowledge helps to improve data quality. It limits the number of possible features by ensuring that these are related to the target variables. It also helps to prevent that irrelevant or redundant data will affect the predictive performance.

Monthly data availability is considered as practical criterion in the feature selection. While interpolation techniques allow for less granular data, for example quarterly or yearly, the inclusion of such data could potentially lead to noise in the models.

I retrieved data for the feature via Refinitiv, a provider of financial markets data and infrastructure. Data is sourced via Refinitiv from the following sources: Eurostat, business information company Wards, the German Association of the Automotive Industry (VDA), S&P Global, the Directorate-General for Economic and Financial Affairs (DG ECFIN), the European Central Bank and the European Banking Federation.

Features are categorized in four types of indicators: Economic Performance, Purchasing Managers' Index (PMI), Confidence and Sentiment, and Price and Market. These categories will be discussed in the upcoming subsections.

4.1.1 Economic Performance Indicators

The first group of indicators are the economic performance indicators. These are listed in table 1. The indicators are industrial production (total and manufacturing), gross value added (automotive, construction and machinery) and the number of cars produced (EU27 and Germany).

Economic Performance Indicators	**Source**
IP Total	Eurostat
IP Manufacturing	Eurostat
GVA Automotive (Gross Value Added in the Automotive Sector)	Eurostat
GVA Construction (Gross Value Added in the Construction Sector)	Eurostat
GVA Machinery (Gross Value Added in the Machinery Sector)	Eurostat
Number of Cars Produced in EU27	Wards
Number of Cars Produced in Germany	VDA

Table 1: Table of Economic Performance Indicators

Industrial production (IP) is a measure of the output of the industrial sector. The market supply of steel products is closely related to production and manufacturing activities in the automotive industry and therefore to the industrial production index. The manufacturing IP index is especially relevant to the automotive industry, due to the fact that market supply of automotive steel products is primarily influenced by manufacturing activities.

Gross value added (GVA) is a measure that represents the value of goods and services produced within a specific industry or sector. It is calculated as the gross domestic product product plus subsidies and minus taxes on products. Therefore it closely reflects the economic performance of particular industries. The Gross Value Added is included for the automotive, construction and machinery sectors. GVA for automotive is a logical choice, as the market supply for automotive steel products dependents on the overall health and performance of the automotive industry.

The reason to include the Gross Value Added for both construction and machinery is that these GVA's can offer valuable insights into broader economic trends and changes in steel demand and prices. In other words, these can act as leading indicators for economic performance. For example, a rise in GVA for construction might signal a positive impact on infrastructure development, an increase in demand for automotive vehicles and an increase in market supply for automotive steel products.

Moreover, by considering Gross Value Added for multiple industries, we gain a more comprehensive reflection of the GDP and overall economic performance of a region, in this case the EU27. While the automotive sector is more consumption driven and the construction and machinery sector are more economic driven, they are closely interconnected and changes in one of these industries can have an impact on the others.

The number of cars produced is included for the EU27 region, because it serves as an indicator of the demand for automotive goods including automotive steel products. Additionally, the number of cars produced in Germany is included. The reason is that Germany is a major user of European steel products and it also has the largest automotive market compared to other EU27 countries. Cars produced tend to better reflect the EU27 market supply of automotive steel goods than cars sold, primarily due to import and export activities. Therefore, the number of cars sold is not included.

4.1.2 Purchasing Managers' Indices

The second type of indicators are the Purchasing Managers' Indices (PMI's). The Purchasing Managers' Indices are economic lead indicators derived from monthly questionnaires sent to experts. The indicators that are included, are presented in table 2.

Purchasing Managers' Indices (PMI)	**Source**
Manufacturing PMI - Output	S&P
Manufacturing PMI - New Orders	S&P
Manufacturing PMI - Input Prices	S&P
Manufacturing PMI - Output Prices	S&P
Construction PMI	S&P
Services PMI	S&P

Table 2: Table of Purchasing Managers' Indices (PMI) Indicators

I include PMI Output, PMI New Order, PMI Input Prices and PMI Output Prices for the manufacturing sector. PMI Output measures the level of production in the manufacturing industry, while PMI New Orders measures the demand for goods in this industry. PMI Input Prices represents the cost of manufacturing and tracks the costs of raw materials and components including steel. PMI Output Prices indicates how effectively companies can pass on cost increases to consumers.

While the automotive sector is primarily related to the manufacturing sector, I also include the overall Purchasing Managers' Indices for the construction and services sectors. PMI Construction indicates the economic performance of the construction industry and PMI Services represent the economic performance in the services industry. These overall Purchasing

Managers' Indices are included, because they play a significant role in representing the economy and are closely interconnected with other industrial sectors, including manufacturing. To illustrate this, the service industry represents sub-sectors such as transportation, wholesale trade and retail trade. These are of interest, because they play a crucial role in supply chains and the delivery of goods to consumers. For example, if transportation costs increase, it could impact the import and export of components necessary for vehicle production, consequently influencing the demand for steel.

4.1.3 Confidence and Sentiment Indicators

The third kind of indicators are confidence and sentiment indicators. These indicators measure the confidence and/or sentiment of consumers and business in the European Union regarding their economic prospects. The indicators are based on surveys that inquire about consumers' and business leaders' expectations for future economic conditions, investments and more. The included indicators are presented in table 3.

Confidence and Sentiment Indicators	**Source**
Economic Sentiment Indicator Volatility	DG ECFIN
Overall Industrial Confidence Indicator	DG ECFIN
Services Confidence Indicator - Total	DG ECFIN
Consumer Confidence Indicator - EU	DG ECFIN
Retail Trade Confidence Indicator - Overall	DG ECFIN
Construction Confidence Indicator - Overall	DG ECFIN
Industrial Confidence Indicator - Motor Vehicles, Trailers and Semi-Trailers	DG ECFIN
Industrial Confidence Indicator - Machinery and Equipment	DG ECFIN

Table 3: Table of Confidence and Sentiment Indicators

I include Overall Industrial Confidence, Services Confidence, Consumer Confidence, Retail Trade Confidence and Construction Confidence. Furthermore, I included the indicator for Industrial Confidence in Motor Vehicles, Trailers and Semi-Trailers and for Industrial Confidence in Machinery and Equipment. These indicators are designed to offer insights into the future

economic outlook and are therefore considered leading or forward-looking indicators.

Other leading indicators are the Economic Sentiment Indicators. These indicators are composite in nature, combining various confidence indicators into one overall measure of economic sentiment in the EU. Sentiment indicators provide a comprehensive insight into how consumers and businesses perceive the current and future economic situation.

Sentiment concerning volatility can significantly influence investment decisions, consumer behavior, supply chain disruptions, and exchange rates. Thus, sentiment regarding volatility is likely to be linked to the outlook of the automotive market, consequently impacting the demand and supply of related steel products.

4.1.4 Price and Market Indicators

The fourth category of indicators included in this practical analysis include price and market indicators. These are presented in table 4.

Price and Market Indicators	**Source**
CPI (Consumer Price Index)	Eurostat
PPI (Producer Price Index)	Eurostat
Unemployment Rate	Eurostat
US$/EUR (US Dollar to Euro Exchange Rate)	Refinitiv
Ten-year Bond (Yield on Ten-Year Government Bonds)	ECB
Euribor (Euro Interbank Offered Rate)	EBF

Table 4: Table of Price and Market Indicators

Firstly, we have the Consumer Price Index, which tracks changes over time in prices consumers pay for goods and services. An increase in the CPI could lead to a decrease in purchasing power, potentially impacting automotive demand.

Secondly, we have the Producer Price Index, which is an indicator for changes in prices that producers receive for their output. An increase in Producer Price Index can signal rising prices at the producer level. When production costs increase, these higher expenses may be passed on to consumers in the form of higher vehicle prices. Consequently, this could impact consumer demand for automobiles. Moreover an indirect effect may occur,

as higher prices for other goods could reduce consumers' available funds for investing in automotive products.

Both the Consumer Price Index and Producer Price Index are commonly regarded as lagging indicators, because they reflect past behavior of prices. They offer insights into the current economic condition and are valuable for assessing historical trends and economic developments.

In addition to Consumer Price Index and Producer Price Index, the unemployment rate in the EU is included. The unemployment rate is considered a lagging indicator, reflecting past labor market conditions and economic performance. Changes in the unemployment rate typically follow broader economic trends, such as economic downturns or recoveries. Labor market conditions are closely tied to consumer spending and purchasing power. Given the consumer-driven characteristic of the automotive market, these indicators offer valuable insights.

I also include Ten-Year Bond Yields as an indicator. This reflects the current expectations of the bond market and provide information on interest rates. Consequently, it holds relevance for predicting future economic conditions, savings, and investment expenditures.

The last indicator is Euribor, the Euro Interbank Offered Rate. Euribor provides information on interbank lending rates and serves as a benchmark for the large euro money market, especially for short-term interest rates in the Eurozone. Changes in Euribor can impact borrowing costs, thereby affecting economic activity.

4.2 Data Processing

Data for the demand and features is achieved in September 2023, covering the period from April 2008 till January 2023. Recent data beyond this period is omitted due to potential revisions that could impact the evaluation. For financial indicators such as interest rate and exchange rate, there are no publication lags and limited data revisions. Sentiment indicators and purchasing managers' indices show limited lags and minimal data revisions. However, macro-economic indicators such as GDP and GVA have significant publications lags and data revisions occur more often. Consequently, these could affect the evaluation.

To prevent issues with dimensionality and model complexity, I use feature selection techniques to only select the most informative subsets of features for each of the steel products. Before applying feature selection techniques,

it is important to process and evaluate the data. Data processing involves data evaluation and cleaning. Furthermore, I test the data on stationarity and I apply both scaling and transformation techniques. This is necessary for the feature selection and helps to improve the performance of the models.

Data evaluation and cleaning is the process of identifying and fixing incorrect or missing values in the data set. This can occur, for example due to issues with the reporting of the data. The process of cleaning is essential, because incorrect or missing values can lead to unreliable prediction results. I have found no incorrect or missing values in the data set of the features. I have also found no incorrect or missing values in the data set of the steel products.

The order in which techniques such are scaling and transformation are applied, effect the data and therefore the modeling outcome. It is important to know that data is used to train the model, validate the model and test the model. Model validation is required for hyperparameter tuning and model testing is required to evaluate its forecasting accuracy on unseen data. To obtain reliable performance measures, it is important that there is no overlap in the data that is used for training, validating and testing. Data should be split to obtain non-overlapping subsets. To prevent data leakage or spill-over effects it is important to split data before scaling or transformation processes. The characteristics of the data set changes throughout time. Time-based validation is a cross-validation techniques whereby the data set is divided into several folds. This helps to assess the model's performance across different time periods and yields a more statistically robust model evaluation.

