School of Computing & Communications | Lancaster University

Machine Learning Applications to Throughput Analysis

MSc Thesis

Kieran Molloy
MSc Data Science

2023

Lancaster University

Abstract

Investigating the relationships between the throughput of a production line and various measurements of line setup (ingredients count, time of day) along with various post-run statistics (average station completion time, average time stopped, missed ingredients) by using multiple popular regression techniques. The main regression techniques used are Ordinary Least Squares (OLS) Regression, Support Vector Regression (SVR), Regression Trees along with Random Forests (RF) and cumulatively, Kernel Random Forests (KeRF). The performance of these algorithms is examined and evaluated against multiple test cases, via cross-validation, for standard random sub-setting, random complete weeks and the final complete weeks.

The case study analysis is based upon data generated automatically by the operations team of the factories over 1+ years. The outcome is interpretable relationships between the predictor and response variables. The main contributory factors to high throughput are defined in addition to the factors leading to low throughput.

Furthermore, the dataset is analysed as a time series, and changepoint analysis is performed using Pruned Exact Linear Time (PELT) using a CROPS penalty term to determine the location and magnitude of significant changes for various aggregations of the dataset.

Combining these approaches suggests that some shifts consistently have higher throughput, and in general, the average number of stations visited is a strong predictor of throughput.

Acknowledgements

The completion of this could not have been possible without the expertise and guidance of Dr Kanchan Mukherjee, my dissertation adviser - thank you for our in depth discussions, shaping the direction of both the dissertation and my personal learning as a whole. Additionally, for taking the time to proof read multiple sections *multiple times*, it really helped to better shape my academic writing.

A considerable debt is owed to my industrial supervisor at the UK Food Box Company, not only for allowing me this opportunity to work with you but also for allowing me to explore my own original ideas. Your consistent input and guidance is clear to see within this work, we had great discussions.

An incredibly massive thank you to my mother, Carole, for proofreading another academic report, as well as Angela for painstakingly reviewing the document with me, your contributions are hugely appreciated.

To Niamh, my highly supportive partner, thank you for listening to all my ramblings and allowing me the luxury of creating this work, I understand you might not have always understood our chats, but talking the problems through always helped me.

Finally, I would like to thank my father, Simon, my grandparents and anyone else that guided me throughout this journey; this would not have been possible without any and all of you.

Contents

List of Figures

List of Tables

List of Algorithms

CHAPTER 1

Introduction

The introduction of automated data collection has triggered a widespread shift of industry mindsets towards the benefits of wholly utilising data. The collection and application of datasets still contain significant untapped potential as computation becomes more efficient and alternative methods are conceived, collecting more data alongside contextual knowledge. Academics well understand the use of data, and in recent years, more so in industry, with applications ranging from machine learning powered search engines to recommendation systems to self-driving cars. While machine learning is already a significant part of our lives and companies worldwide, creating models face challenges such as complexity, dimensionality, variability, and bias. Extracting information from raw data is a complex task and requires a multi-faceted approach balancing computational approach, statistical knowledge and industry expertise.

1.1 Current Foundation

The UK Recipe Box Company's current modelling is limited to linear regression; this gives multiple avenues of exploration. The regression modelling performed previously has mainly been linear using forms of stepwise regression to construct models with the coefficient of determination (R^2) to evaluate performance. The insights that can be drawn from these kinds of models is limited and, due to the use of the R^2 value, models cannot be compared across different datasets; in addition to the known dimensionality problems that R^2 exhibits [1]. The use of stepwise methods is somewhat contested with some critics suggesting that stepwise regression is equivalent to p-hacking or being used as a substitute for subject area expertise [1, 2]. There is significant room for improvement with careful use and execution of appropriate methods.

1.2 Aims and Objectives

The scope of the current project is to identify applicable methods and create a reproducible pipeline to reliably predict the throughput of a production line given the setup of the production line and identify the patterns in the setup that affect the throughput. Furthermore, the exploratory changepoint analysis seeks to extract

[1] The Adjusted R^2 does fix this but has not been used by the UK Recipe Box Company

further contextual information in conjunction with the regression model. This becomes two distinct questions;

Can regression models be used to predict throughput?

This is the critical research area and the subject of the majority of this dissertation. The UK Recipe Box Company breaks this down into three further questions. (1) What is the impact of the menu on throughput? (2) What is the impact of operations on throughput? (3) What is the impact of differences between lines/factories on throughput?. To date, the company have only investigated simple linear models. The task encompasses the thorough exploration, processing and evaluation of multiple approaches and discusses the advantages and disadvantages.

What is the most accurate model to predict throughput based upon factory setup?

An extension of the previous objective is to ensure reliability within the results. In order to achieve this, a thorough evaluation and comparison of various pipelines is required - where a pipeline encompasses the processing of data through to the automated processed behind model performance evaluation. Equally, to ensure reliable results, as a whole, multiple variants at each stage must be considered to find the most accurate approach. The UK Recipe Box Company stated they have previously used R^2; however, this has several problems stated in Chapter 5. For results to be valid, multiple metrics and some form of cross validation are also required.

Can time series analysis be used to gain greater insight into the regression modelling results?

The secondary objective is to model the data as a time series across various aggregations; modelling time-ordered data is well-used in the form of control charts, but the focus is on changepoint analysis.

1.3 Outline

The rest of the text is organised as follows:

Chapter 2 defines the inductive problem-solving approach, states the research strategy, briefly discusses data collection methods and how the analysis will be presented.

Chapter 3 presents a short introduction to some key literature in four focused areas; supply chain performance evaluation, throughput analysis, regression modelling and changepoint analysis.

Chapter 4 introduces the modelling concepts and explains the purposes of each modelling approach and some essential concepts.

Chapter 5 performs a case study of a UK Recipe Box Company, applying the discussed concepts in an inductive process to give the reader a clear picture of the problem and the outcomes.

Chapter 6 discusses and suggests potential improvements or further research topics and concludes the research.

CHAPTER 2

Methodology

Several research methods exist for different problems; selecting one, or several appropriate methods is essential for a persuasive result. This chapter presents the methodological issues focusing on the three key themes to the problem at hand.

2.1 Research Strategy

When deciding upon a research strategy, one can consider two main approaches; an empirical or theoretical base. Empirical research is founded within observation and measurement, whilst theoretical research is hypothetical and does not require data. This work aims to analyse practically the operations with a foundation in theoretical studies. Hence, the core analysis will focus on empirical research with a solid theoretical background. Firstly a selection of crucial prior literature regarding logistics and modelling systems will be reviewed to define the theoretical background. Based upon this background, the focus shifts to applying the theory to the empirical data observed to create an in-depth description of how a company operates its production lines in practice and how this information can be leveraged to make informed decisions.

2.2 Case Study

The purpose behind a case study is the philosophy that only by looking pragmatically at real-life instances can a complete picture be obtained of the genuine interaction of events. The case study has two distinct features; establishing valid and reliable evidence and building a narrative description of the studied situation. Constructing a narrative description represents a research piece in its own right and adds value by its existence. With this context, we find that the case study is the most appropriate approach for this research. The study field of Machine Learning applied to analysing throughput is relatively unexplored, as is the dataset presented by the UK Recipe Box Company. With this provided dataset, the models presented in Chapter 4 are applied and illustrated.

2.3 Methods of Data Collection

Data collection, and data mining, is a growing field, traditionally relying on multiple data sources. There are six critical sources, according to Yin, documentation archival

records, interviews, direct observation, participant-observation, and physical artifacts [3]. Generally, there are two types of data; primary and secondary.

Primary Data

The primary data collection was conducted both as an automated monitoring process and through personal discussions regarding the interaction system. Working closely with the UK Recipe Box Company led to the opportunity to ask questions as they arose. Additionally, publicly available information regarding the company background and some descriptions of critical algorithms give a clearer picture of the business process.

Automated Monitoring Process

Automation must be at the heart of a longitudinal study for longer than a year; it is infeasible to collect statistics by a site visit or discussions with specialists manually. As an automated process, there must be questions about its validity and what is tolerable - ultimately, it must be accepted that there could be problems, and they must be dealt with before modelling. The UK Recipe Box Company performed the collection of data.

Interviews

The nature and complexity of the problem and the company's rapid growth, there is no combined source of information for all required details. Personal discussions provide an excellent platform for collecting information, and it provides a comprehensive view of how the qualitative data fits together. These are unstructured, with a selection of questions prepared and the flexibility to ask the pertinent questions blocking the research problems. The discussions are conducted with an experienced data analyst, with significant experience in operations and supply chains, and invaluable subject knowledge and practical experience.

Site Visits

Both qualitative and quantitative data can be collected from site observations. Visiting the operations in person provides unparalleled qualitative data on how the processes function individually and as an operation, allowing the opportunity to reinforce the finding from discussions and validate the automated data collection. The site visit is crucial to gain a comprehensive understanding of the research questions.

Secondary Data

Secondary data consists of research papers, textbooks and articles amongst other formal literature. These are primarily from Lancaster University Library Services and are summarised in Chapter 3 and form the basis of the theoretical foundation.

2.4 Validity and Reliability

For the results presented in this research to be helpful, they must be trustworthy; hence their validity and reliability must be discussed.

Validity

As mentioned earlier in this chapter, validity is an essential quality for this research to have. Validity quantifies the extent to which the results correspond with reality, and it consists of internal and external validity. Internal validity concerns the research itself and the connection between the theoretical framework and empirical studies (sufficient data and the interviews conducted with experts). External validity is based upon the possibility to generalise the results or findings from the research. The research is translatable to similar problems with the exact theoretical requirements in making future decisions.

Reliability

The question of whether an investigation is reliable depends on whether a future investigation could follow the same procedures as described and the same results would be found. This is especially important for the case study to be repeatable, so the same findings and conclusions are always attainable. Reliability can depend on the accuracy of measuring techniques and randomness. To provide highly reliable conclusions, all code is created within docker containers (ensuring complete reproducibility from the operating system upwards) and always defining a random seed before running any non-deterministic functions - this allows reproducing all results from one command. In addition, the interviews were conducted with experts in the subject area who have extensive knowledge of the system and the associated theoretical background.

CHAPTER 3

Literature Review

3.1 Supply Chain Performance Evaluation

Supply Chain Management (SCM) handles the entire production flow, from raw components to delviery of final product to the consumer. There are two well-renowned reviews of the area by Ross and Simchi-Levi [4, 5]. In the book Managing the Supply Chain [6], it defines supply chain management as

> "Supply chain management is a set of approaches utilised to efficiently integrate suppliers, manufacturers, warehouses and stores, so that merchandise is produced and distributed at the right quantities, to the right locations and at the right time, to minimise system-wide costs while satisfying service level requirements."

The definition of a supply chain varies from author to author, and there is no definitive. [7, 8, 9, 10]

Many recent studies have defined the tools required to make data-informed supply chain decisions; research such as that conducted by Zhong has shown that big data can be used to plan logistics, production, and scheduling [11]. Tan demonstrated that a big data analytics approach based on graph theory could enhance innovation capabilities [12]. Extending this, Shukla introduced a fuzzy rough sets-based multi-agent model for configuring supply chains in dynamic environments [13]. This was practically implemented by Dutta, where they managed a big data project for a cement supply and logistics network [14]. Additionally, Singh proposed a cloud computing framework for reducing the carbon footprint of a supply chain [15]. Waller argued that data science, predictive analytics, and big data could help logistics managers meet internal needs and adjust to changes in the supply chain more rapidly [16, 17]. A comprehensive review of the current literature was conducted by Govindan, where they presented various methods to improve the use of big data analytics and applications for logistics and supply chain management; furthermore, it was noticed that utilising big data could dramatically transform supply chains to become more efficient [18].

3.2 Throughput Analysis

The measurement and analysis of throughput is crucial for the operation, management, and design of production systems. Jingshan produced a detailed review of the current topic space and outlined future areas [19]. There are many different approaches to

measuring production systems' throughput, with many variants on how systems are set up - parallel lines, split and merge, closed-loop and assembly/disassembly. Due to the stochastic nature of production systems (processing time, station breakdowns, ingredient status), the expectation of the number of products produced is a random variable - throughput. Numerous studies have previously reviewed this area [20, 21, 22, 23, 24, 25, 19], notably the Proceeding Series of Conferences on the Analysis of Manufacturing Systems (1997, 1999, 2001, 2003, 2005, 2007). The exact analytical result only exists for the two-machine-one-buffer system and systems without a buffer or with infinite buffer capacity. Li compared eight varying two-machine-line models and showed that a similar performance in throughput is observed for all models - but that for long systems, the analytical results of the two-machine-line systems can be a building block for evaluating the approximate system performance [26].

3.3 Regression Modelling

Linear Regression is a well-defined problem with well-defined solutions. The goal is to estimate the relationship between the explanatory and response variables from a set of training samples [27]. There are also extensions from linear to non-linear applications [28, 29].

Tree-based models are widely known for their ease of interpretation and simplicity. Breiman demonstrated the use of trees within classification and regression [30], which was later expanded to create the well known ID3 algorithm [31]. Further notable improvements upon the decision trees are the introduction of bagging [32], gradient boosting [33] and random forests [34]. Barros (2012) carried out a survey of evolutionary algorithms for decision tree creation [35].

Some observations have been 'corrupted' by significant observation errors in many datasets, known as outliers [36]. This caused the failure of Ordinary Least Square estimation [37] and brought robust estimation into the frame. Robust regression aims to accurately model in the presence of outliers [38, 39, 40]. There are two assumptions about outliers

1. Outlier entries are often of larger magnitude than inlier entries.

2. The number of outliers in the dataset is smaller than those that are inliers.

The are many methods that perform the first assumption, notably M-estimation [40], among others. The second assumption usually takes the form of fitting the majority of data. The Least Median Squares (LMedS) method introduced by Rousseeuw minimises the median of squared residuals instead of the mean [41]. Rousseeuw improved the efficiency with Least Trimmed Squares [38, 39]. Tukey proved that in the presence of outliers, or contamination, a mean based error term performs poorly, whereas a metric based upon the median performs significantly better [42, 43].

3.4 Changepoint Analysis

The final section of this research is applying changepoint analysis, identifying multiple change points within time-ordered data. This is a problem that increases combinatorially as the size of the dataset increases. There are multiple search algorithms proposed to solve this, binary segmentation [44], segment neighbourhood

[45] and more recently, the Pruned Exact Linear Time (PELT) algorithm [46]. Single points of change can be calculated using traditional likelihood-based approaches; however, multiple points poses a more complicated challenge; before Killick's PELT algorithm, the most common approach was to minimise some cost function (sometimes negative log-likelihood) for a segment.

