

Process Alarmistic Tool to Monitor Key Performance Indicators in E-commerce Operations

Inês da Costa Mariz

Master's Dissertation

Supervisor at FEUP: Prof. Luís Gonçalo Rodrigues Reis Figueira

Supervisor at Farfetch: Luís Ribeiro Ferreira

U. PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia e Gestão Industrial

2023-06-26

Abstract

The highly competitive e-commerce industry is rapidly growing, which brings challenges in handling not only logistics services, such as warehousing and shipping, but also payments processing and fraud detection. Businesses need to monitor key performance indicators (KPIs) to achieve operational excellence, but this increasing complexity leads to the need for better resource and time allocation to find and correct any issues. Tools have been developed to monitor operational KPIs and support decision-making, providing insights through, e.g., drill-down analysis, anomaly detection, target comparison, amongst others. However, there is still not a standardised approach for monitoring KPIs in the context of e-commerce marketplaces, which are websites that sell goods online from a variety of providers.

One of the main luxury fashion companies, Farfetch, operates an online marketplace and currently lacks a systematised way to prioritise which KPIs' behaviours should be paid attention to on a weekly basis. To create this standardised approach, a tool was developed which, through a Slack App, informs stakeholders about the critical KPIs they should focus on for the marketplace. It does so by providing insights based on three main criteria:

1. determining if a KPI has significantly deviated from its target;
2. if it has exhibited recent outlier behaviour;
3. if its actual has fallen within the prediction interval calculated by its corresponding forecasting model.

To determine the best forecasting model for each KPI, both exponential smoothing models and Prophet were tested with cross-validation. The selection was based on the Mean Absolute Scaled Error (MASE). The outputs of each of the criterion are red, yellow and green lights, which are categories differentiated by adjustable thresholds. By combining the results of the aforementioned criteria, a KPI's behaviour is classified as an "alarm" or an "attention", being that an "attention" is a less severe version of an "alarm". If the KPI worsened compared to last week, it is classified directly as either "bad alarm" or "bad attention". For good alerts, i.e., KPIs which improved and which were classified as "alarm" or "attention", a different division is made: they are categorised as "recovering" if they are still off target and "over-performer" if they are already on target. When a KPI is neither of these, if the forecasted value for the following week is significantly off target, the alert is "future attention".

The "MAD-Delta" approach is also introduced to detect which dimension groups are the best and worst contributors for each KPI as a whole, depending on whether it improved or worsened, respectively, in comparison to the previous week.

To evaluate the tool's accuracy, thirty scenarios which were categorised as different alert types were presented to two stakeholders, who independently classified them. The results show substantial agreement with the tool's classifications, highlighting its quality, whilst also having led to the implementation of improvements in the tool. However, variations in stakeholder perceptions underscored the challenges of creating a unanimous classification system.

Resumo

A altamente competitiva indústria do *e-commerce* está a crescer rapidamente, o que traz desafios no que respeita não só a serviços de logística, como *warehousing* e *shipping*, mas também ao processamento de pagamentos e à deteção de fraude. As empresas precisam de monitorizar os seus *Key Performance Indicators* (KPIs) para alcançarem excelência operacional, mas esta complexidade crescente leva à necessidade de uma melhor alocação de recursos e tempo para detetar e resolver quaisquer problemas. Na literatura, foram desenvolvidas ferramentas para monitorizar KPIs operacionais e apoiar a tomada de decisões, fornecendo informações através de, por exemplo, análises *drill-down*, deteção de anomalias, comparação com *target*, entre outros. No entanto, ainda não existe uma abordagem normalizada para monitorizar os KPIs no contexto dos *marketplaces* de *e-commerce*, que são *websites* que vendem bens de uma variedade de fornecedores.

Uma das principais empresas de moda de luxo, a Farfetch, opera um destes *marketplaces* e não dispõe de uma forma sistematizada de priorizar os comportamentos dos KPI, semanalmente. Para criar esta abordagem standardizada, foi desenvolvida uma ferramenta que, através de uma aplicação no Slack, informa os *stakeholders* sobre os KPIs críticos nos quais se devem concentrar para o *marketplace*. Esta fornece informações com base em três critérios principais:

1. determinar se um KPI se desviou significativamente do seu *target*;
2. se foi considerado um *outlier* face ao comportamento das semanas recentes;
3. se se enquadrou no intervalo de previsão calculado pelo modelo de previsão correspondente.

Para determinar o melhor modelo de previsão para cada KPI, modelos de *exponential smoothing* e o modelo *Prophet* foram testados através de *cross-validation*, sendo que a seleção se baseou no *Mean Absolute Scaled Error* (MASE). Os *outputs* de cada um dos critérios são luzes vermelhas, amarelas e verdes, que são categorias diferenciadas por *thresholds* ajustáveis. Combinando os resultados dos critérios anteriormente mencionados, o comportamento de um KPI é classificado como um "*alarm*" ou um "*attention*". Se o KPI piorar em relação à semana anterior, é classificado diretamente como "*bad alarm*" ou "*bad attention*". Para os bons alertas, ou seja, os KPIs que melhoraram e que foram classificados como "*alarm*" ou "*attention*", é feita uma divisão diferente: são classificados como "*recovering*" se ainda estiverem fora do *target* e como "*over-performer*" se já estiverem dentro do *target*. Quando um KPI não é nenhum destes, se o seu valor previsto para a semana seguinte estiver significativamente fora do *target*, o alerta é "*future attention*".

A abordagem "MAD-Delta" também é introduzida para detetar que grupos de dimensões são os melhores e piores contribuidores para cada KPI como um todo, dependendo de se este melhorou ou piorou, respetivamente, em comparação com a semana anterior.

Para avaliar a qualidade da ferramenta, trinta cenários que foram classificados como diferentes tipos de alerta foram apresentados a dois *stakeholders*, que os classificaram de forma independente. Os resultados mostram uma concordância substancial com as classificações da ferramenta, tendo também conduzido à implementação de melhorias nesta. No entanto, as discordâncias nas perceções dos *stakeholders* sublinharam os desafios da criação de um sistema de classificação de alertas unânime.

Acknowledgments

The last five years have been full of learning and self-discovery. I have finished this journey more confident than ever and there is a lot to be thankful for.

I thank my parents, my brother, my grandfather and my aunt for not only consistently dealing with my frustrations and often unfair moodiness, but for also supporting me through it and making everything a bit easier for me. Without your support, I would be nothing.

I also thank my friends, as I am lucky to say I have such great ones. Thank you for my Póvoa friends, for still staying in touch and supporting me after all the times we have followed different paths. Thank you for the friends that FEUP has given me and a special thank you for those who started this adventure with me, I could not have done this alone. You made every challenging, but enriching part of university a little bit better, and every good part of it unimaginably better. I look forward to seeing you follow your dreams. Finally, thank you to my Berlin friends. You were alongside me for one of the best experiences of my life and have made me grow immensely. How lucky am I to know a bit of my heart is spread throughout the world.

I thank FEUP for being the propeller of all these powerful memories, for always being my second home and for making me trust in my abilities. Thank you to all the Professors who have taught me, I really appreciate the clear effort and passion you put into teaching, as I have learnt much more than I could have ever imagined. A sincere thank you to Professora Vera Miguéis for all the clarification and advice regarding forecasting, my learning was truly supported by your readiness to help. A special thank you to my supervisor at FEUP, Professor Gonçalo Figueira, the constant encouragement and feedback was key for the accomplishment of this project.

Finally, I would like to thank Farfetch. I could not have asked for a more welcoming or better work environment. I did not know a workplace could feel this inviting, supporting and encouraging. There was never a question I did not ask as there were never dumb questions, which made me learn so much. I would like to highlight the support my team members gave me, as their feedback and openness were crucial for my project and for my overall well-being inside the company. Thank you, as well, to Ana Marques for the support and for enabling such an exciting and friendly environment in the team. Last, but certainly not least, I must thank the endless support given by my supervisor at Farfetch, Luís Ferreira. You always helped me to stay level-headed and confident about my work, while still encouraging me to go the extra mile, with more than enlightening feedback and constant availability. Thank you for all the effort you put into this and for always making me comfortable to discuss my thoughts.

"How one shouts into the woods, so it echoes back out."

German saying

Contents

Abbreviations	ix
1 Introduction	1
1.1 Farfetch	2
1.2 Project Framework, Motivation and Goals	3
1.3 Methodology	5
1.4 Structure of the Dissertation	6
2 Literature Review	7
2.1 Industrial Alarm Systems	7
2.2 Performance Measurement	8
2.3 Business Process Management	9
2.3.1 Business Activity Monitoring	9
2.3.2 Process Monitoring Models in the Literature	9
2.3.3 Process Monitoring Marketed Tools	10
2.3.4 Outcome-Oriented Predictive Process Monitoring	11
2.4 Monitoring Visualization Tools	12
2.5 Time Series Analysis	13
2.6 Outlier Detection	14
2.6.1 Types of Outliers	14
2.6.2 Techniques	15
2.6.3 Applied Methods	16
2.7 Forecasting Model Evaluation	18
2.7.1 Cross-Validation	19
2.7.2 Performance Metrics	20
3 Problem Description	21
3.1 Operations Workflow	21
3.1.1 Order Processing	21
3.1.2 Post Order Procedures	23
3.2 Key Performance Indicators	24
3.2.1 Supply Chain – Fulfilment KPIs	25
3.2.2 Supply Chain – Delivery KPIs	26
3.2.3 Fintech Operations – Payment KPIs	26
3.2.4 Fintech Operations – Fraud KPIs	27
3.3 Operations Monitoring	27
3.3.1 Weekly Operations Tactical Meeting	27
3.3.2 Order Fulfilment	28
3.3.3 Payment Processing & Fraud Prevention	28

3.3.4	Order Delivery & Return	29
3.4	Project's Challenge	30
4	Methodology	31
4.1	A Traffic Light Framework	31
4.1.1	Prediction Interval Criterion	32
4.1.2	Recent Observations Criterion	35
4.1.3	Target Criterion	36
4.1.4	Future Target Criterion	38
4.2	Dimension Analysis	38
4.2.1	Payment Metrics	39
4.2.2	Fraud Metrics	40
4.2.3	Delivery Metrics	40
4.2.4	Fulfilment Metrics	41
4.3	Alert Types	41
4.3.1	Alarm	41
4.3.2	Attention	42
4.3.3	Future Attention	42
4.4	Information Displayed for each Alert	42
4.5	Connection between BigQuery, R and Slack	43
4.6	Code Structure Overview	43
5	Results	44
5.1	Prediction Interval Criterion	44
5.1.1	Cross-Validation Results	44
5.1.2	Assumptions Assessment	45
5.2	Tool Assessment	45
5.2.1	Evaluation Process	45
5.2.2	Evaluation Results	47
5.2.3	Feedback Summary	50
5.3	Final Tool Results	50
6	Conclusion and Future Work	52
A	Decomposition Plots	60
B	Code-Overview	65
C	Cross-Validation Results	66
D	Prediction Intervals Assumptions Testing	73
D.1	Residuals' Histograms	73
D.2	Results of the Kolmogorov-Smirnov Normality Test for Residuals	78
D.3	Fitted Values vs. Residuals Plots	78
E	Tool Evaluation Results	83

Acronyms and Symbols

3PL	Third-Party Logistics
AWB	Air Waybill
B2B	Business to Business
B2C	Business to Consumer
BAM	Business Activity Monitoring
BaU	Business-as-Usual
BPM	Business Process Management
BU	Business Unit
EDDA	Estimated Delivery Dates Accuracy
FPS	Farfetch Platform Solutions
FRR	Fraud Refusal Rate
FxFF	Fulfilment by Farfetch
GTV	Gross Transaction Value
KPI	Key Performance Indicator
MAD	Median Absolute Deviation
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
NGG	New Guards Group
nonFxFF	non-Fulfilment by Farfetch
NS	No Stock
PACR	Payments Attempt Completion Rate
PCR	Payments Completion Rate
RTT	Returns Transit Time
SLA	Service Level Agreement
SoS1D	Speed of Sending of 1 Day
SoS2D	Speed of Sending of 2 Days
TITG	Time in Transit Global
TTNS	Time to No Stock
TTPR	Time to Process Returns
WI	Wrong Item

List of Figures

1.1	Operations Organizational Structure	3
1.2	Project's Gantt Chart	6
2.1	Cross-Validation with Rolling Origin in "Svetunkov, Ivan and Petropoulos, Fotios; Old dog, new tricks: a modelling view of simple moving averages; (Svetunkov and Petropoulos, 2018), 2018"	19
3.1	Operations Workflow	24
5.1	Alert Type Classification Overview	48
5.2	"Bad Alarm" Example	50
5.3	"Bad Attention" Example	51
5.4	"Recovering" Example	51
5.5	"Over-Performer" Example	51
5.6	"Future Attention" Example	51
A.1	Decomposition Plot of EDD	60
A.2	Decomposition Plot of FRR	60
A.3	Decomposition Plot of NS	61
A.4	Decomposition Plot of PACR	61
A.5	Decomposition Plot of PCR	61
A.6	Decomposition Plot of RTT	62
A.7	Decomposition Plot of SoS1D	62
A.8	Decomposition Plot of SoS2D	62
A.9	Decomposition Plot of TITG	63
A.10	Decomposition Plot of TTNS	63
A.11	Decomposition Plot of TTPR	63
A.12	Decomposition Plot of WI	64
B.1	Code Overview	65
D.1	Histogram of EDD's Residuals	73
D.2	Histogram of NS' Residuals	74
D.3	Histogram of PACR's Residuals	74
D.4	Histogram of PCR's Residuals	74
D.5	Histogram of RTT's Residuals	75
D.6	Histogram of SoS1D's Residuals	75
D.7	Histogram of SoS2D's Residuals	75
D.8	Histogram of TITG's Residuals	76

D.9 Histogram of TTNS' Residuals	76
D.10 Histogram of TTPR's Residuals	76
D.11 Histogram of WI's Residuals	77
D.12 Fitted Values vs. Residuals Plot for EDD	78
D.13 Fitted Values vs. Residuals Plot for NS	79
D.14 Fitted Values vs. Residuals Plot for PACR	79
D.15 Fitted Values vs. Residuals Plot for PCR	79
D.16 Fitted Values vs. Residuals Plot for RTT	80
D.17 Fitted Values vs. Residuals Plot for SoS1D	80
D.18 Fitted Values vs. Residuals Plot for SoS2D	80
D.19 Fitted Values vs. Residuals Plot for TITG	81
D.20 Fitted Values vs. Residuals Plot for TTNS	81
D.21 Fitted Values vs. Residuals Plot for TTPR	81
D.22 Fitted Values vs. Residuals Plot for WI	82

List of Tables

4.1	Data Points to Assist in the Definition of b	37
4.2	Rules for the Detection of Main Contributors in each Dimension	40
5.1	Best Model Results for each KPI	44
5.2	Tool Evaluation Case Distribution Overview	47
5.3	Tool Evaluation Case Distribution for each Alert Type	47
C.1	Cross-Validation Results for EDD	66
C.2	Cross-Validation Results for FRR	67
C.3	Cross-Validation Results for NS	67
C.4	Cross-Validation Results for PACR	68
C.5	Cross-Validation Results for PCR	68
C.6	Cross-Validation Results for RTT	69
C.7	Cross-Validation Results for SoS1D	69
C.8	Cross-Validation Results for SoS2D	70
C.9	Cross-Validation Results for TITG	70
C.10	Cross-Validation Results for TTNS	71
C.11	Cross-Validation Results for TTPR	71
C.12	Cross-Validation Results for WI	72
D.1	Results of the Kolmogorov-Smirnov Normality Test for Residuals	78
E.1	Tool Evaluation Results	84

Chapter 1

Introduction

The dimensions and complexity of e-commerce keep rising, especially since the Covid-19 pandemic. In 2023, e-commerce sales are expected to correspond to 20.8% of retail purchases and are predicted to grow 10.4%. For companies with business-models based on e-commerce, this means the macro-services which build the operations have to handle the increasing demand. Payment processing and fraud detection are both key services to ensure order fulfilment in this context. USD 41 billion was lost to e-commerce fraud in 2022 and USD 48 billion is anticipated to be lost in 2023 (Forbes, 2023). Moreover, the logistics services necessary to ensure the order fulfilment and delivery, such as warehousing, shipping and packaging are also key. In 2022, the global market of e-commerce logistics was valued at USD 315.82 billion and, from 2023 to 2030, it is projected to grow at a compound annual growth rate of 22.3% (Research, 2023).

These services need to be constantly monitored and improved, with the purpose of maximizing the value delivered to the customer. However, as e-commerce keeps expanding, not only is monitoring KPIs challenging, but so are the decisions regarding resource allocation when dealing with problems revealed by the poor results of the KPIs.

There have been tools developed to monitor operational KPIs, although most of these control tools refer to an industrial environment. The tools used for business monitoring are meant to empower the optimization of operations, to make informed decisions and to improve overall performance. They do this through drill-down analysis of KPIs, anomaly detection to identify outliers and their causes, target comparison and other analyses, which culminate in insights transmitted with proactive notifications.

Since e-commerce marketplaces started to arise in the 1990s and continue to grow exponentially (Nanehkaran, 2013), there is still not a standard methodology to monitor KPIs in this context. Furthermore, different marketplaces face different challenges in different sectors, such as in the grocery sector, where the items must arrive to the customer's residence in optimal conditions to be consumed (Erdmann and Ponzoa, 2021); in the vinyl sector, where the records must arrive free of any damages, including those caused by heat and mould (Kneese and Palm, 2020); or in the fashion sector, where there is a significant pressure on logistics due to the common returns (Hjort and Lantz, 2016). This thesis addresses the case of Farfetch, a luxury fashion marketplace.

1.1 Farfetch

Farfetch believes in empowering individuality and its goal is to become the premier global platform for luxury apparel by establishing connections among creators, curators and customers. It was founded in 2007 by José Neves and it is a luxury fashion e-retail platform that connects consumers with a premium selection of fashion items from boutiques and brands worldwide.

The company started in 2008 by operating solely as a marketplace, through which the company still charges a commission on each sale made through its platform. Farfetch sources its products from more than 1200 boutiques and over 3000 brands. Additionally, the company generates revenue through advertising, delivery fees and other ancillary services. Farfetch's platform provides several benefits to both buyers and sellers. To buyers, it offers a convenient and secure way to shop for luxury fashion items, with access to a wide range of products from around the world. To sellers, Farfetch provides a platform to reach a global audience, increase brand awareness and benefit from the company's marketing and fulfilment capabilities.

In its early years, it was able to follow a zero-stock policy, as products were sent directly from the boutique to the customer through third-party logistics (3PL) partners. Due to the company's quick expansion and the acquisition of some boutiques, it was necessary to keep certain inventory in privately owned warehouses to fill orders from these Farfetch-owned businesses. However, the dropshipping model continues to be the main method of order fulfilment.

In 2015, it became Farfetch Group, as it acquired Browns – a renowned British fashion and luxury goods boutique – and it started offering Farfetch Platform Solutions (FPS). FPS gives luxury brands and retailers access to Farfetch's e-commerce and technology capabilities, which can be deployed independently or combined. These include, among other services, the creation of their own website, the creative content production of items, in-store innovations like the Farfetch retail suite of technology products and Fulfilment by Farfetch (FxFF), which is a warehousing service where Farfetch manages partners' stock in its warehouses.

Both through Marketplace and FPS, the company operates as a platform, leveraging its existing scale and infrastructure to create new and innovative businesses. It operates like an ecosystem, connecting with its communities and collaborating with them to produce, exchange and co-create value within the platform.

An important distinction to make is between store and tenant. Stores are often also referred to as partners and they sell through the Marketplace, whilst tenants, termed also as clients, sell through their own website, which they acquired through FPS. Both partners and clients then sell to the final consumers. Furthermore, first-party (1P) partners and tenants are owned by Farfetch and third-party (3P) ones are not.

Farfetch has been experiencing constant growth and, in 2018, it became a publicly listed company. In recent years, it has acquired three major stakeholders: the sneaker and streetwear marketplace Stadium Goods in 2018, the brand platform New Guards Group (NGG) in 2019 and luxury beauty products retailer Violet Grey in 2022.

Farfetch serves its businesses with a pure service mindset and works to enable an amazing experience to tenants and their customers, by delivering world-class operations services while driving retention and efficiency. Its operations services are at the core of its platform and are a flying wheel that services the three business units:

- **Marketplace:** the business unit that manages all the business to consumer (B2C) channels, including Farfetch.com, Browns, Stadium Goods and Violet Grey;
- **Farfetch Platform Solutions:** the business to business (B2B) unit dedicated to offering end-to-end e-commerce solutions that meet the specific needs of Farfetch's clients;
- **New Guards Group:** the home of multiple international luxury brands, which adds a brand platform layer to Farfetch's technology, logistics and data platform.

Operations sits within its platform layer and provides many different services to its different businesses. This organizational structure, visible in Figure 1.1, allows the group to benefit from the synergies created by the platform areas, delivering more scalable and efficient services and contributing to its mission.

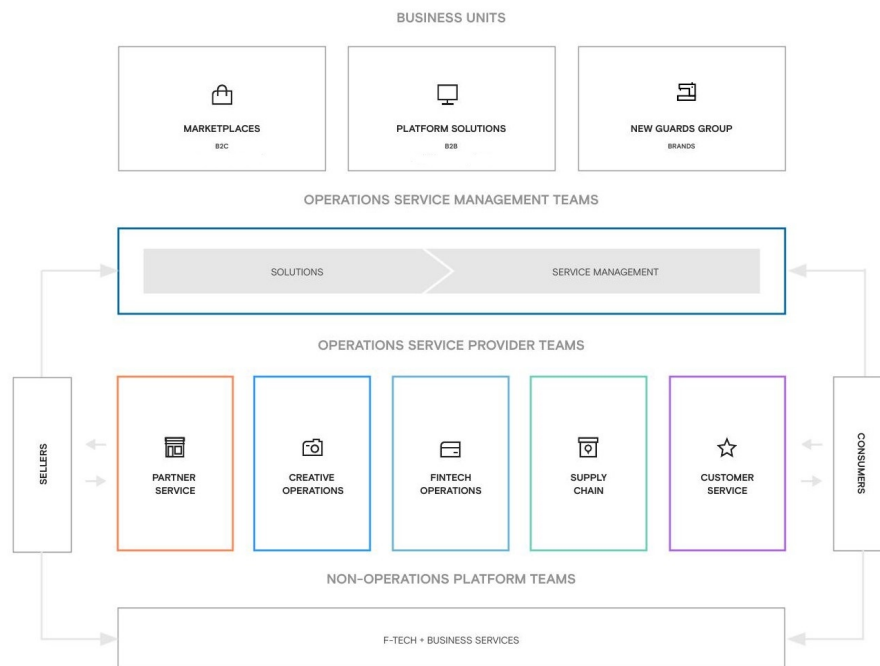


Figure 1.1: Operations Organizational Structure

1.2 Project Framework, Motivation and Goals

The project was integrated into the Operations Performance & Optimization team, which serves as an intermediary between the multiple Operations teams and Finance. The former provides the latter with support in planning and result optimization, taking into account both the client and

the partner perspectives. Additionally, it also supports the service managers regarding operational performance.

For the last few years, Farfetch has been going through a period of intense growth, due to the emergence of new business units and models, as FPS and NGG are both relatively recent, so some business-as-usual (BaU) processes are still being defined and optimized. Additionally, this growth is also due to the increasing trend in the number and size of tenants and partners, which impacts order volume and, therefore, pressures all areas regarding Operations. With the goal of facing the underlying difficulties, Farfetch has recently gone through internal restructuring, to optimize the synergies between each business unit (BU).