4.2.1 Stationarity and Transformation Techniques

After cleaning and splitting, the data is tested for stationarity. Stationary means that seasonality and trends in the data are removed. Machine learning methods such as Long Short-Term Memory are not robust to non-stationarity in the data. Therefore, I test the data for stationarity before applying the data to the models.

While the features are already seasonally adjusted, the time series might still show trends over time. To obtain a better understanding of the time behaviour of data, python libraries such as *statsmodels.tsa* with sub-module *seasonal_decompose* can be used to decompose and visualise data trends on a monthly, quarterly or yearly basis. Furthermore, auto-correlation plots and partial auto-correlation plots can be used to obtain a better understanding

of correlation in the data and to determine the number of lags. These lags represent the differences between months that could help to make the data stationary. A lag of one month represents the difference between the current and next month, a lag of two months represents the difference between the current and second next month and so on. After applying the determined number of lags for each of the features, I use both the augmented Dickey-Fuller test and the Kwiatkowski-Phillips-Schmidt-Shin test to evaluate whether features are stationary or not. For the non-stationary features, I increase the number of lags till these features are stationary.

Thereafter, I continue with transformation techniques. Transformation techniques are used to make the data more suitable for modeling by improving the distribution and other data characteristics. There are several transformation techniques that can be used. One technique is the log transformation, which is used to stabilize variance and reduce skewness in the data. It can be particularly useful for centering data with exponential characteristics. Other techniques include the Box-Cox and Yeo-Johnson transformations, also known as power transformations. These methods serve a dual purpose. They both stabilize variance and make the data more Gaussian-like by adjusting its distribution. After applying these methods, quantile-quantile plots can be used to show that the data follow a Gaussian distribution.

After transformation I apply scaling techniques to make sure that all features have a similar scale. Scaling techniques focus on the magnitude of the data. This is necessary, because some feature selection techniques and machine learning methods are sensitive to the magnitude of the data. Target variables and features are given in different units, leading to significant differences in the magnitude. For example, the demand for steel products is given in kilotons, the number of cars produced in Europe is given in thousands and the Purchasing Managers Index is given as value between 0 and 100.

There are different scaling techniques that can be applied. These include min-max scaling, normalization and standardization. Min-max scaling adjusts the data to a specific range by subtracting the minimum and dividing by the difference between the maximum and minimum value. Normalization is a variation of min-max scaling in which the range is set to 0-1. Standardization, also known as z-score scaling, scales the data such that it has a zero mean and unit variance. It subtracts the mean value of the data from each individual data point and subsequently divides the data points by the standard deviation.

In line with Raju et al. (2022), I perform the Yeo-Johnson power transfor-

mation with implicit data standardization using the Scikit-learn implementation of PowerTransformer (method= "Yeo-Johnson", , standardize=True). The Yeo-Johson power transformation can be described mathematically using

$$y^* = \begin{cases} \frac{\left((y+1)^\lambda - 1\right)}{\lambda}, & \text{if } \lambda \neq 0, y \geq 0, \\ \log(y+1), & \text{if } \lambda = 0, y \geq 0, \\ -\frac{(y+1)^{(2-\lambda)}-1}{(2-\lambda)}, & \text{if } \lambda \neq 2, y < 0, \\ \log(-y+1), & \text{if } \lambda = 2, y < 0, \end{cases}$$

where y^* is the transformed value, y is a list of n strictly positive numbers and λ serves as hyperparameter. The hyperparameter is used to control the nature of the transformation and takes on the value that best transforms the distribution of the data to a Gaussian probability distribution. For features where the data already follow a Gaussian distribution, λ can be set to be 1.

In the following subsections I will dive in the feature selection techniques. These techniques help to obtain the most relevant features and to reduce to overall dimensionality in the data.

4.3 Feature Selection

After the data is processed it is important to consider a more in-depth selection of feature for each of the target variables. Domain knowledge helps to obtain a relevant selection, but still leads to 27 possible feature. Removing the least informative features can prevent uncertainty and improve the predictive performance of the model (Kuhn and Johnson, 2013).

In the following subsections we will look at dimensionality reduction using Principle Component Analysis and we will look into target-based feature selection techniques.

4.3.1 Dimensionality Reduction

Various techniques exists to reduce the dimensionality of the data. A popular technique is Principal Component Analysis, which is also used by Raju et al. (2022) in their analysis for steel demand. Principle Component Analysis is used to reduces the dimensionality in the data set, while preserving as much variance as possible. Principle Component Analysis identifies the principal components in which the data exhibits the most variation. These principal components are formed through linear combinations of features. In these

linear combinations, weights indicate the importance of their corresponding features.

The first principal component explains the most variance in the data, the second principal component explains most of the variance in what is left once the effect of the first component is removed, and so on. These principal components helps to increase the interpretability of the data while keeping the maximum amount of information. Furthermore, this makes it possible to visualise high-dimensional data.

Feature selection based on Principle Component Analysis, relies on the idea that for each principal component the feature with the highest corresponding weight should be kept. I apply this idea to the set of features to obtain an initial idea of the feature importance. Since Principle Component Analysis is sensitive to the scale-size of features, I use the set of features after data transformation and scaling.

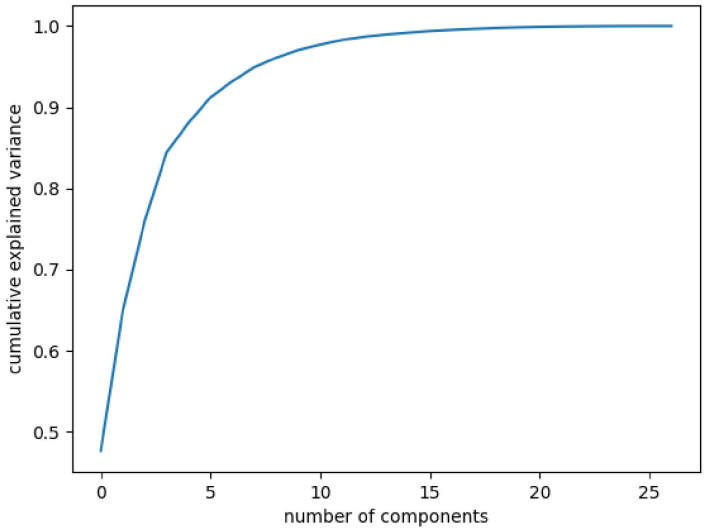

Figure 2: Principal Component Analysis on the Features

Plot 2 shows that the first eight principal components, or linear combinations of features, help to explain 96% of the variance in the data set. After the first fifteen principal components there is a convergence at 99% explanation of the variance. This indicates that it is possible to reduce dimensionality

while still capturing the most significant patterns and variations in the data.

To reduce the dimensionality of our data set, I fit fifteen principal components to the data and subsequently select per principal component the feature with the highest weight. This results in the following features: 'Overall Industrial Confidence Indicator', 'CPI', 'Ten-year bond', 'Number of cars produced in EU', 'GVA Construction', 'US$/EUR', 'GVA Automotive', 'Construction PMI', 'Services PMI', 'Industrial Confidence Indicator - Motor Vehicles, Trailers, and Semi-Trailers', 'Unemployment Rate' and 'Retail Trade Confidence Indicator - Overall'. These features appear to be the most important in explaining the patterns and variation in the data. Therefore, it might be expected that these features are highly represented in further (target-based) feature selection processes.

Note that the number of features is less than the number of principal components that were fitted to the data. This can be explained by the fact that some principal components assign the highest weight to the same feature. Furthermore, it is crucial to emphasize that this method is solely based on the variation in the set with features and helps to reduce the dimensionality in this set of data. It provides insights on the number of features that should be included in a forecasting model. It does not take into account any of the target variables. Since the importance of features is highly depended on the target variable, it is important to consider other (target specific) techniques as well.

4.3.2 Target Based Feature Selection

To forecast demand I use multivariate forecasting models. These models use multiple features to predict one specific target variable. Therefore, each target variable has its own model and relies on its own sub set of features.

Kuhn and Johnson (2013) show that several types of feature selection techniques exists. These techniques can be categorized as wrapper methods or filter methods. Wrapper methods evaluate the model performance of multiple models at a time, using algorithms that add or remove predictors to find the optimal feature combination. Filter methods evaluate the relevance of features outside the predictive performance and include the features that pass specific selection criteria, e.g. a certain level of correlation with the target variable.

I use Recursive Feature Elimination, a technique that is based on an iterative process. Initially, a prediction model is trained on the complete

set of features. Subsequently, the least informative features are removed till the pre-set number of features is reached. Recursive Feature Elimination is therefore known as a wrapper method that basis the feature selection process directly on model performance. The model performance is determined based on the accuracy of the predictions and the relevance of features is determined by their impact on the performance of the model. Recursive Feature Elimination can be applied to different prediction models. I use a standard linear regression model and the ensemble model Gradient Boosting. These models are different than to models in the evaluation, to prevent bias in the feature selection. To make the feature selection process more robust I use Recursive Feature Elimination with Cross-Validation. The Cross-Validation technique switches alternates between different subsets of the data when validating the model. This provides a more robust assessment of the model performance with less sensitiveness to underlying time-trends. Based on the results from both Recursive Feature Elimination with Cross-Validation and Principle Component Analysis, I test the models on different subsets of the features. The features that yield the best overall performance are used in the model evaluation. The overall performance is measured by the Residual Mean Squared Error, Mean Absolute Error and Mean Absolute Percentage Error.

5 Methodology

In the literature review, the Long Short-Term Memory model stands out for macroeconomic forecasting, making it the primary focus of this analysis. To construct a Long Short-Term Memory model, the lstm_nowcast library is used, which is provided by Hopp is his paper on Economic Nowcasting with Long Short-Term Memory Artificial Neural Networks (Hopp, 2021). To determine the effectiveness of this model, I will compare it with Support Vector Regression and Random Forest Regression.

Support Vector Regression and Random Forest Regression are two of the models used in the study by Raju et al. (2022) on demand forecasting in the steel industry. Raju highlights that Support Vector Regression is a popular choice due to its strong capability to generalize to new data and its consistent performance irrespective of the number of features. Random Forest Regression is selected because it is an ensemble bagging model that might provide robust and efficient demand forecasts. More on ensemble learning,

specifically bagging, can be found in Appendix C.2.

In addition to these models, I compare the Long Short-Term Memory model to a new ensemble model, Ensemble 1. This model combines Support Vector Regression and Random Forest Regression through a simple average of the predictions. This approach may help to mitigate weaknesses from both individual models and therefore yield more reliable predictions.

Long Short-Term Memory models are based on Recurrent Neural Network models, which model time dependencies within a time-series using sequential information processing. The disadvantage of Recurrent Neural Network models is that they struggle with long-term dependencies, making them less effective in capturing longer-term relationships within time series. Long Short-Term Memory models are designed to overcome this disadvantage, making them particularly interesting for predicting more complex time-series.