CHAPTER 4

Background

This chapter will attempt to establish a base of previous studies in the relevant areas upon which to build and integrate within the project, whilst also to familiarise the reader with the subject areas. Initially, an introduction to the basics of the logistics behind recipe boxes is vital to underpin the purpose of the current research; with this, the core concepts of appropriate regression modelling techniques shall be addressed as the main output from this project. Additionally, a review of previous related research will be presented and critically analysed for relevance to the current research questions. It will provide an excellent starting point and refer to which methods are more effective in similar problems.

4.1 Learning Paradigms

Supervised Learning

Supervised learning is the most straightforward conceptually; it considers some dataset $\mathcal{D}$ consisting of (X, y) pairs, where X is made up of n columns of predictor variables and, y is the response variable, the goal is usually to learn some function f, such that $f(X) = y$. The name 'supervised' comes from the fact that there are pairs of values (X, y) where the ground truth is known.

An excellent example of supervised learning is binary classification, given some data X_i and the associated response variable y_i where y_i is one of 2 target 'classes'.

$$f(x) = \begin{cases} 1 & \text{if item is erroneous} \\ 0 & \text{if item is not erroneous} \end{cases}$$

Modelling Considerations

There are a wide range of supervised learning algorithms with their specific advantages and disadvantages - hence there is no algorithm that is always the best in any given situation, and the same applies to this project. There are some considerations to make when making the choice regarding which models should be used.

Bias-Variance Dilemma One of the most prominent issues is balancing the bias and variance, known as the bias-variance dilemma, and this is the conflict in attempting to minimise both of these items (as they are sources of error that prevent out of set generalisations). The variance is an error caused by fluctuations in the training set, and overly high variance causes overfitting (modelling the underlying

noise). Conversely, the bias error is caused by incorrect assumptions in the modelling algorithm and leads to underfitting (modelling the less-important relationships) [47, 48].

Dimensionality If the input feature space, the X predictor variable matrix, has high dimensionality, the ability to learn the problem can be complex - even if most features are not helpful for the model - due to these extra dimensions effectively confusing the model, inducing high variance. Some modelling approaches can deal with this by tuning the method to have low variance and high bias, but a better solution is to remove irrelevant features prior to modelling, which benefits increasing accuracy and decreasing computation time.

Data Availability The amount of available training data is of genuine concern, as from the book 'In All Likelihood', Yudi Pawitan states:[49]

> The inductive process raises two problems: one is that it tends to increase the stochastic uncertainty, since, by splitting the original observations into smaller explanatory groups, we are bound to compare smaller sets of numbers. The other is deciding where to stop.

Moreover, this concern is compounded by the relative complexity of the underlying data generation function, arbitrarily designated $f(x)$, and attempting to fit the data too aggressively leads to overfitting. Furthermore, this can also occur when there exists no stochastic noise but the function required is too complex, and the model becomes "corrupted", this phenomenon is known as deterministic noise.

Noise If there exists some stochastic noise in the output values, then there will be some problems fitting the data to some function $f(x)$, and attempting to fit the data too aggressively leads to overfitting. Furthermore, this can also occur when there exists no stochastic noise but the function required is too complex and the model becomes "corrupted", this phenomenon is known as deterministic noise.

Unsupervised Learning

The term unsupervised is derived from this methodology extracting information devoid of labelling, and this collection of methods often exploit the natural patterns present in the data.

The widely known unsupervised techniques include; clustering, which is not the focus of this project; however, an extension of clustering is anomaly detection, which identifies rare events due to their difference with regards to the majority of data [50]. In statistical modelling, the assessment of “being an anomaly” is generally based upon a parametric model of the data, identifying those objects that do not fit well to the modelled distribution as outliers.

4.2 Data Processing

Various well-defined procedures must be used to translate the problem into something interpretable numerically. For example, how would a mathematical model understand the categories "red", "blue", and "grey" if the most prevalent car colour was to be modelled? Furthermore, multiple problems must be statistically addressed at this stage so that the results are reliable.

Cleaning

The processing of data must begin with the cleaning of the dataset. The first stage of cleaning is doing some exploratory analysis to better understand the relationships and distributions of the dataset. The most accessible approach to exploring the dataset is to generate plots of all variables against others and the distributions of each variable. With these plots, a targeted exploration of specific relationships can be performed. This stage varies significantly for the needs of the dataset; some common steps are:

- Missing Value Imputation
- Handling Excess Noise
- Numerical Scaling (Transformation)
- Categorical Encoding
- Time Encoding
- Variable Selection

Each of these steps is explored further in the following subsections.

Missing Value Imputation

Missing data is expected in statistical analysis; there are various approaches to dealing with missing data, ranging from simplistic and crude to comprehensive, yielding varying success.

There are three general strategies for imputing multivariate data:

- Monotone data imputation - for monotone missing data patterns, imputations created by a sequence of univariate methods
- Joint modelling - for general patterns, imputations created by forming a multivariate model fitted to the data
- Fully conditional specification - for general patterns, a multivariate model is implicitly specified by a set of conditional univariate models. imputations created by forming iterated conditional models

Monotone data imputation and Joint modelling are well defined in other literature [51]

Fully Conditional Specification (FCS) imputes missing data on a variable-by-variable basis [52] [53].[53]. The method requires a specification of an imputation model for each incomplete variable and iteratively creates imputations per variable. In contrast to joint modelling, FCS specifies the multivariate distribution $P(Y, X, R|\theta)$ through a set of conditional densities $P(Y_j|X, Y_{-j}, R\phi_j$. This conditional density is used to impute Y_j given X, Y_j and R. Starting from simple random draws from the marginal distribution, imputation under FCS is done by iterating over the conditionally specified imputation models

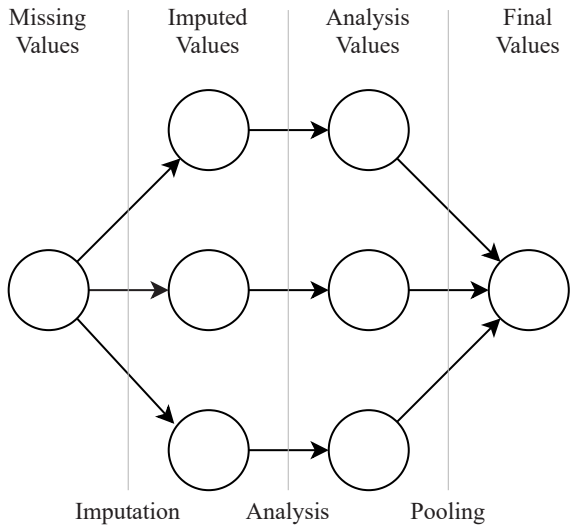


Figure 4.1: Illustration of Multiple Imputation by Chained Equations (MICE) with $m = 3$ [51]

Method

The imputation method should be capable of accounting for (1) the origin process of the missing values. (2) preserving the underlying relations in the data, and (3) preserving the uncertainty about these relations. The imputations are statistically correct by adhering to these points, as Rubin proved for a wide range in Q [54, 55]. As with any complex method, some problematic points can occur within multivariate missing data imputation.

Multiple Imputation creates $m > 1$ complete datasets that are analysed and pooled into a final estimate which Figure 4.1 illustrates for $m = 3$ datasets. These datasets are expected to be identical except the imputed values, and the magnitude of these differences should reflect the uncertainty about the imputed values. Letting Y be the $n \times p$ data matrix, where Y_j is the j^{th} column, and Y_{-j} is the complement - all columns in Y except for the j^{th}. We consider the *missing data pattern* of Y. It is helpful to consider causes of missing data; in some instances, missing data can be intentional (the reason by which is domain-specific); however most missing values are unintentional.

Most imputation models for Y_j are based upon Y_{-j}; this is because the state of Y_{-j} preserves the relationships with Y_j in the observed data. Some problematic points

- For some Y_j, the predictor variables Y_{-j} may also be missing
- There can exist circular dependencies where Y_h^{mis} depends on Y_j^{mis} and Y_j^{mis} depends on Y_h^{mis} due to Y_j and Y_h being correlated.
- Variables are often of different types making the application of theoretically convenient models theoretically inappropriate
- Especially with large p and small n, collinearity or empty cells can occur,
- The ordering of rows can be meaningful, e.g. for longitudinal data

- The relationship between Y_j and predictors Y_{-j} can be complex, e.g. nonlinear, or subject to a censorship process
- Imputation can create impossible combinations (such as negative Poisson values, or the typical example that is pregnant fathers)

For these reasons, the output from imputation should be checked and properly validated to ensure imputed results are valid or sensible.

Erroneous Values

It is to be expected that some (likely unknown) amount of noise will exist in the dataset - this could appear in various ways depending on the distribution of the data (i.e. extreme variability in what is known to follow uniform distribution or extremely heavy tails in a normal distribution). The method to handle these errors depends on how they are introduced; introduced via data collection, or uncontrollable data processing vs unknown origin. However, a standard solution is robust modelling, which prevents the problematic points from having a significant effect or outlier/anomaly detection.

Alternative options include winsorization, which has various well-documented, extremely undesirable pitfalls [43].

Transformations

Often, numerical data is not conducive to be directly input into a regression model; some models have a prerequisite that the data lies within the interval $[0, 1]$ or $[-1, 1]$. Learning algorithms often benefit from standardisation, but doing this in the presence of outliers is complex. The most common approach to standardise a dataset, assuming normally distributed data, is scaling to zero mean and unit variance. An increasingly common alternative approach is to ignore the distribution and simply transform the dataset to the mean then scaling the values by dividing by their standard deviation, producing the same effect. Support Vector Machines require this standardisation as they assume all features are centered on zero. If this was not the case, the estimator would either converge slowly or not at all. Alternative scaling approaches can be helpful in specific problems such as Minimum Maximum Scaling or Absolute Maximum Scaling. However, if the data contains any outliers greater than the expected maximum, all these scaling methods will not produce a well-scaled dataset, so using a more robust scaling approach can remedy these issues by using IQR or median-based scaling.

Categorical Encoding

Datasets usually include some form of categorical data; dealing with these non-continuous values is relatively simple, and there are various strategies to convert these to numerical data - but the choice of strategy can lead to problems when modelling. Categorical features must be encoded to be understood by regression models. The most common approach is one-hot encoding, whereby each categorical feature with n categories is converted into n binary columns whereby one of them is 1, and the others are all 0 representing which category the row represents. Some models can take categorical data as input, for example, some implementations of

decision trees. If the categorical data is ordinal, this may not work very well, as the model cannot understand the ordinal nature of the feature; in this case, it would be better to assign an ordinal numeric value for each category in a single feature.

4.3 Regression Models

Regression Models are a set of models that estimate the relationships between a dataset and a target variable. This section defines the relevant models to build the required understanding for the case study.

Linear Models

The simplest of approaches is simple linear regression

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where y is a vector of observed values, here-on referred to as the response variable.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

and X is an n-dimensional column-vectors, referred to as the explanatory variables. β is a parameter vector of size $(p+1)$-dimensions, where β_0 is the intercept term (hence the $(p+1)$). In simple linear regression, $p = 1$, and this parameter is known as the regression slope. ϵ is a vector of error terms.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Some assumptions about the predictor variables, the response variables and their relationships must hold.

- **Linearity** The mean of the response variable must be a linear combination of the regression coefficients and predictor variables.
- **Homoscedasticity** The variance of the errors does not depend on the values of the predictor variables.
- **Error Independence** The errors of the response variables are uncorrelated with each other. [1]
- **No perfect multicollinearity** There must exist no perfect multicollinearity within the predictors.

[1] True statistical independence is desired, but this assumption is far more robust than simply lack of correlation and is extremely rare to achieve in real datasets - this is often not required, but if it is known to be truly statistically independent, this can be exploited

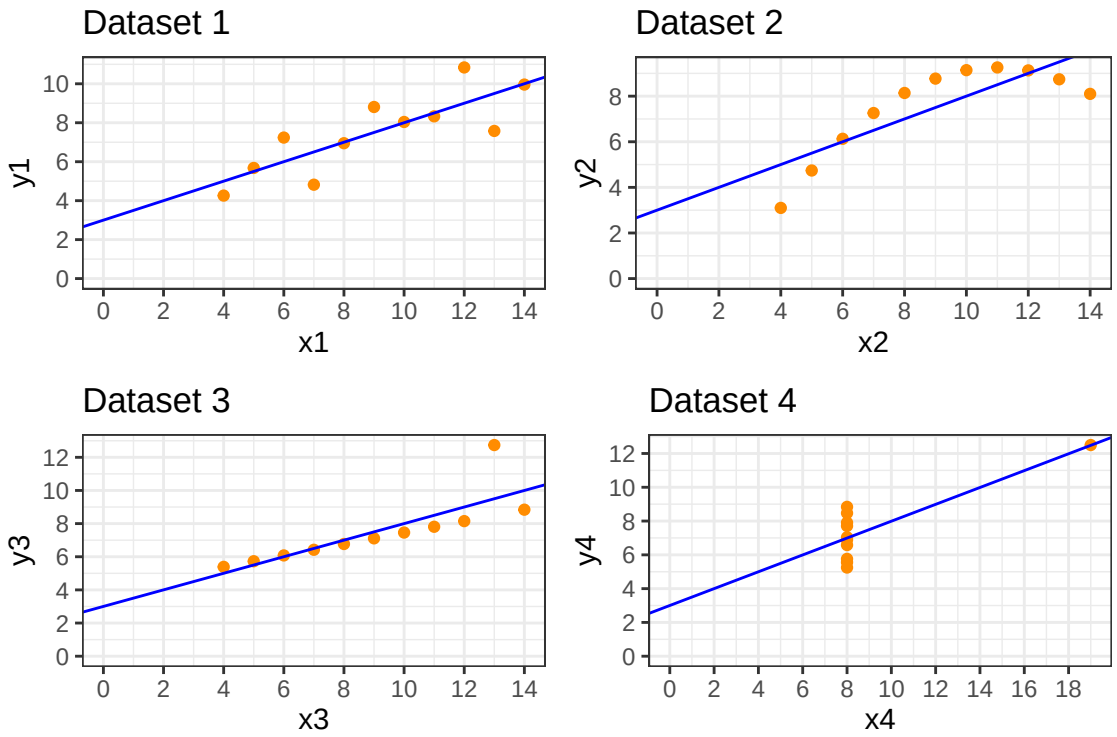


Figure 4.2: Anscombe Quartet: 4 differing datasets with the same summary statistics

- **Weak Exogeneity** The predictor variables can be treated as fixed values instead of random variables.