Operations can be divided in multiple ways. For example, an analysis on Operations can be done, on a higher-level position, for, e.g., Fulfilment by Farfetch (FxFF) and non Fulfilment by Farfetch (nonFxFF), or for Marketplace and FPS. However, further division can help to better understand how to improve performance.

The way Operations as a whole is monitored has also been recently changed in order to be aligned with the internal restructuring. The operational results regarding the previous week are presented and discussed in a tactical meeting with responsible employees from all the operations service provider teams, including some in higher-level positions. As the meeting is short and there are many topics to be discussed, there is a need for a tool that can not only analyse multiple operational areas and dimensions, but also convey which KPIs' incidents the stakeholders should focus on and with which priority. Since multiple operational areas are involved in the meeting, each one monitors their KPIs in a personalised manner and there is not a common methodology. Finding a fair fashion to do this that can be applied to any KPI is one of this project's goals. Moreover, although this tool would be especially useful in the weekly tactical meeting's context, the overview it would provide of the operations performance through systematised criteria would be practical for the Operations teams.

Hence, the specific objectives of this project are:

- To create a tool which uses established criteria to empower a systematised approach for the prioritisation of which KPIs' behaviours should be paid attention to, each week. This focused view can not only assist in increasing the productivity of the tactical meeting, but also lead to further discussions outside of this context;
- To predict which KPIs will be significantly off target in the following week, which expands the tool's approach from a *post-mortem* one to a more proactive one.

This helps raise a better awareness of operational performance throughout all the teams and throughout the higher-level stakeholders responsible for these areas, as a summary of the main possible topics is provided without any extra work on their part. All these objectives are regarding operational performance in the Marketplace context. However, in the future, this tool can be expanded to KPIs in other BUs.

1.3 Methodology

The Gantt chart shown in Figure 1.2 shows the steps that were defined for this project. There was an overlap between most of the stages, as some of them were related and to incorporate some buffers.

- **Business and Operations overview, understanding of the main tools and databases, problem comprehension:** getting familiar with both Farfetch and the project, regarding the monitoring processes throughout the Operations teams and the dynamics of the weekly tactical meeting;
- **State of the art study for outlier detection and business process monitoring:** understanding the research on both of these areas to incorporate this knowledge on the decisions ahead;
- **Conceptualisation of possible criteria and alert types, definition of KPI scope:** deciding which are the base criteria to set off alerts, the alert type classification and which KPIs would be included;
- **Definition of the forecasting models to evaluate and model testing for each KPI:** choosing the most appropriate forecasting models according to the problem's context and the research done in step 2; testing through cross-validation, in order to understand which is better for each KPI;
- **Decision on the approach for each of the criteria:** characterising the methodology for each criteria;
- **Development of the final tool:** continuously incorporating features in the tool as more of its design is being defined;
- **Definition of how to conduct dimension analysis and its scope:** delineating the dimension analysis approach and the most fitting dimension for each KPI;
- **Impact analysis, definition of future work:** inquiring not only about how relevant the alert types are, but also about the relevance of the information shown by the tool.

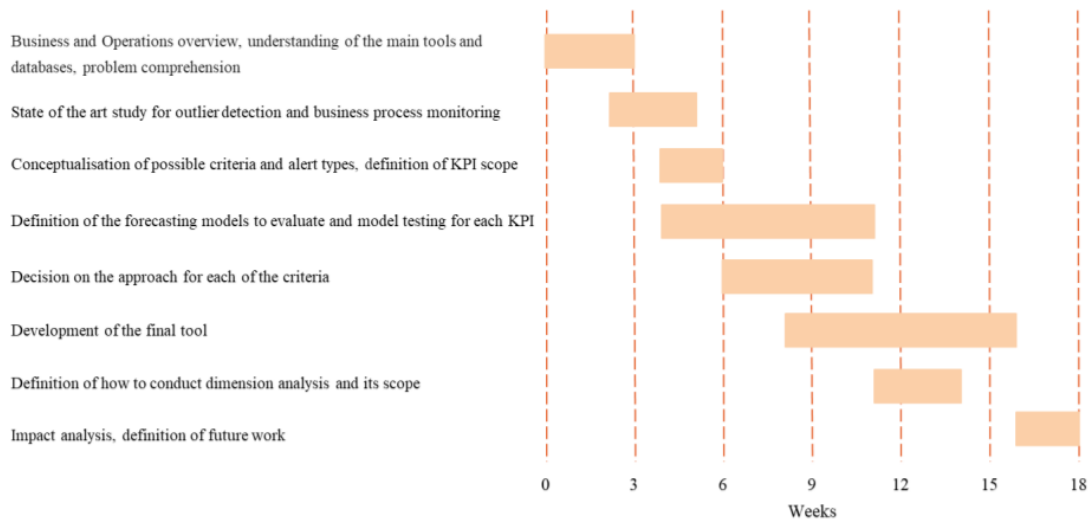


Figure 1.2: Project's Gantt Chart

1.4 Structure of the Dissertation

The dissertation is divided into five additional chapters.

Chapter 2 includes the literature review of the major themes regarding this thesis, such as performance measurement, business process management and outlier detection, also providing a background on time series forecasting and forecasting model evaluation.

The following chapter explains the context in which the problem is integrated in, which entails understanding the order processing workflow, the KPIs regarding fulfilment, delivery, payments and fraud, and how their monitoring is done as preparation for the weekly tactical meeting.

The Fourth Chapter describes how the tool was developed, namely in which criteria it is based on, how dimension analysis is performed and which alerts can be set off associated with which information.

Chapter 5 shows the results of the forecasting models, which are used for one of the tool's criteria. Moreover, the quality of the alert type classification is analysed, which led to some changes in the tool.

Finally, the Sixth Chapter brings together all the topics discussed in the preceding chapters, the key conclusions and potential future courses of action.

Chapter 2

Literature Review

The core of an enterprise's value production is represented by its business processes. Consequently, a complete management of business process performance offers a significant contribution to business success (Hammer and Champy, 2009). Business monitoring primarily works to spot changes or trends that point to opportunities or issues that call for managers to take preventative or corrective action (Nesamoney, 2004).

This chapter first provides a process monitoring overview, with the introduction of key concepts and examples of some tools which have been developed in the literature and some that have been commercialised. Furthermore, outlier detection methods and the time series forecasting models explored in this thesis are explained, along with some background theory.

2.1 Industrial Alarm Systems

Process monitoring is a key activity in industrial settings, both in discrete and process manufacturing. Discrete manufacturing processes create individual items in small batches, which are generally homogeneous and have low variability. Unexpected events can arise, such as safety risks, malfunctions and shortage of available agents, which are notified by signals and alerts (Chen and Jin, 2006). On the other hand, process manufacturing is a method of producing items in which components or ingredients are combined and then put through a series of procedures until the product is finished and in its final state. Manufacturers frequently use quality specifications, which include standards with acceptable tolerance limits and a list of the attributes to be checked, in order to control the process variability. Quality problems typically result from a failure of the input at some point along the batch manufacturing process rather than from a single defective machine or operator (Fisher et al., 2020).

There are important differences when transposing process monitoring in an industrial scenario, where the traditional alarm systems are placed in, to the context of e-commerce operations. When an alarm happens in the traditional context, an operator often intervenes immediately by detecting its cause and trying to take corrective actions (Izadi et al., 2009). Even though e-commerce operations monitoring can be done on a daily-basis, e-commerce companies with a business-model

based on a marketplace, in which suppliers sell their products directly to buyers (Hagiu and Wright, 2015), acquire proportions which make monitoring hard to accomplish live. Thus, a *post-mortem* approach is applied, which is not continuous, but done with a certain frequency instead, and is divided throughout the main macro-services. Moreover, in the industrial context there are clearly false alarms (Xu et al., 2011), whereas when monitoring operations in e-commerce, there is not a clear distinction between what the business wants to be alerted to, as it is not a binary truth. Finally, in industrial contexts, there is often an attempt to define a limit to the maximum number of alarms an operator should receive per hour or day, especially in discrete manufacturing, so they do not become too overwhelming (Wang et al., 2015), whereas in e-commerce, as the monitoring is done throughout multiple teams responsible for their own KPIs and not in a continuous way, there is not a need to establish a maximum number. All these differences are related to the fact that those interested in receiving the alarms in one scenario are operators, responsible for live monitoring variables which either depict the quality of the industrial process or are essential in process safety, whereas the ones in an e-commerce context are business analysts and people in higher management positions, reporting about the processes behind multiple macro-services.

2.2 Performance Measurement

Performance measurement is essential for developing a diagnosis, controlling and measuring results, developing strategies and disseminating them throughout the company (Wouters, 2009).

The most important issues that need to be addressed to evaluate corporate performance are both why and what should be measured, which cannot be judged separately (Lebas, 1995). To carry these tasks on, the concept of key performance indicator was developed.

A KPI is a quantifiable performance indicator — also referred to as "metric" in some literature — that represents a significant performance determinant for an organization's present and future success (Parmenter, 2015). It is assessed within a specific analysis period and it has a target value that should be attained or maintained during that time to reflect the accomplishment of set business objectives (Wetzstein et al., 2008). When KPIs are correctly established and executed, they provide insights on exactly where to intervene as a means to boost performance (Weber and Thomas, 2005). Considering these purposes, a KPI should be defined according to its attributes, links to other performance indicators and ties to other formalized ideas like roles, procedures and goals (Popova and Sharpanskykh, 2010). Moreover, for a metric to be effective it should be actionable, aligned with the company's strategy, owned, standardized and context driven, amongst other aspects (Eckerson, 2010). Thus, to stay competitive, a corporation must identify its pertinent indicators, how they relate to its set goals and how they depend on the activities carried out (Popova and Sharpanskykh, 2010). The measurement of process performance constitutes an integral component of Business Process Management (BPM), enabling organizations to systematically evaluate the effectiveness and efficiency of their processes in achieving desired outcomes and organizational objectives.

2.3 Business Process Management

Elzinga et al. (1995) defines BPM as a methodical, structured approach to process analysis, improvement, control and management with the goal of raising service and product quality.

There are multiple proposed BPM lifecycle models, but a crucial component present in all of them is the constant supervision of business objectives (de Morais et al., 2014). Business activity monitoring (BAM) technology often supports this by providing continuous, close to real-time monitoring of processes built on an event-driven architecture (Jeng et al., 2003).

2.3.1 Business Activity Monitoring

While BPM provides a comprehensive framework for process design, implementation and improvement, BAM complements these efforts by delivering real-time visibility and monitoring functionalities that enable performance measurement, bottleneck identification and proactive decision-making. BAM is used to increase the efficiency of business operations by monitoring them and promptly bringing issues to the responsible stakeholders' attention (Jiang et al., 2007). Monitoring and control have the purpose of finding and addressing not only the technical causes of anomalous business process behavior, but also the operational issues that arise when carrying out business processes (Zur Muehlen, 2001).

Both push (active monitoring) and pull (passive monitoring) technologies can be used for business-related monitoring (Zur Muehlen and Rosemann, 2000). BAM enables the close to real-time active monitoring of process instances and the computation of business process KPIs (Golfarelli et al., 2004). It uses strategies such as statistical process control, which calls for the establishment of an upper and lower bound for the KPIs. These bounds are obtained from historical runtime data or are explicitly entered into the system (Gillot, 2008). When these boundaries are breached, this type of BPM solution either alerts users or takes action to resolve the issue on its own. This method can be applied to management-by-exception, i.e., when an issue is handled as it arises, or to foresee and prevent unfavourable scenarios (Grigori et al., 2004). On the other hand, passive monitoring entails asking for data on active process instances, for instance, to verify the progress of a specific order.

Depending on the interest of the intended information recipient and the quality/quantity of the information provided, monitoring information is best absorbed when it is provided at different levels of detail. While some users may only be interested in high-level information, others may wish to dig deeper and get more specific data (Leymann et al., 2002). This highlights the need for a tool like the one developed for this thesis, as currently at Farfetch there is not an automated way of receiving summarised information with the key monitoring takeaways.

2.3.2 Process Monitoring Models in the Literature

In Wetzstein et al. (2008), a KPI monitoring model is developed in which only KPIs calculated based directly on runtime data of the respective processes are analysed. This means financial KPIs

are not considered. Moreover, permitted ranges of KPI values are defined by the business analysts, around the target value, to sound alarms to the proper KPI owners when its actual value falls outside these parameters. These can then be translated, for example, into an email or a message on the dashboards depicting the KPIs' behaviours. This tool entirely relies on the business analysts' ability to define proper ranges for each KPI, which can lead to ineffective alarms, as the reasoning behind the range definition may be faulty and not based on the data history. Moreover, the ranges being based on the target values may also be unsuccessful, as these targets may have been poorly defined.

Another example of a business monitoring model is the platform known as intelligent Business Operations Manager (iBOM). It makes it possible to manage and optimize business operations in an automated, intelligent and process-oriented manner in accordance with organizational objectives. iBOM offers an in-depth understanding of the business and of its operations. It also detects actual or expected anomalies at the business and IT levels, provides an explanation for these anomalies and suggests how to adjust some process elements to prevent insufficient executions. Metrics are specified through its Metric Definer, which enables the user to set some parameters for each KPI, like the values corresponding to green, yellow and red ranges. This system considers the range as acceptable, the yellow one as troublesome and the red one as worthy of being alerted. A requisite that business owners asked for in iBOM was for it to provide not only explanations, but also forecasts for the performance of each KPI, without the need for the tool users to have any data mining knowledge. Thus, provided historical data and certain parameters for each metric, the factor analysis and prediction engine uses data mining techniques to recognize patterns that can be explanatory of previous behaviour and predictors of future one. Through these predictions, which are always associated with a confidence interval, users can better understand root-causes of certain abnormal behaviours and act accordingly to attempt to reduce or eliminate the harm these predicted values with negative impact can potentially result in. The predictions are made through decision trees algorithms as they are easily interpretable by business users. Additionally, with these trees and always considering the confidence intervals, they gain a better understanding of if the KPIs will soon be out of range or violate any target commitments established in Service Level Agreements (SLAs) (Castellanos et al., 2005). Giving the user the responsibility of defining the values for green, yellow and red ranges may, once again, be risky and it may require frequent re-definition, as the reality keeps changing. If this tool is applied to multiple KPIs, the user may have difficulties in maintaining updated ranges and this feature may become obsolete. Additionally, neither of these tools explicitly distinguishes between beneficial and detrimental alerts.

2.3.3 Process Monitoring Marketed Tools

There are many Business Process Management products available on the market. Those which are worth mentioning in the context of this thesis include Business Activity Monitoring features.

IBM Business Process Manager provides comprehensive features for tracking and evaluating process metrics in real-time. When a KPI breaches a defined threshold, the tool generates alerts to promptly notify stakeholders about the issue. These alerts can be delivered through various

channels, including email, SMS or integration with external notification systems. Additionally, IBM Business Process Manager supports escalations paths and rules which can be defined to automatically escalate process instances or tasks to higher-level users when specific conditions or delays occur. The tool stores historical process data, enabling retrospective analysis and tracking of process performance over time. Reports and visualizations of historical trends help identify patterns, bottlenecks and areas for improvement. This data-driven analysis supports informed decision-making for process optimization (IBM, 2021).

Oracle BAM empowers organizations to effectively monitor and analyse KPIs in real time. Users can create personalized alerts that proactively notify them via email or SMS, allowing prompt actions to be taken. By delving into specific details, users can uncover valuable insights and identify the root causes of deviations. The system triggers alerts whenever a metric surpasses a threshold defined by the user, ensuring proactive management of issues and maintaining performance at desired levels. Oracle BAM provides customisable dashboards and reports that present real-time data and insights with visualizations. Reports can also be generated and shared to provide detailed analysis and historical trends for informed decision-making. Moreover, users can effortlessly slice and dice data using flexible querying and filtering options based on different dimensions or criteria (Oracle, 2023).

Both tools have a steep learning curve due to their extensive feature set and both require significant computing resources. Moreover, neither offers sophisticated forecasting capabilities, although Oracle BAM can provide insights into historical and current data.

DataGenie also presents BAM features, but these are greatly supported by AI. It continuously and autonomously monitors KPIs, recognises outliers and the factors that caused them and identifies other KPIs that were impacted, amongst other features. The user must define the metrics to be tracked, the dimensions they need to be tracked across, a population threshold below which they do not need to be tracked (e.g., countries with less than 100 orders per week do not need tracking), if within a dimension only specific values need to be tracked (e.g., not all the countries, but only a specific set), amongst a few other parameters. The users can easily ingest all the insights as they are presented in an easy-to-use StoryCard format, which can also be sent through internal communication platforms like Slack. Its outlier detection process consists of analysing deviations from the predictions calculated by their forecasting models. These models do not require any programming expertise from the user and can be adapted to any metric and granularity (DataGenie, 2023). Although it is built on a strong analytical component, it lacks a target comparison for each KPI.

2.3.4 Outcome-Oriented Predictive Process Monitoring

Besides the approaches already mentioned, there is also a group of methods known as outcome-oriented predictive process monitoring which concentrate on foretelling whether a case will have desired or undesirable repercussions. The predictions, which should be accurate, can be used by the user to decide whether to take action in an effort to avert these outcomes or mitigate their harmful effects.

In addition to making predictions, another group of methods entitled prescriptive process monitoring advise on when and how to intervene in a case that is already in progress to maximize a specific utility function (Teinemaa and Depaire, 2019). The contribution of Teinemaa et al. (2018) was to expand on these methods by adding an alarm-generating mechanism that informs the users when it is time to act on the prediction. The framework created is equipped with a parameterized cost model that captures the trade-off between the price of an intervention and the price of an unfavourable result. The decision to raise an alarm or not is then made by the alarming mechanism based on this cost model and the prediction made by a forecasting model. This can be challenging in some contexts where these prices cannot be straightforwardly determined.

A technology called the Business Process Cockpit, created by HP, enables users to build and track business KPIs on top of execution data for processes (Sayal et al., 2002). When combined with the methodologies described in Castellanos et al. (2004), the cockpit enables users to additionally have explanations about why process measurements have specific values and about the expected value of such metrics, in a close to automatic fashion, by interacting with the reporting engine's graphical user interface. The most appropriate data mining approach is chosen automatically considering the kind of problem and metric addressed. However, for these approaches to provide meaningful results the business must have a significant amount of good quality historical data.

There are multiple situations in which, through the help of predictions, important information can be retrieved, such as if, for example, there will be an abnormal deviation in a KPI, which could violate SLAs or which could lead to a deadline not being met. If this is noted in time, the system can alert an authority who could then take the necessary action and mitigate the impact of the exception (Castellanos et al., 2004).

2.4 Monitoring Visualization Tools

In most firms, dashboards are the primary tools for monitoring business performance and processes. By measuring current performance against a target intended to fulfil corporate goals, the inclusion of KPIs in dashboards promotes quick and precise information, which makes them also helpful for decision-making (Peral et al., 2017).

Both practical and visual-appealing elements should be present in a dashboard's design: functional elements include drill-down capabilities, presentation flexibility, scenario analysis or automated alerts, whereas cosmetic features should allude to how information is delivered to senior management and how data is visualized (Janes et al., 2013). Furthermore, the complexity of information, reflected, for example, in the number of decision cues, and accessibility of dashboards must be balanced, though, as they may have a detrimental impact on productivity and morale. Since there are different ways to display measurements and trends on a dashboard, such as tables and graphs, those designing them must also deal with the issue of presentation format. They should keep in mind that the link between the amount of information provided and choice accuracy is an inverted U-shaped curve, showing that accurate decisions can only be made when the

amount of information is optimal (Yigitbasioglu and Velcu, 2012). Due to this difficult balance and other reasons explained at the end of Chapter 4, they were not the chosen visualization tool for the project presented in this thesis, but they are, nevertheless, an important part of the KPI monitoring process in the company in which it was developed.

Effective data visualization aids in the comprehension of data relationships, which can be done through nominal comparison, time series evolution, ranking, part of a whole, deviation, distributions, and correlation (Archambault et al., 2015).

2.5 Time Series Analysis

A time series can be simply explained as a succession of observations taken sequentially in time. It is typically thought of as a sample realisation from an infinite population of time series created by a stochastic process, which can be stationary or non-stationary (Box et al., 2015). The fundamental premise of time series analysis is that some elements of the historical pattern will persist into the future. Thus, as it is the case in this thesis, it is frequently assumed that time series forecasting is based on historical data for the primary variable rather than explanatory variables that could have an impact on the variable being studied (Yu et al., 2014).

Through time series analysis, one can comprehend the reliance and structure of time series. As a result, the knowledge gained from time series analysis can be used for forecasting, process control, outlier detection, among other applications. Time series analysis is the investigation of a temporally distributed sequence of data or the creation of a prediction model where time is an independent variable (Box et al., 2015).

When analysing a time series, it is a common practice to decompose it into four components (Hyndman and Athanasopoulos, 2018):

1. **Trend:** reflects an increase or decrease in the data on the long-term, although it does not need to be linear;
2. **Seasonal:** present when the time series behaviour is influenced by fixed time periods, such as months or days of the week;
3. **Cyclic:** occurs when the time series displays rises and falls which do not follow a fixed frequency. These are often justified by economic conditions;
4. **Remainder:** represents the residue after the other components have been subtracted.

There are two types of decomposition: additive and multiplicative. Using an additive decomposition, a time series is modelled as:

$$y_t = T_t + S_t + R_t \quad (2.1)$$

Whereas with the multiplicative decomposition, it is modelled as:

$$y_t = T_t \times S_t \times R_t \quad (2.2)$$

Where y_t is the original time series, T_t is the trend and cyclic components, S_t is the seasonal component and R_t is the remainder component, at period t .

When multiplicative seasonality is present, unlike the additive one, there are significant changes to widths or heights of seasonal periods over time (Hyndman and Athanasopoulos, 2018).

When modelling business time series behaviour it is common, although not mandatory, to find a set of challenges, which often include (Taylor and Letham, 2017):

1. **Trend:** subtle and sometimes unpredictable changes in trend may occur, such as the impact of market changes or of the entry of new products, leading to piece-wise trends;
2. **Seasonality:** it is common for multiple seasonal effects to be present, such as weekly and annual, as they are representative of human behaviours;
3. **Outliers:** the presence of outliers is common throughout any type of time series, but especially expected in business related ones, as they are easily affected by calendar holidays and by their own promotional events. Each of these are expected to produce somewhat similar impacts each year, which makes them useful for building forecasts, but not intuitive to incorporate since some holidays are floating ones.

2.6 Outlier Detection

Although there is not a perfect characterization of an outlier, Hawkins (1980) has defined it as an observation that differs so greatly from others that it raises the possibility that it was created by a distinct mechanism.

2.6.1 Types of Outliers

There are multiple ways to categorise outliers, the following classification was created according to an analysis on the literature on time series made by Blázquez-García et al. (2021):

1. **Point outliers:** describes a situation when a point datum, compared to either all the remaining values in the time series or to its closest neighbours, exhibits anomalous behaviour at a given time instant. Point outliers can either affect one or more time-dependent variables, depending on whether they are univariate or multivariate;
2. **Subsequence outliers:** this term describes a series of time points that exhibit unexpected behaviour together, even if each observation on its own may not necessarily qualify as an outlier point. Additionally, subsequence outliers might be global or local and they might affect one or multiple time-dependent variables;
3. **Outlier time series:** when a full time series contains anomalies, but this is only possible when the input data is a multivariate time series.

The forecasting models used in this thesis are used for univariate business time series and the purpose of outlier detection in this context is to find univariate point outliers. Outlier time series are outside of the thesis' scope. Subsequent outliers analysis is not a necessary step, since the criteria used by the tool which are not based on forecasting models are responsible for finding the patterns the business wants to be notified of. In other words, if a KPI does present a constant improving pattern, Farfetch does not need to be alerted each week, whereas if it presents a constant worsening pattern, it will be alerted by the developed tool in the week in which it becomes significantly off target, and the future target criterion will likely even alert for this before it happens. The criteria in which the tool is based will be further explained in the Methodology chapter.