Long Short-Term Memory models rely on an underlying algorithm that relies on many components including layers, neurons, time-steps, batches and training epochs. The algorithm is optimized using an optimizer. I will use Adam, which is an efficient adaptive optimizer that works well in practice compared to other stochastic optimization methods (Kingma and Ba, 2015). The optimizer is effected by different rates, namely the learning rate, decay rate and dropout rate. I will shortly go into these components and rates.

The layers consists of an input layer, an output layer and one or more hidden layers. Increasing the number of layers, which are also known as gates, result in deeper models which are capable of capturing more complex patterns. Each layer consists of a set of neurons which are responsible for learning and retaining patterns in sequential data. Increasing the number of neurons could help to grasp complex patterns in the data, but could also lead to issues with overfitting. This would mean that the model fits perfectly on the training data, but performs poorly on new data.

Long Short-Term Memory models process input data in sequences. The length of these sequences is determined by the number of time-steps. Increasing the number of time steps could help to capture longer-term dependencies. The disadvantage is that it could lead to computationally intensive models. This also holds true for including many neurons or layers.

To reduce computational requirements, data can be divided in multiple batches. Each batch contains a subset of the training data. After the procession of a batch of data, the model gets updated. While smaller batches require less memory, they might lack information on underlying patterns in the data, resulting in noisier updates of the model.

Furthermore Long Short-Term Memory models have a learning and decay rate. The learning rate determines the number of steps taken till the model's weights are updated. The decay rate is used to reduce the learning rate over time, which can help improve the optimization process. The number of training epochs specifies how many times the model will be exposed to the entire data set. Too few epochs lead to underfitting, while too many epochs can lead to overfitting. Lastly a dropout rate can be used to prevent overfitting. The dropout rate determines the faction of input units that are randomly set to zero in each batch. This helps to generalize the model, by relying less strongly on the training data.

The components and rates are known as the hyperparameters of the model. Determining the optimal values for these hyperparameters is essential for achieving the best model performance. Grid Search is used to do so. This is a systematic method for testing predefined combinations of hyperparameter values. The best combination, which depends on the underlying data, is used to obtain the results.

The Random Forest model, built with scikit-learn's RandomForestRegressor, is an ensemble technique that consists of multiple decision trees. Each tree has nodes representing decisions, branches for decision outcomes, and leaf nodes for predictions. Decisions are based on features, for example whether the number of cars produced in Germany exceeds a threshold. The predicted output, an average of values in the reached leaf nodes, corresponds to the target variable. The model's hyperparameters, including the number of trees, the maximum number of decisions or depth per tree, and the minimum data points required in a leaf node to make a new decision, are determined through Grid Search. The optimal combination is used for training. In time series, each tree learns patterns from historical data, which helps to improve the model's performance.

The Support Vector Regression (SVR) model, built with scikit-learn's SupportVectorMachines, works by finding a so-called hyperplane that best represents the relationship between features and the target variable. It identifies support vectors, the most important data points, and optimizes their location in the feature space. This helps to to minimize the difference between the actual values and the predictions of the target variable. Hyperparameters such as cost, gamma, kernel, and epsilon determine, among other things, the shape of the plane and its tolerance for prediction errors. Fitting precisely the training data could lead to low prediction errors, but it could result in poor generalization and worse predictions for the test data. The

best combination of hyperparameters is again determined using Grid Search.

The Ensemble 1 model combines both the Random Forest model and the SVR model using simple average. It takes the average of the predictions of both models. Combining these predictions allow for a more diverse modeling approach, using the strengths of both models. Furthermore, averaging can help to mitigate the impact of biases or poor performance from individual models, leading to more robust and reliable results. Additionally, it can improve the model's ability to generalize well to unseen data.

To evaluate the models, I use different performance metrics. These are the Root Mean Squared Error, the Mean Absolute Error and the Mean Absolute Percentage Error. Using multiple performance metrics in model evaluation provides a more comprehensive understanding of the model's performance, with each metric capturing different aspects of model behaviour.

The Root Mean Squared Error represents the square root of the mean squared differences between the actual and predicted values in the test set. It is sensitivity to outliers and heavier penalty for larger errors make it valuable for assessing the overall model's accuracy. The Mean Absolute Error represents the average absolute differences between actual and predicted values. It is less sensitive to outliers compared to Root Mean Squared Error and provides a straightforward interpretation of the average magnitude of errors. The Mean Absolute Percentage Error is expressed as a percentage and represents the average percentage difference between the predicted and actual values.

6 Results and Analysis

Both Recursive Feature Elimination with Cross-Validation and Principle Component Analysis are used to determine the best subsection of features. For each product, I select the features that provide the best overall performance among the models. This overall performance is determined by aggregating multiple performance metrics, the Root Mean Squared Error, Mean Absolute Error and Mean Absolute Percentage Error. These performance metrics provide a comprehensive evaluation of the model's performance.

For each product, the selected feature are (in random order):

- **Hot Rolled Narrow Strip:** 'Number of Cars Produced in EU', 'PPI', 'Industrial Confidence Indicator - Machinery and Equipment', 'GVA

Machinery', 'CPI', 'Consumer Confidence Indicator - EU', 'Unemployment Rate', 'Manufacturing PMI - Input Prices'.

- **Lengths Cut From Hot Rolled Wide:** 'Number of Cars Produced in EU', 'Industrial Confidence Indicator - Machinery and Equipment', 'Manufacturing PMI - new orders', 'Manufacturing PMI - Output', 'Industrial Confidence Indicator - Motor Vehicles, Trailers & Semi-Trailers', 'Economic Sentiment Indicator Volatility', 'Overall Industrial Confidence Indicator', 'Retail Trade Confidence Indicator - Overall'.
- **Cold Rolled Sheets:** 'Construction PMI', 'GVA automotive', 'Services PMI', 'Number of Cars Produced in EU', 'Overall Industrial Confidence Indicator', 'Manufacturing PMI - Input Prices', 'Unemployment Rate', 'Manufacturing PMI - New Orders'.
- **Hot Dipped Metal Coated Sheets:** 'Industrial Confidence Indicator - Motor Vehicles, Trailers and Semi-Trailers', 'Number of Cars Produced in Europe', 'Construction Confidence Indicator - Overall', 'Manufacturing PMI - New Orders', 'Manufacturing PMI - Output', 'Number of Cars Produced in Germany', 'Unemployment Rate', 'Euribor'.
- **Electrolytically Metal Coated Sheets:** 'Number of cars produced in EU', 'Number of Cars Produced in Germany', 'Manufacturing PMI - New Orders', 'Euribor', 'Services PMI', 'Overall Industrial Confidence Indicator', 'Economic Sentiment Indicator Volatility', 'Industrial Confidence Indicator - Machinery and Equipment'.
- **Grain Non-Oriented:** 'Retail Trade Confidence Indicator - Overall', 'Manufacturing PMI - Input Prices', 'Number of Cars Produced in EU', 'Construction PMI', 'CPI', 'GVA Construction', 'Industrial Confidence Indicator - Motor Vehicles, Trailers and Semi-Trailers', 'Manufacturing PMI - New Orders'.

The feature selection shows that the 'Number of Cars Produced in the EU' appears in the feature selection of all the different products, which indicates its broad impact on steel production. This is in line with the expectation, considering the focus on steel products within the automotive industry. Furthermore, we observe that both the Purchasing Managers' Indices (PMI) and the Confidence and Sentiment Indicators appear frequently in the selected

features. This suggests that these leading indicators play a significant role in predicting steel demand.

From the category Price and Market Indicators, we observe that 'Unemployment Rate' and 'Euribor' are used multiple times. The 'PPI' is only used to predict the demand for Hot Rolled Narrow Strip. Moreover, 'CPI,' 'US$/EUR,' and 'Ten-year Bond' are not used at all. This suggests that the contribution of these indicators to the prediction models of these products is not significant and indicates a relatively weak relationship between these indicators and the steel products

After the features are selected, we can continue with the modeling performance. I start with with Hot Rolled Narrow Strip. I obtain the best performing Long Short-Term Memory model for Hot Rolled Narrow Strip using Grid Search. This technique test the model on different subsets of hyperparameters. It appears that the best performing model consists of five hidden layers, 100 neurons, 12 time steps, 5 batches, and 10 training epochs. The learning rate is 0.01, the decay rate is 0.001 and the dropout rate is 0.2. This model forecasts the demand of Hot Rolled Narrow Strip for the test period January 2021 till January 2023. The results are shown in figure 6.

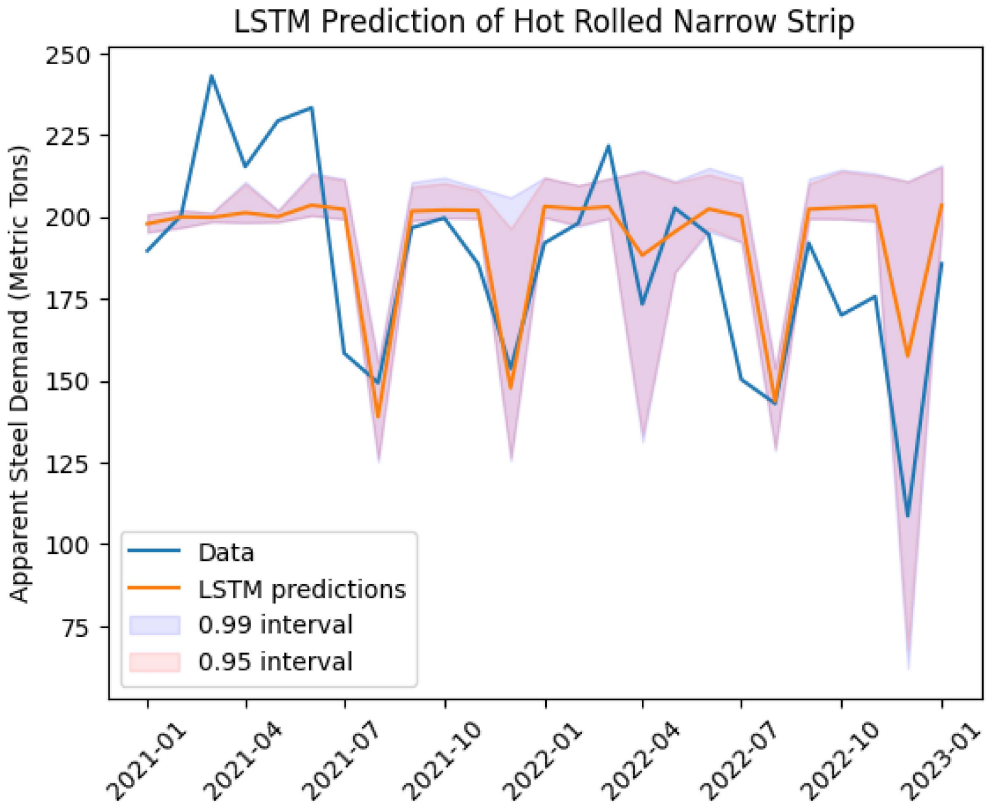

Figure 3: Predictions of Hot Rolled Narrow Strip using the best performing Long Short-Term Memory model

We observe that the model's predictions are close to the data during some periods of time, but not at others. For example, during the first half of 2021 there is a significant mismatch between the data and predictions. During the second half of 2021 we see that the data fall within the 95% confidence interval of the predicted values for most of the months, with exception for August and November. In both these months the predictions peaked too high. Especially in November, the predictions showed an increase in demand, which is in high contrast with the actual decrease. This resulted in a significant mismatch between the data and predicted values, making the model at this point impractical. For practical purpose, a difference within plus or minus 10 metric tonnes between data and predicted values is desired.