Interpretation

The fitted linear function, for $i = 1, \ldots, n$

$$\hat{y}_i = \sum_{k=0}^{K} b_k f_k(X_{i,1}, X_{i,2}, \ldots, X_{i,p})$$

where $\hat{y}_i$ is the estimated response variable with the fitted coefficients b_k. These are estimated attempting to minimise the objective function, which the standard is mean squared error between the predicted $\hat{y}$ and true response vector y. This method is also known as the least squares. For a linear regression problem, the coefficient b_k represents the impact of a one-unit change in the predictor variable X_j on the mean of the expected response $E[y]$ whilst holding all other predictor variables constant.

The Anscombe Quartet, presented in Figure 4.2, demonstrates the importance of graphing or visually inspecting data before analysing it to understand the effect and influence of observations prior to the statistical summary [56].

Robust Regression

Standard statistical approaches begin with a parametric model and develop optimality using maximum likelihood theory. However, these are just approximations, and when the data does not precisely follow the model there can be problems. For example, a standard modelling assumption is Gaussian errors (based upon mean and variance), here there are two scenarios (1) the data is entirely not normal (2) most data is

normal, but some is not. These few non-normal data points are denoted outliers. It is well known that for scenario (1), methods that assume Gaussian distribution are inefficient estimators whilst they could be invalid for scenario (2). Robust statistics attempt to remedy these issues and do this intending to attempt to fit the majority of the data whilst for scenario (1) fitting the centre of the data well and putting lower emphasis on tails or for scenario (2) limiting the influence of outliers.

Above the definition for least-squares estimator was given, and despite the beneficial properties, different types of outliers have differing effects on the estimator

- Vertical outliers shift the regression line.
- Good leverage points can lead to better models, but bad leverage points make the model useless.

Thus, as a result, regression models based upon least squares can be highly misleading in the presence of outliers. Many alternative methods are robust against these types of outliers, such as M-estimators, S- and MM-estimators or Least-Trimmed Squares. These are effective in lower dimensions, but in high dimensional data (especially where $p > n$), some form of dimensionality reduction with robust estimation is required. For a detailed discussion, see, for example, [37, 38, 40].

Support Vector Machines

Support Vector Machine (SVM) analysis is a common machine learning tool, mainly for classification but also regression [57]. SVM Regression is an extension of the classification method and relies on kernel functions. Kernel functions will pose a vital factor of this analysis; as such, the formulation of the SVM is essential. This implementation is the epsilon-insensitive regression method (ϵ-SVR), whereby the objective is finding a function, or hyperplane, that sees deviations lower than ϵ for each training point x. A linear SVM in 2-dimensions with two sets is demonstrated in Figure 4.3; a decision surface can represent any further dimensions.

Formulation

Given the set X and y, as defined in Section 4.3, for the linear function

$$f(x) = X\beta + b$$

the SVR optimises

$$J(\beta) = \frac{1}{2}\beta'\beta$$

subject to all residuals having a value less than ϵ. To handle the possibility that no such function exists to satisfy this condition for all x, the slack variables ξ and ξ^* are introduced to deal with otherwise infeasible constraints.

Adding this to the objective function leads to the primal formulation [57]

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{i=1}^{N} \xi_i + \xi_i^*$$

subject to

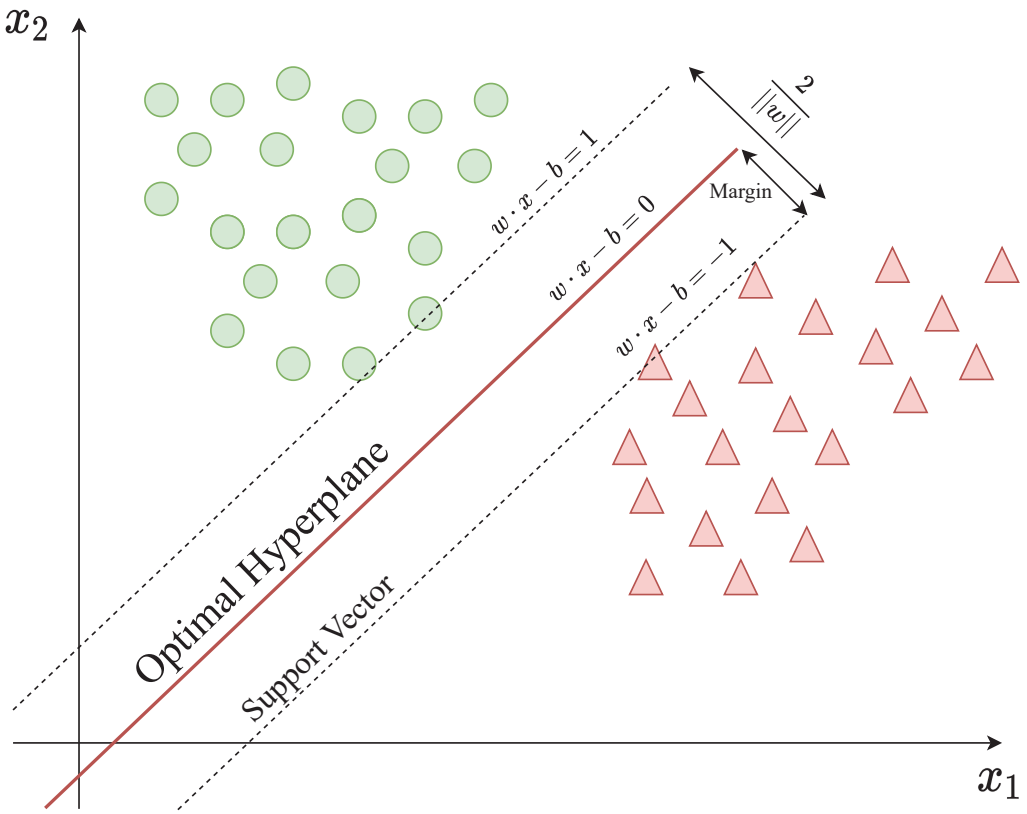


Figure 4.3: Linear Support Vector Machine with two clear sets

$$\forall n : y_i - (X\beta + b) \leq \epsilon + \xi_i$$
$$\forall n : (X\beta + b) - y_i \leq \epsilon + \xi_i^*$$
$$\forall n : \xi_i^* \geq 0$$
$$\forall n : \xi_i \geq 0$$

Where the constant C is the box constraint, some strictly positive penalty value impacting solely observations beyond the ϵ margin and prevents overfitting.

The linear ϵ-insensitive loss function ignores error terms within ϵ distance of the observed value by considering them as 0; therefore, the loss is the distance between observed value y and the boundary, ϵ. Formally this is,

$$L_\epsilon = \begin{cases} 0 & \text{if } |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & \text{otherwise} \end{cases}.$$

This optimisation problem is somewhat computationally expensive, so it is the norm to use the Lagrange dual formulation instead.

The solution to the dual problem provides a lower bound to the solution of the primal problem. The optimal values of the primal and dual problems are not necessarily equivalent, and so, the difference is called the "duality gap". Nevertheless, when the problem is satisfies a constraint qualification condition and convex, the value of the objective function of the optimal solution to the primal problem is equivalent to that of the dual problem [58].

Constructing a Lagrangian function from the primal by introducing non-negative multipliers α_i and α_i^* for each x_i to obtain the dual formula.. This leads to

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)x_i'x_j + \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) \quad (4.1)$$

subject to the constraints

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$$
$$\forall n : 0 \leq \alpha_n \leq C$$
$$\forall n : 0 \leq \alpha_n^* \leq C$$

The β parameter can be completely described as a linear combination of the observations using the equation

$$\beta = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)x_n$$

The function to predict new values depends solely on the support vectors

$$f(x) = \sum_{n=1}^{N}(\alpha_n - \alpha_n^*)(x_n'x) + b \quad (4.2)$$

The Karush-Kuhn-Tucker (KKT) complementarity conditions are optimisation constraints required to obtain optimal solutions. For linear SVM regression, these are

$$\forall n : \alpha_n(\epsilon + \xi_n - y_n + x_n'\beta + b) = 0$$
$$\forall n : \alpha_n^*(\epsilon + \xi_n^* - y_n + x_n'\beta + b) = 0$$
$$\forall n : \xi_n(C - \alpha_n) = 0$$
$$\forall n : \xi_n^*(C - \alpha_n^*) = 0$$

These conditions indicate that all observations strictly inside the epsilon tube have Lagrange multipliers $\alpha_n = 0$ and $\alpha_n^* = 0$. If either α_n or α_n^* is not 0, then the corresponding observation is a support vector.

Often real-world problems cannot be modelled linearly, as previously mentioned, and motivates the extension to non-linear regression; this extension is relatively simple and is based upon the Lagrange dual formulation. By replacing the dot product $x_j'x_k$ with a kernel function $G(x_j, x_k)$.

Hence, the dual formulation for the non-linear SVM regression is as follows

$$L(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)G(x_i'x_j) + \epsilon\sum_{i=1}^{N}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) \quad (4.3)$$

subject to the constraints

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0$$
$$\forall n : 0 \leq \alpha_n \leq C$$
$$\forall n : 0 \leq \alpha_n^* \leq C$$

And the function to predict new values is

$$f(x) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*) G(x_n, x_k) + b \tag{4.4}$$

And the KKT complementary conditions are

$$\begin{aligned}
&\forall n : \alpha_n(\epsilon + \xi_n - y_n + f(x_n)) = 0 \\
&\forall n : \alpha_n^*(\epsilon + \xi_n^* + y_n - f(x_n)) = 0 \\
&\forall n : \xi_n(C - \alpha_n) = 0 \\
&\forall n : \xi_n^*(C - \alpha_n^*) = 0
\end{aligned}$$

Algorithm 1 Support Vector Machine

Require: A dataset X of size S, F (samples, features), with y labels
  C =randint()
  while S =True **do**
    for x_i, y_i, x_j, y_j **do**
      Optimise α_i and α_j
    end for
  end while
  return trained support vector machine

Hence SVM Regression can be solved using ordinary quadratic programming techniques; however, this can be computationally expensive. In addition, high dimensionality can lead to memory problems. Solving the computation and memory problems, the decomposition approach is more commonly used. The decomposition methods separate all observations into two disjoint sets: the working set and the remaining set. A decomposition method modifies only the elements in the working set in each iteration. Therefore, only some of the data is required in each iteration, which reduces the required storage for each iteration.

Sequential Minimal Optimisation (SMO) is the most popular approach for solving SVM problems [59]. SMO performs a series of two-point optimisations; in each iteration, a working set of two points are chosen based upon a selection rule that uses second-order information. Then the Lagrange multipliers for this working set are solved analytically using the approach from [60]. In SVM regression, the gradient vector ∇L for the active set is updated after each iteration. The decomposed equation for the gradient vector is

$$(\nabla L)_n = \begin{cases} \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) G(x_j, x_n) + \epsilon - y_n, n \leq n \\ -\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) G(x_j, x_n) + \epsilon + y_n, n > N \end{cases} \tag{4.5}$$

The solver algorithm iteratively computes until one of the convergence criteria are met. There are three main options

1. Gradient Difference - If the difference between consecutive iterations ∇L values is greater than a set value

2. Largest KKT violation - If the KKT violation size is greater than a set value

3. Feasibility Gap - If the difference between consecutive iterations is greater than a set value. Where the feasibility gap is defined as

$$\Delta = \frac{J(\beta) + L(\alpha)}{J(\beta)_1}, \tag{4.6}$$

where $J(\beta)$ is the primal objective, and $L(\alpha)$ is the dual objective.

There are various extensions to SVR that can increase its speed of convergence, or increase the accuracy. However, the effectiveness for a kernel SVM depends on the selection of the kernel, the kernel hyperparameters and the margin parameter C. Hence a drawback of the model, is for the results to be helpful, comparable to alternatives, a significant amount of hyperparameter searching is required.

Tree-Based Models

Decision Trees is a statistical model that maps decisions into branches, whilst they are more complex in theory than other linear and non-linear models. There are some critical motivations for using decision trees; the ease of interpretation poses a substantial benefit, as a decision tree is a flowchart style with clear decision boundaries and paths to outcomes. This is a primary target for the UK Recipe Box Company; many other benefits make tree models desirable, such as handling multiple data types, the cost of creating trees and the ability to validate the outcome. However, tree models can be unstable with the presence of minor variations, causing overfitting (which pruning or hyperparameter tuning can attempt to remedy but at the cost of additional computation). An example binary decision tree is presented in Figure 4.4; the nodes are highlighted for convenience, the yellow (top) node is considered the root with a condition, here $x < 1$. The left path is where this condition is true, and the right is false. The green node represents a subtree containing a right child node solely; this would be removed when pruning the tree. The blue nodes represent subtrees containing both left and right child nodes. All four red 'leaf' nodes represent the final bin categories.