2.6.2 Techniques

Point outliers can be detected through multiple techniques (Blázquez-García et al., 2021):

1. **Model-based:** a point that greatly deviates from its predicted value is the most common and intuitive definition for the term "point outlier", thus a point, at time t , can be classified as an outlier in a univariate time series if its distance from its expected value exceeds a certain threshold. With this method, historical data are used to create a forecasting model that is then applied to predict future values. The primary challenge with outlier detection using prediction is establishing the value of the threshold that should be applied:
 - (a) The methodology falls under the category of estimation model-based methods if the anticipated value is determined utilizing prior and subsequent observations to the current value;
 - (b) In contrast, the methodology falls under prediction model-based methods if the anticipated value is determined only by prior observations made in relation to the current value.
2. **Density-based:** a point with less than τ neighbours - i.e., when less than τ objects are located within a distance R of those points – is an outlier;
3. **Histogramming:** this approach is based on finding the points from the univariate time series whose removal yields a histogram representation with less error than the original, even when the number of buckets is decreased to allow for the separate storage of these points.

A model-based approach was chosen for detecting point outliers, as this approach, being based on predicting values, is the most easily interpretable for the stakeholders who will receive the alerts given by the developed tool.

When choosing a suitable methodology for an outlier detection system, there are two primary factors to take into account (Hodge and Austin, 2004):

1. Adopting an algorithm that can effectively describe the data distribution and appropriately highlight outlying points. Additionally, the method needs to scale to the size of the data sets to be analysed;

2. Selecting an interesting neighbourhood for an outlier, which is not simple. Numerous algorithms create thresholds and build limits around normality as they process data. However, these methods frequently rely on user-specified criteria, such the number of clusters, or they enforce a particular distribution model.

2.6.3 Applied Methods

The reasons why each of the following methods were chosen will be explained in the Methodology chapter of this dissertation.

2.6.3.1 Naïve Method

The naïve method is typically used as a benchmark, given its simplicity. For naïve forecasts, all forecasts are simply set to be the value of the last observation. That is (Hyndman and Athanasopoulos, 2018):

$$\hat{y}_{T+h|T} = y_T \quad (2.3)$$

2.6.3.2 Exponential Smoothing

Some of the most effective forecasting techniques were motivated by Brown and Holt's work in the 1950s with their development of the exponential smoothing forecasting method. This method can be described as a weighted average of historical observations with weights that exponentially decrease as observations age. Depending on the value of the coefficient α , the weights decrease exponentially: the prior observations are completely disregarded if α equals 1 and the present observation is completely disregarded if it equals 0. This model's ability to create relatively accurate forecasts rapidly and for a variety of time series made it especially relevant for industry use (Hyndman and Athanasopoulos, 2018).

The single exponential smoothing model was the base for the development of two others Kalekar et al. (2004):

1. **Single**: the model assumes that there is no trend or regular pattern of growth and that the data varies around a relatively steady mean;
2. **Double (Holt)**: the only difference between it and simple smoothing is that trend must also be updated each period. The coefficient for the trend smoothing is specified by β ;
3. **Triple (Holt-Winters)**: it should be used when the data exhibits not only trend, but also seasonality. This leads to a need to provide a third parameter to account for seasonality, which can be either additive or multiplicative. The coefficient for the seasonal smoothing is specified by γ and the coefficients describes for the previous models are also used.

For this thesis, it is important to understand how **prediction intervals** are calculated for single, double and triple exponential smoothing models. They are based on the residuals' estimated standard deviation, while assuming a normal distribution with constant variance for the residuals. For double exponential smoothing, the prediction intervals calculation not only considers the estimated error distribution, but also the estimated standard deviations of the trend component. Regarding triple exponential smoothing, the calculation considers, in addition, the estimated standard deviation of the seasonal component. The prediction intervals incorporate all these error components to reflect the uncertainty in future data (Chatfield, 2001).

2.6.3.3 Prophet

Prophet is a time series forecasting model built to handle the aforementioned typical characteristics of business time series. It was developed by Facebook's core Data Science team, having been released in 2017 by Sean J. Taylor and Benjamin Letham. This decomposable model has four primary elements: trend, seasonality, holidays and error, reflected through the following equation:

$$y_t = g_t + s_t + h_t + \varepsilon_t \quad (2.4)$$

The trend function g_t models non-periodic changes in the time series' values, s_t models periodic changes and h_t models the effects of holidays. Any idiosyncratic changes that the model cannot account for are represented by the error term ε_t .

Similar to Holt-Winters, Prophet can also take on additive or multiplicative seasonality, such that a log transform can be used to achieve the latter seasonality, where the seasonal effect is a factor that multiplies g_t .

For each time series, the definition of the main elements that configure the Prophet model is the following (Taylor and Letham, 2017):

- Two **trend models** can be applied by Prophet according to the type of forecasting problem: when saturating growth is present, a logistic saturating growth model is used, whereas when it is not then a piece-wise linear model with changepoints is employed, since a piece-wise constant growth rate allows for a frequently useful and modestly resource-intensive model. The trajectories of real-time data frequently exhibit rapid variations. By default, Prophet will detect these changepoints and adjust the trend accordingly, but they can also be provided by the analyst based on established dates of product releases and other growth-altering instances;
- Regarding **seasonality**, standard Fourier series are used to capture multi-period seasonal effects;
- With respect to **holidays and events**, even though they can be viewed as foreseeable disruptions that affect several business time series, their effects cannot adequately be represented

by a smooth cycle, as they mostly do not behave in a periodic fashion. To tackle this issue, one of Prophet's possible inputs is a list of past and future occurrences, identifiable by the unique name of the events or holidays, which can either be worldwide or countrywide. Since the impacts of these events are not exclusive of the day of, it is common to encompass the effect of the surrounding days.

There are various opportunities in the Prophet model formulation in which users can adapt it according to their experience and external knowledge without having to comprehend the underlying statistics, such as capacities, changepoints, smoothing parameters, holidays and seasonality. This type of modelling, which was named **Analyst-in-the-Loop Modelling**, allows analysts to add their discretion using a small number of model parameters, while also keeping the flexibility to, when needed, rely on completely automated statistical forecasting. For example, the τ parameter was incorporated to positively or negatively adjust the trend flexibility and the σ to tune the magnitude of the seasonality component (Taylor and Letham, 2017).

When generating **prediction intervals**, Prophet follows a simulation-based approach, i.e., it first creates multiple future trajectories by considering the uncertainty in the trend and seasonality estimates and then it samples from the posterior distribution of the model parameters to generate multiple future trends and seasonality patterns. Once the trend and seasonality trajectories are obtained, Prophet combines them to compute prediction intervals. Prophet summarizes the range of the simulated future values at each time point to determine the upper and lower bounds of the intervals.

2.6.3.4 Median Average Deviation

The median is a measure of central tendency - a characteristic it shares with the mean -, but it has the advantage of being very insensitive to the existence of outliers. The median is the location estimator with the highest breakdown point, with it being 0.5. Furthermore, the MAD is completely unaffected by sample size. Huber Huber (2011) describes it as the single most helpful accessory estimate of scale, because of these two qualities.

To compute the median (M), the observations need to be ordered in an ascending fashion to find the mean rank of the statistical series and the value associated with that rank. Afterwards, all the absolute deviations from the median must be calculated and then the MAD is simply the median of those (Leys et al., 2013). An observation is categorised as an outlier if it is outside of an area defined by a lower and upper intervals, which are calculated by the median subtracted of or added to a multiple of the MAD. Specific examples are provided in Chapter 4.

2.7 Forecasting Model Evaluation

Model evaluation is a crucial step given the large array of available forecasting models and corresponding parameters. It does not only contribute to choosing the optimum model and parameter

setup, but it also shows exactly how accurate and precise it is. Multiple performance error metrics that can be used to assess them, as well as methodologies for their validation are following detailed.

2.7.1 Cross-Validation

When testing multiple forecasting models to determine, for each time series, which one better fits their behaviour, cross-validation is better than a simple train-test split. However, cross-validation in time series has more particularities than standard cross-validation, as the order of the observations is essential for the data to stay meaningful, which makes the random selection of data for each train and test set unreliable. Therefore, in this thesis, the rolling origin technique was applied.

The following figure, extracted from Svetunkov and Petropoulos (2018), depicts the logic behind this method. Initially, the first train set – depicted by the white area in the first row – is chosen, which, for this thesis, was approximately 60% of the total data. After the forecasted and the residual values for the first test and train set are calculated and stored, the data from a single time period t is added to the train set, whilst the size of the test set, coloured in grey, remains constant each time, as it simply moves one time period t ahead, as well. This is repeated until the three-steps-ahead projections correspond to the last three available data points. The number of times this process is done corresponds to the total number of origins. The number of origins used for each KPI depended on how far back its data history went: if it started in 2018 then more origins were implemented than if it started in 2021.

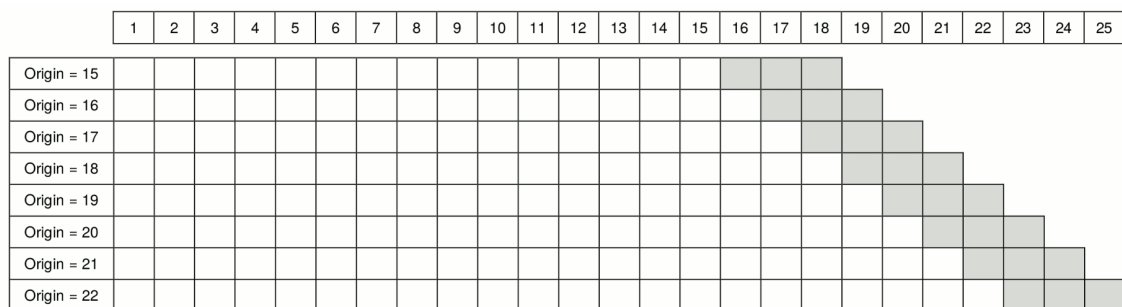


Figure 2.1: Cross-Validation with Rolling Origin in “Svetunkov, Ivan and Petropoulos, Fotios; Old dog, new tricks: a modelling view of simple moving averages; (Svetunkov and Petropoulos, 2018), 2018”

The method is called rolling origin as the start of the test set keeps moving forward as more train and test splits are done. This way, data are not randomly split, losing their chronological order, and it simulates the reality of how the tool designed in this thesis will forecast. In other words, the tool uses all available data, except the week being forecasted, to train the data and it predicts the value for the week being analysed, repeating this process each week. In the example represented in the figure, the forecast corresponds to the following three weeks, which, for future explanatory purposes, can also be mentioned as forecasts for $h=1$, $h=2$ and $h=3$. As the number

of weeks simultaneously forecasted increases, the harder it is to achieve accurate results, i.e., for $h=3$ the model performance metrics calculated will probably be worse than for $h=1$.

2.7.2 Performance Metrics

The purpose of calculating performance metrics with the results of cross-validation is to determine the best forecasting model. Performance metrics evaluate the quality of the model and are based on the errors obtained in cross-validation for each model. There is not an all-encompassing rule for which performance metrics are better to make this judgement, as all of them have advantages and disadvantages, which means the selection of the most appropriate ones depends on each forecasting problem.

There are the different types of performance metrics (Hyndman and Koehler, 2006):

- **Scale-dependent error metrics:** the scale of this metrics is determined by the data's scale. Thus, they are helpful when contrasting various models used on the same set of data, but they should not be applied, for instance, when evaluating data sets with various scales. Some examples are Root Mean Square Error (RMSE) and Mean Absolute Error (MAE);
- **Percentage error metrics:** they can be used to compare forecast accuracy across various data sets, which means they have the advantage of being scale-independent. Some of the most typical measures are Root Mean Square Percentage Error (RMSPE) and Mean Absolute Percentage Error (MAPE);
- **Relative error metrics:** these, unlike the percentage error metrics, are scaled by splitting each error by the error acquired using a different standard method of predicting, which makes them have a statistical distribution with undefined mean and infinite variance. Moreover, their calculation is only possible when there are multiple forecasts on the same series, i.e., they are not adequate for a single forecast horizon. An example is the Mean Relative Absolute Error (MRAE);
- **Scale-free error metrics:** these were created to handle the problems with using relative error metrics. The error is scaled based on the in-sample MAE from a naïve forecasting method. The mean absolute scaled error (MASE) is an example of this kind of metrics.

Chapter 3

Problem Description

This chapter provides a comprehensive understanding of Farfetch's operations, the order processing and the post order procedures, as the KPIs analysed in this thesis are related to these. The order processing steps are stock validation, fraud check, product packaging, shipping preparation, parcel pickup and delivery, whereas the post order possible steps are product return and payment refund. It is key to comprehend all of these steps, in order to understand the KPIs which measure their success. Then, for each KPI, the monitoring routines which occur as a preparation for the weekly tactical meeting are described.

3.1 Operations Workflow

The following processes concern orders placed in Farfetch's Marketplace, as it is the focus of this thesis. Other marketplaces like Browns or other BUs like NGG have different needs and capabilities and, therefore, slightly different processes.

3.1.1 Order Processing

Each step during the order processing only starts once the previous one is finished, except for the first two steps which take place at the same time and independently of each other.

Step 1 - Stock Validation: After a successful checkout and order allocation, the order starts its journey. Partners check if they physically have the product requested online by the customer in stock. If they do not have it, Farfetch checks if an alternative partner is able to fulfil the boutique order. In case this is not a solution, the order must be cancelled and the customer refunded.

Step 2 - Fraud Check: The payment process is analysed to evaluate the risk associated with the order and then accept or refuse the payment. To do this, Farfetch relies on external service providers to support it in detecting fraudulent behaviour. This first analysis is done automatically and can have three possible outcomes:

- Order accepted: when it shows non-suspicious behaviour;
- Manual revision: when it shows suspicious behaviour;
- Order cancelled: when it shows very suspicious behaviour. The Farfetch fraud team evaluates the legitimacy of the customer, assessing the risk of accepting the order or not.

Step 3 - Product Packaging: The partner starts by selecting the packaging option in which the items will be shipped to the customer, out of the thirteen box sizes and three envelope sizes. It makes this decision knowing Farfetch's recommendation, which is made according to the order dimension, the best customer experience and delivery costs. When a partner does not accept this suggestion, it creates an exception and it needs to justify the reason. The packaging supply chain is composed of many suppliers worldwide, which deliver the boxes to a warehouse managed by 3PL. The warehouse is responsible for managing partners' packaging requests. Some products can also be shipped directly by one of the FxFF warehouses or stock points to optimize shipping costs and customer experience by shortening delivery times.

Step 4 - Shipping Preparation: the products are packed and have all the necessary information to generate the shipping labels. Partners order the creation of the shipping label in the platform via the integration with the chosen carrier for the order, which should be done almost instantaneously. Farfetch then generates the Air Waybill (AWB) with the details of the carrier and the address of the customer. When shipping problems happen, like an error in the shipping details or a legal restriction, the Delivery Support team is automatically notified and works to solve them.

Step 5 - Parcel Pickup: the boutiques prepare the order to be picked up by the carrier and the order becomes ready to send. The partner prints all the documents, inserts them inside and outside of the box, closes the box and updates the order status to let the carrier know when to pick up the package. If the partner faces any problem with the carrier or the shipping, it can create an exception to avoid being penalized for their performance regarding speed of sending in the operational excellence program in place. This program is an incentives and penalties scheme, which considers some KPIs, including this speed, which is designed to help partners achieve excellent service levels to positively impact the customer experience, while also improving profitability and sales.

Step 6 - Delivery: this step starts when the package is picked up at partner location by the carrier and it is on its way to its final customer destination. This process is managed by Farfetch, partnering with more than fifteen carriers globally, like DHL, UPS and FedEx. Farfetch combines multiple delivery capabilities to ensure all orders arrive at their final destination at the estimated delivery window. The most common shipping options are standard and express delivery, having each different estimated shipping times.

3.1.2 Post Order Procedures

3.1.2.1 Product Return

After receiving an order, customers evaluate if they want to keep the product or return it for free, making sure the returned items arrive at their origin within fourteen days after order delivery. Some of the reasons that may make them want to return a product are:

- They received a wrong item;
- The item was not as described, for example, colour-wise;
- They changed their minds;
- There was a fitting problem;
- The ingredients, for beauty products, were not as expected.

They can start a return automatically through the website or they can request it through the customer service team. After scheduling the pickup, the carrier collects the parcel and returns it back to the partner.

3.1.2.2 Payment Refund

After receiving the returned product, the partner has two days to analyse the returned product and accept or contest the return, otherwise it will be accepted automatically.

If the partner accepts the return, the customer receives the refund at the partner's cost and the return process is closed. In case the partner contests the return, the partner service team analyses the situation and negotiates with it, which can lead to three outcomes:

- Partner and Farfetch do not accept the return: no refund is processed and the item is shipped back to the customer;
- Partner does not accept the return: customer is refunded at Farfetch's cost and the product is sent to the Farfetch Boutique to be recovered and resold;
- Partner accepts the return: the customer is refunded at the partner's cost.

The operations workflow is summarised in Figure 3.1.

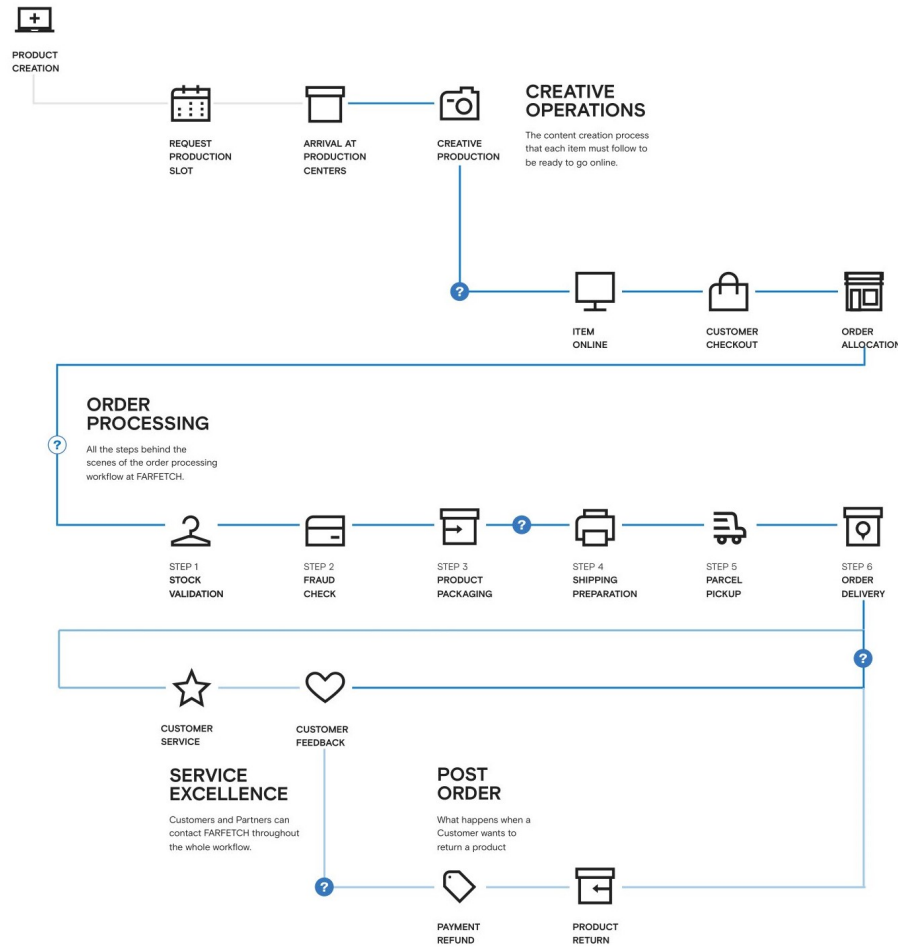


Figure 3.1: Operations Workflow

3.2 Key Performance Indicators

The KPIs chosen to be analysed regard Operations. Due to the complexity of this domain, it is divided into building blocks, which are:

- **Service Excellence**, including all service areas, from service to the sellers on the platform, service to the final customers of the platform (customer service) and service to the platform customers (platform services);
- **Supply Chain**, including delivery and fulfilment;
- **Content Creation within Creative Operations**, which is responsible for all the imagery and product written information necessary for an item to go online. This includes styling, photographing, photo editing and producing merchandise descriptions and measurement information;
- **Fintech Operations**, which includes payments and loss prevention, including fraud.

To monitor the partners' performance regarding the multiple key Operations areas, KPIs suitable for this purpose have been created by Farfetch. The ones studied in this thesis regard Supply Chain, both the delivery and the fulfilment areas, and Fintech Operations.

3.2.1 Supply Chain – Fulfilment KPIs

1. **No Stock (NS) [% Qty]**: Ratio between items in an order that were identified as not having stock divided by all the items that were processed. It includes the cases in which having no stock happens due to Farfetch's fault. This metric is reflective of the negative impact of a no stock situation on the customer experience, as it often leads to a lost sale, which has led Farfetch to be extremely demanding on this KPI's performance;
2. **Time to No Stock (TTNS) [days]**: Days elapsed between the order creation and the order "no stock" evaluation, for no stock net orders;
3. **Speed of Sending (SoS) [days]**: Measures the time between the moment the order was placed and the moment when it was picked up by the courier and it only includes the steps that are of the partner's responsibility. Thus, it represents the time between the creation of the order and the scan of the package at the end of step 5, with the deduction of the time spent in all steps that are not controlled by the boutique (steps 2 and 4), the time spent on weekends and holidays and the time where the order may be held by factors external to the partner. The following are the steps included in between those moments:
 - **Step 1 - Stock Validation**: time between the order creation and the stock confirmation, spent verifying the available stock, excluding weekends and bank holidays;
 - **Step 2 - Fraud Validation**: time between the stock confirmation and the fraud check (if the fraud check is performed before the stock confirmation, then this time is 0). It does not count for Speed of Sending because it is part of Farfetch's responsibility. Moreover, when a partner has a type of integration that only allows it to validate its stock after the fraud check, the Speed of Sending starts on the moment the fraud validation is done;
 - **Step 3 - Decide Packaging**: time spent during the packaging step, between the stock-/fraud check and the packaging selection;
 - **Step 4 - Shipping Label**: time spent creating the shipping label, between packaging being selected and the label being created. After an order is placed and everything goes according to Farfetch's standards regarding payment, fraud and stock management, there is an important step where the AWB is created. This is normally an automatic step, but if the system is not able to match the address/zip-code, the order will stop in Step 4 and request manual work to correct the problem. An order with these characteristics is identified as a "Step 4 Order";
 - **Step 5 - Waiting for Courier**: time the partner spends waiting for the courier since the AWB was created.

- (a) **Speed of Sending of One Day (SoS1D) [%]**: percentage of boutique orders sent in less than one day, without weekends, holidays and exceptions;
 - (b) **Speed of Sending of Two Days (SoS2D) [%]**: percentage of boutique orders sent in less than two days, without weekends, holidays and exceptions.
4. **Wrong Item (WI) [% Qty]**: ratio between the number of returns with the return reason being having received a wrong item and the total quantity of returns. The current standard is for this to be measured according to the date in which the order was sent, but for the system that was created in this thesis the date used to calculate WI refers to the return being processed. This means that, currently, this KPI does not have a useful interpretation when observing the most recent data, since Farfetch is still receiving wrong item returns for recent orders. Therefore, WI is currently being monitored by Farfetch with a delay of three weeks. By analysing WI according to the date in which the return is processed, the alerts given by the final system happen before the business detects any problems;
 5. **Time to Process Returns (TTPR) [days, gross]**: time that partners take to process returns after having received them.