The predictions for the first half of 2022 seem to capture the trend of

the data, however the predictions themselves are too low and should be adjusted upwards. A possible explanation for this deviation could be the strong economic recovery efforts which lead to a higher actual supply of Hot Rolled Narrow Strip than predicted.

In the second half on 2022, this goes the other way around. The predictions are too high compared to the data. The highest deviations during this time are in October and December. For October the predictions showed a strong increase, which is in high contrast with the decrease in demand as shown by the data. For December we see that the predictions capture a downward trend but in much less extent than the data. This suggests that the forecasting model tends to make predictions on the higher side.

A potential reason for the developments in the second half of 2022 could be the ongoing geopolitical conflicts and a rise in inflation, which could have resulted in a lower actual demand of Hot Rolled Narrow Strip compared to the predictions. However, it should also be emphasized that forecasting too far into the future can yield unreliable results due to increased uncertainty and other challenges associated with longer-term predictions. The fact that training and validation of the prediction model is solely based on data till January 2021, may have contributed to the mismatch between data and predictions in the second half of 2022.

It is important to highlight that the 95% and 99% confidence intervals are extremely broad at certain points in time. For example, the predictions for April 2022 fall within the 99% interval of 130 to 210 metric tonnes of steel. Such a broad interval introduces uncertainty, making the predictions less reliable compared to smaller intervals.

When looking at the performance metrics, we observe that the Residual Mean Square Error indicates, on average, a 22.70 metric ton difference between the predicted and the data. The MAE indicates, on average, an 17.61 Mean Absolute Error between the predicted and data. The Mean Absolute Percentage Error suggests that the predictions have an absolute percentage error of 10.41%. This means the model's predicted values deviates from the data by an average of 10.41% in absolute terms.

We continue with the benchmark models Random Forest Regression, Support Vector Regression and Ensemble 1. The results for these these models are shown in figure 4.

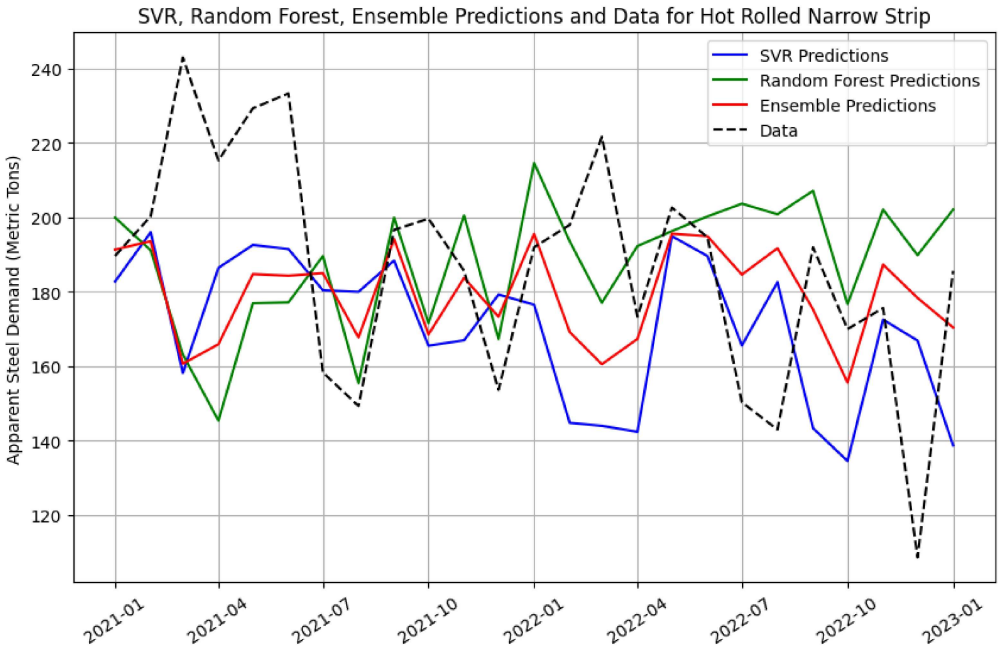

Figure 4: Predictions of Hot Rolled Narrow Strip using Random Forest Regression, Support Vector Regression and Ensemble 1

Figure 4 shows that the predictions of all three models deviate significantly from the data. Similar to the Long Short-Term Memory model, we see that the deviations between the predictions and data in the benchmark models are the largest in the first half of 2021 and the second half of 2022. In between these periods, the predictions are somewhat closer to the data. While Random Forest Regression provides the best results between August and November 2022, we see that Support Vector Regression provides the best results for the periods between April and June 2021 and between June and August 2023. The ensemble model provides the second best predictions in both periods. The simple average approach of this model makes it less sensitive to extreme predictions, which on one hand is a strength of the model. On the other hand it shows to be a disadvantage in periods whereby the increase or decrease in the demand is also extreme.

Table 5 provides an overview of the different performance metrics, Residual Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These performance metrics correspond

the Long Short-Term Memory model and benchmark models for each of the steel products.

		RMSE	MAE	MAPE
Hot Rolled Narrow Strip				
	LSTM	**22.70**	**17.61**	**10.14**
	RFR	38.43	29.05	16.74
	SVR	37.49	29.63	16.32
	ENS1	33.45	24.05	13.65
Lengths Cut From Hot Rolled Wide				
	LSTM	45.33	32.90	14.23
	RFR	**33.97**	**27.19**	**12.59**
	SVR	42.93	34.33	15.63
	ENS1	35.22	29.43	13.51
Cold Rolled Sheets				
	LSTM	157.6	**123.0**	**21.06**
	RFR	172.5	146.4	23.72
	SVR	**163.1**	142.1	22.48
	ENS1	164.9	139.5	22.47
Hot Dipped Metal Coated Sheets				
	LSTM	364.6	278.1	14.81
	RFR	360.6	265.8	13.57
	SVR	**328.1**	**247.9**	**12.56**
	ENS1	341.7	256.1	13.02
Electrolytically Metal Coated Sheets				
	LSTM	**67.76**	**58.02**	**50.29**
	RFR	80.19	76.09	63.29
	SVR	71.36	68.90	58.53
	ENS1	74.96	72.49	60.91
Grain Non-Oriented				
	LSTM	**27.42**	18.15	**17.70**
	RFR	23.04	18.73	20.85
	SVR	22.28	**17.90**	19.24
	ENS1	22.25	18.08	19.80

Table 5: Results from Long Short-Term Memory (LSTM), Random Forest Regression (RFR), Support Vector Regression (SVR) and Ensemble 1 (ENS1) for each product type (whole data set).

For Hot Rolled Narrow Strip, there are minimal differences in the performance metrics among the models. The Long Short-Term Memory model only performs better than the benchmark models in terms of the Mean Absolute Percentage Error.

For Lengths Cut From Hot Rolled Wide, the Long Short-Term Memory model yields the highest Residual Mean Squared Error. This indicates a greater deviation of predicted values from the data compared to the benchmark models. Random Forest Regression outperforms the Long Short-Term Memory model, Support Vector Regression and Ensemble 1 across all three performance metrics. Ensemble 1 outperforms Long Short-Term Memory model and the Support Vector Regression model across all three performance metrics.

For both Cold Rolled Sheets and Electrolytically Metal Coated Sheets, we observe that the Long Short-Term Memory model performs significantly better than the benchmark models across all three performance metrics. For Cold Rolled Sheets, the Residual Mean Squared Error is 157.6 metric tons, the Mean Absolute Error is 123.0 metric tons and the Mean Absolute Percentage Error is 21.06 metric tons of steel. For Electrolytically Metal Coated Sheets, the Residual Mean Squared Error is 67.76 metric tons, the Mean Absolute Error is 58.02 metric tons and the Mean Absolute Percentage Error is 50.29 metric tons of steel. While the Long Short-Term Memory model outperforms the benchmark models, the performance metrics are rather high given an average of 745.6 metric tons of steel each month for Cold Rolled Sheets and an average of 185.0 metric tons of steel each month for Electrolytically Metal Coated Sheets.

Conversely, for Hot Dipped Metal Coated Sheets the Long Short-Term Memory model performs worse than the benchmark model across all three performance metrics.

The performance metrics for Grain Non-Oriented are relatively close among the models. Long Short-Term Memory has the highest Residual Mean Square Error, but the lowest Mean Absolute Percentage Error compared to the benchmark models. No model outperforms the others on all of the performance metrics.

In general, the Long Short-Term Memory model shows competitive results, but does not consistently outperform the benchmark models. It performs reasonably well for products such as Cold Rolled Sheets and Electrolytically Metal Coated Sheets, but faces challenges with others. Regarding the benchmark models, Ensemble 1 demonstrates robust performance. The un-

derlying models, Support Vector Regression and Random Forest, perform well on some products but not as effectively on others. It cannot be concluded that one of the models consistently performs better than the others.

6.1 Forecasting Without the COVID-19 Pandemic

The COVID-19 period had a significant impact on economic conditions and steel demand. Therefore, it is interesting to examine whether the model's performance differs when the COVID-19 period is excluded. In this section, I use data up to January 2020 to analyze the model's performance. The models are tested on the last 22 months, starting at April 2018. Data after January 2020 is completely discarded. Assessing the model's performance without the influence of the COVID-19 period provides helps to get a better understanding of how well the model can make predictions under less extreme economic conditions. If the model's performance significantly improves when the COVID-19 period is excluded, it suggests that the model is possibly sensitive to extreme events. Such weaknesses should be identified to improve forecasting and risk management approaches. Also, knowing how well the model generalizes to different economic conditions helps companies to plan for more typical business scenarios. While the COVID-19 period impacted the recent years significantly, it might have less influence in the upcoming years. If the model's work well under less extreme circumstances, they might still be useful for planning in the long-term.

I use the same feature selection as I have for including the COVID-19 period. This provides consistency in the analysis. It should be noted that the relevance of features may have shifted during the COVID-19 period. For example, economic and industrial performance indicators such as the number of cars produced and the purchasing manages' index might have shown unique patterns during the pandemic. While the features are not as pandemic-specific as for example the number of COVID-19 cases or hospitalization rates, their relevance in a scenario without COVID-19 may differ. New feature selection which does not rely on the extremes of the COVID-19 period, might lead to models that are better in capturing demand dynamics under more stable economic conditions. This could potentially lead to improved forecasting accuracy.