Formulation

Given the dataset $mathcalD$, consisting of predictor matrix X and label vector y, a decision tree partitions the feature space to group similar labels. For a binary tree, let the data at node m be represented by Q_m with N_m samples. For each split value $\theta = (j, t_m)$, consisting of a feature j and threshold t_m, the data partitioned into $Q_m^{\text{left}}(\theta)$ and $Q_m^{\text{right}}(\theta)$ subsets [30]. This becomes

$$\begin{aligned} Q_m^{\text{left}}(\theta) &= \{(x, y) | x_j \geq t_m\} \\ Q_m^{\text{right}}(\theta) &= Q_m \setminus Q_m^{\text{left}}(\theta) \end{aligned}$$

To measure the value, or quality, of a splitting value for node m, the value of the split computed using a loss function $H(x)$); the function $H(x)$ depends on both the task at hand (either regression or classification) and the distribution of the dataset.

$$G(Q_m, \theta) = \frac{N_m^{\text{left}}}{N_m} H(Q^{\text{left}}(\theta)) + \frac{N_m^{\text{right}}}{N_m} H(Q^{\text{right}}(\theta))$$

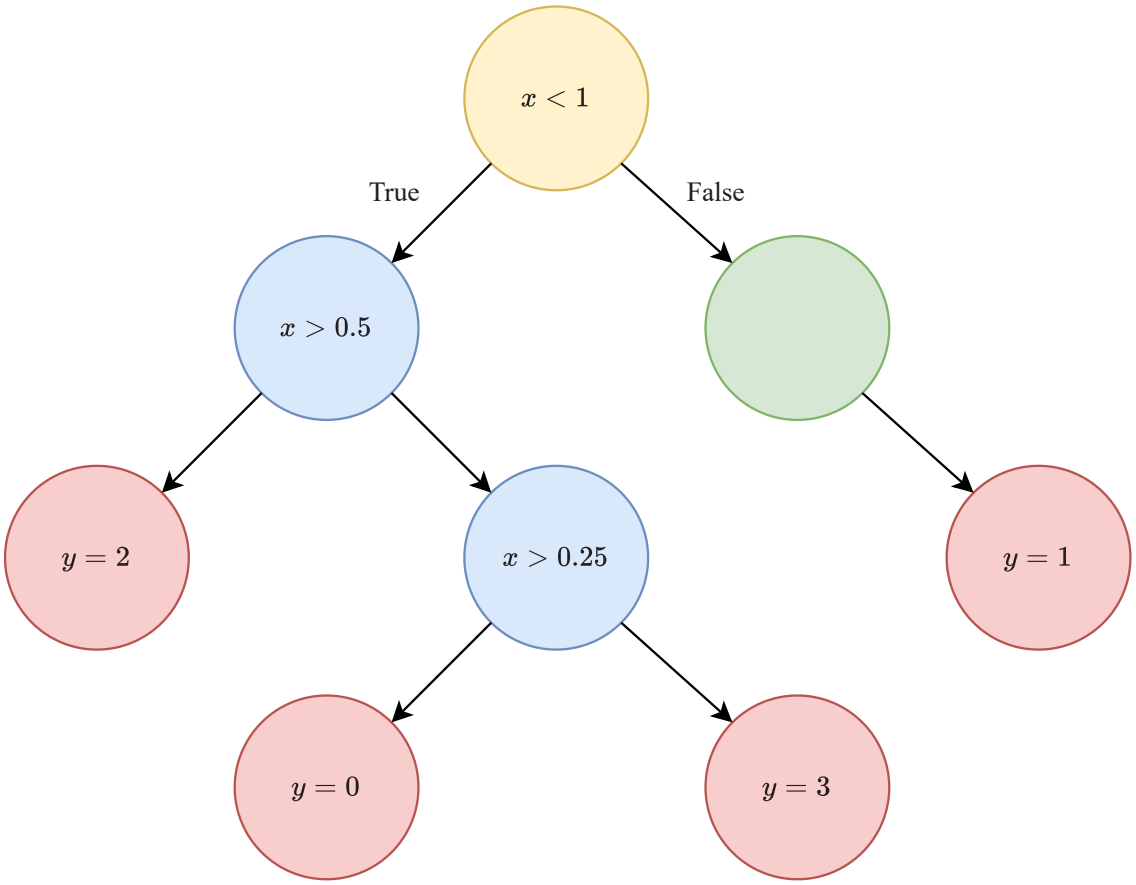


Figure 4.4: Classification Decision Tree depth $d = 3$

This is then minimised by

$$\theta^* = \text{argmin}_\theta G(Q_m, \theta)$$

This is recursive for both left and right subsets until the stopping condition is met - usually the maximum allowable depth; however, the stopping condition of trees is a significant enough topic to have an entire chapter [30].

Algorithm 2 Decision Tree

Require: A dataset X of size S, F (samples, features)
  if S = True **then**
    $leaf$ = node()
    return $leaf$
  end if
  $root$ = node()
  $value = H(x, v)$
  $V = v \in value$
  for $v \in V$ **do**
    $S_v = s \in v = x \in X$
    $child = node()$
  end for
  return trained decision tree

Splitting Criterion

The task of splitting the tree, referred above to as $H(x)$, impacts the whole quality of trees. Three common criteria are minimised to determine split locations.

Mean Squared Error is usually the default method of libraries and performs strongly; this evaluates splits by setting the predicted value of a terminal node to the learned mean value $\bar{y}_m$

$$\bar{y}_m = \frac{1}{N_m} \sum_{y \in Q_m} y$$
$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y - \bar{y}_m)^2$$

Half Poisson Deviance can result in better splits; hence, better trees if the target variable follows a Poisson distribution (i.e., a count or count per unit). Nevertheless, the drawback is that this method is far more computationally expensive, which is exacerbated within ensemble functions.

$$H(Q_m) = \frac{1}{N_m} \sum_{y \in Q_m} (y \log \frac{y}{\bar{y}_m} - y + \bar{y}_m)$$

Mean Absolute Error is aimed at splitting in a more robust manner; instead of using the mean $\bar{y}_m$, the median value $\text{median}(y)_m$ is not as strongly influenced by outliers; however as with the Half Poisson Deviance, this has higher computational cost than the Mean Squared Error.

$$H(Q_m) = \frac{1}{N_m} |y - \text{median}(y)_m|$$

Random Forests

Random Decision Forests combine multiple decision trees or an ensemble - hence the term ensemble learning method. The method of combination is often the focus of the differing random forest algorithms. The default method for classification is majority voting, and regression is the average of all trees. This sounds simple in practice, but trees are not necessarily the same, leading to more complicated tree structures that are often hard to interpret. This leads to the motivation for the use of a kernel trick random forest.

The trees are grown to maximum depth, to the point of extreme overfitting. This allows low bias at the cost of extreme variance. These can be scaled back to keep the underlying patterns intact. The method of averaging the trees reduces the variance with the trade off being a slight increase in bias and loss of interpretability.

Bootstrap Aggregating

Bootstrap **agg**regat**ing**, or bagging as it is more commonly known, is an algorithm that improves the accuracy and stability of machine learning algorithms. Given training set D of size n, bagging generates m new training sets D_i, each of size n^*, by sampling from D uniformly and with replacement (so some observations may be repeated in each D_i). If $n' = n$, then for large n, the set D_i is expected to have $1 - \frac{1}{\exp}$ of the unique samples of D, the rest being duplicates [61]. This is

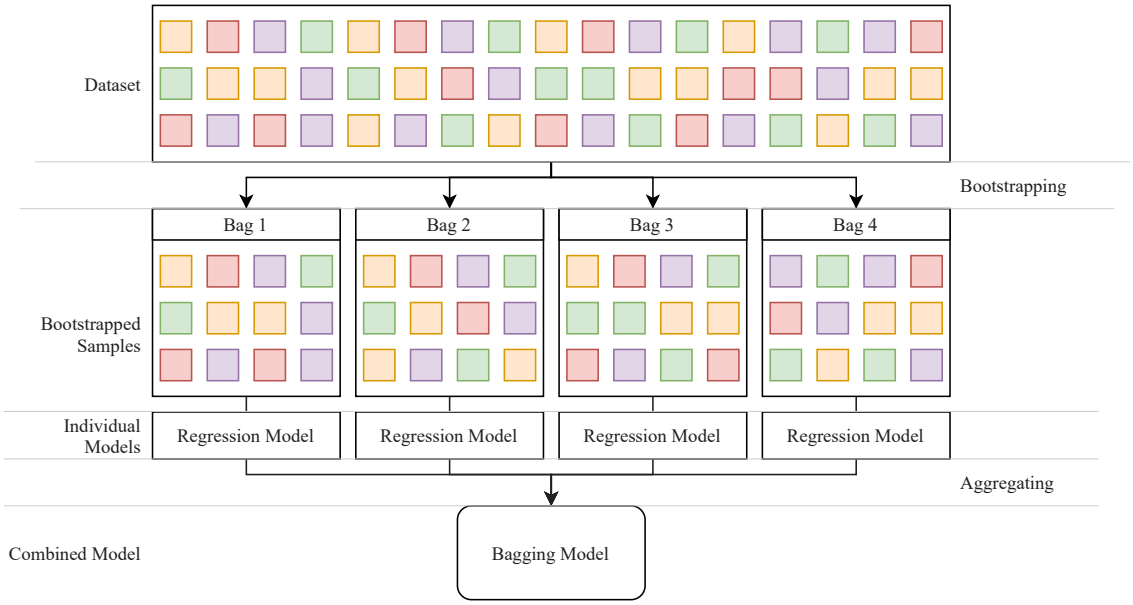


Figure 4.5: Bootstrap Aggregating with $m = 4$

known as a bootstrap sample; sampling must be performed in a replacing manner to ensure bootstraps are independent. The m models are then fitted individually using the bootstrap D_m and combined with whichever desired algorithm (averaging for regression and voting for classification). This is illustrated in Figure 4.5.

According to Brieman, bagging improves unstable procedures, which is especially important for regression trees [32]. Additionally, it increases model performance due to decreasing model variance without increasing bias. This is due to bootstrapped trees independence and thus reducing noise sensitivity. An estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all individual trees.

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B} B(f_b(x') - \hat{f})^2}{B-1}}$$

The optimal number of samples, m, can be found using cross validation of multiple m values. In conjuction, by comparing out-of-bag error, this being the mean prediction error on each training sample x_i, using only the trees which did not contain x_i in their sample [34].

Kernel Random Forest (KeRF)

Random forests are effective due to their combination procedures. They possess strong practical performance especially in high dimensional datasets, there are various studies that link the capability of kernels applied to random forests. The full detail of connecting this theory is presented by Scornet, which states the benefits are interpretability [62]. It states that KeRF estimates compare favourably to random forest estimates. Random forest estimates satisfy, for all $x \in [0,1]^d$

$$m_{M,n}(x, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{j=1}^{M} (\sum_{i=1}^{n} \frac{Y_i \mathbf{1}_{x_i}}{N_n(x, \Theta_j)}),$$

where $A_n(x, \Theta_j)$ is the cell containing x, designed with randomness Θ_j and data set $\mathcal{D}_n$ and

$$N_n(x, \Theta_j) = \sum_{i=1}^{n} \mathbf{1}_{x_i},$$

is the number of data points falling in $A_n(x, \Theta_j)$. Note that the weights $W_{i,j,n}(x)$ of each observation Y_i defined by

$$W_{i,j,n}(x) = \frac{x_i}{N_n(x, \Theta_j)},$$

depend on the number of observations $N_n(x, \Theta_j)$. Thus, the cells with higher density provide a higher contribution than those in less populated cells. This is particularly true for non-adaptive tests since the number of observations in each cell cannot be controlled. Giving necessary weights to observations that are in low-density cells can potentially lead to rough estimates. An extreme example of this is trees of non-adaptive forests containing empty cells leading to a substantial misestimation - as the prediction is empty cells is by default set to 0. To improve the random forest methods and compensate for the misestimation induced by random forest weights, the natural idea is to consider Kernel Random Forest (KeRF) estimates defined for all $x \in [0, 1]^d$ by

$$\tilde{m}_{M,n}(x, \Theta_1, \ldots, \Theta_M) = \frac{\sum_{j=1}^{M} \sum_{i=1}^{n} Y_i \mathbf{1}_{x_i}}{\sum_{j=1}^{M} N_n(x, \Theta_j)},$$

where $\tilde{m}_{M,n}(x, \Theta_1, \ldots, \Theta_M)$ is equal to $\overline{Y_i}$ of the observations containing x somewhere in the forest. Thus, each observation is weighted by the number of times it appears in the trees of the forest. Consequently, in this setting, an empty cell does not contribute to the prediction. The proximity between KeRF estimates $\tilde{m}_{M,n}$ and random forests is well proved in [62], the summary of this is for all $x \in [0, 1]^d$ we have that

$$\tilde{m}_{M,n}(x, \Theta_1, \ldots, \Theta_M) = \frac{\sum_{i=1}^{n} Y_i K_{m,n}(x, X_i)}{\sum_{\updownarrow=1}^{n} K_{m,n}(x, X_\ell)},$$

where

$$K_{M,n}(x, z) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_z.$$

This states that KeRF estimates have a more interpretable form than random forest estimates since their kernels are the connection functions of the forests. This is the key benefit of the the KeRF over an RF, and is likely to always return lower accuracy. This connection function can be seen as a geometrical characteristic of the cells in the random forest. Fixing X_i, the quantity $K_{M,n}(x, X_i)$ is solely the empirical probability that X_i and x are in the same cell in the finite random forest M. Hence, the connection function is a natural pathway to build kernel functions from random forests, a fact that Breiman noticed and Davies proved to have the property of being positive semi-definite [34, 63].

4.4 Evaluation Strategy

The methods by which these modelling approaches are evaluated is critical to both measure performance of each modelling approach and ensure the methods can be compared directly. In general, the highest accuracy is the most desirable property for regression models, but the best method to measure this accuracy is up for debate.

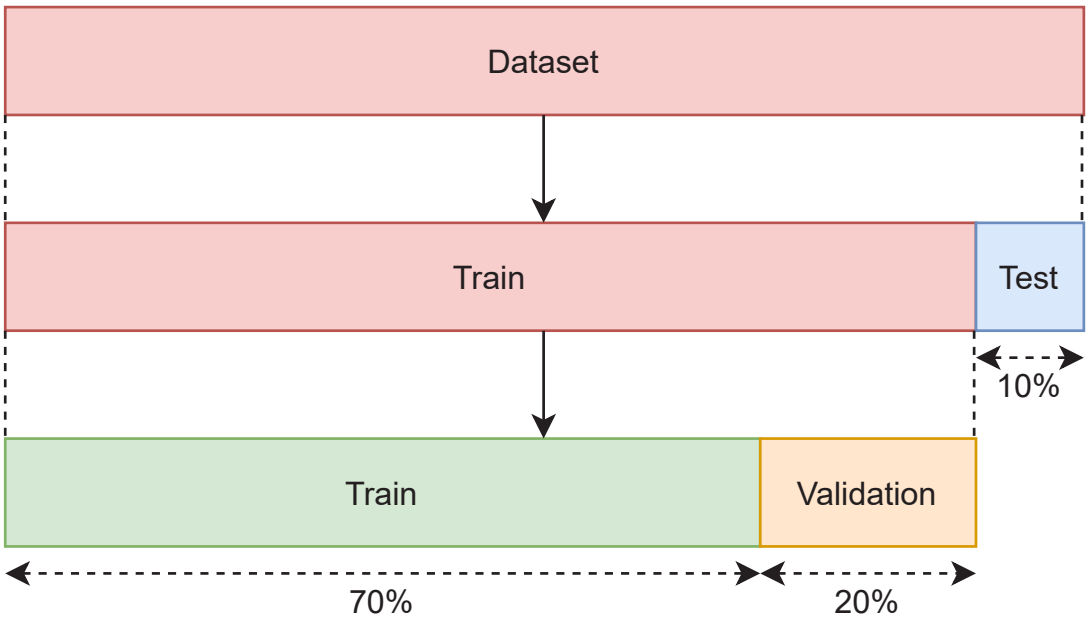


Figure 4.6: Test Train Validation Split

Cross Validation

Cross Validation (CV) is a selection of techniques that seek to validate statistical analysis results when they are applied to an independent dataset. The objective function of cross validation is to measure a model's ability to predict unseen data (data that was not used to train the model) to detect modelling problems such as selection bias or overfitting [64]. Whilst there are multiple types of cross validation, they can be categorised into exhaustive and non-exhaustive; Exhaustive methods will train and test every possible combination to subset the data to obtain full coverage. Non-exhaustive will compute only a few subsets. . In most real-world datasets, performing exhaustive cross validation is extremely unlikely due to the enormous computational cost; furthermore, the non-exhaustive method can probably detect any issues within the model.