3.2.2 Supply Chain – Delivery KPIs

1. **Time in Transit Global (TITG) [days, gross]**: measured in the step 6 of the order process as the time between in-store pickup and the final delivery to the customer. It can be divided into two other KPIs: Time in Transit Standard and Express, but this division was not considered by the tool developed in this thesis;
2. **Returns Transit Time (RTT) [days]**: time in transit, for returns;
3. **EDD Accuracy – Checkout (EDDA) [%]**: measures the share of orders that were delivered within the Estimated Delivery Dates (EDD) given to the customer in checkout or that were too early, i.e., delivered before the EDD.

3.2.3 Fintech Operations – Payment KPIs

1. **Payments Attempt Completion Rate (PACR) [%]**: ratio between the number of completed payment attempts and the total number of payment attempts. A payment attempt is defined as the customer action of trying to pay. A payment attempt is completed only when all its instruments are completed;
2. **Payment Completion Rate (PCR) [%]**: ratio between the number of completed payment sessions and the total number of payment sessions. A payment session aggregates all the customer interactions within the same tenant, customer, calendar day and with the same total order price.

3.2.4 Fintech Operations – Fraud KPIs

Fraud Rejection Rate (FRR) [% GTV]: ratio between the sum of the amount of USD pre-authorized or captured from orders which were cancelled due to fraud concerns and the total amount of Gross Transaction Value (GTV) from that period. Although the targets established for this KPI refer to it not surpassing certain thresholds, it is not as simple as that. There needs to be an equilibrium between the KPI not having results that are too low or too high, as when FRR is suspiciously low it may indicate the fraud team is not being successful at detecting fraudulent behaviours.

3.3 Operations Monitoring

3.3.1 Weekly Operations Tactical Meeting

Every Monday, a one-hour-long meeting takes place between operational KPI owners, i.e., those accountable for monitoring the metrics behaviour and more senior staff responsible for Operations. The purpose of this meeting is to analyse how successfully Operations ran the week before and to understand what should the Operations teams pay attention to, by trying to find the root causes and fixing them if they are within Farfetch's agency and not a sporadic event like strikes or natural disasters. The macro-services represented in this meeting are Product Creation & Catalogue Management, Payments Processing & Fraud Prevention, Order Fulfilment, Order Delivery & Return and Partner & Customer Care. These macro-services are then divided into areas, which allows for better responsibility allocation and for a more comprehensive data exploration. The meeting tends to be fast paced, as the goal is to cover all the areas within the established timeframe and as Farfetch's Operations scope is extensive. Some deep dives into uncovered issues and further discussion often occur after the meeting, as some information is not relevant enough for most of the stakeholders in the meeting.

Tactical meetings should be a place for problem and impactful situations sharing, so each team can gain access to information that is not directly related to their field, but which may also affect them. It acts as an instrument to diminish the organizational silos that inevitably exist due to the Operations team being divided into further teams. It facilitates the information flow between them and it clarifies their interdependencies, which may trigger otherwise non-existing discussions.

Each area requires people to prepare slides in a collaborative PowerPoint presentation which is shared with interested stakeholders with the main highlights and lowlights about the KPIs' behaviour in the previous week. There is also someone responsible in each area for presenting their slides during the meeting and for clarifying any questions during and after it.

Previously to each meeting, those responsible for monitoring each KPI have different approaches on how to do so, as each metric requires a personalised process according to which dimensions it should be viewed through and with which scopes. Furthermore, these people are distributed by multiple teams and, therefore, have distinct routines regarding KPI monitoring. How they carry their analyses will be explained next, specifically in the Marketplace context as it

is the focus of the thesis, although the processes done for FPS are relatively similar. However, a common denominator between all these approaches is the lack of a systematized methodology on what classifies a topic as worthy of being mentioned in the tactical meeting.

According to the results of their previous data exploration, each team discloses their main takeaways – highlights and lowlights, including updates on ongoing issues – in the tactical meeting. Visually, it is either done through a Looker dashboard, which is also a useful data analytics tool, or through the corresponding slides in the shared PowerPoint presentation.

It is important to mention that what is next described as the monitoring routines for each area corresponds only to the basis of the KPI analysis since, depending on the initial conclusions, further deep dives are often made.

3.3.2 Order Fulfilment

The areas in which this macro-service is divided into are FxFF and nonFxFF. All fulfilment KPIs are monitored by the two teams, but some metrics, like Time to No Stock, are not discussed in the tactical meeting due to lack of time.

For **nonFxFF**, the KPI analysis is done between the Partner Success Management team and there is a deeper assessment at the partner level. This team is in charge of monitoring the fulfilment performance of Farfetch's partners which are not fulfilled by Farfetch.

This team makes the following analysis for **NS, TTNS, SOS1d, SOS2d, WI** and **TTPR**:

- Deviation week over week;
- Comparison with the data of the previous eight weeks;
- Partner analysis: for each partner, the team analyses the numerator variable (e.g., in NS it is the quantity of NS items), the NS itself and the impact of the store on the overall NS (which corresponds to the delta calculation, explained in the next chapter, which unveils the negative and positive contributors. This analysis is only not done for TTNS, as the team solely reviews its actuals per partner.

The operations **FxFF** team meets weekly to discuss the metrics which are part of their scope of responsibilities, which are the aforementioned ones, with the goal of better understanding their performance, possible root causes and next steps. To prepare for this meeting, all metrics are:

- compared to the monthly target and to the deviation from last week;
- reviewed by warehouse group (warehouses are aggregated by country and product type), by partner and by stock point, comparing their performance with the previous four weeks, with the main offenders being highlighted. Each step in the speed of sending process as a whole is also analysed to see if there were any abnormalities.

3.3.3 Payment Processing & Fraud Prevention

The areas in which this macro-service is divided into are payments and fraud. The payments team and the fraud team meet each week before the tactical meeting to share how the KPIs' behaviours

were and to discuss possible root causes and solutions. An analysis is carried out, both for **PCR** and **PACR**:

- Absolute deviation regarding last week and absolute deviation regarding target. The way **demand mix** affects the KPI is also considered, i.e., the impact of the demand in each shipping region and the demand distribution fluctuations from week to week on the KPI's results;
- **Customer mix**, according to two customer types: existing and new customers. Their impact on the metrics according to their share is examined and these shares are also compared to their targets;
- **Performance impact**, which simply represents how the regions performed overall against their targets and how this impacted the metrics as a whole;
- The performance of each payment method and of each channel (e.g., the Farfetch iOS App or the website), when compared to target and the previous week.

For **FRR**, the following aspects are discussed:

- Absolute deviation regarding last week and absolute deviation regarding target;
- Relative deviation in GTV from promotions and in average rejected order value;
- Region analysis:
 - Regions with the biggest relative deviations from target;
 - Main drivers for the overall metric being off or on target;
 - Region top offenders regarding chargebacks, i.e., payments which are contested by the customer directly with the financial entity responsible for processing the payment.

At the end of these meetings, if there any inquiries about the analysis, they are discussed and suggestions are often made regarding further analysis about new and reoccurring issues.

3.3.4 Order Delivery & Return

The areas in which this macro-service is divided into are Transit Time and EDD Accuracy.

Regarding **EDD Accuracy**, before the tactical meeting, the analyst responsible for analysing this metric does a comprehensive overview of the metric's last week's behaviour. The orders delivered last week as a whole are divided into "on time", "too early" and "too late" deliveries. The analyst investigates these shares and compares them with the ones from the previous week. An overall target comparison is also made. EDDA is also looked at through the following dimensions:

- **Customer region**: comparison with the previous week and with the established targets. There is also a "on time", "too early" and "too late" analysis;

- **Service (express and standard):** comparison with the previous week and with the same week, but in the previous year. This is also done for each service in each customer region;
- **Carrier:** not only a week over week comparison, but also an analysis regarding “on time”, “too early” and “too late”. Moreover, the volume order by carrier is also considered.

TITG's actual is compared to the previous four weeks considering:

- Volume of orders delivered;
- Service type distribution;
- Distribution of transit time gross range (e.g., 0-3 days, 3-5 days);
- Carrier distribution.

RTT's actual is compared to the previous four weeks considering:

- Volume of returns delivered;
- Returns delivered type distribution (direct and indirect);
- Distribution of returns delivered pickup accuracy (e.g., “on time”, “not on time – before”);
- Carrier distribution.

3.4 Project's Challenge

Simple time and target based criteria are used to analyse the KPIs across these macro-services and there is not a systematic way to decide which topics should be mentioned in the tactical meeting. Moreover, there is not a prioritization concerning which order to discuss the KPIs' behaviour in the previous week. This prioritization could add value to the contents of the meeting, since sometimes not all macro-services are talked through for lack of time. The solution to this problem does not, however, necessarily entail discontinuing the way in which the analyses prior to the meeting are done, since deep-dives and extensive examination are necessary, nevertheless. The tool developed in this thesis serves as an add-on to the work currently done and as an answer to these issues regarding the tactical meeting.

Chapter 4

Methodology

The classification of a KPI's behaviour as an alert depends on each criterion's analysis of a KPI, which can either output a green, yellow or red light, according to certain thresholds. This chapter starts with the introduction of this traffic light framework, which through the combination of four criteria defines different types of alerts. There are three main criteria and a fourth one which is only tested in certain conditions. Both the criteria and alert types are characterised. In the context of this thesis, an alert alludes to a message given by the tool regarding a KPI's behaviour in a certain week, which reflects a situation relevant enough, according to the criteria, for the user to be notified of. An alert is, therefore, a general term for a situation which the tool finds worth noting. Alerts can, then, be classified in different types, which will be later explained.

In addition, to better the detail of the tool's outputs, dimension analysis for each KPI was implemented and is also explained in this chapter. The final topics regard the work behind setting up the tool, which is embodied in a Slack App.

4.1 A Traffic Light Framework

After being familiarized with the current situation and understanding the multiple ways in which the KPIs are monitored, it was possible to develop a common framework for setting alerts according to the KPIs' weekly performance.

The three final main criteria, in no particular order, are, according to the KPI's weekly result:

1. Is the KPI inside a certain prediction interval according to the forecasting model?
2. Is the KPI an outlier according to the recent past?
3. Is the KPI significantly off target?

The combination of these three criteria was defined with the goal of being all-encompassing of situations that could be of interest for the business.

When a KPI's behaviour is judged with the outlier and the prediction interval criteria, the result is either a green, yellow or red light. On the other hand, with the target criterion it can only be a

green or a red light. Since targets are defined by humans and often have an error associated with them, the tool created for this thesis is not as conservative towards them. Thus, it did not make sense to create a yellow light on such a volatile parameter, but instead only trigger a red light when the actual is significantly off-target and not simply off-target. The combination of these lights will output different types of alerts, but when a KPI is not considered to be any alert, then a fourth criterion is tested, the future target criterion, which, similarly to the target criterion, only outputs a green or a red light. To establish the rules for each criterion regarding what sets the difference between each colour, thresholds needed to be set.

4.1.1 Prediction Interval Criterion

Regarding this criterion, the rule which makes the KPI's behaviour a yellow light is for it to be outside the 80% prediction interval, but inside the 95% prediction interval. Whereas for it to be a red light, it needs to be outside of the 95% prediction interval. Green light corresponds to being inside the 80% prediction interval. These thresholds were chosen as they are common standards in literature for prediction intervals (Hyndman and Athanasopoulos, 2018). Nevertheless, these parameters are programmed in a straightforward way, so that Farfetch can easily calibrate them.

This criterion was chosen with the underlying purpose of predicting what the KPIs' behaviour one-week-ahead will be. If the best forecasting model for a certain KPI encompasses a seasonality component, this characteristic influences the predicted value. By already considering this criterion, if the tool was to additionally compare the actual of the week being analysed with the actual of the corresponding week in the previous year, as it is currently done for some KPIs, it would be redundant. This year-on-year comparison may even not be relevant if the metric's behaviour does not exhibit yearly seasonality, which is why this comparison was not used as one of the criteria. However, this comparison is, in a way, still implicitly considered in models with a seasonality component when calculating the prediction intervals.

To obtain a reliable forecasting model for each KPI, four different models were tested: simple exponential smoothing, double exponential smoothing, Holt-Winters and Prophet. For each KPI, to calculate the prediction intervals and the forecasted values, the corresponding best forecasting model was trained with the available data regarding the weeks prior to the week being analysed.

To visually analyse the possible trend and seasonal components in each of the KPI time series, decomposition plots were first devised. These graphs decompose a time series into a sum of seasonal, trend and irregular components using moving averages. A fair indicator of the quality of this decomposition is the randomness of the remainder component. However, this was done merely as an initial overview of their behaviour and not an infallible method for determining the presence of seasonality or trend. The decomposition plots are shown in Figures A.1 to A.12 in Appendix A.

Some forecasting models have already been developed at Farfetch, including for some fulfilment KPIs, but they are monthly based. Although the same code was not adapted for a weekly model, Prophet with yearly seasonality was still used as a base for the Prophet forecasting. Producing accurate weekly forecasts is inherently harder than monthly ones, as weekly data tend to

be more volatile and subjected to noise compared to monthly data. Temporary events and random fluctuations can affect weekly patterns, making it challenging to distinguish real trends. Furthermore, they are more susceptible to factors like weather, holidays or shifts in consumer behaviour. Prophet has multiple advantages: it is generally reliable, it has parameters easily interpretable by someone without a deep knowledge in data science, it is trained quickly, it is especially useful for time series which show seasonality, it can handle multiple seasonality frequencies at once (for example, weekly and annual), it is robust to outliers and it does not require the data input to have regular time intervals. Contrariwise, all exponential smoothing models are sensitive to unusual events or outliers and cannot handle missing data.

Holt-Winters' seasonality component is not as powerful as Prophet's, as it does not allow for multiple frequencies. Moreover, using it for yearly seasonality is challenging, as the number of days in a year throughout the years is not constant and both the R and the Python Holt-Winters' function do not allow non-integer numbers for the frequency parameter, which is a mandatory input. Therefore, there is not a proper direct way of considering yearly seasonality, as whatever integer is chosen there will always be a misalignment between the model and the calendar year. However, Holt-Winters has a parameter not intrinsic to Prophet, which is α . Each forecasted value is the weighted average of the prior observations, with weights decreasing exponentially according to this parameter. Thus, it is capable of understanding whether to value more ancient observations or not, whereas Prophet attributes the same weight to each one of them.

Since these two methods are solely applicable to seasonal time series, the simple and the double exponential smoothing models were also tested.

4.1.1.1 Parameter Tuning

The Holt-Winters parameters, i.e., α , β and γ were determined by minimizing the squared prediction error. Similarly, the corresponding arguments were tuned through the same method in the simple and double exponential smoothing models. For Holt-Winters, the seasonality mode, which can be additive or multiplicative, was also optimized.

Regarding Prophet, multiple components can be tuned, such as the seasonality prior scale, which controls the magnitude of the seasonality fluctuation; the seasonality mode; the type of growth, which can be either linear or logistic; the changepoint range, which indicates the percentage of historical data that allow a trend change; amongst others. Due to the considerable computer power necessary to tune the entire list, it was decided to tune seasonality mode, similarly to what was done with Holt-Winters, through cross-validation. Moreover, Facebook claims in Taylor and Letham (2017) that its default parameters are appropriate for most forecasting problems.

4.1.1.2 Cross-Validation

The chosen timespan for the data used to build the forecasting models, if there was data available, was the period between the beginning of 2018 and May of 2023, but if not, it was the following longest timeframe, which was either since 2019 or 2021 for some KPIs.

While doing cross-validation, for each origin, the three forecasted values and the three corresponding error values were stored, so that performance metrics could be calculated separately for $h=1$, $h=2$ and $h=3$. The purpose is not only to evaluate the models on how well they predict for $h=1$, which will be the practical case for the system, but also on how they handle forecasting for bigger timespans. Therefore, for each h , the average of the chosen performance metrics was calculated. Although the forecast period for the thesis' use case is $h=1$, to understand how robust each model is, the performance metrics were also calculated for $h=2$ and $h=3$.

By basing the model choice on the average of the performance metrics as a result of cross-validation, instead of solely calculating them based on a single train and test set, the models' performance can be better judged and, hence, choosing one which deals with overfitting can be prevented.

4.1.1.3 Performance Metrics

The performance metrics calculated for each model tested for each KPI were the Mean Absolute Scaled Error (MASE) and the Mean Absolute Percentage Error (MAPE).

MASE: this metric evaluates how much better the model performs compared to a naïve model, i.e., how much better its Mean Absolute Error (MAE) is compared to the MAE of the corresponding naïve forecast. This is reflected in Equations 4.1 and 4.2. Hence, MASE is a scale-free error metric. If MASE is below 1, it evidences that the forecasts are better, on average, than the in-sample one-step forecasts from the naïve model (Hyndman and Koehler, 2006).

$$MASE = \frac{MAE}{MAE_{in-sample,naive}} \quad (4.1)$$

, where:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (4.2)$$

The selection process was solely based on the models' average MASE. The best model, for each KPI, was considered to be the one with the best average MASE for $h=1$, $h=2$ and $h=3$, that is, the average MASE for a three-day simultaneous forecast. It was also mandatory for MASE to be less than 1 for $h=1$, as it is the practical use case for this project and, thus, the model has to be meaningful for that. Nevertheless, the models with the best average MASE not only consistently fulfilled this last criterion, but also always had the best average MAPE.

MAPE: MAPE is a percentage error metric, which means it is scale independent and, thus, suitable for comparing forecast performance between different time series. It is calculated through:

$$MAPE[\%] = \frac{100}{n} \sum_{t=1}^n \frac{|A_t - F_t|}{A_t} \quad (4.3)$$

, where A_t are the actual values and F_t the corresponding forecasts.

MAPE was not directly used as a criterion to choose the best model, due to its inherent bias towards underforecasts. If a model is optimised based on minimising its MAPE, it will likely lead to smaller forecasted values (Kolassa and Martin, 2011). However, as its definition is easily understandable, it was calculated as an extra measure to show the stakeholders the quality of the models.

4.1.2 Recent Observations Criterion

As the criteria for the final tool were being pondered, a problem arose: if the forecasting model performs well enough, it will predict peaks and lows which set themselves apart from recent behaviour, i.e., the actual would likely still be within the 80% and 95% prediction intervals. Therefore, if there was to be an alert it would not be due to the prediction interval criterion, even though the business would still like to be aware of these situations, regardless of the forecasting model's ability to anticipate them. As stakeholders have an interest on being informed of unusual behaviour when compared to recent history, there was a need for a criterion that covered this. Hence, a method to detect deviations in the recent past was needed.

When employing the mean as the central tendency indicator, three issues can be observed (Miller, 1991). First, it presumes that the distribution is normal. Second, outliers have a significant impact on the mean and standard deviation. Third, as indicated in Cousineau and Chartier (2010), it is extremely unlikely that this method will find outliers in small samples. As a result, this measure is essentially flawed: although it is intended to facilitate outlier detection, the presence of outlying results undermines the indicator itself.

Although it is slightly susceptible to them, the MAD is very insensitive to the presence of outliers. Furthermore, it is totally immune to the sample size (Dave and Varma, 2014). In the event of outliers, the MAD is the most resilient dispersion measure in univariate statistics (Leys et al., 2013).

When business analysts at Farfetch compare the week being analysed to the recent past it is usually regarding the previous four weeks, as described in the previous chapter. However, this is too brief to obtain meaningful MAD calculations, so eight weeks was the established timeframe.

Finally, the outlier rejection threshold must be set, which was suggested by Miller (1991) to be 3 (extremely conservative), 2.5 (moderately conservative) and 2 (poorly conservative). Based on these recommendations, two decision thresholds were implemented in this thesis, which led the red light in this criterion to be defined by:

$$M - 3 \times MAD > x_i > M + 3 \times MAD \quad (4.4)$$

And the yellow light to be defined by:

$$(M - 2 \times MAD < x_i < M + 2 \times MAD) \wedge (M - 3 \times MAD > x_i > M + 3 \times MAD) \quad (4.5)$$

, where x_i is the KPI's actual in the week being analysed and M is the median.

Finally, green light corresponds to the actual being inside the $2 \times MAD$ region.

These thresholds can be easily adjusted by the business, according to its needs.

4.1.3 Target Criterion

Unlike the two other criteria, the target criterion only classifies a KPI's actual as a green or a red light. It only triggers a red light when the actual is significantly off-target, which means a KPI will still be associated with a green light regarding this criterion, if it is only slightly off-target. This brought the question: how off-target can the business tolerate a KPI to be without raising concern? At which threshold does the business want to be alerted regarding this criterion?

Before defining the threshold, the first step was to decide between absolute and relative target deviations. Even though a relative one may seem to create a fairer approach across all metrics, the following example shows its nuances. A 5% deviation from the target entails a 0.1pp absolute deviation for an actual of 2%, which is close to the values that metrics like WI and NS usually acquire, whereas for a KPI with actuals closer to 80%, this would entail a 4pp absolute deviation. Therefore, simply basing this criterion on relative deviation is not a bulletproof method: a 5% deviation is not equally as concerning for WI as it is for PCR. In other words, the smaller the actual, the greater should the threshold be for classifying it as a red light.

This acknowledgment led to two conclusions:

- The threshold for the relative deviation from the target had then to be defined with a function, since it should be high when the target is low and vice-versa. This type of relationship can be translated into the rational function:

$$f(x) = \frac{b}{x} \quad (4.6)$$

, where b is a constant and x is the target;

- Regardless of the target, the threshold should, at least, be $a=3\%$. This value was defined as the minimum threshold for there to be a red light in this criterion, based on a discussion with some analysts considering practical examples of some KPIs with high targets. These were the KPIs leading the discussion, because if the minimum threshold would already be

too high for the KPIs which should have the lowest thresholds, it would mean the minimum was poorly defined.

This meant the threshold function is defined as:

$$Threshold(target) = a + \frac{b}{target} \quad (4.7)$$

It is key to realise that the KPIs being analysed have targets which tend to range from 1 to 5 and from 60 to 97. Hence, the threshold function should be adequate for these values. Thus, the two poles were tested: 1 and 97, which are usual target values for No Stock and Speed of Sending of 2 Days, respectively. For each, analysts were asked from what value would they start getting concerned due to too much deviation from target, as this marks a red light regarding the target criterion. The results are shown in the following table:

Table 4.1: Data Points to Assist in the Definition of b

KPI	Target	Red Light Value	Relative Deviation from Target [%]
No Stock	1	1.15	15
Speed of Sending of 2 Days	97	94	3.09

The threshold function should output the proper relative deviation from target, when the actual is off target, which would trigger a red light, depending on the target value. Therefore, having these two datapoint references and only one unknown variable in the function, b was determined to be, approximately, 12.

The function then became:

$$Threshold(target) = 3 + \frac{12}{target} \quad (4.8)$$

Parameters a and b can be regulated by the business, if, based on the analysts' interpretation of the targets, this function starts to no longer match the thresholds they want the tool to base this criterion on.

To apply this threshold in an adequate manner, it was necessary to classify each KPI, according to what was termed "target type", i.e., whether the defined target was meant to be reached or avoided. A KPI with the former aim was said to have a "minimum" target type and a KPI with the latter goal a "maximum" target type. This idea is key to understand if the metric is off target in the first place. The underlying concept behind these target types is that an incremental improvement week over week for KPIs with a "minimum" target type corresponds to an increase, whereas for the ones with a "maximum" target type it corresponds to a decrease. This is true regardless of the KPI being on or off target.

Furthermore, targets for the KPIs in the considered scope are set monthly. However, for WI, as explained in the previous chapter, it was not calculated according to the current standards with

reference to date, so the monthly targets could not be applied to the criterion. Instead, the target which was used was the yearly one, as it is valid regardless of the date used for the calculations. Ideally, monthly targets with reference to the return processed date would exist and be used.