Furthermore, it should be noted that the size of the data set is even more limited when the COVID-19 period is excluded. By omitting the last 24 months, more than 10% of the data set is lost. This leads to a reduction

in training data, which potentially results in less accurate and robust models. Furthermore, the feature selection is biased towards the data set that includes the COVID-19 period. However, despite these potential biases and limitations, understanding how the models perform when the COVID-19 period is excluded is still interesting for developing forecasting strategies. If the same models consistently outperform the others, it is likely that these models that outperform are robust and suitable to use under different economic conditions. If the performance of a model widely varies between the data sets, it implies that the model is not so robust and does not adapt well to changing market dynamics. This should be taken into account in forecasting strategies for the long term.

First we look at Hot Rolland Narrow Strip, as predicted by the Long Short-Term Memory model, for the time period without COVID-19.

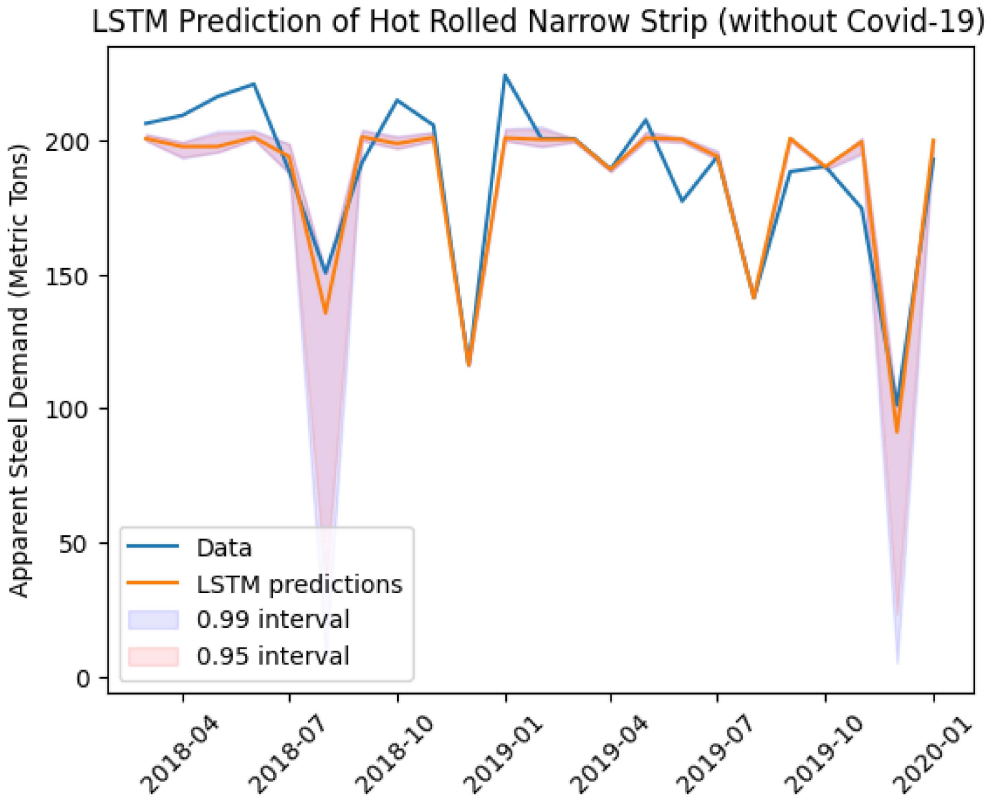

Figure 5: Predictions of Hot Rolled Narrow Strip using SVR, RFR and ENS1 without COVID-19

Compared to the model for the period including COVID-19, this model shows a better fit. We see that the predictions are quite close matching the data. Especially during the troughs, or anti-modes, in July and November 2018 as well as in August and December 2019, there is not much deviation between the predictions and the data. It almost seems like the model is overfitting the data during this periods. However, we do see around the first peak in July 2018 and the last peak in December 2019 that there is quite a large confidence interval. This makes it difficult for steel-making companies to fully rely on these predictions when it comes to the tactical or strategical plan. Furthermore, we see in the first half of the time frame, till April 2019, that the predictions tend to be lower than the data. It shows that the model has difficulty with predicting when it comes to steel demand above

200 metric tons. Surprisingly, we have also seen this for the model for the period including COVID-19. The deviations around June 2018 and January 2019 are more than 10% from the data. In the second half of the training data, the predictions tend to be higher than the data. Around April 2019 and November 2019 these deviations are around 20 metric tons compared to the data. In conclusion, this model seems to fit quite well but steel-making companies should be cautious due to these limitations.

We continue with Hot Rolled Narrow Strip, as predicted by Random Forest Regression, Support Vector Regression and Ensemble 1, for the time period without COVID-19.

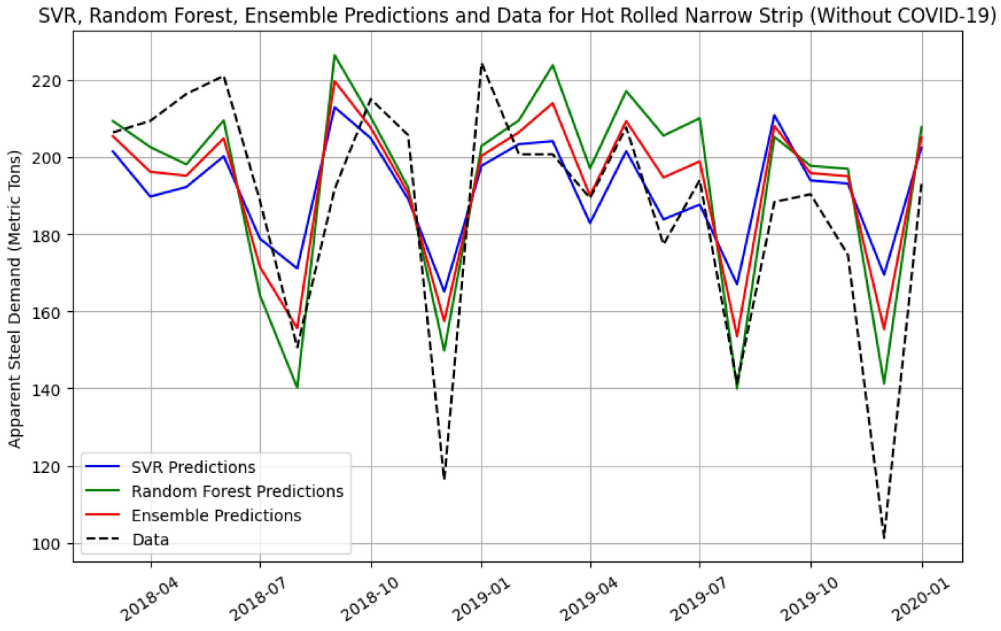

Figure 6: Predictions of Hot Rolled Narrow Strip using SVR, RFR and ENS1 without COVID-19

Figure 6 shows that the predictions of all three models follow follow a similar trend as the data. In the first few months, till the anti-mode in July 2018, the predictions of each model tend to be lower than the data. From August 2018 onward, the predictions of each model tend to be higher than the data for almost every month. It is difficult to say which of the models performs the best. During the first few months, from April 2018 to

June 2018, the Random Forest Regression model performs the best while the predictions of the Support Vector Regression model are deviating the most. The similar is true for the period from October 2019 to December 2019. On the other hand, between February 2019 and July 2019 the Support Vector Regression performs the best, while the prediction of the Random Forest Regression model are deviating the most. During other time periods in the test set, the best performing model seems to alternate more often. The Ensemble 1 model provides a prediction based on the average between the Random Forest Regression Model and Support Vector Regression model. In some months, where one of these two is over-predicting and the other is under-predicting the demand, the Ensemble 1 model yields the most accurate predictions. This is true in for example August 2018 and April 2019. Furthermore, due to averaging, the deviations from the data seem to be less extreme. Unfortunately, all three models have troubles with capturing the large anti-modes during December 2018 and December 2019. Even the Random Forest Regression model, which is the best predicting model for these months, differs in its predictions with 35 to 40 metric tons of steel from the data. This model does however provide accurate predictions for the other two smaller anti-modes. It shows deviation of only 10 metric tons during August 2018 and a deviation close to 0 metric tons during August 2019.

Random Forest Regression seems to predict more extreme values, often higher or lower than both Suport Vector Regression and Ensemble 1.

Table 6 provides an overview of the different performance metrics, Residual Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). These performance metrics correspond the Long Short-Term Memory model and benchmark models for each of the steel products, modelled without the COVID-19 period.

		RMSE	MAE	MAPE
Hot Rolled Narrow Strip				
	LSTM	24.17	18.68	11.02
	RFR	**19.43**	16.44	9.814
	SVR	23.11	17.49	11.29
	ENS1	19.91	**15.50**	**9.727**
Lengths Cut From Hot Rolled Wide				
	LSTM	39.81	28.97	15.49
	RFR	**33.97**	**27.19**	**12.59**
	SVR	40.97	36.13	17.73
	ENS1	44.86	38.61	18.91
Cold Rolled Sheets				
	LSTM	118.6	76.46	10.87
	RFR	**75.27**	63.78	**8.965**
	SVR	90.30	68.06	10.39
	ENS1	77.89	**63.16**	9.258
Hot Dipped Metal Coated Sheets				
	LSTM	**341.8**	**295.5**	**14.32**
	RFR	376.5	311.8	15.08
	SVR	414.3	337.5	15.92
	ENS1	390.0	322.5	15.33
Electrolytically Metal Coated Sheets				
	LSTM	**44.68**	**41.16**	**26.94**
	RFR	62.95	58.44	38.51
	SVR	56.72	53.63	36.50
	ENS1	59.22	56.04	37.50
Grain Non-Oriented				
	LSTM	16.85	12.09	13.54
	RFR	**12.32**	**9.409**	**10.33**
	SVR	14.71	11.84	14.36
	ENS1	13.05	10.01	11.57

Table 6: Results from Long Short-Term Memory (LSTM), Random Forest Regression (RFR), Support Vector Regression (SVR) and Ensemble 1 (ENS1) for each product type (data without COVID-19-19 period).