The dataset is split into new sets, namely train, test and validation (an example of which is in Figure 4.6), these sets are used in different stages of the modelling process, see Figure 4.7, to be able to statistically quantify a models performance with a degree of certainty as to how it performs on average for any observations.

Metrics

There are various methods to quantify a models success, and these are primarily based upon some metric that assesses prediction error. The choice of metric is vital for the given problem definition as most metrics have advantages and disadvantages. For example, regression metrics like explained variance will not be helpful for scoring classification problems.

Coefficient of Determination, often denoted as R^2, represents the proportion of variance explained by the model's independent variables. It provides a decent indication of goodness of fit and can measure how well unseen samples are likely to be predicted.

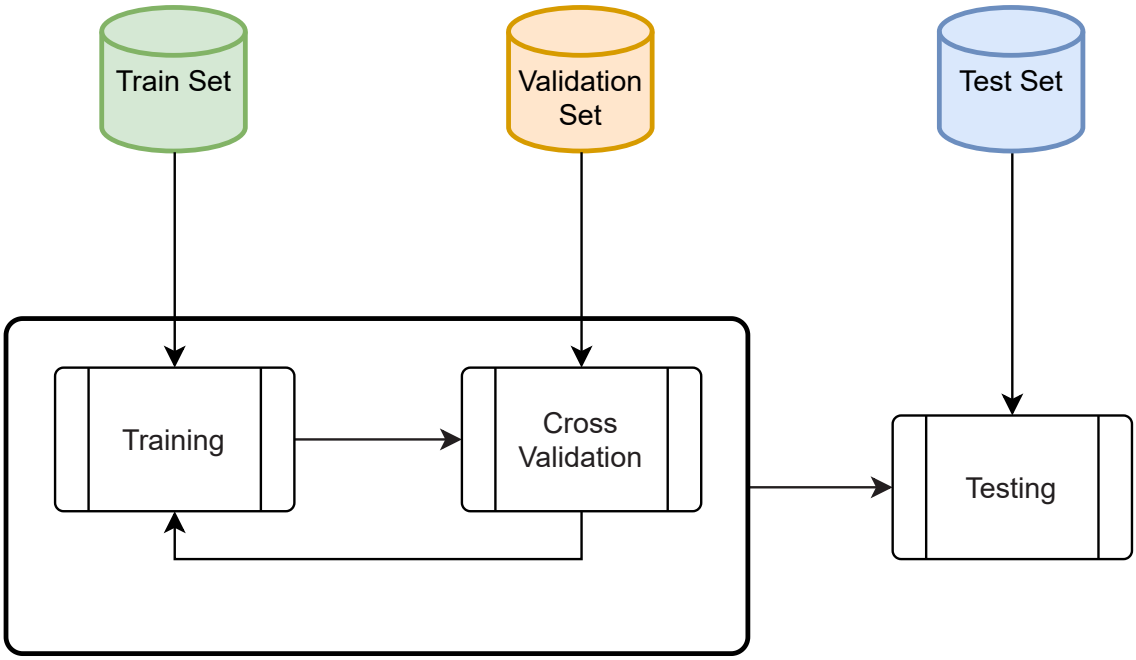


Figure 4.7: Test Train Validation Operations

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{(y_i - \hat{y}_i)^2}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.

This has a problem whereby the R^2 value increases when additional explanatory variables are added to the model. The standard fix, proposed by [65], defines the Adjusted R^2, or $\bar{y}^2$.

$$R^2(y, y^*) = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the total number of explanatory variables. The R^2 value lies within the range $(-\infty, 1]$, where 1 is a perfect fit and anything < 0 is where a model is arbitrarily worse than a horizontal hyperplane. R^2 metric is standard and is also one of the target outcomes for the UK Recipe Box Company as they have previously used it internally to evaluate model performance - however, it has many problems, which is the purpose of the inclusion of more metrics.

Root Mean Squared Error(RMSE) is a risk metric representing the expected value of the quadratic error,

$$\mathrm{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n-1} (y_i - \hat{y}_i)^2}.$$

If a model had an RMSE value of 0, that would be a model with perfect accuracy; however, this is mainly impossible in real-world datasets. RMSE has been criticised for being used as default in all applications by Berger, who stated MSE may not always be appropriate to use [66]. Furthermore, MSE heavily weights outliers, whereby more significant errors will be impacted more than smaller ones. The RMSE metric is widely used to represent regression model performance and be the

foundation of ANOVA and Ordinary Least Squares. RMSE values can be directly compared as they measure how well a model can explain a given set of observations.

Mean Absolute Percentage Deviation(MAPD) is designed to be sensitive to relative errors, and so is not affected by scaling of the target variable,

$$\text{MAPD}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n-1} \frac{|y_i - \hat{y}_i|}{\max(\epsilon, |y_i|)}.$$

where ϵ is a positive small number to prevent undefined results in the case $y = 0$.

This metric can be considered an improvement upon the Mean Absolute Error; for example, Mean Absolute Error (MAE) will ignore (or effectively mask) minor magnitude errors in the presence of significant magnitude errors.

Median Absolute Error(MedAE) Many statisticians state the benefits of the median over the mean, the error term is robust to outliers as the loss is calculated via the median,

$$\text{MedAE}(y, \hat{y}) = \text{median}(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|).$$

4.5 Changepoint Analysis

Changepoint Analysis for time series data is used to detect changes in the mean and variance across time. It seeks to determine the location and magnitude of these changes and provide confidence intervals for each change.

Single Changepoint

The detection of a single changepoint is constructed as a hypothesis test,

$$\begin{aligned} H_0: &\quad \text{No Changepoint} \quad m = 0 \\ H_1: &\quad \text{One Changepoint} \quad m = 1. \end{aligned}$$

A test statistic can be constructed which is used to decide whether a change has occurred. The likelihood ratio method requires the calculation of the maximum log-likelihood under both the null and alternative hypotheses. For the null hypothesis, the maximum log-likelihood is $\log p(y_{1:n}|\hat{\theta})$, where $p(.)$ is the probability density function associated with the distribution of the data and $\hat{\theta}$ is the maximum likelihood estimation (MLE) of the parameters. Under the alternative hypothesis, consider a model with a changepoint at τ_1, with $\tau_1 \in 1, 2, \ldots, n-1$. Then the maximum log-likelihood for a given τ_1 is,

$$L(\tau_1) = \log p(y_{1:\tau_1}|\hat{\theta}_1) + \log p(y_{(\tau_1+1):n}|\hat{\theta}_2).$$

Given the discrete nature of the changepoint location, the maximum log-likelihood value under the alternative is simply $\max_{\tau_1} L(\tau_1)$, where the maximum is taken over all possible locations. Thus, the test statistic based upon the deviance such that the null hypothesis is rejected if $\lambda > c$, where,

$$\lambda = 2\left[\max_{\{\tau_1\}} L(\tau_1) - \log p(y_{1:n}|\hat{\theta})\right].$$

Choosing an appropriate threshold c remains an open question, with some authors devising p values and other information under different types of changes [67].

Multiple Changepoint

The single changepoint methodology can be extended to multiple changes by simply summing over all m segments. However, this problem expands combinatorially. One commonly used approach to identify multiple changepoints is to minimise,

$$\sum_{i=1}^{m+1} [\mathcal{C}(y_{\tau_{i-1}+1):\tau_i})] + \beta f(m).$$

Where $\mathcal{C}$ is some cost function for a segment and $\beta f(m)$ is a penalty to guard against overfitting [46]. This can be solved with a brute force approach; however, there are 2^{n-1} solutions, which reduces to $(\frac{n-1}{m})$ solutions if m is known. The most widely used search method is Pruned Exact Linear Time (PELT), which improves the binary segmentation and segment neighbourhood algorithms. The binary segmentation method is an approximate minimisation, but the computational cost is considerably lower than the segment neighbourhood algorithm, which is exact. PELT improves on both of these, being exact whilst having lower computation costs than binary segmentation (due to leveraging dynamic programming and pruning), resulting in an $\mathcal{O}(n)$ search algorithm subject to some not particularly onerous assumptions.

PELT is a modification of Optimal Partitioning; it removes values of τ that could never be minima from the search space at each iteration of the partitioning algorithm [46]. Killick proved that if there exists some constant K such that $s < t < T$,

$$\mathcal{C}(y_{(s+1):t}) + \mathcal{C}(y_{(s+1):T}) + K \leq \mathcal{C}(y_{(s+1):T}),$$

and for $t > s$, if

$$F(s) + \mathcal{C}(y_{(s+1):t}) + K \leq F(t),$$

then at some point, $T > t$, s can never be the optimal final changepoint before T; hence t can be removed from the set of possible values.

Algorithm 3 PELT

inputs

A dataset of the form $y_{1:n} = (y_1, y_2, \ldots, y_n)$

A cost function $\mathcal{C}$ dependent on the data

A penalty constant β, and a constant K

Let $c(p) = 0$, $rescp(0) = 0$, $F(0) = 0$, $m(0) = 0$ and $R_1 = 0$
for $t \in 1, \ldots, n$ **do**
  Calculate $F(t) = \min_{s \in R_t}[F(s) + \mathcal{C}(y_{(s+1):t}) + \beta]$
  Let $cp(t) = \arg\min_{s \in R_t}[F(s) + \mathcal{C}(y_{(s+1):t}) + \beta]$
  Let $m(t) = m(cp(t)) + 1$
  Set $rescp(t) = [rescp(cp(t)), cp(t)]$
  Set $R_{t+1} = s \in R_t : F(s) + \mathcal{C}(y_{(s+1):t}) < F(t)$
end for
return changepoints of the optimal segmentation

CHAPTER 5

Case Study: UK Recipe Box Company

The core idea behind the UK Recipe Box Company is to provide high quality simple recipes delivered directly to consumers. Sustainability is a key goal, and this goes hand-in-hand with optimising the supply chain to reduce costs and simultaneously increase production capacity to supply the increasing demand.

Whilst the process does not begin at the factory, the region of interest for this project starts from there (see Figure 5.1). The factory serves as a central point for ingredients, and functions similarly to a traditional grocery store. The recipes are then dispatched in boxes with the relevant ingredients directly to the consumer via courier service to ensure ingredients stay fresh and chilled ingredients do not expire.

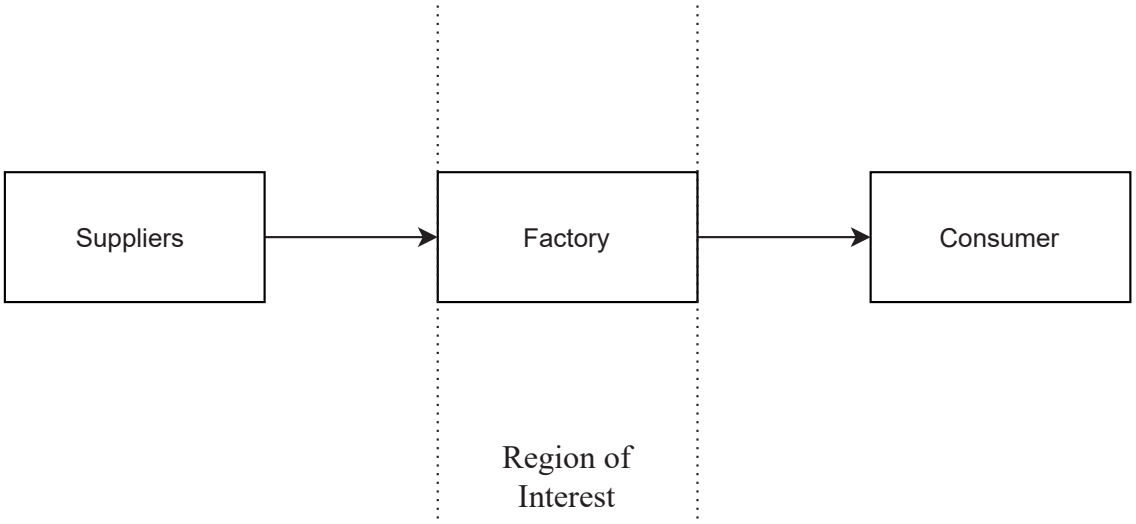


Figure 5.1: Supply Chain Region of Interest

A production line is central to the operation as all boxes travel down the production line to pick up the order. Ingredients are classed into Stock Keeping Units (SKUs), and these are stored in 'Pick Stations' located by the side of the main production line; each station is crewed by a human operator and stores a different variety of SKUs. As the box travels along the production line, it will pass pick stations; if the box needs to stop at a station, it will come off the mainline and join a station queue. At a visited station, the required ingredients are selected by the

operator and placed into the box, and it then re-joins the central conveyor onto its next destination station - see Figures 5.3 and 5.4.

There are sections of the production line which are at ambient temperature; these will stock fresh fruit and vegetables as well as dried goods such as pasta or herbs, and a refrigerated section which stocks fresh meat and dairy – the refrigerated goods are put into a recyclable compartment of the Recipe box which contains cooling aids. An example layout is shown in Figure 5.2.

A random selection of Recipe boxes is subject to quality control measures at the end of the line, whereby the contents are manually checked to ensure that the entire contents are present and not damaged.

Throughput is the measure of boxes completed per hour, scaled by the up-time of that hour. This relationship is highlighted as a source of problematic assumptions.

Customer Process

From a customer perspective, customers can log in to a web portal where they select 2-4 recipes for 2 or 4 people and then select a day where their box will be delivered. The box is delivered with all the recipe cards and all necessary associated ingredients to create their selected recipes. As a recurring process, the customer must select their next week's recipes three days before delivery; otherwise, a random selection is sent - customers can, however, opt-out at any point. There are two sizes of boxes, based upon how many ingredients they can hold; this is considered when planning the boxes to be put onto the production line at any specific hour.

Production Line Considerations

There are two critical algorithms at play within the production lines:

1. Order Routing Algorithm (ORA)
2. Pickface Optimisation (PFO)

The pickface optimisation algorithm dictates which SKUs are stocked in stations. The goal of this algorithm is to spread common SKUs across the production line to prevent pooling. Once this has been calculated, the next stage is the Order Routing Algorithm which allocates which stations any specific box will visit.

This algorithm is known to show high positive correlation with throughput, but evaluating the impact is difficult as it is affected by various uncontrollable and immeasurable factors, i.e. workforce efficiency [1]. The status of these algorithms is unknown, both how they operate and when they change?

Workforce Considerations

There are four shift teams for each production line, labelled by colour; an example 24 hour period is shown in Figure 5.5. The skill and efficiency of an individual worker is immeasurable with the provided dataset, so alternatively, the workforce as a whole must be considered. The only bias that could exist is between day and night, as it

[1] However there are some variables within the dataset that could measure these; they are not independent variables and are affected by multiple sources

could suggest algorithmic differences in the production line algorithms. Any bias introduced by a specific workforce always working the same shift is removed by the pattern by which they operate, shown in Figure 5.6, where the two nights shifts rotate on an eight-week cycle, as do the day shifts, which are effectively mirrored every four weeks.