4.1.4 Future Target Criterion

This criterion is only checked when, according to the others, there is not an alert. Similarly to the target criterion, this one can also only output a green or a red light.

Based on the historical data thus far, including the week being analysed, the KPI's result in the following week is predicted and, if the KPI becomes significantly off target, according to the function explained for the target criterion, it results in a red light.

4.2 Dimension Analysis

Having had the criteria properly defined, the information that could be useful in case of an alert had to be defined. As it is currently done while monitoring each KPI, the business is interested in analysing certain dimensions. Only the dimension considered to be the most meaningful, for reasons explained for each KPI, was assessed, although it may be relevant to add more dimensions in the future.

Before deciding which one to analyse for each metric, the final KPIs had to be selected. One of the tool's goals is to create a distinction between how alarming the KPIs' behaviour was in the week being analysed, for the user to understand which ones they should first focus on. Therefore, it was important to choose KPIs which were already being monitored, even if not discussed in the tactical meeting, and which, when viewed together, could create a reasonable perception of the business operations. An additional condition for the data used to train the forecasting models was that there should not have been a significant change in the way the metric is calculated, since it will negatively impact the forecast quality.

Regarding dimension selection, it was based on the MECE principle, which is a grouping principle for dividing a set of objects into mutually exclusive (ME) and collectively exhaustive (CE) subsets (Chia, 2019). There are multiple ways of achieving this. The process was to first recognise what the current dimensions that the business goes through are and if these show a clear magnitude difference, then the most high-level one is selected. This rule is only applicable to the fulfilment metrics since, when analysing them, there is first a division between FxFF and nonFxFF, which does not exist in the other KPIs. This is why for the fulfilment KPIs this division was done, instead of a less magnitude one, such as dividing them according to the warehouses' performance.

The goal of the dimension analysis is to give more detailed insightful information about the KPIs which were either labelled as alarms or attentions. It was done through two different methods for different KPIs: for the fulfilment metrics a more straightforward way and for the rest of the KPIs an approach which, for simplicity reasons, will be referred to as the "MAD-Delta approach".

4.2.1 Payment Metrics

The dimension chosen was payment method, because the main takeaways of these KPIs are often associated with this dimension in tactical meetings, which is aligned with the fact that the payment process is highly dependent on the method.

4.2.1.1 The MAD-Delta Approach

This approach was used for both payment metrics and to explain it, its application on Payments Attempt Completion Rate (PACR) will be detailed as an example. It is pertinent to note not only that the dimension chosen for this KPI is payment methods, but also that this metric is the result of the ratio between the number of completed charge operations and the overall number of charge operations.

To verify, when there is a negative alert, which payment methods were the main contributors to the poor performance, the analysts mostly focus on mentioning which had, in absolute terms, the worst actuals and which had the most significant absolute deviation to target. However, this does not encompass the full picture, as a payment method can have the lowest PACR or the most significant deviation from its target, but not be the worst offender. The reasons for this are that, on one hand, targets can be inadequately defined and, on the other hand, only assessing the PACR is incomplete as the payment methods distribution, i.e., how many of the payments each method is responsible for, is also a factor in determining the worst offenders.

To determine the payment methods with the most significant impact on the overall metric on a certain week, the share regarding the number of completed charge operations and the share regarding the number of charge operations should be calculated for each. These shares are regarding the entirety of the payment methods. Then, still for each method, the difference between the latter and the former should be calculated, which results in an output which was named “delta” (Δ). Its general formulation is:

$$\Delta \text{ (for each Dimension)} = \text{Share of Denominator} - \text{Share of Numerator}$$

Its application to PACR is:

$$\Delta \text{ (for each payment method)} = \text{Share of Number of Charge Operations} - \text{Share of Completed Charge Operations}$$

Therefore, a payment method can have the worst PACR in comparison to the others, but not be the worst offender. The worst offender would be the one with the most negative impact. This translates into the most positive delta, since a positive delta means its share of completed charge operations is lower than its share of number of charge operations. In other others, its contribution to the number of completed charge operations would be lower than its contribution to the number of charge operations. To conclude this idea, not only would it be lower, but it would be the payment method with the most pronounced detrimental difference between its contributions.

A parallel way of thinking can be done to determine the best contributor – which should be mentioned when there is a positive alert –, which would be the one with the most negative delta, i.e., the one with the most pronounced beneficial difference between its contributions.

It is key to remark that PACR is a KPI which goal is to be maximised, that is why, regarding a payment method, its share of completed charge operations being greater than its share of charge operations shows it is a beneficial contributor to the metric as a whole. However, if the objective is for a KPI to be as low as attainable, the logic switches: it is desirable for the share regarding the numerator to be inferior to the share regarding the denominator, which means a beneficial contributor, dimension wise, would have a positive delta. On the contrary, a prejudicial contributor would have a negative delta.

As each dimension has multiple options – e.g., there are more than twenty payment methods available overall at Farfetch – it is not useful to show all negative or positive contributors each time there is an alert regarding a detrimental or a favourable behaviour, respectively. In order to filter the most impactful ones, the MAD technique for outlier detection was used within the data referring to each dimension’s delta. The rules for detecting the main contributors are summarised in the following table:

Table 4.2: Rules for the Detection of Main Contributors in each Dimension

KPI Goal	Alert Type	Rule (select dimensions with:)
Increase	Good	$\Delta < M - 2 \times \text{MAD}$
Increase	Bad	$\Delta \geq M + 2 \times \text{MAD}$
Decrease	Good	$\Delta \geq M + 2 \times \text{MAD}$
Decrease	Bad	$\Delta < M - 2 \times \text{MAD}$

These rules select the best and worst influencers, dimension wise, for good and bad alerts, respectively. The MAD multiplier, in this case 2, can also be adjusted by the business, depending on how conservative it wants to be regarding which contributors to show.

4.2.2 Fraud Metrics

FRR was analysed through the customer region dimension, because its performance is highly influenced by the customer behaviour and not as significantly by the chosen payment method. The approach for this further assessment was also the MAD-Delta one, considering FRR’s formula and the fact that the goal, according to target, is to minimize this metric.

4.2.3 Delivery Metrics

The dimension for all the delivery metrics was customer region, since it has been a constant deep-dive in tactical meetings, which shows the relevance of its impact on the overall metric.

For EDD, the MAD-Delta method was used, taking into consideration its formulation and its objective of being as high as reachable.

Regarding TITG and RTT, since they correspond to average times, the delta formula used was:
 Δ (for each Customer Region) = Share of Transit Time – Share of Total Time

Then, the MAD-Delta method was once again used for taking into consideration that the goal is to minimize these two KPIs.

4.2.4 Fulfilment Metrics

As explained before, the dimension for these KPIs is a division between FxFF and nonFxFF.

When looking through the **FxFF** lens, the original idea was to, unlike the other KPIs, include not only Marketplace, but also FPS, since this is how the monitoring is being conducted at the moment, as the goal is to analyse warehouse performance, which is Farfetch's full responsibility, regardless of the business unit. However, these are separate business units – the former being B2C and the latter B2B – which are composed of different entities, with different needs, behaviours and share distributions. This means, for example, that the impact in the overall KPIs when a new partner starts selling through Marketplace is much less significant than when a new tenant joins FPS, because there are more than one thousand Marketplace partners and less than fifty FPS tenants. This made time series forecasting for FxFF as a whole much harder and, hence, the multiple model results for the fulfilment metrics overall were not accurate enough for them to be relevant. The decision was then made to just consider Marketplace when analysing FxFF.

Regarding **nonFxFF**, it can be divided into Marketplace and FPS, i.e., the analysis is not currently done with them together. However, for the same reason that FxFF as a whole was not considered in the final tool, nonFxFF was only analysed through Marketplace.

The approach for analysing both FxFF and nonFxFF within Marketplace was to compare the KPI in each of these scopes with the corresponding target and to check whether the week-on-week performance was favourable or not, i.e., if it became more on-target or not. The nonFxFF parcel is much more significant on the overall KPIs, as it has always taken on a significantly greater share of order volume.

4.3 Alert Types

For each KPI's performance, the types of alerts are: alarm, attention and future attention.

4.3.1 Alarm

A KPI which is responsible for an alarm is characterised by having, at least, two red lights regarding the three main criteria (prediction interval, recent observations and target criteria). As the name suggests, it represents a situation in which the KPI is showing considerable abnormal behaviour in, at least, two of the criteria. An exception happened for FRR, as it is only judged based on two criteria, since none of the forecasting models were accurate enough to consider the prediction interval criterion. In that specific case, it is only an alarm if it shows two red lights or one red light and one yellow light.

An alarm can either be, as it was labelled, "good" or "bad". A good alarm means that, week over week, the KPI became more on target, which depending on the KPI it can mean it increased or decreased. It is crucial to distinguish between being on target and being more on target: more on target does not indicate that it is on target. It just conveys that it acted according to the KPI's main objective, i.e., that it is moving in the desired direction.

4.3.2 Attention

An attention is characterised by only having one red light or by having two yellow lights. An attention is, putting it simply, a less severe version of an alarm, i.e., the performance shown does not correspond to the norm, but it is not as disruptive as an alarm. Once again, for FRR what characterises this type of alert is slightly different: it solely occurs when there is only one red light, as it is not possible for it to have two yellow lights, since the target criterion cannot output that colour.

Similarly to an alarm, originally an attention could also be entitled as “good” or “bad”, according to the same rationale. However, to only deliver the key takeaways for each week of the stakeholders’ interest, it was decided that good attentions were not relevant enough to be presented by the final tool. Thus, every attention depicted is a bad attention. In the Results Chapter, this decision was later changed.

4.3.3 Future Attention

The reasoning behind this alert has similarities with outcome-oriented predictive process monitoring. Moreover, a red light in the future target criterion corresponds to a future attention alert, recalling that red lights in this criterion only occur when there is no other alert.

For this final alert, there was no distinction between it being “good” or “bad”, as all future attentions are inherently bad attentions. If in the week analysed the KPI did not set off an alert, it means it is not concerningly off-target, otherwise it would be, at least, an attention. Thus, if in the following week it is predicted to be significantly off target, it is becoming more off target.

4.4 Information Displayed for each Alert

Every text associated with an alert starts by showing a basic characterization of the KPI’s performance, by mentioning its actual and its main goal, which can either be to stay below or to reach the corresponding monthly target.

Information displayed which is directly related to the three criteria is solely about the criteria which contributed to it being an alarm or an attention, i.e., when the criteria were labelled as yellow or red. If one of the criteria is considered a green light, although the KPI’s behaviour can even be classified as an alarm, it will not be mentioned, since only the criteria that contributed for it to have this classification will be referred. However, texts do not only reveal what the criteria have concluded, as they also present details that contribute to a better overview of the metrics’ behaviour according to what the criteria have detected:

1. When the recent observations criterion contributes to the alert, information is also shown regarding the last four week’s overall trend, i.e., if it is increasing or decreasing, according to the positive or negative signal of the slope of the linear regression that was fitted into the KPI’s behaviour during this timeframe;

2. Furthermore, if this aforementioned criterion is either a red or a yellow light or if the target criterion is a red light, the user is informed not only about the absolute and relative deviation from the week prior to the week being analysed, but also if the KPI became more on-target during this time span. Both the absolute and relative deviation should be mentioned as they provide a clearer picture than solely one of them.

Regarding dimensions, the information that is extracted about them from the data, which is different for fulfilment and non-fulfilment KPIs, as explained before, is also shown. Examples of the information displayed for the multiple alert types are shown in Figures 5.2 to 5.6 in Chapter 5.

4.5 Connection between BigQuery, R and Slack

Google BigQuery, a web-based data warehouse that allows querying through ANSI SQL, is the data warehouse where Farfetch stores all its business data.

In order to extract the necessary data for this project, it had to be imported from BigQuery, through the `dbGetQuery()` function in R, which takes a BigQuery connection and a SQL string. For the connection, not only the corresponding BigQuery project ID and dataset name are needed, but it is also necessary to go through an authentication process as the data is restricted to Farfetch.

While designing the tool for this thesis, it was key to choose a format to deliver the extracted information that would allow it to be easily interpreted. Although dashboards are a useful instrument to convey and filter information, the business already possessed the required ones for monitoring KPIs. By sending straightforward texts through Slack, the user does not have to take any initiative to get a clear picture of the weekly operational performance in Marketplace.

To send texts from R to Slack, the function used in R was `slackr_msg()`, which requires the corresponding text, the channel and an authentication token bearing required scopes. The texts were sent to each stakeholder through the app, thus the channel was their Slack username.

Additionally, the authentication token is a string that the OAuth client uses to make requests to the resource server. OAuth client refers to the specific application or service responsible for initiating the OAuth process, which, in this case, is the app.

To create an authentication token, another concept which was important to grasp was the bot token scopes, which are the scopes that govern what the app can access. Thus, when creating it, the appropriate scopes had to be requested, as the app's capabilities and permissions are governed by them. The necessary scope was "chat: write", which allows the app to send texts to the channels which would then be mentioned in the `slackr_msg()` function. To start an app with this scope, a formal request was sent and then approved by the employees responsible for Farfetch's Slack.

4.6 Code Structure Overview

Figure B.1 in Appendix B describes the overall structure of the code behind the final system. To not make the schema visually overwhelming, the processes regarding dimension analysis and the extraction of useful information that is not directly related to the criteria were not included.

Chapter 5

Results

The prediction interval criterion showed results from two perspectives: the quality of the best models for each KPI and the verification of the assumptions regarding these intervals. Furthermore, the final tool was tested through an assessment made by two stakeholders and its outputs were analysed, which led to some feedback being implemented.

5.1 Prediction Interval Criterion

5.1.1 Cross-Validation Results

For each KPI, the four aforementioned models were evaluated through cross-validation for $h=1$, $h=2$ and $h=3$, i.e., simultaneously forecasting for one, two and three weeks. The average MAPE and MASE results, for each, are in the Tables C.1 to C.12 in Appendix C.

Based on the rules described in the previous chapter for choosing the best model for each metric, the results were the following:

Table 5.1: Best Model Results for each KPI

KPI	Model	Average MAPE (h=1) [%]	Average MASE (h=1)
EDD	Single Exponential Smoothing	1.581	0.973
NS	Single Exponential Smoothing	12.793	0.930
PACR	Prophet – Multiplicative Seasonality	2.226	0.852
PCR	Prophet – Additive Seasonality	1.580	0.883
RTT	Prophet – Additive Seasonality	9.686	0.991
SOS1d	Prophet – Multiplicative Seasonality	4.435	0.840
SOS2d	Prophet – Additive Seasonality	1.30	0.895
TITG	Prophet – Additive Seasonality	4.273	0.961
TTNS	Prophet – Additive Seasonality	10.947	0.767
TTPR	Prophet – Additive Seasonality	13.747	0.817
WI	Single Exponential Smoothing	6.843	0.789

5.1.2 Assumptions Assessment

5.1.2.1 Assessment of the Normality of the Residuals' Distribution

Only exponential smoothing models require the normality of the residuals' distribution to calculate prediction intervals, so this assumption was only tested for the KPIs which had one of these as their best performing model. For those, a histogram was plotted and a normality test was run. The Kolmogorov-Smirnov test was chosen over the Shapiro-Wilk test, since the data contains more than fifty observations. The null hypothesis for this test is that the data is normally distributed and the p-value threshold for its rejection was considered to be 0.05.

This null hypothesis was rejected both for EDD Accuracy and WI, which means their residuals do not follow a normal distribution. This means this assumption was broken for the prediction interval calculation. However, this was not considered to make them invalid, as the corresponding models behaved well according to the performance metrics tested, nevertheless. Moreover, for NS the null hypothesis was not rejected. The histograms are in Figures D.1 to D.11 and the results of the test are in Table D.1, all in Appendix D.

5.1.2.2 Assessment of the Constant Variance of the Residuals' Distribution

For the calculation of the prediction intervals, both exponential smoothing and Prophet models require the residuals to show constant variance. This was tested through a fitted values vs residuals plot for each KPI: if the spread of the residuals was roughly equal at each level of the fitted values, the constant variance assumption was met, otherwise this assumption was likely violated. The latter option happened for most of the KPIs, as it can be seen by the plots in Figures D.12 to D.22 in Appendix D. Nevertheless, for the same reason as why the prediction intervals were still used considering the non-normality of the residuals' distribution, the prediction intervals for Prophet were still regarded as valid.

5.2 Tool Assessment

5.2.1 Evaluation Process

The quality of this tool can be judged through two different perspectives:

1. Is the information in each alert useful and easy to understand? Are there any key points missing?
2. Are the alerts given the ones the business wants to receive? Does their classification match the reality of how relevant they are?

In regard to the first perspective, the tool was presented in a meeting to the Operations Performance & Optimization team. The feedback was overall positive, but a few suggestions were made:

- The text, for each alert, was, at the time, all together in one paragraph. It is now separated by: a straightforward weekly characterization, the main analysis and the dimension analysis;
- The KPIs' names were displayed as acronyms, but writing their full name makes it clearer for every stakeholder;
- In the "About" section of the App, it should include some explanation behind the forecasting models used, in order for the stakeholders to have a better understanding of the predicted values. This has not yet been done, but it should be implemented in the future, as the key takeaways for each algorithm are already explained in this thesis.

On the other hand, to test the quality of the classification in alert types, an inquiry was done with two stakeholders. According to the tool results for multiple weeks in the current year, ten "bad alarm", ten "good alarm" and ten "bad attention" situations were selected. Ideally, this study would have been done with a greater number of participants, who would need to be aware of the Marketplace context and of the analysed KPIs, and with a greater number of cases for each alert type, to achieve more meaningful results.

Every KPI was approximately equally represented, except for FRR, which considering that it is, exceptionally, only assessed through two criteria and that it has been consistently stable and on target this year, no alerts arose for it. Then, for each of these thirty cases, which refer to a KPI's behaviour in a certain week, the following information was stored:

- KPI's last eight weeks datapoints, which was conveyed through a chart;
- The target throughout these eight weeks;
- The forecasted value by the tool created for the week being analysed.

For each KPI in a certain week, these data were depicted in an organised manner in an Excel sheet, which meant an Excel file with thirty sheets was created to portray thirty alert-worthy cases according to the alert tool.

Two stakeholders who possess a comprehensive awareness and understanding of the Marketplace reality throughout the entire Operations context participated in the evaluation. They were asked, for each case, to classify the KPI behaviour as "no alert", "attention" or "alarm", according to how relevant it would be for them to be notified of it. There was no distinction between good and bad alerts, since this classification does not depend on the stakeholders' perception, it is solely based on the improvement or worsening of the KPI, week over week. They were not asked to judge future attentions, because they can be measured by the quality of the forecasting models. They were also informed about the differences in which Wrong Item was being monitored in comparison to the current standard. Furthermore, they were not told what classification the tool gave for the thirty situations, so they would not be biased, and they did their evaluation individually, without interacting with each other.

5.2.2 Evaluation Results

The results of the test were classified in a scale of 0 to 3. This scale reflects the consensus between the two participants and the tool regarding the classification of the behaviour of the KPIs in multiple weeks:

- 0: the participants and the tool all disagreed;
- 1: the participants agreed with each other, but disagreed with the tool;
- 2: only one of the participants agreed with the tool;
- 3: the participants and the tool agreed.

The results given by the tool and of each of the stakeholders' judgements are presented in Table E.1 in Appendix E. The following table summarizes them according to the aforementioned scale:

Table 5.2: Tool Evaluation Case Distribution Overview

Case Type	Number of Cases	Share (%)
0	4	13.33
1	8	26.67
2	7	23.33
3	11	36.67

The most frequent situation was number 3, which is a good indicator of the quality of the tool. Types 0 and 2 indicate the situations in which the participants did not agree with each other, which represent approximately 37% of the total cases. This reflects how there is not an absolute truth regarding what constitutes any of these alert types and what the business wants to be alerted to, as different stakeholders have different views and needs. Situations of type 2 were not deeply analysed, as they already partially reassure the quality of the tool.

To perceive which alert types the tool classified more accurately, the previous results were decomposed by alert type, as presented in Table 5.3:

Table 5.3: Tool Evaluation Case Distribution for each Alert Type

Case Type	Bad Alarms	Good Alarms	Bad Attentions
0	10%	20%	10%
1	10%	30%	40%
2	10%	30%	30%
3	70%	20%	20%

"Bad alarms" showed positive results, as for 70% of them the two stakeholders had the same classification as the tool. However, 30% of "good alarms" and 40% of "bad attentions" were poorly classified according to both stakeholders.

Based on the results, a conversation with the two stakeholders was held to discuss the main discrepancies between their opinions and the tool’s results, with the goal of finding the correct adjustments necessary to improve the rules in which the tool is built on.

Although the words “attention” and “alarm” have sometimes a negative connotation, these terms alone did not intend to be perceived as either negative or positive when building the tool. For example, what differentiates a “good attention” from a “good alarm” should be the intensity of how good its behaviour in the week being analysed is. However, the feedback from the stakeholders was that they struggled to differentiate between “good attentions” and “good alarms”. The suggestion they made was to use more business-oriented terms and, therefore, instead divide good alerts into the categories “recovering” and “over-performers”. Acknowledging that good alerts, according to the previously given definition, are KPIs which become more on-target in the week being analysed, then “recovering” would correspond to the good alerts which are, nevertheless, still off target, and “over-performers” the ones which are already on target. This suggestion was implemented and the denominations “good alarms” and “good attentions” were abolished from the App.

5.2.2.1 Type 1 Cases

Since these distinction between “good attentions” and “good alarms” was no longer considered from the user’s point of view, the type 1 situations regarding good alerts were not further discussed with the stakeholders, as they are now being displayed by the tool with the differentiation they asked for. However, to make this decision, it was essential to first notice that there were not any type 1 situations in which a good alert was classified as “no alert” by the participants. If that was the case, it should have been analysed further, since the distinction between “no alert” and having an alert is still considered by the tool. Only the KPIs with a good alert will be divided into “recovering” or “over-performer”. This new classification is represented in Figure 5.1.

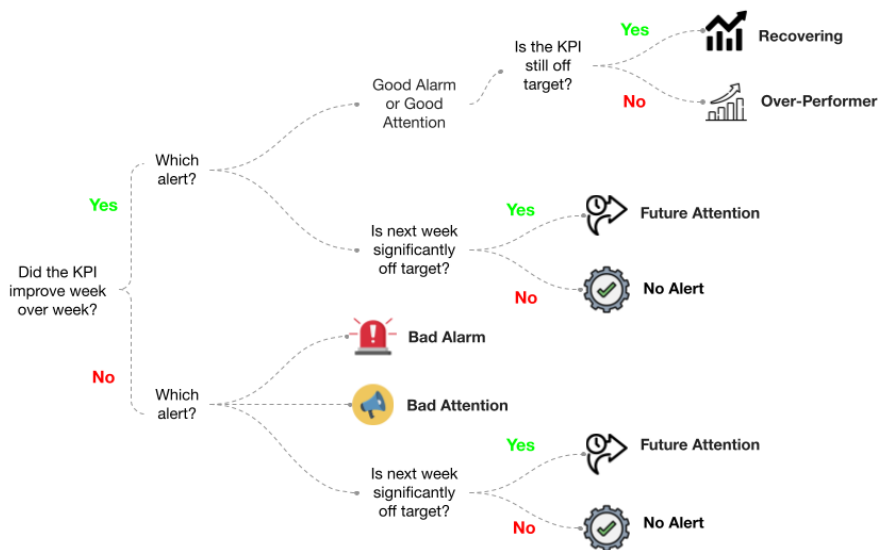


Figure 5.1: Alert Type Classification Overview

The remaining type 1 situations, which were only five, were discussed with the stakeholders to understand how to improve the tool in order to minimize the amount of these cases.