For Hot Rolled Narrow Strip it becomes clear that the Random Forest Regression and Ensemble 1 outperform both the Long Short-Term Mem-

ory model and Support Vector Regression model across all three performance metrics. Random Forest Regression achieves the lowest Residual Mean Squared Error, which is 19.43 metric tons of steel. Ensemble 1 achieves the lowest mean Absolute Error, which is 15.50 metric tons of steel and the lowest Mean Absolute Percentage Error, which is 9.727%. For Lengths Cut From Hot Rolled Wide, we see that Random Forest Regression outperforms all other three models on each of the performance metrics. Long Short-Term Memory is the second best performing model, outperforming both Support Vector Regression and Ensemble 1 on all performance metrics. There is a significant difference between the scores of the best and second best performing model. For instance, Random Forest Regression has a Residual Mean Squared Error of 33.97 metric tons of steel, while the Long Short-Term Memory model has a Residual Mean Squared Error of 39.81 metric tons. This indicates that Random Forest Regression is clearly the superior choice. For Cold Rolled Sheets we see that Random Forest Regression and Ensemble 1 are the best performing models. Random Forest Regression has the lowest Residual Mean Squared Error (75.27 metric tons of steel) and lowest Mean Absolute Percentage Error (8.965%). Ensemble 1 has the lowest Mean Absolute Error of 63.16 metric tons of steel. Surprisingly, the Long Short-Term Memory model is outperformed by all of the three other models on all performance metrics, indicating that it is the most inferior choice. It is remarkable that the Long Short-Term Memory model outperforms the other three models on all three performance metrics. This shows that the Long Short-Term Model performs relatively well under certain situations. Furthermore, we see that for Hot Dipped Met Coated Sheets, the second best model is Random Forest Regression, the third best model is Ensemble 1 and the fourth best model is Support Vector Regression. For Electrolytically Metal Coated Sheets this is the other way around. Support Vector Regression is the second best model, while Ensemble 1 is the third best model and Random Forest Regression is the fourth best model. This shows that the performances of Random Forest Regression and Support Vector Regression are highly sensitive to adjustments in the data. While one model performs the best on one set of data, the other model performs the best on an other set of data. Ensemble 1 shows a more robust performance, which might be an advantage even-though it is not the best performing one. For Grain Non-Oriented we see that Random Forest Regression is the best performing model. With a Residual Mean Squared Error of 12.32 metric tons of steel, a Mean Absolute Error of 9.409 metric tons and a Mean Absolute Percentage Error of 10.33%

it performs significantly better than the other three models. In general, evaluating the performance on all six products, it can be said that both Long Short-Term Memory and Random Forest Regression perform well compared to the other models. However, with a remark that the performance of the models is highly dependent on the underlying data. This makes it difficult to select one model with the best overall performance and shows that steelmaking companies should not rely on one single model. It should also be note that these errors are quite large compared to the average apparent steel demand of these products each month. For instance, the average demand for Hot Rolled Narrow Strip is 190,43 metric tons of steel. The Residual Mean Squared Error of the best performing model is 19.43 metric tons of steel, which is more than 10% of the demand. The models for the data set including COVID-19, perform even worse due to increased volatility and high fluctuations in demand. Although these models provide guidance for tactical and strategical plans, there is still need to search for better options.

7 Conclusion

In conclusion, this study explores the application of machine learning techniques to improve the accuracy of demand forecasting in the steel industry. Accurate demand forecasts help steel companies to optimize sales allocation in profitable market segments and regions. We focus on steel demand in the automotive industry in the EU 27 region. We specifically look at six different steel products that are mainly used in this industry; Hot Rolled Narrow Strip, Lenghts Cut From Hot Rolled Wide, Cold Rolled Sheets, Hot Dipped Metal Coated Sheets, Electrolytically Metal Coated Sheets and Grain Non-Oriented. We use market supply data from these products as most reliable indicator of apparent steel demand and argue that apparent steel demand provides the most valuable insights for tactical and strategic planning purposes.

To improve the demand forecasts, we include a variety of features. These features are selected based on discussion with industrial experts from Tata Steel Netherlands. These features are categorized in four categories, namely Economic Performance Indicators, Purchasing Managers' Indices, Confidence and Sentiment indicators, and lastly Price and Market Indicators.

Differencing techniques are used to make the data stationairy. To validate this, we perform the augmented Dickey-Fuller test and the Kwiatkowski-

Phillips-Schmidt-Shin test. Furthermore, Yeo-Johson power transformation with implicit standardization is used to account for differences in scale size and distribution functions among the data.

For forecasting, we use the Long Short-Term Memory model and compare its performance with other machine learning techniques, which served as benchmarks. These benchmarks are Random Forest, Support Vector Regression and an ensemble model based on the simple average of those two.

Different subsets of features are tested based on results from principal component analysis and recursive feature elimination with cross-validation. The overall score is determined by aggregating multiple performance metrics, including Root Mean Squared Error, Mean Absolute Error, and Mean Absolute Percentage Error. These metrics provide a comprehensive evaluation of the model's accuracy, precision, and reliability. For each product, the feature selection with the highest overall performance is included in the results.

The results for all six steel products show that the performance metrics are relatively close among the models. In general, the Long Short-Term Memory model shows competitive results, outperforming the Support Vector Regression, Random Forest Regression and Ensemble 1 on all performance metrics in the predictions for two of the six products. Long Short-Term Memory did not outperform the benchmark models on the other four products. Furthermore, it should be noted that for both the Long Short-Term Memory model and the benchmark models, performance was not consistent during the whole test period. This raises questions about the practical usefulness of these models and underscores the need to explore model improvements or alternatives.

Analyzing the performance on all six products excluding the COVID-19 period, shows that both Long Short-Term Memory and Random Forest Regression perform well compared to Support Vector Regression and Ensemble 1. In general, all models perform better when the COVID-19 period is excluded. However, even-though the data set excluding COVID-19 contains less fluctuations, the errors are relatively high compared to the average apparent steel demand of these products each month. Although these models provide guidance for tactical and strategical plans, there is still need to search for better options.

Lastly, the differences in modeling performance among the steel products indicate that the models are highly sensitive to changes in the data. This sensitivity is likely due to the limited data that is used for training. It highlights a disadvantage of using machine learning models when only limited,

low-frequency data is available. This underscores that steel-making companies should not rely on only one single model and that it might be of interest to maintain traditional modeling techniques as well.

8 Research Limitations and Further Recommendations

The results indicate that the Long Short-Term Memory model and benchmark models have room for improvements. To start with data preparation, exploring different transformation and scaling techniques might potentially result in improved model performance. For instant, considering Box-Cox power transformation instead of Yeo-Johson transformation and normalization or min-max scaling instead of standardization could be beneficial.

Furthermore, including extra features in the feature analysis and using alternative feature selection methods may improve the predictive ability of the features that were included in the prediction models. Additionally, I have now selected the features for each product that yielded the best overall performance in the models. Model improvement could be achieved by allowing each of the models to have their own feature selection.

Another option is to integrate external forecasts, such as GDP projections, which extend beyond the current time frame. This approach can improve the model's adaptability to changing economic conditions, resulting in a more dynamic and forward-looking modelling approach.

It might also be possible to improve model performance by including publications lags, accounting for data vintages and accounting for data revisions. Currently, the reliance on the most recent available data may be a limitation. Considering the time it takes for information to be released or revised can help to better capture the dynamics of the data, leading to improvements in the model's performance.

Moreover, it is possible to include features with quarterly or yearly data. Currently, the focus is on features that have monthly data available. Introducing features that are only available on a quarterly of yearly basis may provide new information, possibly improving the predictive capabilities of the model. However, when including such features, it is crucial to test different interpolation methods to align these features with the monthly time frame. Improper interpolation can introduce additional noise to the data, which can

negatively affect the forecasting accuracy.

A method to improve the current methods is to conduct a more comprehensive exploration of hyperparameter tuning. Grid Search can be expanded to consider a wider range of hyperparameter combinations. Alternatively, other methods can be used such as Random Search. This approach, implemented in the Scikit-learn library in Python, makes use of a different algorithm to select the best set of hyperparameters.

Furthermore, the analysis focuses on multivariate models with one target variable and multiple features. Spatial patterns among steel products or across different market regions could offer valuable insights. Currently, each steel product within the EU27 region is addressed individually. However, given the connections between these products, it could be interesting whether the relations between these products could contribute to improvements in modeling performance.

It is also recommended to explore more effective methods for constructing ensemble models. The benchmark model Ensemble 1 is based on only two underlying models and works on the basis of a simple average. Including more or different models and using other methods than simple average may lead to an improvement in the results.

When exploring other individual methods, it could be interesting to consider using N-Beats, Facebook Prophet and Neural Prophet. These methods have gained significant popularity in recent forecasting studies. It is worth noting that Python's Darts library offers implementations of these models and other interesting models such as Temporal Convolution Network and Fast Fourier Transform. This library might also be useful for more extensive data exploration and for testing model combinations.

As for model combinations, it might be interesting to explore combinations of Long Short-Term Memory models with traditional models. These traditional methods are less data-hungry and perform relatively well on small time series. Insights from studies combining Fuzzy Neural Networks with ARIMA could provide valuable ideas. Additionally, exploring combinations of Long Short-Term Memory models with other Neural Network architectures such as Convolutional Neural Networks might be of interest. Such combinations could bring together the strengths of multiple Neural Network models, possibly leading to an improvement in the predictive capabilities.

Lastly, it is advisable to look more into the handling of extreme events such as the COVID-19 pandemic. Given the exceptional nature of this period, there is a possibility that it might negatively impact predictions. It is

recommended to explore the use of rolling time windows to capture changes in the data over different time spans. It might also be interesting to evaluate the model performance after excluding the COVID-19 period from the forecasting analysis. This could provide valuable insights into the effects of such an extreme event on current predictions.

9 Acknowledgements

I want to thank Tata Steel Netherlands for the collaboration. I would like to thank specifically Ronald de Haan, director Markets & Pricing and Neelotpal Kundu from the Markets & Pricing team for the valuable discussions and insights into the steel industry. I also want to thank Guido Joosen, Senior Economist at Tata Steel Netherlands, for helping me with obtaining the data that is used in this thesis. Furthermore, I want to thank my supervisor Betrand Achou from the University of Groningen for the valuable meetings and detailed feedback that helped me to write this thesis.

References

Al Hajj Hassan, L., Mahmassani, H. S., & Chen, Y. (2020). Reinforcement learning framework for freight demand forecasting to support operational planning decisions. *Transportation Research Part E: Logistics and Transportation Review*, *137*, 101926. Retrieved from https://www.sciencedirect.com/science/article/pii/S1366554519315169 doi: https://doi.org/10.1016/j.tre.2020.101926

Alpaydin, E. (2021). *Machine Learning.* The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/13811.001.0001 doi: 10.7551/mitpress/13811.001.0001

Association, W. S. (2015, March). *Indirect trade in steel* (Tech. Rep.). Author. Retrieved from https://worldsteel.org/wp-content/uploads/Indirect-trade-in-steel-March-2015.pdf (Report on indirect trade of steel from 2000 to 2013)

Azadeh, A., Neshat, N., Mardan, E., & Saberi, M. (2012, 03). Optimization of steel demand forecasting with complex and uncertain economic inputs by an integrated neural network–fuzzy mathematical programming approach. *The International Journal of Advanced Manufacturing Technology*, *65*. doi: 10.1007/s00170-012-4221-1

Baroyan, A., Kravchenko, O., Prates, C., Vercammen, S., & Zeumer, B. (2023). *The resilience of steel: Navigating the crossroads.* Retrieved 2023-09-24, from https://www.mckinsey.com/industries/metals-and-mining/our-insights/the-resilience-of-steel-navigating-the-crossroads

Basu, S., & Michailidis, G. (2013, 11). Estimation in high-dimensional vector autoregressive models.