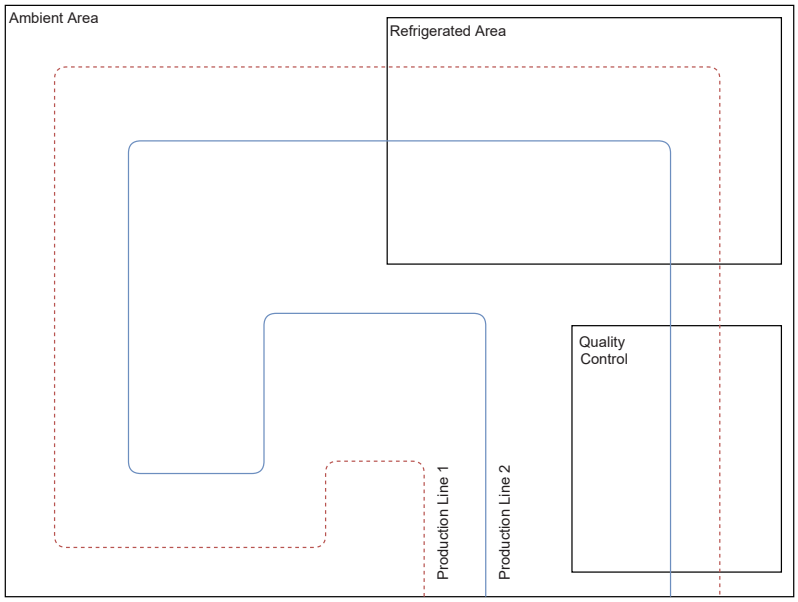


Figure 5.2: Example layout of a single factory with 2 production lines

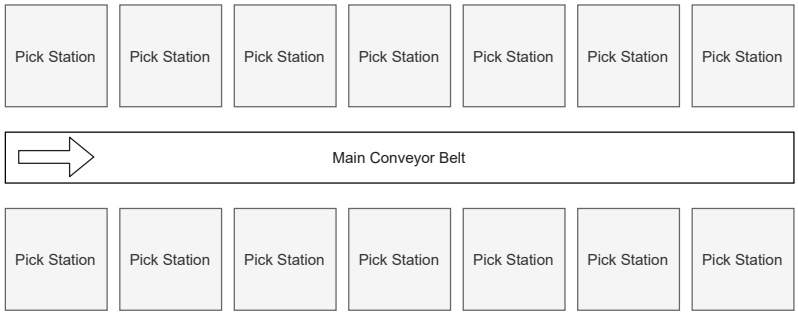


Figure 5.3: Demonstrative layout of a production line

Briefly exploring two of the relationships that likely highly influence throughput are demonstrated in Figure 5.7. The relationship appears to be non-linear, with multiple relationships that depends on both run time and boxes completed - modelling this is difficult to achieve.

5.1 Processing

The original dataset contains 36 predictor variables and one desired target variable; some of this data is related to the setup, but most are summary statistics from the hour of run time. These are a mix of categorical, time and numeric types, which must all be handled differently.

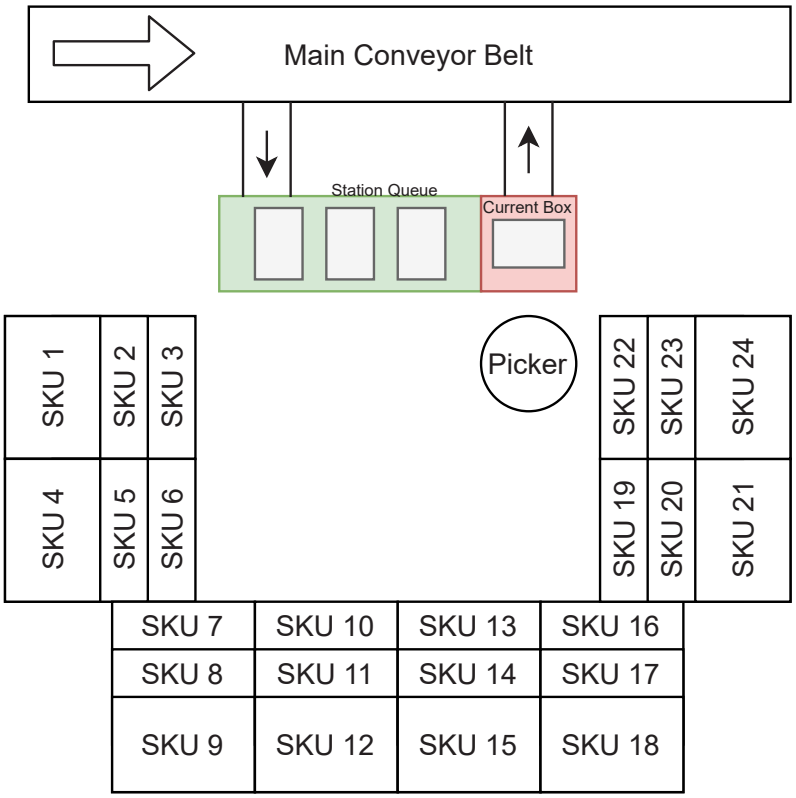


Figure 5.4: Example configuration of a 'pick station' with a queue size of 3 and 24 SKU's.

The data must be inspected for any irregularities that must be dealt with; whilst this depends on the dataset at hand, this includes checking for duplicates, zero values, outliers/noise and any required transformations.

In order to deal with any irregularities in the dataset, it can be very informative to understand how they have been introduced. Figure 5.8 shows a density plot of duplicate values against menu week; it shows a relatively constant number of duplicate values across production lines 1 and 2 throughout the study period, with production lines 3 and 4 introducing roughly equal in the few weeks in which they were operational. The righthand plot of Figure 5.8b shows the counts of individual menu week duplicate occurrences. It is acceptable to assume all production lines have introduced duplicates according to the same distribution; any stochastic element is unknown with these two figures. It is common to deal with duplicates by simply removing them, as there are no other easy options.

Outliers can be handled in two ways, as discussed in a previous chapter, either by using a model that is capable of accounting for the presence of outliers or by removing values that are 'decided/classified' as outliers.

It could be argued that classifying a value as an outlier is simple; any value outside of a specified range (i.e. interquartile range, or two standard deviations) is an outlier. However, these methods require assumptions about the data which cannot be guaranteed, namely the symmetric nature of these methods.

An alternative is to decide thresholds by which values should be removed; Figure 5.9 presents the distribution of the Average Station Visits statistic in its original form, and with values above 20 cut off, this removes only 448 values; however, this threshold was manually decided by visually examining and comparing Figure 5.9b

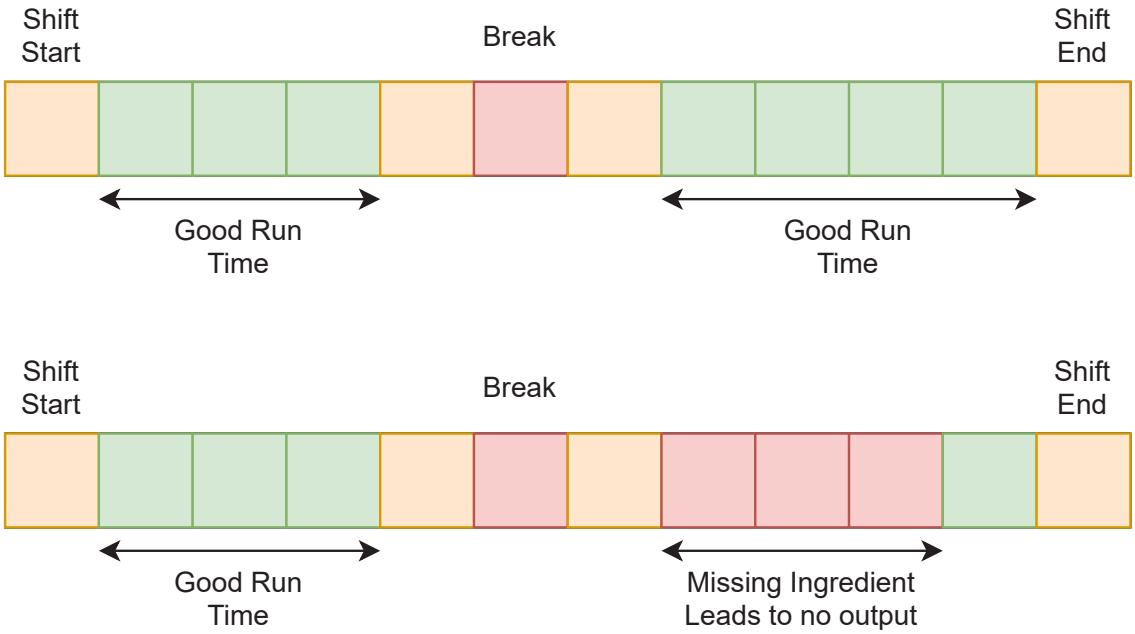


Figure 5.5: Example 12 hour running period of 2 shifts

with 5.9d. Figures 5.9b and 5.9a suggest an extremely heavy tail for Average Station Visits. It is known from interviews and site visits that whilst Average Station Visits does have a heavy tail; it is not the magnitude represented by the dataset.

The method of determining the best threshold value is contentious; it is a trade-off between the dataset being too clean and thus not modelling the process, or including outliers that are erroneous and thus not modelling the process. This method is equivalent to one-sided $n\%$ truncation and is similar to using a trimmed estimator - these methods are based upon the core concept that is using alternative measurements such as interquartile ranges or the median instead of mean derived values [2]. An alternative suggested by the UK Recipe Company is winsorization; the key difference between winsorization and truncation or trimming is that data is not excluded but censored. Figure 5.10 shows winsorization performed at four levels, $0.05, 0.025, 0.01, 0.001$; these levels represent the typical significance intervals, and effect this has on the distribution of average station visits. These show an undesirable property which is the clear and well-defined boundary at the x^{th} percentile. This effect is more pronounced with lower percentiles, such as 95%, compared with higher such as 99.9%; however, this also retains more extreme outliers. Additionally, it is undesirable that the lower bound is modified due to the location of the mean, $\mu \approx 5$, being relatively close to 0, and whilst it is very unlikely that boxes can complete their journey in a single visit, it is proportionally more likely than visiting 50, which is the 95% level; in other terms, the distribution is not symmetric and definite positive. This is the motivation for investigating 1-tailed winsorization, a relatively unknown and unused method; the exact percentiles performed for 2-tailed, are shown for 1-tailed in Figure 5.11.

However, it is well documented that winsorization has significant problems that cannot be dealt with via transformations and so will not be used further.

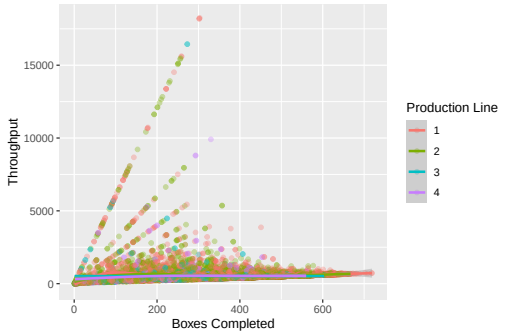
[2] whilst it is noted there are 'trimmed' means, these require more information regarding the distribution and unless the distribution is symmetric, and is likely not an unbiased estimator of the mean or median

	Monday		Tuesday		Wednesday		Thursday		Friday		Saturday		Sunday	
	Night	Day	Night	Day	Night	Day	Night	Day	Night	Day	Night	Day	Night	Day
Week 1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 403						Menu Week 404							
Week 2	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 404						Menu Week 405							
Week 3	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 405						Menu Week 406							
Week 4	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 406						Menu Week 407							
Week 5	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 407						Menu Week 408							
Week 6	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 408						Menu Week 409							
Week 7	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 409						Menu Week 410							
Week 8	0	1	2	3	4	5	6	7	8	9	10	11	12	13
	Menu Week 410						Menu Week 411							

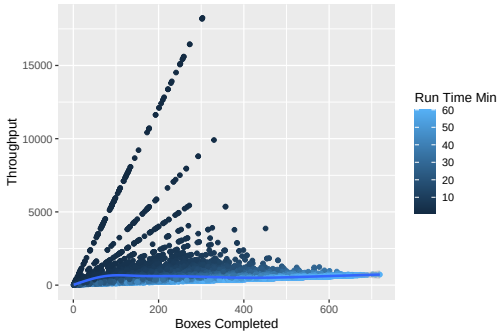
Figure 5.6: Example 8 week shift breakdown by shift colour

On the subject of manual removals, in real-world datasets, there are problems in the creation of data that cause the introduction of anomalies. For these situations, it can be conducive to simply remove any affected values. The contextual information removes data known to be incorrect. The UK Recipe Box Company detailed four suggested removals;

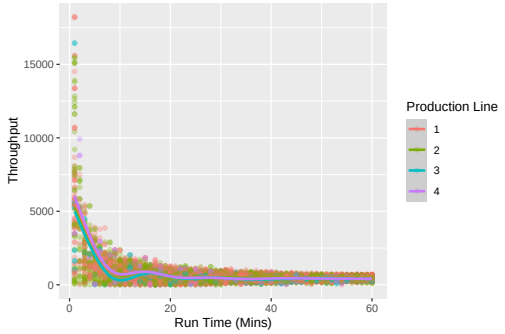
- Shift Sequence 14 - The Wednesday night shift is the most fragmented shift due to being the final segment of a menu week. This redirects production capacity to resetting the production line in preparation for the next menu week.
- Any run time below 35 minutes - The UK Recipe Box Company are mostly interested in the indicator variables of the production lines when they are running at full speed, so anything that is spending at least 25 minutes stopped does not represent that.
- Menu Weeks 405-410 for production lines 1 and 2- There was a plant shutdown (reasons for which are unknown)



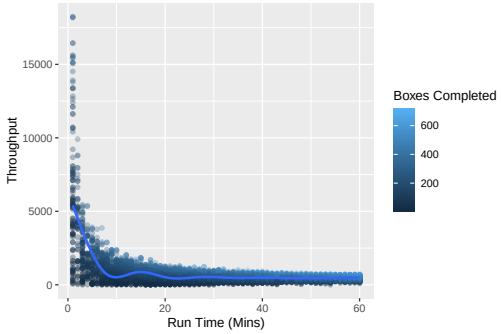
(a) vs Boxes Completed by Production Line