The tool classified both EDD Accuracy on week 15 and Speed of Sending of 2 Days on week 19 as “bad attention”, whereas the participants regarded them as “bad alarms”. For both, only the recent observations criterion was a red light, which meant another red light was needed for the KPIs to be considered an alarm by the tool. The stakeholders mentioned that one of the main reasons for their classification was how off target it was. However, according to the threshold function explained in Chapter 4, the target criterion would output a green light. It is important to note that, in recent years, both of these metrics consistently have targets above 90%, which means the output of the threshold function for them is much more influenced by a , than b . In order to have red lights according to the target criterion for both of these situations, the target threshold function, in which originally $a=3$, was changed to $a=2$, which improved the results of the tool for these situations. This change did not negatively affect the threshold function, as, on one hand, it made it more consistent with the business’s opinion and, on the other hand, for the metrics with the smallest targets the variation regarding absolute threshold induced by the change in the function translates into a maximum of a couple of hundredths, which is negligible.

For Time in Transit Global in week 15, the tool classified it as “bad attention”, as there were two green lights and the recent observations criterion was a red light, even though the stakeholders classified it as “no alert”. The aforementioned criterion was one tenth away from being a yellow light, which would instead make the KPI a “no alert”. Therefore, this result seems to point to making the threshold less conservative, but it would be necessary to obtain more similar cases to support this decision.

For Speed of Sending of 1 Day in week 12, the adjustments in the criteria thresholds in order for it to be a “bad attention” as the participants suggested and not a “bad alarm” would need to be significant and would affect negatively the classification of other metrics. Hence, this case was accepted as one in which the tool would not have given the exact classification the participants desired.

For No Stock in week 16 the tool showed “bad attention”, whereas the participants both replied “bad alarm”. Although the metric was significantly off target, this had been a pattern since the fifth week of the year, which is why there was not a red light regarding the recent observations criterion. Nevertheless, No Stock is one of the most important metrics for the business to keep track of due to its significant impact on customer experience and financially, which is why the stakeholders also felt an added layer of concern regarding the behaviour.

5.2.2.2 Type 0 Cases

Regarding type 0 cases, there were only two bad alerts. For No Stock in week 19, the tool had classified it as “attention”, whereas one of the stakeholders classified it as a “bad alarm” and the other as “no alert”. After some discussion about it, both stakeholders changed their minds and agreed on the “attention” classification.

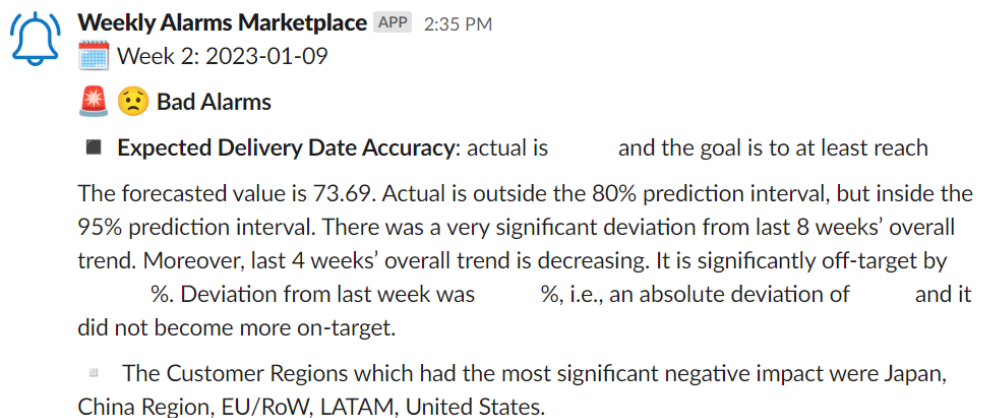
For Time to No Stock in week 20, the tool deemed it as a “bad alarm” and the participants as the two other alert types. For them, this metric is not one of the most important, since it does not have a direct impact financially, which, therefore, influenced their perception of the alert type.

5.2.3 Feedback Summary

Besides modifying the threshold function and the division that should be made within good alerts, the main add-on to the tool, according to the stakeholders’ feedback, would be for it to also consider how impactful a metric is when categorising it into alert types. Thresholds could be inherently less or more conservative according to the priority of the KPI. A standardised way to do this across all metrics should be defined and incorporated into the classification process.

5.3 Final Tool Results

The order in which alerts are shown are: “bad alarms”, “bad attentions”, “recovering”, “over-performers” and “future attentions”. The following figures are examples of alerts the Slack App gave for each alert type. They are examples from multiple weeks.



Weekly Alarms Marketplace APP 2:35 PM
 Week 2: 2023-01-09

Bad Alarms

- **Expected Delivery Date Accuracy:** actual is \dots and the goal is to at least reach \dots .
 The forecasted value is 73.69. Actual is outside the 80% prediction interval, but inside the 95% prediction interval. There was a very significant deviation from last 8 weeks’ overall trend. Moreover, last 4 weeks’ overall trend is decreasing. It is significantly off-target by \dots %. Deviation from last week was \dots %, i.e., an absolute deviation of \dots and it did not become more on-target.
- The Customer Regions which had the most significant negative impact were Japan, China Region, EU/RoW, LATAM, United States.

Figure 5.2: "Bad Alarm" Example

Bad Attentions

- **No Stock:** actual is 1.00 and the goal is to stay under 0.50 .

It is significantly off-target by 100% . Deviation from last week was 100% , i.e., an absolute deviation of 1.00 and it did not become more on-target.

- nonFxFF: actual was 1.00 , it was not on-target and it did become more on-target.
- FxFF: actual was 1.00 , it was on-target and it did not become more on-target.

- **Time In Transit Global:** actual is 1.00 and the goal is to stay under 0.50 .

There was a very significant deviation from last 8 weeks' overall trend. Moreover, last 4 weeks' overall trend is increasing. Deviation from last week was 100% , i.e., an absolute deviation of 1.00 and it did not become more on-target.

- The Customer Regions which had the most significant negative impact were Brazil.

Figure 5.3: "Bad Attention" Example

Good Alerts: Recovering

- **No Stock:** actual is 1.00 and the goal is to stay under 0.50 .

There was a very significant deviation from last 8 weeks' overall trend. Moreover, last 4 weeks' overall trend is decreasing. It is significantly off-target by 100% . Deviation from last week was 100% , i.e., an absolute deviation of 1.00 and it did become more on-target.

- nonFxFF: actual was 1.00 , it was not on-target and it did not become more on-target.
- FxFF: actual was 1.00 , it was on-target and it did become more on-target.

Figure 5.4: "Recovering" Example

Good Alerts: Over-Performers

- **Returns Transit Time:** actual is 4.95 and the goal is to stay under 4.00 .

The forecasted value is 4.95 . Actual is outside the 95% prediction interval. There was a very significant deviation from last 8 weeks' overall trend. Moreover, last 4 weeks' overall trend is decreasing. Deviation from last week was 100% , i.e., an absolute deviation of 1.00 and it did become more on-target.

Figure 5.5: "Over-Performer" Example

Future Attentions

- **Speed Of Sending 1 Day:** actual is 73.15 and the goal is to at least reach 70.00 .

Criteria indicate not concerning behavior, but next week it is predicted it will become significantly off-target. Its forecasted value is 73.15 and the target will be 70.00 .

Figure 5.6: "Future Attention" Example

Chapter 6

Conclusion and Future Work

The fiercely competitive e-commerce market is expanding quickly, posing problems for managing services, such as delivery and fraud detection. To achieve operational excellence, businesses must monitor KPIs, but as the e-commerce's complexity keeps increasing, it becomes more difficult to identify and address problems. Yet, there is still not a standard procedure for monitoring KPIs in the context of online marketplaces. E-commerce marketplaces need to constantly assure the improvement of their operational performance, while guaranteeing a high standard for customer experience. This is even more important in Farfetch's business model not only because its focus is not manufacturing the items which are sold in its platform, so it must assure the purchase experience goes as smoothly as possible, but also because its customers expect services with as much quality as the items they buy.

Hence, the purpose of this project was to create a tool that could rapidly convey to the stakeholders which KPIs they should focus on each week for Marketplace and that could send useful information about their behaviour, by basing this selection on standardised criteria. It should also provide warnings regarding the following week's KPI behaviour, based on the forecasting models' predictions. This knowledge is especially useful in the weekly operations tactical meetings, as the present stakeholders have to cover an extensive scope in just one hour.

The first step of this project was to decide the criteria which the tool would base its judgements regarding the KPIs' performance on. The current monitoring processes of the metrics being considered were studied, in order to extract best practices and to understand which could be the standard methodology across them to evaluate each KPI on a weekly basis.

The proposed criteria are "is the KPI's significantly off target?", "is the KPI an outlier according to the recent past?" and "is the actual inside a certain prediction interval according to the forecasting model?", which were named the target criterion, the recent observations criterion and the prediction interval criterion, respectively. When a KPI is evaluated according to the first criterion, it can either result in a green or red light, whereas when it is assessed according to the two other criteria it can result in a green, yellow or red light. Through the combination of the light results of the three criteria, KPIs' behaviour was either considered to be an "alarm" or an "attention": an alarm is characterised by having, at least, two red lights, whereas an attention is

characterised by only having one red light or two yellow lights. Moreover, an alarm and an attention can either be “good” or “bad”: what makes it good is if the KPI, compared to the week prior, became more on-target, which does not imply that it is on-target, it just improved its performance. Depending on the nature of the KPI, its goal may either be to, at least, reach its target or to stay below it. This detail had implications in the code behind the tool created.

Bad alerts are divided and showcased in the tool as “bad alarms” and “bad attentions”, whereas good alerts are instead divided into “recovering” and “over-performers”. Those recovering are the ones still off target and those over-performing are already on target, both categories sharing the fact that the KPI’s behaviour showed improvement compared to the prior week. Thus, the user is not shown “good alarms” and “good attentions”, as the final division corresponds to more business-oriented outputs.

The target criterion considers a KPI to be significantly off-target when it is off-target and its relative deviation from target is over a threshold defined by a function, since merely being slightly off-target is not enough reason for there to be a red light regarding this criterion. This buffer, which enables a KPI to be slightly off-target, but still be, when judged by this criterion, a green light, exists due to the inherent error to targets due to them being designed by humans, instead of being a characteristic of the data itself.

The recent observations criterion determines if the week being analysed is an outlier when compared to the eight weeks prior to it, through the MAD outlier method, which was chosen due to its robustness when detecting outliers in data regardless of their distribution and when used in small datasets like this one.

The prediction interval criterion assesses if the KPI’s actual is in accordance with the prediction intervals calculated with the corresponding best forecasting model. Basing a criterion on forecasting models added value to how KPIs are monitored, since only a few of the KPIs considered in this thesis are currently forecasted by Farfetch, but on a monthly basis, which means these forecasts were not being considered when monitoring them weekly.

To determine the best forecasting model for each, cross-validation with rolling origin was applied to calculate the average MASE and MAPE when testing single exponential smoothing, double exponential smoothing, Holt-Winters with additive seasonality, Holt-Winters with multiplicative seasonality, Prophet with additive seasonality and Prophet with multiplicative seasonality. All of these were compared to evaluate models with and without a seasonality component. The chosen model for each was the one with the smallest average MASE, which had to mandatorily be inferior to 1 for the one-week-ahead forecast, as this is the use case depicted in this thesis. For this timeframe, most KPIs had an average MAPE lower than 7%, although some had up to 13%. Regarding the average MASE in the same forecast timeframe, it was always lower than 1, but most of the KPIs presented results lower than 0.9. Moreover, the model which was most chosen was Prophet with additive seasonality.

When according to all these criteria, there is not an alert for a KPI, the future target criterion is tested. This criterion is similar to the target criterion, the only difference is that it compares the

forecasted value with the target regarding the following week, instead of the week being analysed. The name of the alert when there is a red light in this criterion is “future attention”.

A dimension is a way to divide the main KPI in further logical parts. Each KPI was analysed through a dimension when there was an attention or an alarm. In order to determine which dimension groups are the best and worst contributors to each KPI overall, depending on whether it improved or worsened relative to the prior week, the "MAD-Delta" approach is introduced.

Texts with the main takeaways about the KPIs' performance are sent through a Slack App, in which they are shown divided by alert type.

Regarding results, the quality and relevance of the information shown for each alert type has been confirmed by the feedback from the Operations Performance & Optimization. However, the other aspect that needed to be assessed was if the alerts that were sent were relevant or not and if they were properly classified as “attention” or “alert” according to the stakeholders' standards. Therefore, thirty scenarios with ten “bad alarms”, ten “good alarms” and ten “bad attentions”, i.e., thirty situations in which certain KPIs in a certain week showed these alerts were selected, having mostly a fair representation of all the KPIs analysed. The recent KPI data for the situation, the target and the forecasted value for the week in which the KPI had an alert were shown to two different stakeholders who are well familiarised with this set of KPIs in the Marketplace context. They then individually classified them as alarms or attentions and an analysis of the results was made. Through this evaluation, the main conclusions were that for, approximately, 37% of the situations the stakeholders both agreed with the tool's classification and that for, approximately, 23% of them one of the stakeholders agreed with it. This not only shows the quality of the tool, but also how hard it is to create a tool that classifies the alerts properly, since not all stakeholders would categorise them in the same way. Moreover, “bad alarms” was the alert type which was more accurately depicted, as in 7 of the 10 cases both stakeholders agreed with this classification and that the then division of good alerts into “good alarms” and “good attentions” was not intuitive. To improve the quality of the tool, the new division of “recovering” and “over-performers” for these alerts was implemented and the target threshold function was slightly changed.

Having a tool that enables stakeholders to receive alerts from multiple service areas has created a way to prioritize KPIs to be discussed inside and outside of the tactical meeting and has included some that may be neglected as they are never discussed in the meeting for lack of time. Moreover, the thresholds for each criterion are adjustable, which enables Farfetch to adapt them according to its needs. The traffic light framework allows for an easy understanding of the tool, which facilitates this customisation.

Hence, this project led to the creation of a tool which enables businesses to have an overview of their weekly operational performance, through the prioritization of their KPIs' behaviour in a detailed, but not overwhelming fashion. This tool contributes to the literature by combining machine learning with simple, but effective techniques, taking into account both historical data and business requirements, to deliver alerts with business-oriented classification and content.

Regarding future work, for the Farfetch use case, in order for the prediction intervals criterion to still have meaningful outputs, cross-validation to test the multiple models should be done peri-

odically for each KPI, including the most recent data, to ensure the models used are still the most adequate. Additionally, the prioritization of the KPIs' behaviour each week should be assisted by how impactful the deviation from the week before is, regarding customer experience and/or financially. Deviation in all the metrics was considered to be equally as impactful, even though the negative impact of a client receiving his order in more than two days is not as severe as the one of being sent a wrong item. Furthermore, an add-on to this tool could be a dependency model, where the underlying dependencies between the multiple KPIs could be determined and taken into consideration for the definition of future attentions. Moreover, to improve the performance of the forecasting models, events as holidays and past promotions according to the region could be considered. Finally, although the thresholds for each criterion can be adjusted, further studies can be done not only about the values these should take on, but also regarding which light combinations should lead to an "attention" or an "alarm".

Bibliography

- Archambault, S. G., Helouvry, J., Strohl, B., and Williams, G. (2015). Data visualization as a communication tool. *Library Hi Tech News*, 32:1–9.
- Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33.
- Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.
- Castellanos, M., Casati, F., Dayal, U., and Shan, M.-C. (2004). A comprehensive and automated approach to intelligent business processes execution analysis. *Distributed and Parallel Databases*, 16:239–273.
- Castellanos, M., Casati, F., Shan, M. C., and Dayal, U. (2005). ibom: A platform for intelligent business operation management. pages 1084–1095.
- Chatfield, C. (2001). Prediction intervals for time-series forecasting. *Principles of forecasting: A handbook for researchers and practitioners*, pages 475–494.
- Chen, Y. and Jin, J. (2006). Quality-oriented-maintenance for multiple interactive tooling components in discrete manufacturing processes. *IEEE Transactions on Reliability*, 55(1):123–134.
- Chia, A. (2019). Distilling the essence of the mckinsey way: The problem-solving cycle. *Management Teaching Review*, 4(4):355–370.
- Cousineau, D. and Chartier, S. (2010). Outliers detection and treatment: a review. *International Journal of Psychological Research*, 3(1):58–67.
- DataGenie (2023). Why datagenie for augmented intelligence. Retrieved from = <https://www.datagenie.ai/overview>. Accessed: May 3, 2023.
- Dave, D. and Varma, T. (2014). A review of various statistical methods for outlier detection. *International Journal of Computer Science & Engineering Technology (IJCSET)*, 5(2):137–140.
- de Moraes, R. M., Kazan, S., de Pádua, S. I. D., and Costa, A. L. (2014). An analysis of bpm lifecycles: From a literature review to a framework proposal. *Business Process Management Journal*, 20:412–432.
- Eckerson, W. W. (2010). *Performance dashboards: measuring, monitoring, and managing your business*. John Wiley & Sons.
- Elzinga, D., Horak, T., Lee, C.-Y., and Bruner, C. (1995). Business process management: survey and methodology. *IEEE Transactions on Engineering Management*, 42(2):119–128.

- Erdmann, A. and Ponzoa, J. M. (2021). Digital inbound marketing: Measuring the economic performance of grocery e-commerce in europe and the usa. *Technological forecasting and social change*, 162:120373.
- Fisher, O. J., Watson, N. J., Escrig, J. E., Witt, R., Porcu, L., Bacon, D., Rigley, M., and Gomes, R. L. (2020). Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Computers & Chemical Engineering*, 140:106881.
- Forbes (2023). Ecommerce statistics: The latest data and future projections [updated]. Accessed: April 30, 2023.
- Gillot, J.-N. (2008). *The Complete Guide to Business Process Management: Business process transformation or a way of aligning the strategic objectives of the company and the information system through the processes*. Lulu. com.
- Golfarelli, M., Rizzi, S., and Cella, I. (2004). Beyond data warehousing: What's next in business intelligence? In *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, DOLAP '04*, page 1–6, New York, NY, USA. Association for Computing Machinery.
- Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., and Shan, M. C. (2004). Business process intelligence. *Computers in Industry*, 53:321–343.
- Hagiu, A. and Wright, J. (2015). Marketplace or reseller? *Management Science*, 61(1):184–203.
- Hammer, M. and Champy, J. (2009). *Reengineering the corporation: Manifesto for business revolution, a*. Zondervan.
- Hawkins, D. M. (1980). *Identification of outliers*, volume 11. Springer.
- Hjort, K. and Lantz, B. (2016). The impact of returns policies on profitability: A fashion e-commerce case. *Journal of Business Research*, 69(11):4980–4985.
- Hodge, V. and Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22:85–126.
- Huber, P. J. (2011). Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer.
- Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.
- Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.
- IBM (2021). Ibm business process manager overview. Retrieved from <https://www.ibm.com/docs/en/bpm/8.5.5?topic=manager-business-process-overview>. Accessed: May 5, 2023.
- Izadi, I., Shah, S. L., Shook, D. S., Kondaveeti, S. R., and Chen, T. (2009). A framework for optimal design of alarm systems. *IFAC Proceedings Volumes*, 42(8):651–656.
- Janes, A., Sillitti, A., and Succi, G. (2013). Effective dashboard design. *Cutter IT Journal*, 26(1):17–24.

- Jeng, J.-J., Schiefer, J., and Chang, H. (2003). An agent-based architecture for analyzing business processes of real-time enterprises. In *Seventh IEEE International Enterprise Distributed Object Computing Conference, 2003. Proceedings.*, pages 86–97. IEEE.
- Jiang, W., Au, T., and Tsui, K.-L. (2007). A statistical process control approach to business activity monitoring. *Iie Transactions*, 39(3):235–249.
- Kalekar, P. S. et al. (2004). Time series forecasting using holt-winters exponential smoothing. *Kanwal Rekhi school of information Technology*, 4329008(13):1–13.
- Kneese, T. and Palm, M. (2020). Brick-and-platform: Listing labor in the digital vintage economy. *Social Media+ Society*, 6(3):2056305120933299.
- Kolassa, S. and Martin, R. (2011). Percentage errors can ruin your day (and rolling the dice shows how). *Foresight: The International Journal of Applied Forecasting*, (23).
- Lebas, M. J. (1995). Performance measurement and performance management. *International journal of production economics*, 41(1-3):23–35.
- Leymann, F., Roller, D., and Schmidt, M.-T. (2002). Web services and business process management. *IBM Systems Journal*, 41(2):198–211.
- Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49:764–766.
- Miller, J. (1991). Reaction time analysis with outlier exclusion: Bias varies with sample size. *The quarterly journal of experimental psychology*, 43(4):907–912.
- Nanehkaran, Y. A. (2013). An introduction to electronic commerce. *International journal of scientific & technology research*, 2(4):190–193.
- Nesamoney, D. (2004). Bam: Event-driven business intelligence for the real-time enterprise. *Information Management*, 14(3):38.
- Oracle (2023). Oracle business activity monitoring. Retrieved from = <https://www.oracle.com/middleware/technologies/business-activity-monitoring.html>. Accessed: May 6, 2023.
- Parmenter, D. (2015). *The Great KPI Misunderstanding*, chapter 1, pages 1–23. John Wiley & Sons, Ltd.
- Peral, J., Maté, A., and Marco, M. (2017). Application of data mining techniques to identify relevant key performance indicators. *Computer Standards and Interfaces*, 54:76–85.
- Popova, V. and Sharpanskykh, A. (2010). Modeling organizational performance indicators. *Information Systems*, 35:505–527.
- Research, G. V. (2023). E-commerce logistics market size, share & trends analysis report. Accessed: April 30, 2023.
- Sayal, M., Casati, F., Dayal, U., and Shan, M.-C. (2002). Business process cockpit. pages 880–883.

- Svetunkov, I. and Petropoulos, F. (2018). Old dog, new tricks: a modelling view of simple moving averages. *International Journal of Production Research*, 56(18):6034–6047.
- Taylor, S. J. and Letham, B. (2017). Forecasting at scale. *The American Statistician*, 72(1):37–45.
- Teinemaa, I. and Depaire, B. (2019). Predictive and prescriptive monitoring of business process outcomes. In *BPM (PhD/Demos)*, pages 15–19.
- Teinemaa, I., Tax, N., de Leoni, M., Dumas, M., and Maggi, F. M. (2018). Alarm-based prescriptive process monitoring. In *Business Process Management Forum: BPM Forum 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings 16*, pages 91–107. Springer.
- Wang, J., Yang, F., Chen, T., and Shah, S. L. (2015). An overview of industrial alarm systems: Main causes for alarm overloading, research status, and open problems. *IEEE Transactions on Automation Science and Engineering*, 13(2):1045–1061.
- Weber, A. and Thomas, R. (2005). Key performance indicators. *Measuring and Managing the Maintenance Function*, Ivara Corporation, Burlington.
- Wetzstein, B., Ma, Z., and Leymann, F. (2008). Towards measuring key performance indicators of semantic business processes.
- Wouters, M. (2009). A developmental approach to performance measures-results from a longitudinal case study. *European Management Journal*, 27:64–78.
- Xu, J., Wang, J., Izadi, I., and Chen, T. (2011). Performance assessment and design for univariate alarm systems based on far, mar, and aad. *IEEE Transactions on Automation Science and Engineering*, 9(2):296–307.
- Yigitbasioglu, O. M. and Velcu, O. (2012). A review of dashboards in performance management: Implications for design and research. *International Journal of Accounting Information Systems*, 13:41–59.
- Yu, Y., Zhu, Y., Li, S., and Wan, D. (2014). Time series outlier detection based on sliding window prediction. *Mathematical Problems in Engineering*, 2014.
- Zur Muehlen, M. (2001). Process-driven management information systems combining data warehouses and workflow technology. In *Proceedings of the International Conference on Electronic Commerce Research (ICECR-4)*, pages 550–566. Citeseer.
- Zur Muehlen, M. and Rosemann, M. (2000). Workflow-based process monitoring and controlling-technical and organizational issues. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pages 10–pp. IEEE.