Chen, D., Clements, K., Roberts, E. J., & Weber, E. (1991). Forecasting steel demand in china. *Resources Policy*, *17*(3), 196-210. Retrieved from https://EconPapers.repec.org/RePEc:eee:jrpoli:v:17:y:1991:i:3:p:196-210

Chu, J., & Kong, M. (2018, 6). A comprehensive survey of steel demand forecasting methodologies and their practical application for the steel industry. *Asian Steel Watch*, *5*.

Crompton, P. (2000). Future trends in japanese steel consumption. *Resources Policy*, *26*(2), 103-114. Retrieved from https://www.sciencedirect

.com/science/article/pii/S0301420700000209 doi: https://doi.org/10.1016/S0301-4207(00)00020-9

Crowson, P. (2018). Intensity of use reexamined. *Mineral Economics*, *31*, 61–70. Retrieved from https://doi.org/10.1007/s13563-017-0113-z doi: 10.1007/s13563-017-0113-z

Dickmann, P. (2009). Die struktur von schlankem materialfluss mit lean production kanban und innovationen. In P. Dickmann (Ed.), *Schlanker materialfluss: mit lean production, kanban und innovationen* (pp. 1–2). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-540-79515-5_1 doi: 10.1007/978-3-540-79515-5_1

Eurofer. (2022). *Eurofer statistical definitions.* Retrieved 2023-09-24, from https://www.eurofer.eu/statistics/about-eurofer-statistics/eurofer-statistics-definitions/

Eurofer. (2023). *Economic and steel market outlook 2023-2024, third quarter.* Retrieved 2023-09-24, from https://www.eurofer.eu/publications/economic-market-outlook/economic-and-steel-market-outlook-2023-2024-third-quarter/

Hamilton, J. D. (1994). *Time series analysis.* Princeton: Princeton University Press. Retrieved 2023-10-01, from https://doi.org/10.1515/9780691218632 doi: doi:10.1515/9780691218632

Hopp, D. (2021). *Economic nowcasting with long short-term memory artificial neural networks (lstm).*

Kayacan, E., Ulutas, B., & Kaynak, O. (2010, 03). Grey system theory-based models in time series prediction. *Expert Systems with Applications*, *37*, 1784-1789. doi: 10.1016/j.eswa.2009.07.064

Kelleher, J. D. (2019). *Deep Learning.* The MIT Press. Retrieved from https://doi.org/10.7551/mitpress/11171.001.0001 doi: 10.7551/mitpress/11171.001.0001

Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization.*

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (1st ed.). Springer. Retrieved from https://doi.org/10.1007/978-1-4614-6849-3 (Published: 17 May 2013) doi: 10.1007/978-1-4614-6849-3

Kumar, V., & Kumar, R. (2023). Development of artificial neural network model for indian steel consumption forecast. *J. Inst. Eng. India Ser. D*. doi: 10.1007/s40033-023-00482-x

Kunapuli, G. (2023). *Ensemble methods for machine learning* (1st ed.). Shelter Island, NY: Manning Publications. (1 online resource (xx, 330

pages) : illustrations, charts)

Labson, S., Gooday, P., & Manson, A. (1994). Adoption of new steelmaking technologies. *Available at SSRN 2554879*.

Lee, H., Padmanabhan, V., & Whang, S. (2004, 12). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*, 546-558. doi: 10.1287/mnsc.1040.0266

Malenbaum, W. (1973). *Material requirements in the united states and abroad in the year 2000: A research project prepared for the national commission on materials policy.* Philadelphia: University of Pennsylvania.

Malenbaum, W. (1978). *World demand for raw materials in 1985 and 2000.* New York: McGraw-Hill.

Mashrur, A., Luo, W., Zaidi, N., & Robles-Kelly, A. (2020, 01). Machine learning for financial risk management: A survey. *IEEE Access*, *8*, 203203-203223. doi: 10.1109/ACCESS.2020.3036322

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning (second edition)* (Second ed.). Cambridge, Massachusetts: The MIT Press.

Oxford Learner's Dictionaries. (2023). *Machine learning.* `https://www.oxfordlearnersdictionaries.com/definition/english/machine-learning`. (Accessed: October 2, 2023)

Pei, F., & Tilton, J. E. (1999). Consumer preferences, technological change, and the short-run income elasticity of metal demand. *Resources Policy*, *25*(2), 87-109. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0301420799000136` doi: https://doi.org/10.1016/S0301-4207(99)00013-6

Peirelinck, T., Kazmi, H., Mbuwir, B. V., Hermans, C., Spiessens, F., Suykens, J., & Deconinck, G. (2022). Transfer learning in demand response: A review of algorithms for data-efficient modelling and control. *Energy and AI*, *7*, 100126. Retrieved from `https://www.sciencedirect.com/science/article/pii/S2666546821000732` doi: https://doi.org/10.1016/j.egyai.2021.100126

Polikar, R. (2006, 10). Polikar, r.: Ensemble based systems in decision making. ieee circuit syst. mag. 6, 21-45. *Circuits and Systems Magazine, IEEE*, *6*, 21 - 45. doi: 10.1109/MCAS.2006.1688199

Raju, S. M. T. U., Sarker, A., Das, A., Islam, M. M., Al-Rakhami, M. S., Al-Amri, A. M., ... Sarfraz, S. (2022, jan). An approach for demand

forecasting in steel industries using ensemble learning. *Complex., 2022*. Retrieved from https://doi.org/10.1155/2022/9928836 doi: 10.1155/2022/9928836

Reddy, Y. C. A. P., Viswanath, P., & Reddy, B. E. (2018). Semi-supervised learning: a brief review. *International journal of engineering and technology*, *7*, 81. Retrieved from https://api.semanticscholar.org/CorpusID:55044284

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second ed.). The MIT Press. Retrieved from http://incompleteideas.net/book/the-book-2nd.html

Torbat, S., Khashei, M., & Bijari, M. (2018, 06). A hybrid probabilistic fuzzy arima model for consumption forecasting in commodity markets. *Economic Analysis and Policy*, *58*. doi: 10.1016/j.eap.2017.12.003

Wang, J., & Chen, Y. (2023). *Introduction to transfer learning: Algorithms and practice*. Singapore: Springer. Retrieved from https://www.springer.com/gp/book/9789811975844 doi: 10.1007/978-981-19-7584-4

Wang, X., Li, C., Yi, C., Xu, X., Wang, J., & Zhang, Y. (2022). Ecoforecast: An interpretable data-driven approach for short-term macroeconomic forecasting using n-beats neural network. *Engineering Applications of Artificial Intelligence*, *114*, 105072. Retrieved from https://www.sciencedirect.com/science/article/pii/S0952197622002299 doi: https://doi.org/10.1016/j.engappai.2022.105072

Wu, Y., & Crompton, P. (2003, 02). Bayesian vector autoregression forecasts of chinese steel consumption. *Journal of Chinese Economic and Business Studies*, *1*, 205-219. doi: 10.1080/1476528032000066703E

Appendix

A Definitions of Steel Demand

Steel demand can be defined in multiple ways. I follow the definitions from the World Steel Association (worldsteel), which is the international industry association for the iron and steel sector. The definitions are shown in figure 7 (worldsteel, 2015).

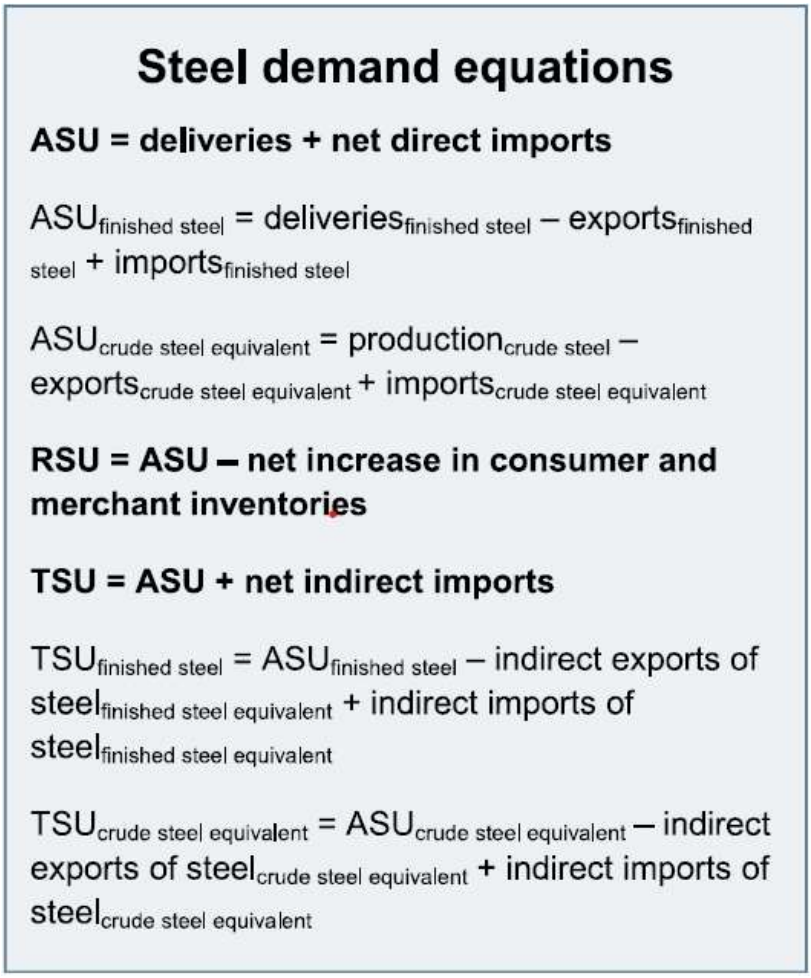

Figure 7: Definitions for apparent steel demand (ASU), real steel demand (RSU) and true steel demand (TSD), by worldsteel (2015)

worldsteel defines market supply as the total steel production in the market. Apparent steel demand, also referred to as apparent steel use, is calculated by subtracting the net export from the market supply. Net export is the export minus import of finished steel products to countries outside the primary market. Therefore, apparent steel demand represents the visible consumption of steel within the market. In this paper, the primary market is the European Union (EU27).

Real steel demand, also referred to as real steel use, is defined as the actual demand of steel as seen from the viewpoint of steel-making companies. It differs from apparent steel demand in the sense that it accounts for production process needs and fluctuations in inventory levels at direct customers. For example, if a car producer increases their steel use by emptying inventory, this change is reflected in the real steel demand but not in the apparent steel demand.