(b) vs Boxes Completed by Run Time

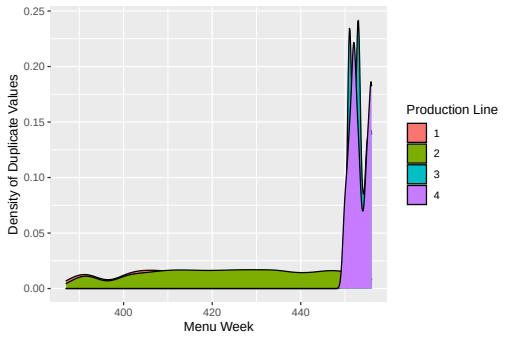


(c) vs Run Time by Production Line

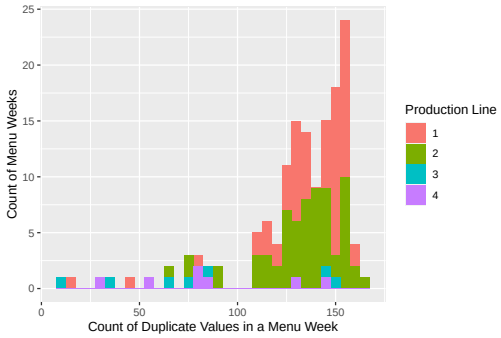


(d) vs Run Time by Boxes Completed

Figure 5.7: Throughput vs Contributory Factors

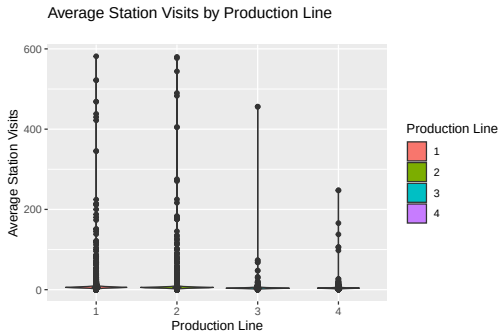


(a) menu week

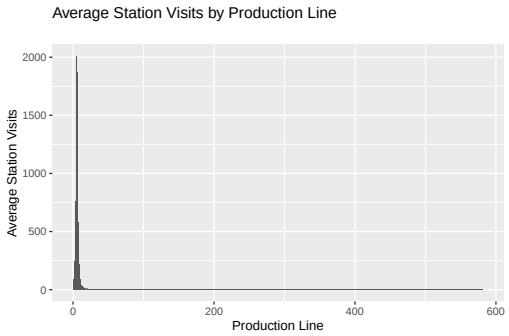


(b) production line

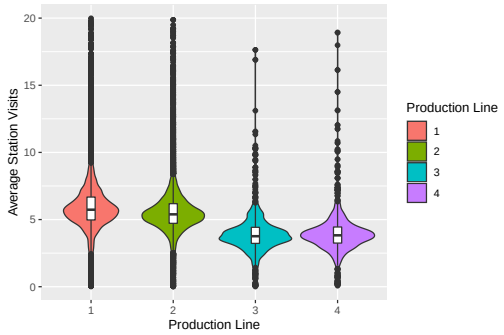
Figure 5.8: Duplicate Values location and density



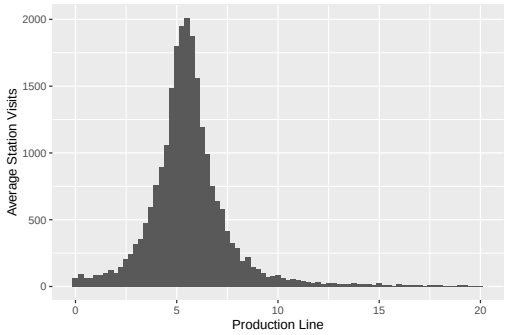
(a) standard violin plot



(b) standard histogram

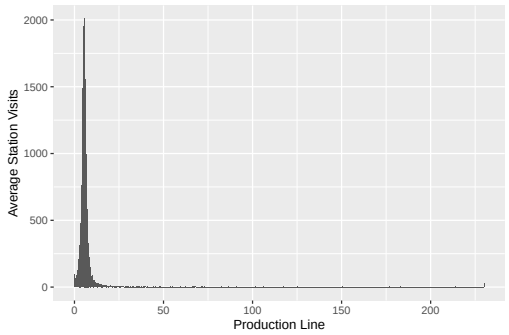


(c) violin plot $x < 20$

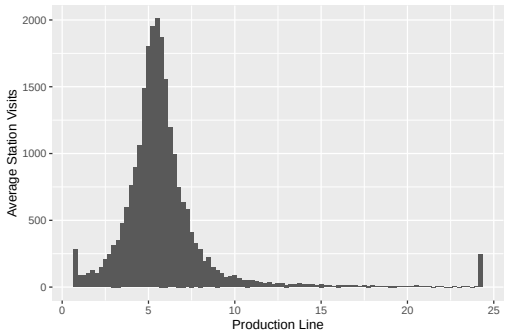


(d) histogram $x < 20$

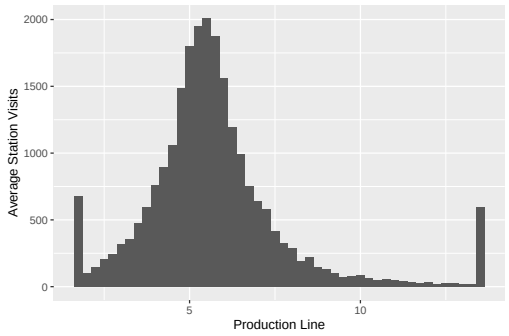
Figure 5.9: Average Station Visits Distribution. Standard and cut $x < 20$



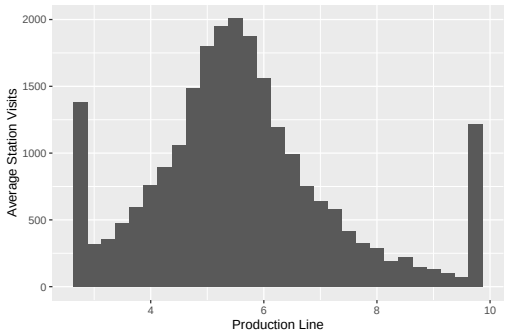
(a) 0.001



(b) 0.01

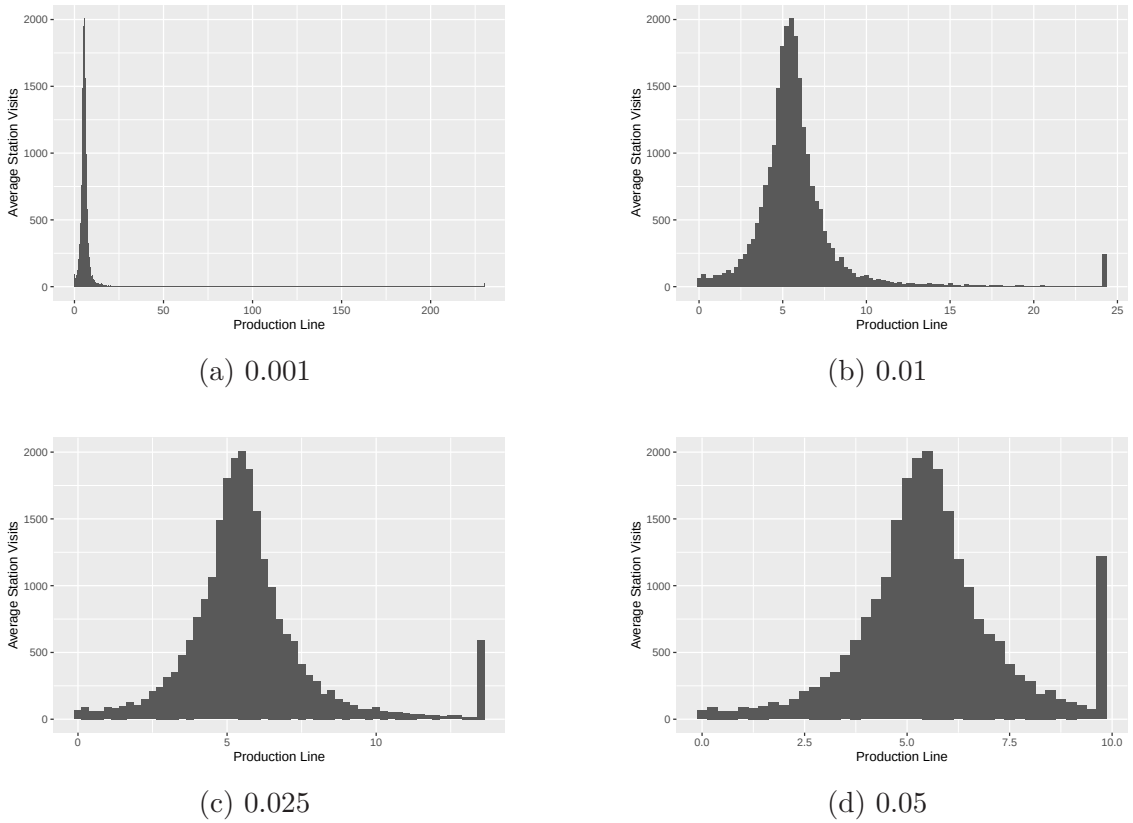


(c) 0.025



(d) 0.05

Figure 5.10: Average Station Visits Distribution 2-tail winsorized at varying quantiles



(a) 0.001

(b) 0.01

(c) 0.025

(d) 0.05

Figure 5.11: Average Station Visits Distribution 1-tail winsorized at varying quantiles

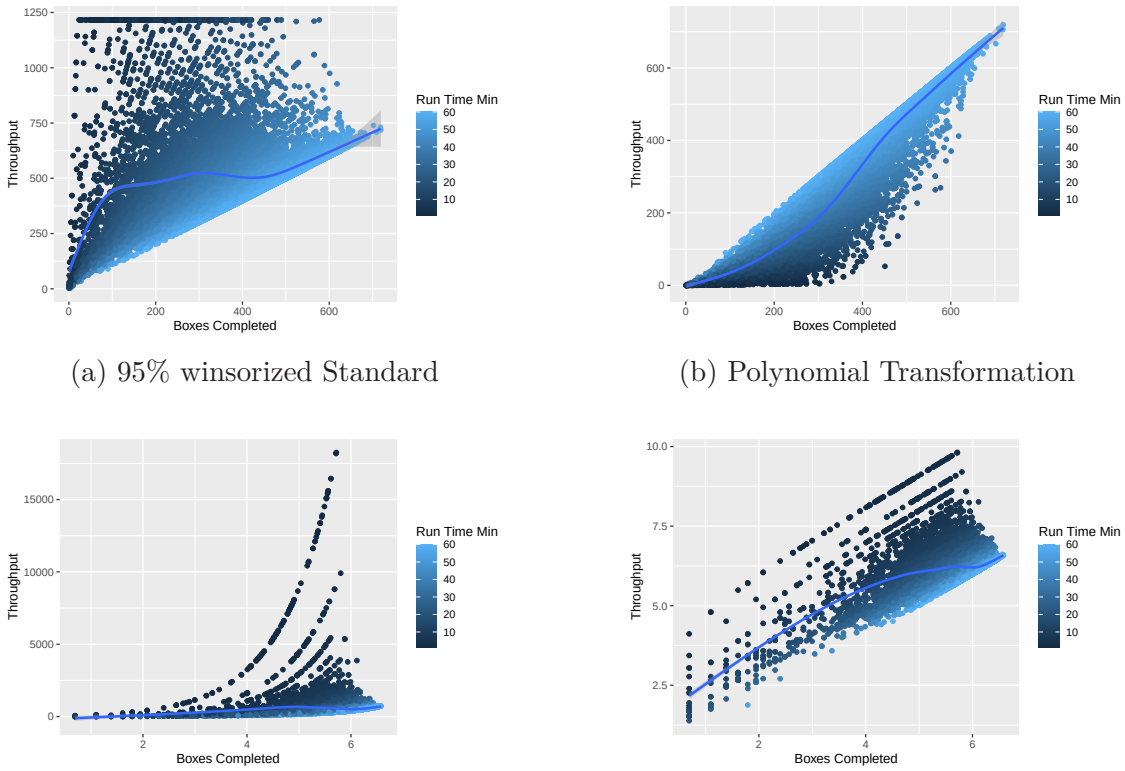
- Menu Weeks > 451 for production lines 3 and 4 - the lines were still facing commissioning problems being a newly built factory.

Heteroscedasticity

When plotted against boxes completed, the prediction variable, throughput, appears to require different models for differing running times, and it also appears that the model is heteroscedastic. A polynomial transformation to the throughput variable leads to that shown in Figure 5.12b; this is a more desirable relationship than the un-transformed; however, the model is still heteroscedastic. As noted previously, this can be corrected by taking the natural log, shown in Figure 5.12c, but it would appear to be a multiple regression problem. Testing this model with the Studentised Breusch-Pagan test for heteroscedasticity [68, 69], a test value of 1636 and a $p < 2.2 \cdot 10^{-}16$ proves that heteroscedasticity is still present.

Initially, the problem statement looked to understand the factors associated with throughput; however, as throughput is directly calculated from the number of boxes completed and the run time, this leads to a non-linear problem where one of these explanatory variables must be removed from the model.