Appendix A

Decomposition Plots

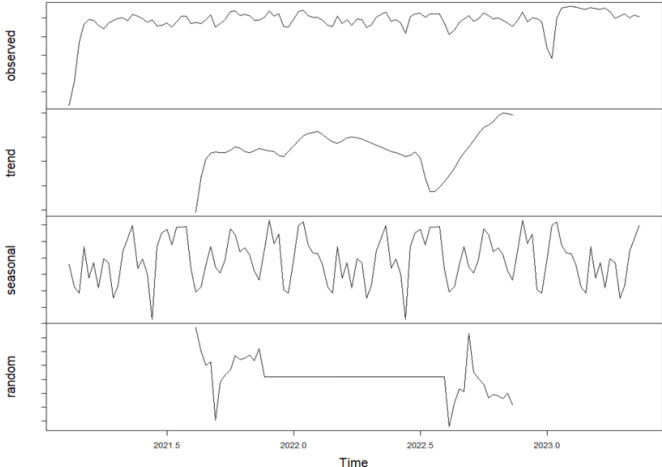


Figure A.1: Decomposition Plot of EDD

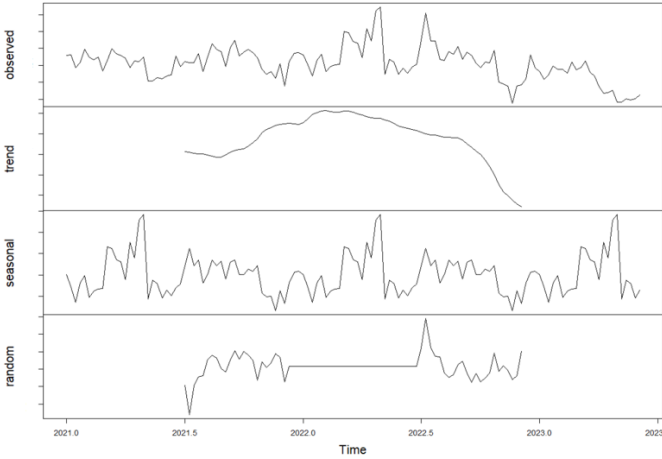


Figure A.2: Decomposition Plot of FRR

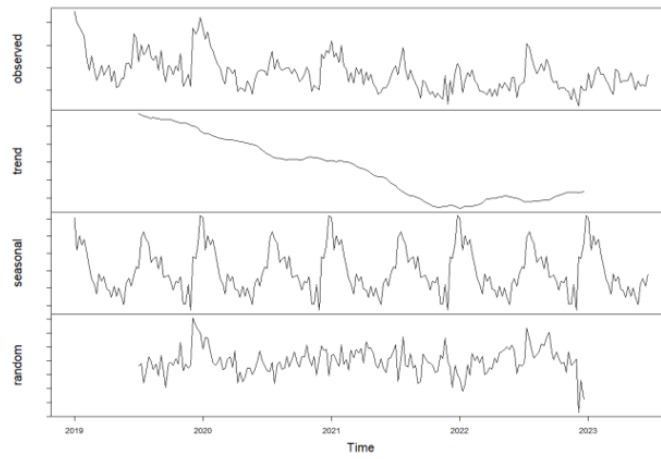


Figure A.3: Decomposition Plot of NS

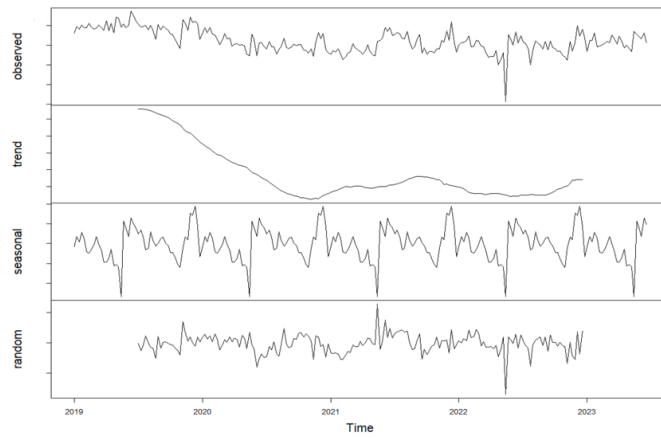


Figure A.4: Decomposition Plot of PACR

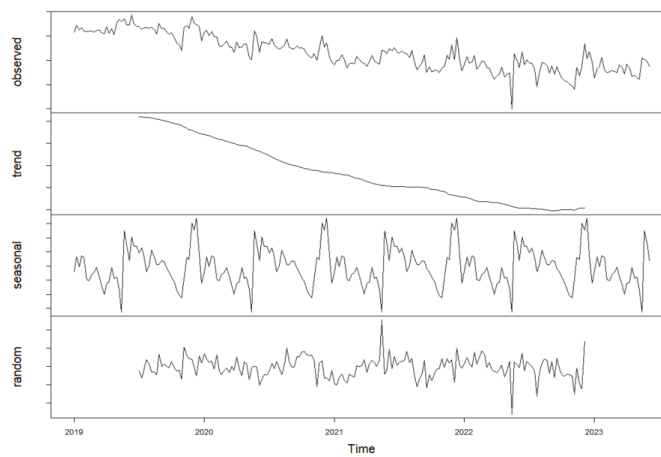


Figure A.5: Decomposition Plot of PCR

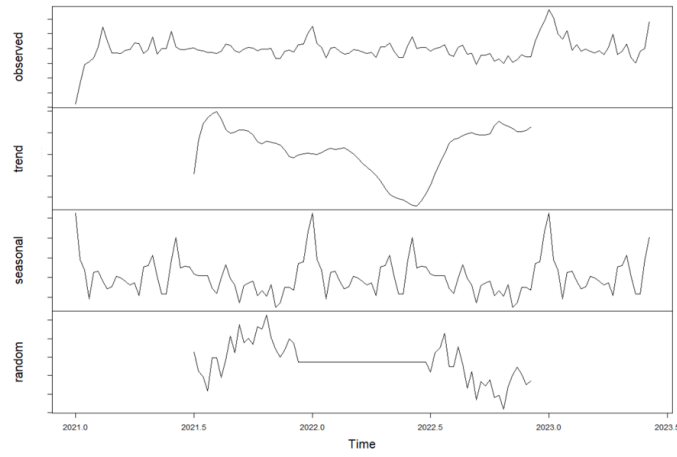


Figure A.6: Decomposition Plot of RTT

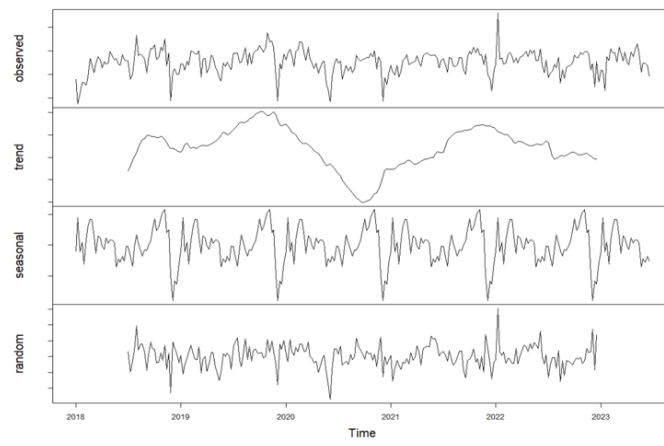


Figure A.7: Decomposition Plot of SoS1D

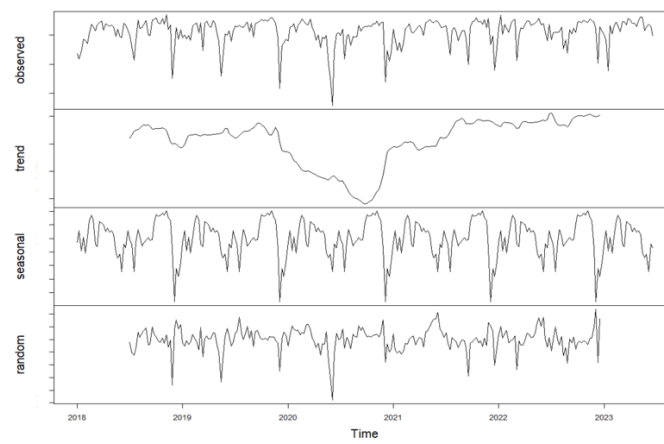


Figure A.8: Decomposition Plot of SoS2D

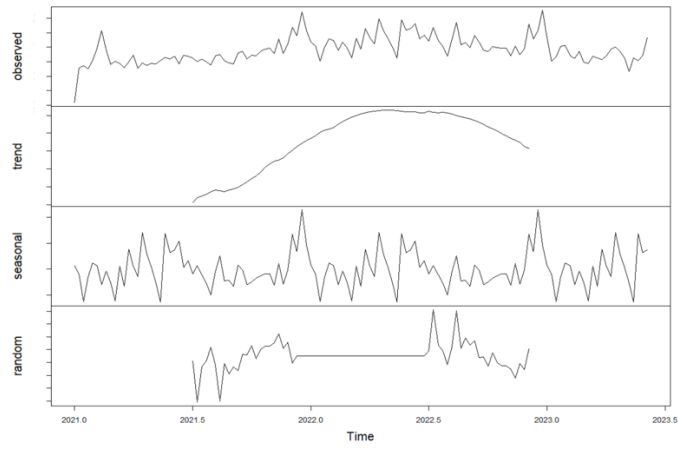


Figure A.9: Decomposition Plot of TITG

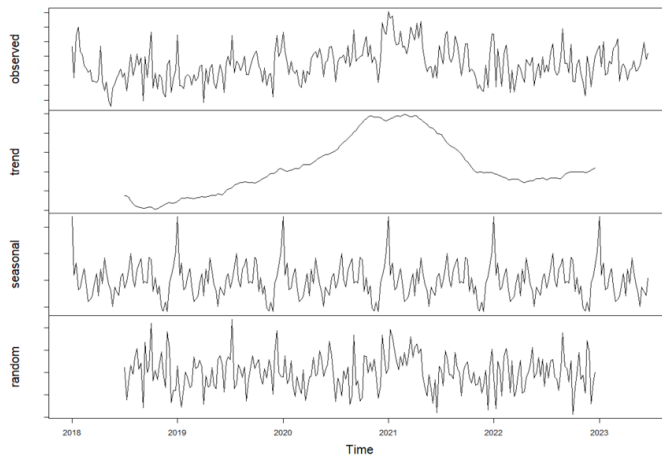


Figure A.10: Decomposition Plot of TTNS

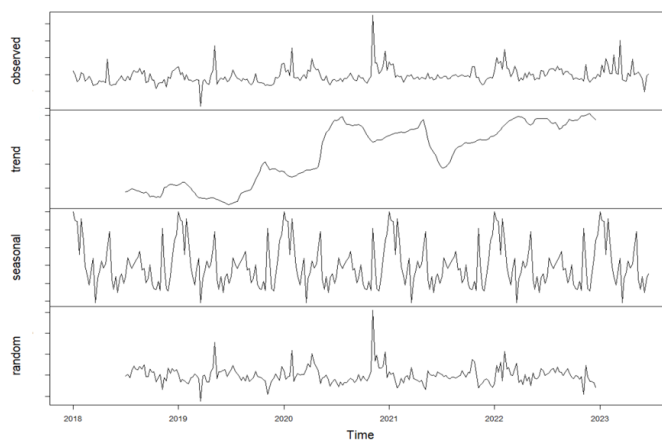


Figure A.11: Decomposition Plot of TTPR

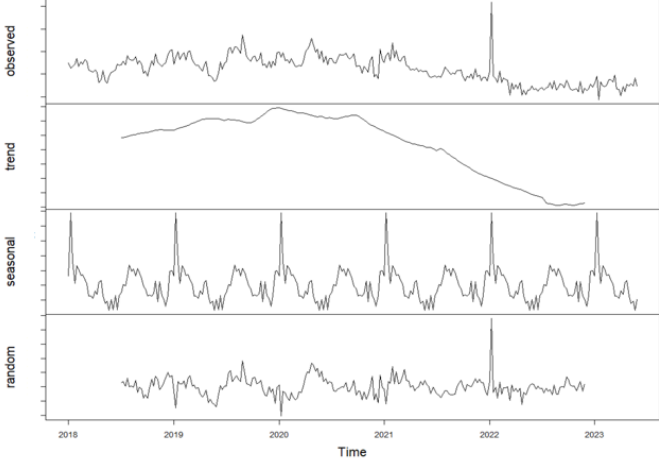


Figure A.12: Decomposition Plot of WI

Appendix B

Code-Overview

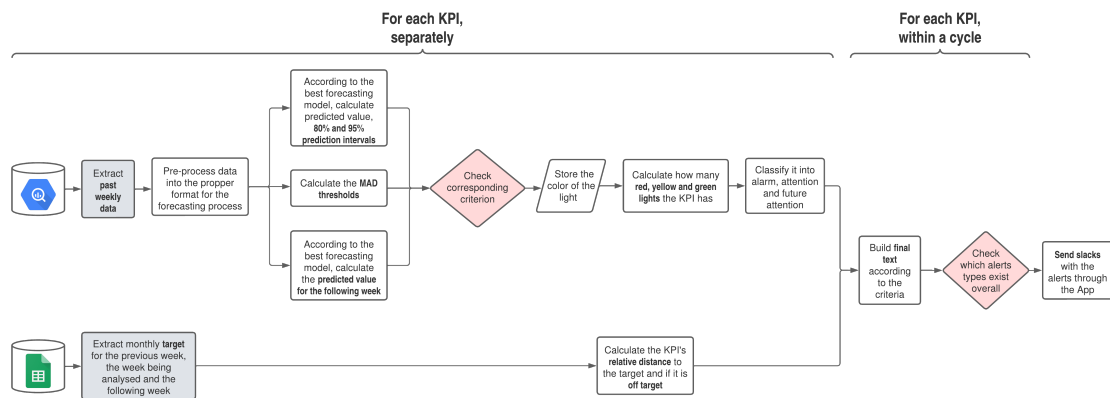


Figure B.1: Code Overview

Appendix C

Cross-Validation Results

Table C.1: Cross-Validation Results for EDD

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.973	1.047	0.975
	MAPE	1.581	1.833	1.790
Double Exponential Smoothing	MASE	1.111	1.244	1.589
	MAPE	1.805	2.166	2.904
Holt Winters – Additive Seasonality	MASE	2.552	2.415	2.564
	MAPE	4.107	4.227	4.663
Holt Winters – Multiplicative Seasonality	MASE	2.7	2.543	2.694
	MAPE	4.342	4.45	4.897
Prophet – Additive Seasonality	MASE	1.248	1.134	1.424
	MAPE	2.002	1.968	2.594
Prophet – Multiplicative Seasonality	MASE	1.248	1.165	1.417
	MAPE	1.999	2.018	2.578

Table C.2: Cross-Validation Results for FRR

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	1.249	1.986	2.712
	MAPE	15.056	20.738	29.292
Double Exponential Smoothing	MASE	1.297	2.160	2.915
	MAPE	15.716	22.498	31.476
Holt Winters – Additive Seasonality	MASE	3.327	6.113	7.224
	MAPE	47.941	72.272	84.504
Holt Winters – Multiplicative Seasonality	MASE	1.582	3.135	4.004
	MAPE	21.137	34.677	44.211
Prophet – Additive Seasonality	MASE	3.342	4.704	4.867
	MAPE	42.178	50.075	53.913
Prophet – Multiplicative Seasonality	MASE	3.073	4.252	4.372
	MAPE	39.121	45.161	48.497

Table C.3: Cross-Validation Results for NS

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.930	1.092	1.247
	MAPE	12.793	15.088	17.226
Double Exponential Smoothing	MASE	0.958	1.166	1.365
	MAPE	12.967	15.923	18.418
Holt Winters – Additive Seasonality	MASE	1.264	1.577	1.825
	MAPE	18.462	22.261	25.418
Holt Winters – Multiplicative Seasonality	MASE	1.222	1.448	1.608
	MAPE	17.607	20.520	22.539
Prophet – Additive Seasonality	MASE	1.460	1.637	1.745
	MAPE	19.638	21.903	23.256
Prophet – Multiplicative Seasonality	MASE	1.399	1.547	1.643
	MAPE	18.937	20.886	22.116

Table C.4: Cross-Validation Results for PACR

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.894	1.028	1.143
	MAPE	2.359	2.722	3.028
Double Exponential Smoothing	MASE	0.903	1.096	1.348
	MAPE	2.383	2.904	3.575
Holt Winters – Additive Seasonality	MASE	1.002	0.967	0.968
	MAPE	2.635	2.547	2.555
Holt Winters – Multiplicative Seasonality	MASE	0.989	0.961	0.955
	MAPE	2.600	2.529	2.519
Prophet – Additive Seasonality	MASE	0.858	0.889	0.913
	MAPE	2.243	2.325	2.392
Prophet – Multiplicative Seasonality	MASE	0.852	0.880	0.917
	MAPE	2.226	2.301	2.404

Table C.5: Cross-Validation Results for PCR

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.951	1.166	1.332
	MAPE	1.708	2.1	2.392
Double Exponential Smoothing	MASE	0.939	1.189	1.480
	MAPE	1.682	2.139	2.651
Holt Winters – Additive Seasonality	MASE	1.013	1.022	0.991
	MAPE	1.812	1.830	1.770
Holt Winters – Multiplicative Seasonality	MASE	0.999	1.007	0.980
	MAPE	1.787	1.803	1.750
Prophet – Additive Seasonality	MASE	0.883	0.956	0.998
	MAPE	1.580	1.715	1.786
Prophet – Multiplicative Seasonality	MASE	0.891	0.962	1.003
	MAPE	1.595	1.726	1.794

Table C.6: Cross-Validation Results for RTT

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.994	1.029	0.963
	MAPE	9.951	9.605	9.249
Double Exponential Smoothing	MASE	1.079	1.201	1.206
	MAPE	10.766	11.170	11.392
Holt Winters – Additive Seasonality	MASE	0.972	1.224	1.287
	MAPE	9.876	11.533	12.082
Holt Winters – Multiplicative Seasonality	MASE	1.118	1.426	1.523
	MAPE	11.265	13.464	14.303
Prophet – Additive Seasonality	MASE	0.991	0.944	1.008
	MAPE	9.686	9.083	9.985
Prophet – Multiplicative Seasonality	MASE	1.284	1.675	1.779
	MAPE	13.035	16.118	17.363

Table C.7: Cross-Validation Results for SoS1D

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.833	0.904	1.064
	MAPE	4.411	4.857	5.624
Double Exponential Smoothing	MASE	0.937	0.978	1.227
	MAPE	4.952	5.241	6.466
Holt Winters – Additive Seasonality	MASE	1.020	1.034	1.037
	MAPE	5.381	5.546	5.488
Holt Winters – Multiplicative Seasonality	MASE	1.003	1.010	1.006
	MAPE	5.301	5.426	5.341
Prophet – Additive Seasonality	MASE	0.847	0.882	0.880
	MAPE	4.472	4.736	4.671
Prophet – Multiplicative Seasonality	MASE	0.840	0.887	0.877
	MAPE	4.435	4.760	4.651

Table C.8: Cross-Validation Results for SoS2D

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.877	0.965	1.062
	MAPE	1.274	1.447	1.572
Double Exponential Smoothing	MASE	0.9	0.994	1.112
	MAPE	1.307	1.490	1.647
Holt Winters – Additive Seasonality	MASE	1.198	1.162	1.128
	MAPE	1.734	1.735	1.666
Holt Winters – Multiplicative Seasonality	MASE	1.152	1.118	1.088
	MAPE	1.669	1.672	1.609
Prophet – Additive Seasonality	MASE	0.895	0.887	0.868
	MAPE	1.301	1.330	1.289
Prophet – Multiplicative Seasonality	MASE	0.901	0.890	0.870
	MAPE	1.309	1.335	1.291

Table C.9: Cross-Validation Results for TITG

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	1.023	1.405	1.126
	MAPE	4.432	6.843	6.848
Double Exponential Smoothing	MASE	1.150	1.778	1.636
	MAPE	4.952	8.578	9.990
Holt Winters – Additive Seasonality	MASE	1.376	1.248	1.215
	MAPE	5.950	5.830	7.215
Holt Winters – Multiplicative Seasonality	MASE	1.328	1.189	1.051
	MAPE	5.758	5.603	6.285
Prophet – Additive Seasonality	MASE	0.961	1.130	0.807
	MAPE	4.273	5.559	4.973
Prophet – Multiplicative Seasonality	MASE	1.001	1.134	0.809
	MAPE	4.444	5.588	5.007

Table C.10: Cross-Validation Results for TTNS

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.870	0.882	0.886
	MAPE	12.835	12.924	12.762
Double Exponential Smoothing	MASE	0.974	1.083	1.134
	MAPE	14.436	15.949	16.450
Holt Winters – Additive Seasonality	MASE	0.801	0.847	0.881
	MAPE	11.871	12.424	12.757
Holt Winters – Multiplicative Seasonality	MASE	0.821	0.873	0.898
	MAPE	12.183	12.820	13.024
Prophet – Additive Seasonality	MASE	0.767	0.763	0.784
	MAPE	10.947	10.768	10.917
Prophet – Multiplicative Seasonality	MASE	0.806	0.809	0.833
	MAPE	11.313	11.209	11.406

Table C.11: Cross-Validation Results for TTPR

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.832	0.881	0.925
	MAPE	13.991	15.156	15.717
Double Exponential Smoothing	MASE	0.856	0.924	0.973
	MAPE	14.405	16.050	16.711
Holt Winters – Additive Seasonality	MASE	0.937	0.962	1.061
	MAPE	15.736	16.337	17.954
Holt Winters – Multiplicative Seasonality	MASE	0.914	0.928	0.946
	MAPE	15.204	15.583	15.632
Prophet – Additive Seasonality	MASE	0.817	0.824	0.850
	MAPE	13.747	14.007	14.270
Prophet – Multiplicative Seasonality	MASE	0.826	0.830	0.856
	MAPE	13.866	14.085	14.357

Table C.12: Cross-Validation Results for WI

Models	Performance Metric	h=1	h=2	h=3
Single Exponential Smoothing	MASE	0.789	0.785	0.857
	MAPE	6.843	6.860	7.702
Double Exponential Smoothing	MASE	0.833	0.831	0.970
	MAPE	7.223	7.263	8.740
Holt Winters – Additive Seasonality	MASE	1.218	1.421	1.506
	MAPE	10.234	12.139	13.265
Holt Winters – Multiplicative Seasonality	MASE	1.025	1.161	1.240
	MAPE	8.681	9.987	11.019
Prophet – Additive Seasonality	MASE	1.138	1.229	1.276
	MAPE	9.532	10.405	11.126
Prophet – Multiplicative Seasonality	MASE	1.068	1.126	1.158
	MAPE	8.878	9.443	9.995