Finally, there is true steel demand, also referred to as true steel use. True steel demand differs from apparent steel demand in the sense that it is adjusted for indirect net exports. The indirect net exports are the net exports of steel products in steel containing goods such as vehicles and machinery equipment. True steel demand offers therefore a more precise reflection of the essential demand for steel in the market.

To be fully complete, it is worth mentioning that the European Steel Association EUROFER applies a more specific definition of steel demand than worldsteel (EUROFER, 2023). According to EUROFER, finished steel products do not include steel tubes and other products that undergo primary transformation from raw steel. Products that only undergo secondary processes, such as coating, are included as finished products. Due to this difference in definition, EUROFER and worldsteel report differently on demand and supply. Steel-making companies often handle multiple models following both the definitions of EUROFER and worldsteel.

B The Bullwhip Effect

Steel-making companies have difficulty to align production with demand. The reason being is that these companies sit in front of the production chains. The order intake of steel-making companies is therefore affected by demand and inventory levels of all other stakeholders in the production chain. Due to lack of information from stakeholders and their inventories, it is difficult to estimate how changes in market demand will lead to changes in incoming

orders. Information distortion in the production chain can lead to amplifications in demand, as shown in figure 8.

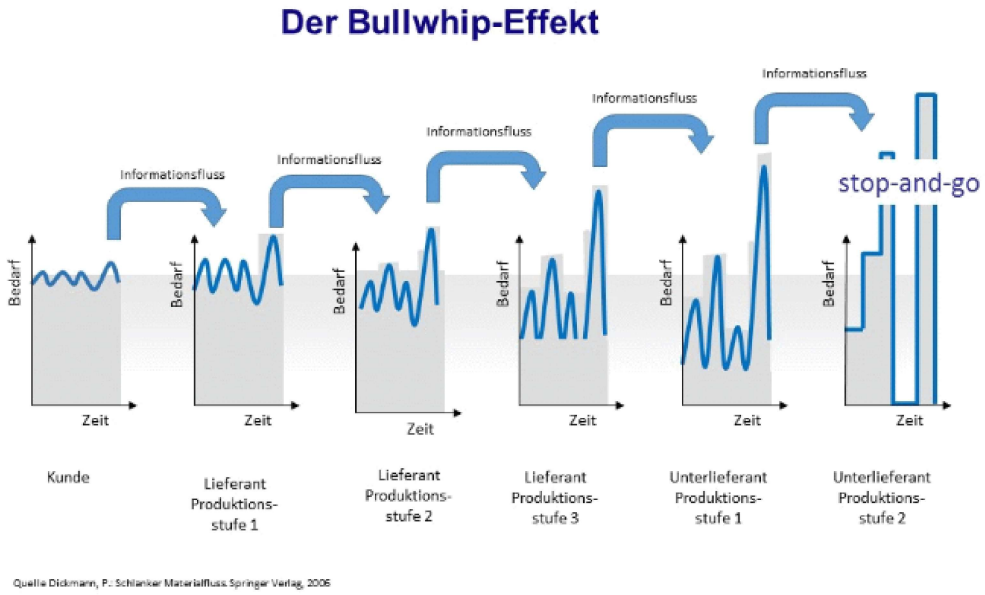

Figure 8: Visualisation of the Bullwhip Effect (Dickmann, 2009)

In this figure we see the need (Bedarf) against time (Zeit). This process begins on the left, starting with the customer (Kunde). The need of the customer is perceived as information by the supplier in the first production stage (Lieferant Produktionsstufe 1). This information influences the need of the supplier which in turn is perceived by the supplier in the second production stage. This process repeats itself till the sub-supplier in the last production stage (Unterlieferant Produktionsstufe 2). This stage is the beginning of the production chain. From customer all the way to the last sub-supplier, the variability in the observed demand continues to increase. In the last production stage, this leads to an inflow of orders that seem to fluctuate in a random fashion. This results in a stop-and-go reaction whereby one time the production stops due to large decrease in demand and the other time the production goes due to large increase in demand. The amplification of the variability in demand from beginning to end in the production chain is known as the bullwhip effect (Lee, Padmanabhan, & Whang, 2004).

Due to the bullwhip effect, steel-making companies observe high fluctuations in real and apparent demand as well as market prices. Since steel-making companies operate in a competitive industry, these prices are set by the market. Fluctuations in demand and prices have significant effect on order book decisions. When an order book is empty and low demand is expected in the upcoming time, it might be more attractive for steel-making companies to take on contracts with low profit margins than no contract at all. To prevent entering contracts with low profit margin, it is important to not only observe real demand, but also to get a better understanding of the apparent demand.

Steel-making companies use production equipment with high fixed costs. For example, iron production with blast furnaces. This is an ongoing process, 24 hours a day, 365 days a year. Furthermore, due to competitiveness in the market, profit margins are fairly limited. Due to high fixed costs and limited profit margins, it is important to have a continuous flow of orders and to fully use the production capacity. Hence, it is crucial to understand demand dynamics and to be prepared for fluctuations in demand.

C Basics of Machine Learning

Before diving into machine learning concepts, it is important to clarify what exactly is meant by machine learning. Following the definition of the Oxford Dictionary, machine learning is a form of artificial intelligence in which computers use extensive data sets to learn how to perform tasks, rather than being explicitly programmed to do so (Oxford Learner's Dictionaries, 2023).

Following the definition of Foundations of Machine Learning (Mohri et al., 2018), machine learning is the use of computational methods that use experience to execute specific tasks. Experience, in this context, is the information that is accessible to the 'learner'. This information usually comes in the form of electronic data. The learner refers to a method such as an algorithm or computational system that is trained on its capability to improve performance or to make accurate predictions.

In line with these definitions, machine learning models can be viewed as models with predictive performance that are trained on data. Contrary to traditional models such as linear regression, machine learning models are based on algorithms that autonomously learn and adapt from data. This autonomy makes machine learning techniques appealing for a range of applications, including demand forecasting.

C.1 Learning Scenarios

Machine learning models can be classified into four distinct categories, also known as learning scenarios. These learning scenarios are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning (Mashrur et al., 2020).

The key distinction between supervised and unsupervised machine learning lies in the data that is used. Data can be labeled or unlabeled. Labeled means that it is clear what the data represents. Supervised learning techniques use data with predefined labels to train models. Unsupervised learning techniques use unlabeled data to train models. Labeled data has predefined labels or categories, while unlabeled data lacks these predefined labels. An example of unlabeled data is photos that do not have any category assigned to it. Supervised learning is used for tasks such as classification or regression, whereby the goal is to predict specific outcomes. Unsupervised learning is used for tasks such as pattern recognition and data clustering, whereby the goals is to uncover patterns or structures in the data. Semi-supervised learning methods make it possible to train models on data sets that includes both labeled and unlabeled data. Semi-supervised learning can be used for classification and clustering (Reddy, Viswanath, & Reddy, 2018).

Reinforcement learning is a form of machine learning that is based on interaction with an environment. This learning scenario involves a model that interacts with an environment, receives feedback from that environment and makes accordingly decisions to achieve long-term goals (Sutton & Barto, 2018). Reinforcement learning can be used in games (training the computer as a chess opponent) and in autonomous vehicles (training self-driving cars), but also in demand forecasting. Hassan et al. (2020) construct a reinforcement learning framework to forecast freight demand. This framework combines reinforcement learning with traditional time series methods. The goal is for predictions to closely track fluctuations and time trends, while minimizing the need for user intervention.

C.2 Deep Learning, Transfer Learning and Ensemble Methods

Deep learning is a subfield of machine learning that is based on the use of neural networks. Neural networks can be applied to supervised, unsupervised and semi-supervised learning but also be combined with reinforcement learning. Neural networks are computational models that are designed to ex-

tract hierarchical features from data. They typically consist of an input layer that receives the data and an output layer that produces the predictions. So called hidden layers can be included between the input and output layer, to serve as intermediate representations of the data. Deep neural networks are neural networks that include one or more of such hidden layers. As noted by Kelleher (2019), neural networks excel at detecting non-linear patterns within the data and are particularly effective when handling large data sets. An example of a recent study on neural networks is provided by Wang et al. (2022). This study uses a specific type of neural networks, N-beats, to offer an interpretable data-driven approach for short-term macroeconomic forecasting.

Hybrid models combine models or machine learning techniques to improve modeling performance. Transfer learning and ensemble learning techniques are such techniques. Transfer learning improves model performance in a specific target domain by using knowledge that is acquired from a different domain (Wang et al., 2023). This techniques allows the use of 'pre-trained models', which are models that have already been trained on large data sets, to improve the learning capabilities of other models that rely on limited data. Transfer learning methods are often applied when the available data makes it challenging to fit sufficiently complex models (Peirelinck et al., 2022). This is valuable when data samples are highly complex or when of observed data is scarce.

The reason to address transfer learning is that the prediction of steel demand relies heavily on macro-economic and industrial indicators. Data on these indicators are often reported on a monthly or quarterly bases. This results in relatively small data sets that may not be suitable for machine learning methods that require large data sets. Transfer learning might therefore be interesting to address this issue.

Another popular method to improve model performance is ensemble learning. This approach combines models which are trained on the same data set, to improve the overall predictive performance. As discussed by Polikar (2006), ensemble models can improve the overall performance compared to individual models. While individuals models might perform well on the training data, their ability to adapt to new data might be weak. In that case the models do not generalize well under new circumstances. Combining multiple models and averaging their predictions, reduces the overall risk of selecting a model with poor generalization performance. Furthermore, due to their reliance on multiple models, ensemble models are less sensitive to specific

changes in the data and therefore less likely to produce outliers. This leads to a more robust performance.

Popular ensemble learning methods include bagging and boosting. Bagging, which is short for bootstrap aggregating, combines complex base estimators. These complex base estimators (strong learners) are similar models that are trained on different subsets of the data. The predictions of these models are aggregated to improve the overall predictive performance. Boosting on the other hand, is a technique that combines simple base estimators (weak learners), which are models that perform slightly better than random guessing. These models are trained sequentially and focus on data that the previous weak learner found difficult to predict. The combination of multiple weak learners provides a more robust model. Interested readers can find more on bagging and boosting in machine learning textbooks, such as those authored by Kunapuli (2023) and Alpaydin (2021).

Another ensemble method is stacking. This method is based on a two step approach with first-level models and a second-level model. The first-level models are trained independently of each other using the original training data. After the first-level models are trained, stacking creates a new data set based on the predictions made by these first-level models. These predictions serve as input features for the second-level model(Alpaydin, 2021). Learning from errors in the predictions of the first-level models can reduce bias and improve model performance of the second-level model. Therefore, this two-step approach could lead to more accurate and robust outcomes.