Appendix D

Prediction Intervals Assumptions Testing

D.1 Residuals' Histograms

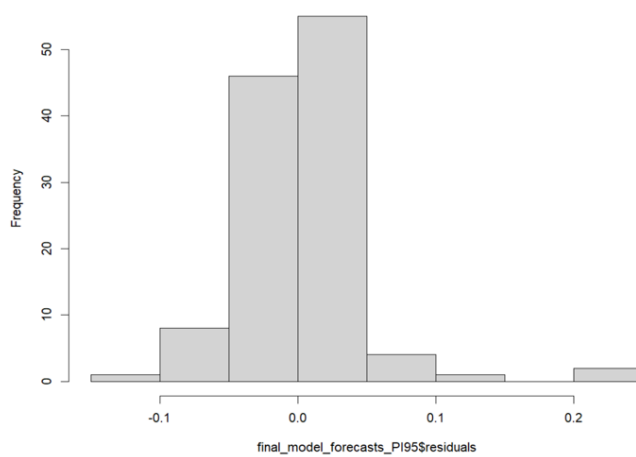


Figure D.1: Histogram of EDD's Residuals

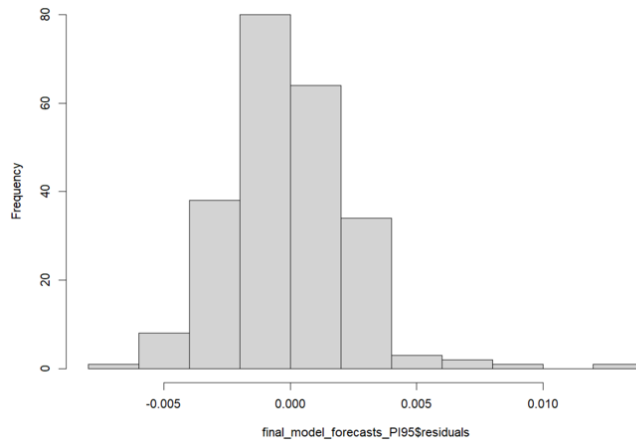


Figure D.2: Histogram of NS' Residuals

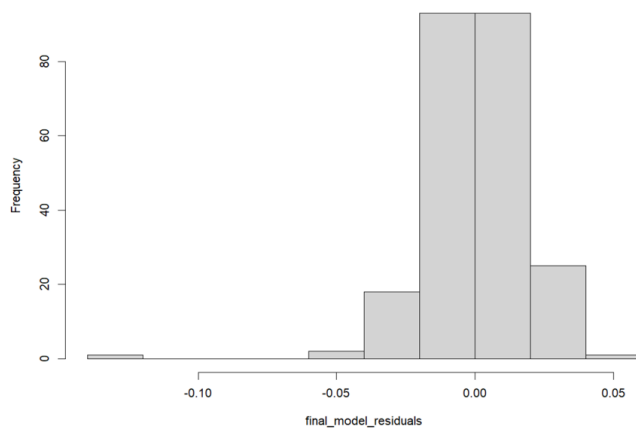


Figure D.3: Histogram of PACR's Residuals

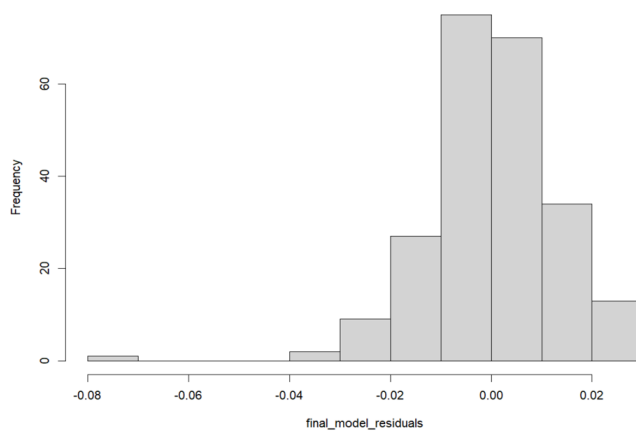


Figure D.4: Histogram of PCR's Residuals

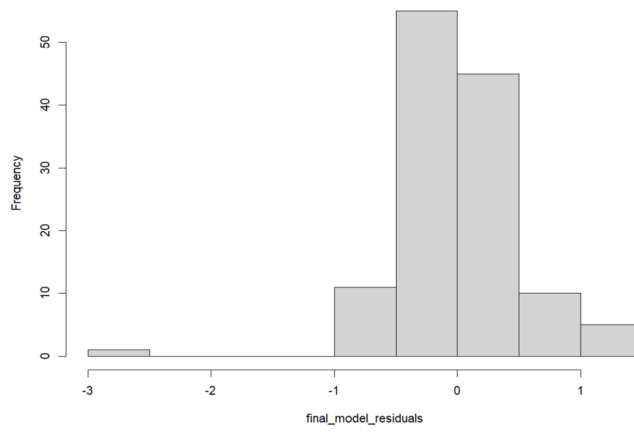


Figure D.5: Histogram of RTT's Residuals

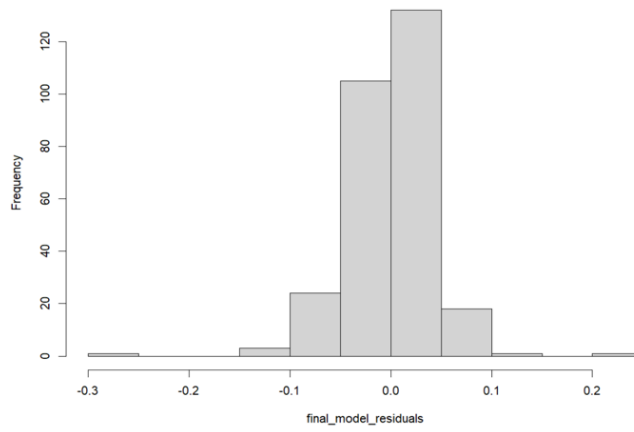


Figure D.6: Histogram of SoS1D's Residuals

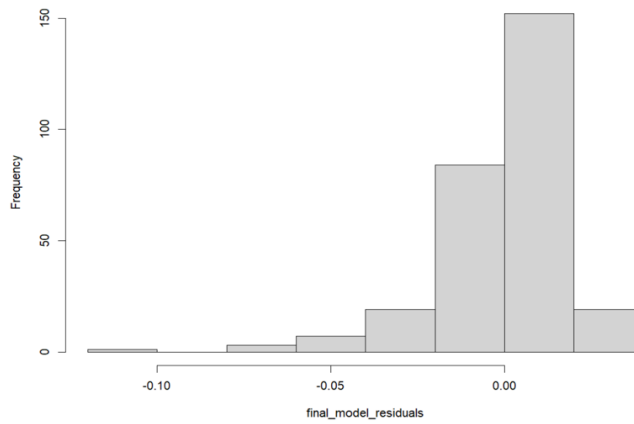


Figure D.7: Histogram of SoS2D's Residuals

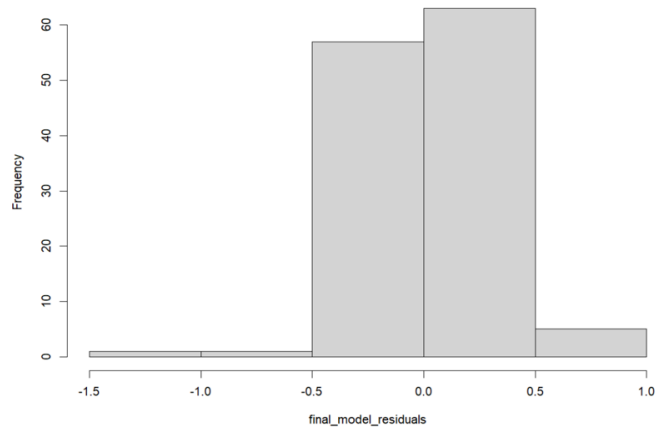


Figure D.8: Histogram of TITG's Residuals

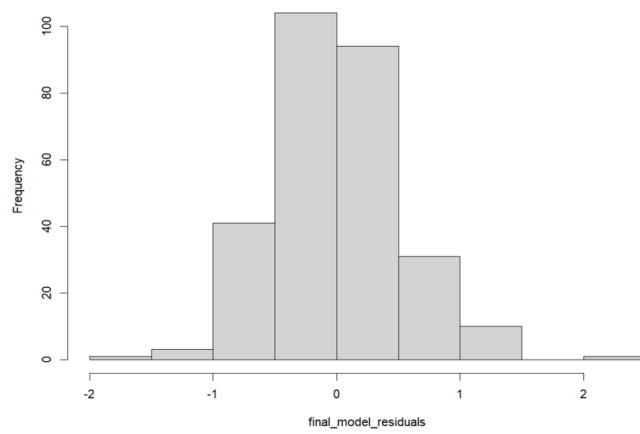


Figure D.9: Histogram of TTNS' Residuals

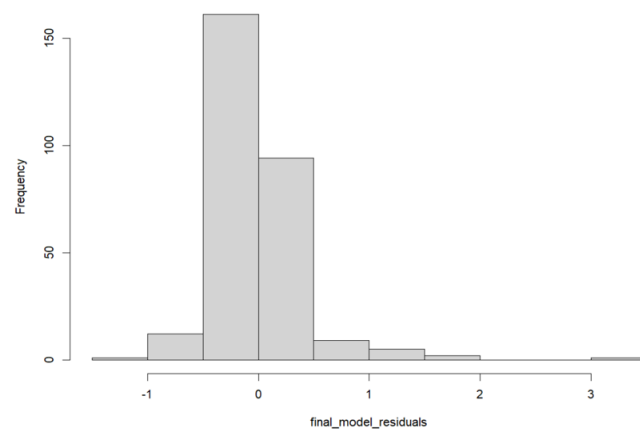


Figure D.10: Histogram of TTPR's Residuals

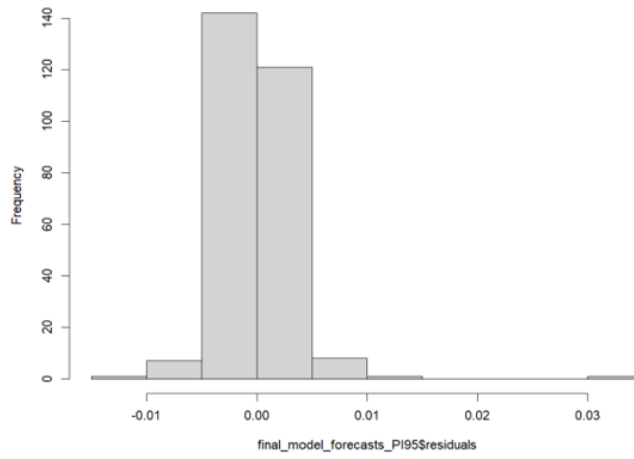


Figure D.11: Histogram of WI's Residuals

D.2 Results of the Kolmogorov-Smirnov Normality Test for Residuals

Table D.1: Results of the Kolmogorov-Smirnov Normality Test for Residuals

KPI	Statistic	P-Value
EDD	0.1425	0.0172
NS	0.0500	0.6072
PACR	0.0565	0.4472
PCR	0.0665	0.2589
RTT	0.0983	0.1720
SoS1d	0.0766	0.0707
SoS2d	0.1352	0.0001
TITG	0.1061	0.1145
TTNS	0.0447	0.6189
TTPR	1.1413	0.0000
WI	0.0964	0.0108

D.3 Fitted Values vs. Residuals Plots

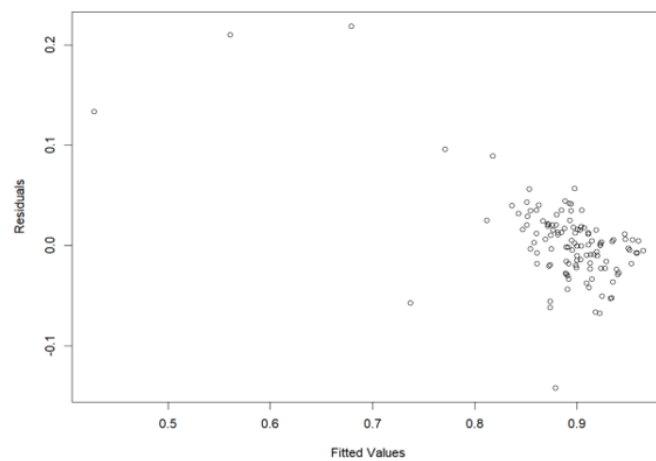


Figure D.12: Fitted Values vs. Residuals Plot for EDD

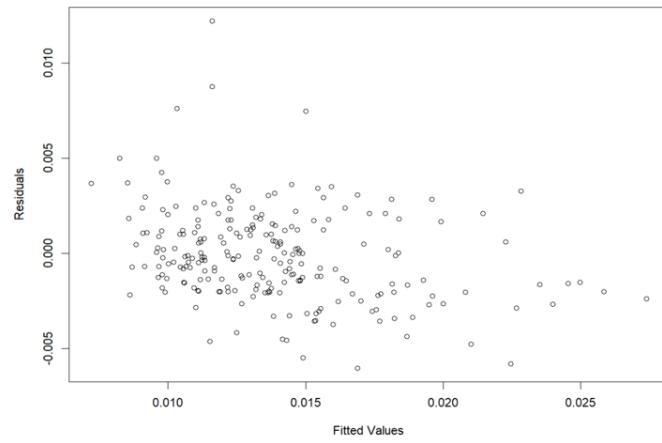


Figure D.13: Fitted Values vs. Residuals Plot for NS

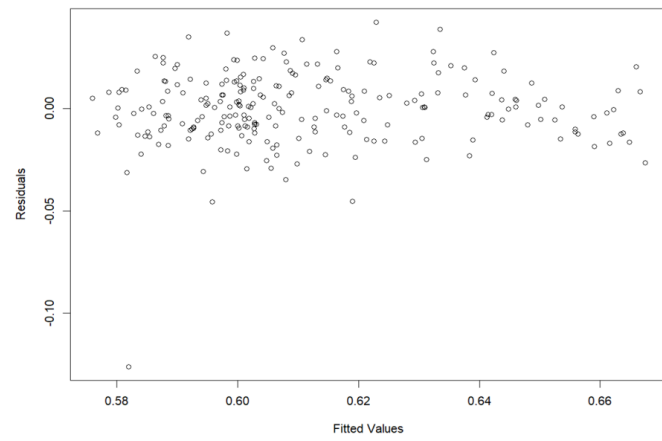


Figure D.14: Fitted Values vs. Residuals Plot for PACR

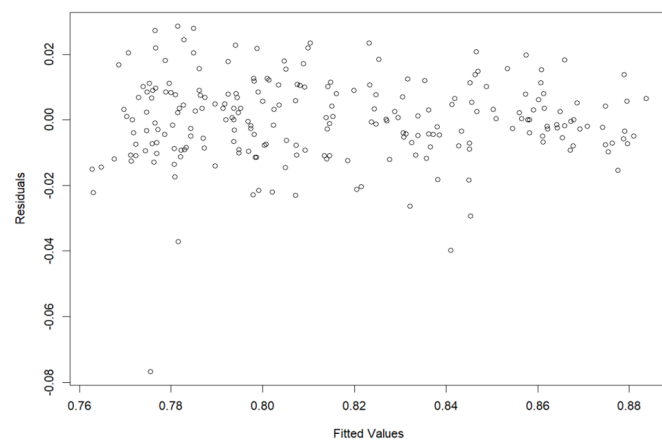


Figure D.15: Fitted Values vs. Residuals Plot for PCR

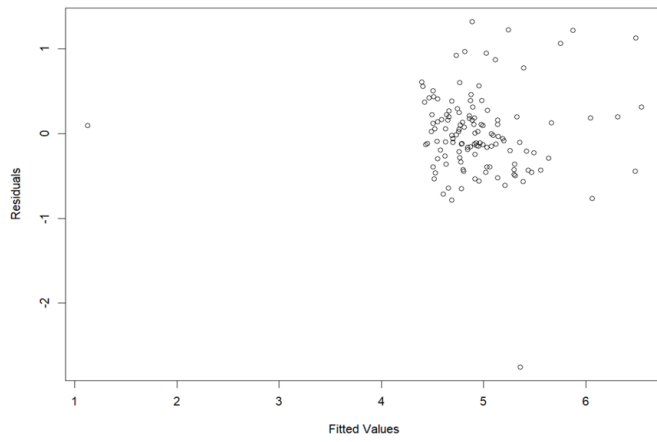


Figure D.16: Fitted Values vs. Residuals Plot for RTT

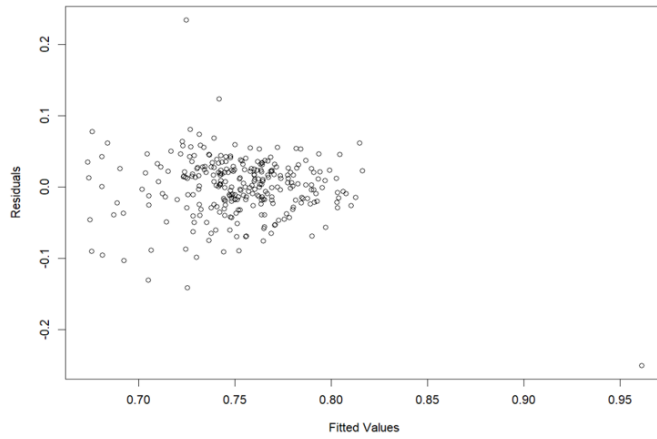


Figure D.17: Fitted Values vs. Residuals Plot for SoS1D

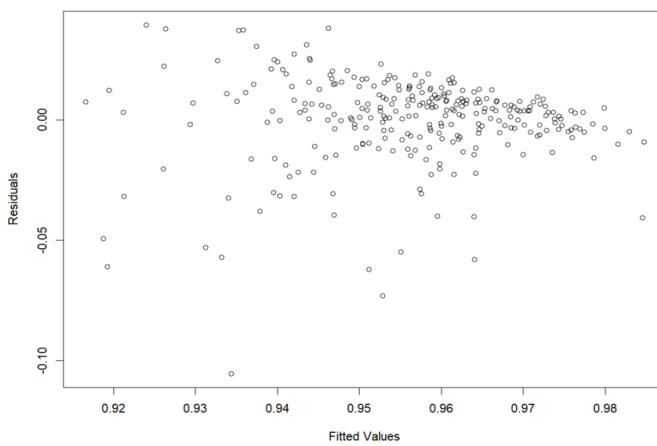


Figure D.18: Fitted Values vs. Residuals Plot for SoS2D

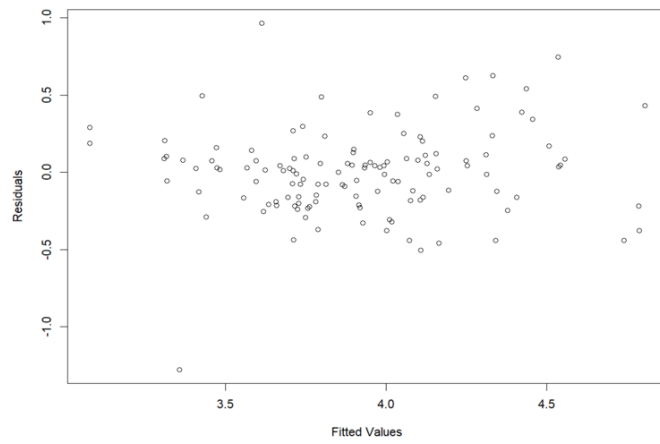


Figure D.19: Fitted Values vs. Residuals Plot for TITG

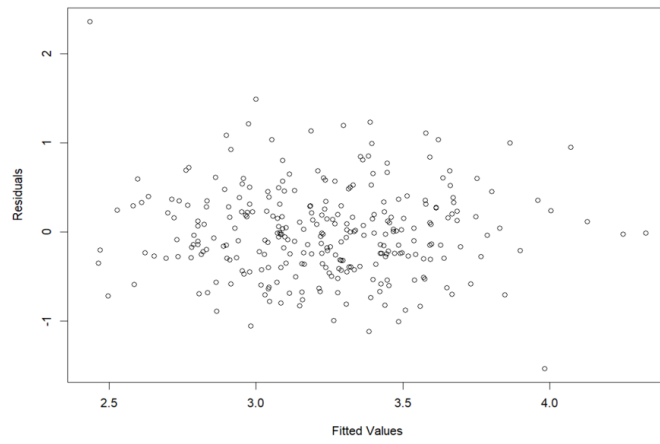


Figure D.20: Fitted Values vs. Residuals Plot for TTNS

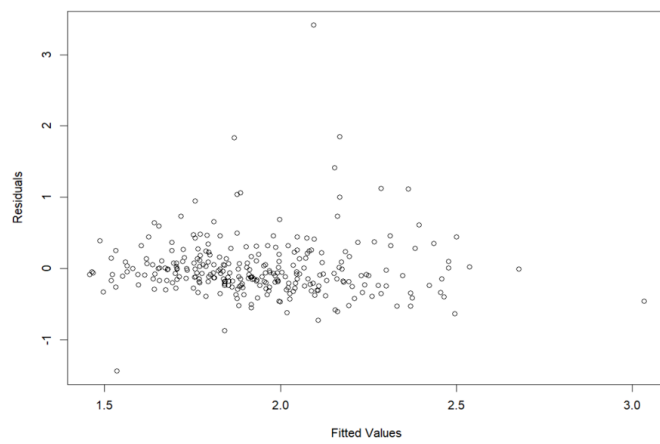


Figure D.21: Fitted Values vs. Residuals Plot for TTPR

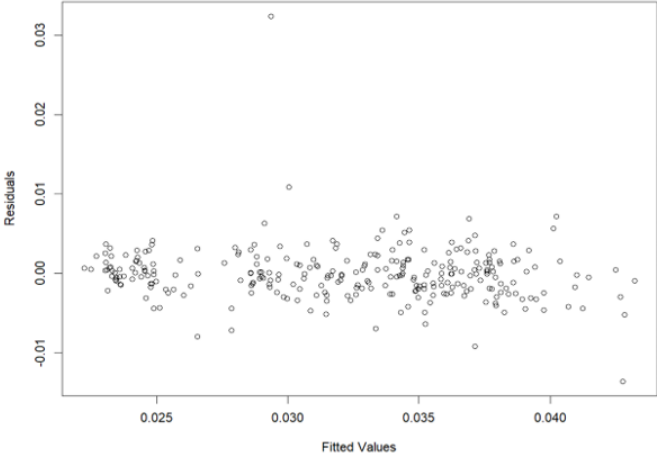


Figure D.22: Fitted Values vs. Residuals Plot for WI

Appendix E

Tool Evaluation Results

Table E.1: Tool Evaluation Results

Situation	Participant 1 Results	Participant 2 Results	Tool Results
Week 20 - TTNS	Attention	No alert	Alarm
Week 20 - NS	Attention	Attention	Alarm
Week 20 - RTT	Attention	No alert	Alarm
Week 19 - SoS1d	Alarm	Alarm	Alarm
Week 19 - WI	Alarm	Alarm	Alarm
Week 19 - PCR	Alarm	Attention	Alarm
Week 19 - NS	No alert	Alarm	Attention
Week 19 - SoS2d	Alarm	Alarm	Attention
Week 18 - PCR	Alarm	Alarm	Alarm
Week 18 - RTT	Attention	Attention	Attention
Week 17 - SoS1d	Attention	Alarm	Alarm
Week 17 - WI	Attention	Alarm	Attention
Week 16 - NS	Alarm	Alarm	Attention
Week 16 - TITG	Attention	No alert	Attention
Week 15 - TTPR	Alarm	Alarm	Alarm
Week 15 - RTT	Alarm	Alarm	Alarm
Week 15 - TITG	No alert	No alert	Attention
Week 15 - EDD	Alarm	Alarm	Attention
Week 14 - EDD	Attention	Attention	Attention
Week 13 - SoS1d	Alarm	Alarm	Alarm
Week 13 - NS	Attention	Alarm	Attention
Week 12 - SoS1d	Attention	Attention	Alarm
Week 12 - PACR	Attention	Attention	Alarm
Week 12 - PCR	Alarm	Alarm	Alarm
Week 10 - NS	Alarm	Alarm	Alarm
Week 9 - TTPR	Alarm	Alarm	Alarm
Week 9 - PACR	Attention	Attention	Alarm
Week 9 - PCR	Attention	Alarm	Alarm
Week 6 - RTT	Alarm	Attention	Alarm
Week 4 - RTT	No alert	Attention	Alarm