The relationship between run time and boxes completed is crucial to the model; hence a proposed remodel to use the number of boxes completed as the predicted variable and remove throughput, as it can be calculated after the fact with confidence intervals. The relationship between boxes completed and run time is shown in Figure 5.13, with the cubic spline regression shown for each production line. This has some positive effects on the modelling; the critical positive is the problem now semi-linear



(a) 95% winsorized Standard

(b) Polynomial Transformation

(c) Semi Log Transformation

(d) Natural Log Transformation

Figure 5.12: Throughput vs Boxes Completed for various transformations with a polynomial regression spline

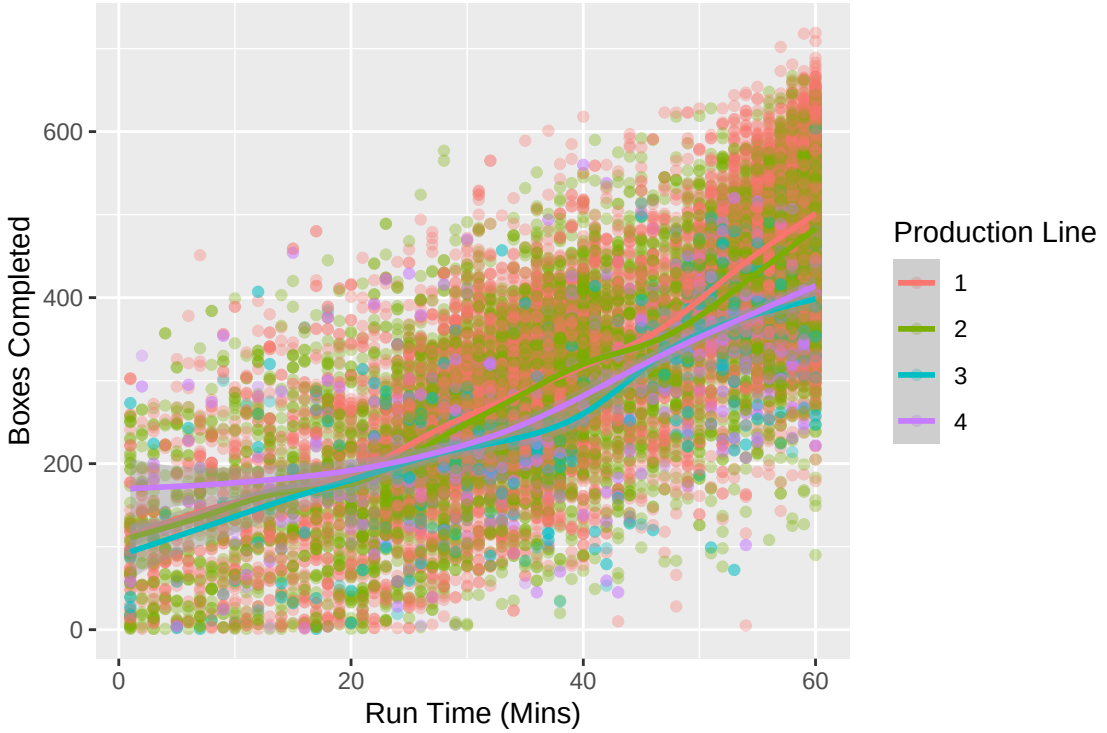
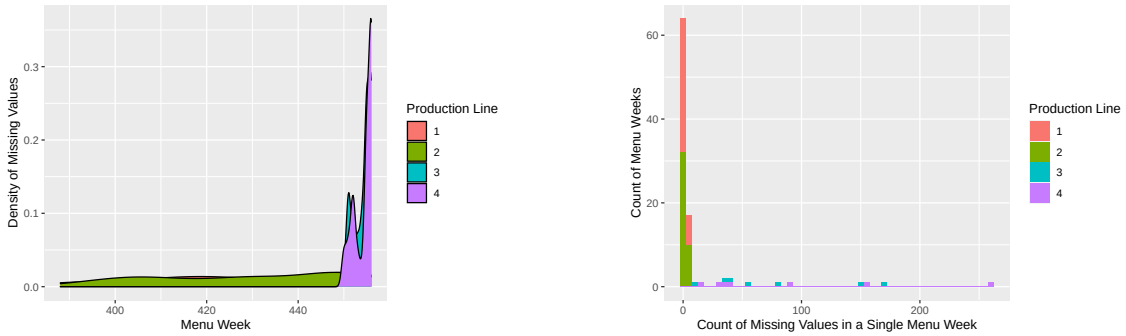


Figure 5.13: Boxes Completed vs Run Time for individual production lines with polynomial regression splines



(a) menu week

(b) counts of occurrences

Figure 5.14: Missing Value vs Menu Week

(linear with a small order non-linear term) with relatively constant error terms. This eliminates any log or box-cox transformation requirement, which allows the model to retain its ease of interpretation. The drawback of this approach is that to return to the original problem requires confidence intervals to estimate throughput.

Missing Values

To analyse the missing values, they are inspected in the same way as duplicated. The missing values are shown in Figure 5.14, a similar pattern for the density (Figure 5.14a), with production lines 1 and 2 containing a relatively constant number of missing values (generally 0, 1 or 2) throughout the whole dataset. In contrast, production lines 3 and 4 produce significantly more, ranging from 12 in menu week 450 to 258 in menu week 456 - this is reflected in the histogram of counts in Figure 5.14b. For the case of missing values, it is not safe to assume these were introduced stochastically or constantly, so dealing with these requires some more complexity, hence requiring the use of MICE (Multiple Imputation by Chained Equations).

The MICE model takes the entire dataset as input and builds individual models for each variable. The algorithm can impute mixes of variable types (continuous, binary, unordered categorical and ordered categorical), so no further processing is required to create the models. The algorithm imputes an incomplete column (the target column) by generating 'plausible' synthetic values given other columns in the data for each column with missing values. If a predictor column has missing values, the most recent generation imputation is used. The method used to calculate the 'plausible' values is random forests, a good balance of efficiency and accuracy, for five iterations. This gives a result similar to cross validation and allows the assumption that, given that the missing values have a stochastic element, the imputation is good enough at 95% confidence.

Running on an AWS EC2 p2.xlarge instance, which consist of high-frequency Intel Xeon E5-2686v4 and high-performance NVIDIA K80 GPUs, this modelling took $\tilde{1}0$ to fill the missing values. This dataset is the dataset used throughout the remainder of this case study.

For regression modelling, the data is converted into types (categorical, integer or float) then split into the X data and y labels, where y is the number of boxes completed. The numeric columns within dataset X are then scaled using a robust scaler which removes the median and scales the data according to the interquartile

range. The centering happens independently on each feature by taking the relevant statistics on the samples in the training set. Any categorical data is One Hot Encoded.

Collinearity

The correlation between two predictor variables can be problematic, as the variables cannot independently predict the value of the dependent variable. A collinearity matrix is shown in Figure 5.15 for all numeric variables. The colour represents the value, and the areas represent the absolute value of corresponding correlation coefficients. Some variables with extremely high correlation exist, such as unit picks per box and SKU picks per box. A standard metric to measure the severity of multicollinearity is the variance inflation factor (VIF). A common rule of thumb is that anything VIF > 10 has high multicollinearity (although 5 is also commonly used).

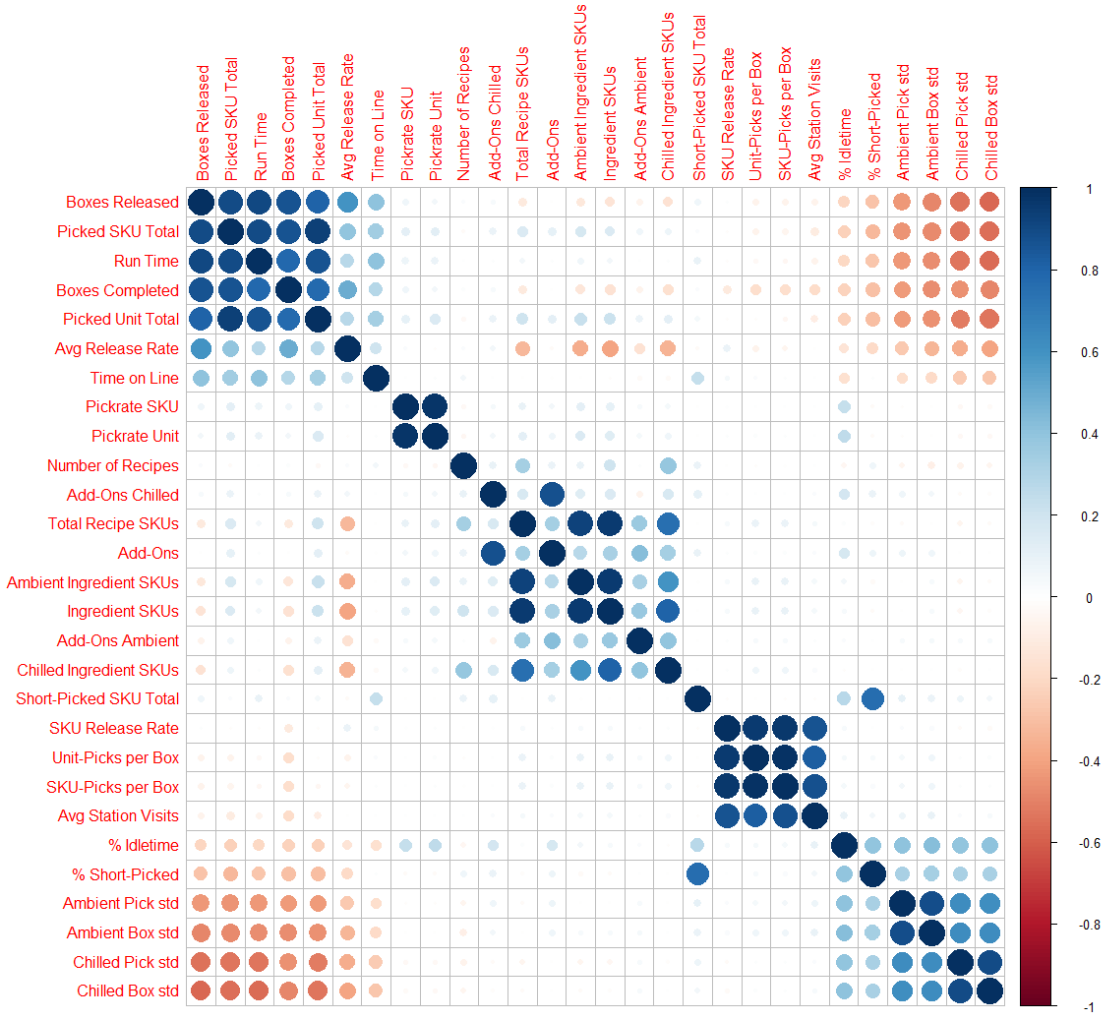


Figure 5.15: Collinearity Matrix Visualised by circle colour and size, ordered by the first principal component order

The variables with VIF > 10 is shown in Table 5.1; most of these values are to be removed due to their significant overlap with other variables in the dataset, such as Add Ons with Add Ons Chilled and Add Ons Ambient. If a VIF threshold of 5 were used, the model would consist of only five numeric variables.

Table 5.1: Variance Inflation Factor VIF > 10

Variable	VIF	Variable	VIF
Ingredient SKUs	7373.49	Unit Picks per Box	37.11
Add-Ons	6263.64	Pickrate Unit Mins	34.89
Add-Ons Chilled	5151.12	Pickrate SKU Mins	33.63
Ingredient SKUs Ambient	4128.94	Run Time Mins	28.07
Add-Ons Ambient	1490.15	SKU Release Rate	25.30
Ingredient SKUs Chilled	941.65	Picked SKU Total	23.35
SKU Picks per Box	58.06	Total Recipe SKUs	20.30
Boxes Released	37.48	Picked Unit Total	15.29

It is known that some relationships are non-linear, and for this reason, all variables that exceed the VIF threshold are not removed. Namely, those that are excluded are only those which are mostly covered or summarised in other variables. This would pose a problem with Ordinary least squares or similar methods, but for methods capable of non-linear modelling the multicollinearity should be tolerable.

5.2 Regression Modelling

The primary modelling to be conducted is regression models to understand the relationships and the magnitude of those relationships within the dataset. All the experiments are performed within individual docker containers running on AWS EC2 p2.xlarge instances. The docker containers allow perfect results reproducibility, and the AWS instance has a powerful GPU to accelerate computational time. This combination allows for methods to be iterated upon quickly and consistently. The dataset is split into a test, train and validation set using a ratio of $[80 : 10 : 10]$ where the train set is used to train the model, validate set is used to score the model and test is used to report the metric values so to entirely remove bias in the testing procedure. This strategy is to give a better picture of how the model performs on unseen data. This is required because of how the data is structured; for any given menu week, there are $2 \times 24 \times 7 = 336$ entries with the same menu setup, which leads to significant effective overfitting and is unavoidable realistically.

The experiments are run using repeated kfold crossover with $k = 20$ folds and $n = 15$ repeats - this allows the assumption that the results are valid for any test split of the data. The metrics used are Mean Squared Error (MSE), Median Absolute Error (MedAE) and R^2. Table 5.2 shows the results for Ordinary Least Squares (OLS), Support Vector Regression (SVR), Huber Loss regression and Theil-Sen estimator. These are all linear models, and the latter two are robust to outliers. The presented results give the mean of all 15 runs, with the standard deviation and the time to perform all repeats. The OLS method is extremely fast to compute and has decent accuracy across all metrics, notably the R^2 value of 0.86.

For random forests to be effective, the hyperparameters must be optimised, an initial random search is performed to efficiently search ample parameter space. The optimisation procedure searches across:

- Number of Estimators
- Maximum Features
- Maximum Depth
- Minimum Samples per Split
- Minimum Samples per leaf
- Bootstrap State

The first approach is randomly searching the parameter space; the defined space allows for 7200 possible combinations, the random search selects 300 of those (at random) and performs $k = 3$ cross validation. This search requires ≈ 76 minutes to perform, and the best model from this is shown in Table 5.3. The random search results form the basis for the grid search, where a search considers all possible combinations - hence the need to make the space smaller. The grid search consists of 225 candidates, each fitted with $k = 3$ cross validation and this search requires $\approx 3 : 30 hours$ to perform, again the best model is presented in Table 5.3. Grid searching is exhaustive, so the need to use multiple methods in conjunction is required. If Grid searching were to be used for the random search space, it would require ≈ 129 hours or five days of continual running, and the combination method proposed here is likely to find the global maxima.

The scoring metrics for these random forests are presented in Table 5.4 for the corresponding hyperparameters presented in Table 5.3. A default random forest can significantly improve ordinary regression trees at the cost of computation for training. The random search leads to 0.47% increased accuracy at the cost of 1 hour of searching, whilst the grid search improves upon the default model the same amount, but it required six times longer to search. A Random Search RF observes decreased mean of the mean squared error and decreased standard deviation. The grid search marginally improves the maximum error whilst using fewer estimators and a lower depth - this could be favourable for this tree to be translatable. However,

Table 5.2: Summary Results for Standard Testing using repeated kfold $k = 20,\ n = 15$ for multiple metrics and methods

Metric	Score	OLS	SVR	Huber	Theil Sen
MSE	Mean	2620.90	3266.06	33636.47	130589.96
	Std	531.74	234.16	60976.70	249259.13
	Time	00:02.48	15:02.56	00:38.93	05:21.92
MedAE	Mean	30.23	38.38	21.05	11.16
	Std	1.48	2.12	1.08	0.52
	Time	00:01.80	10:54.36	00:28.65	04:28.68
R^2	Mean	0.86	0.83	-0.67	-5.62
	Std	0.02	0.01	2.80	12.75
	Time	00:01.76	16:01.83	00:34.55	04:44.68

Table 5.3: Random Forest Hyperparameter Optimisation Search Optimal Parameters

	Default RF	Random Search RF	Grid Search RF
Estimators	100	1200	900
Max. Depth	None	171	160
Min. Leaf Samples	1	2	2
Min. Split Samples	2	2	4
Max. Features	n features	n features	n features
Bootstrapped	True	True	True
Search Time	0	01:16:53.36	03:30:31.42

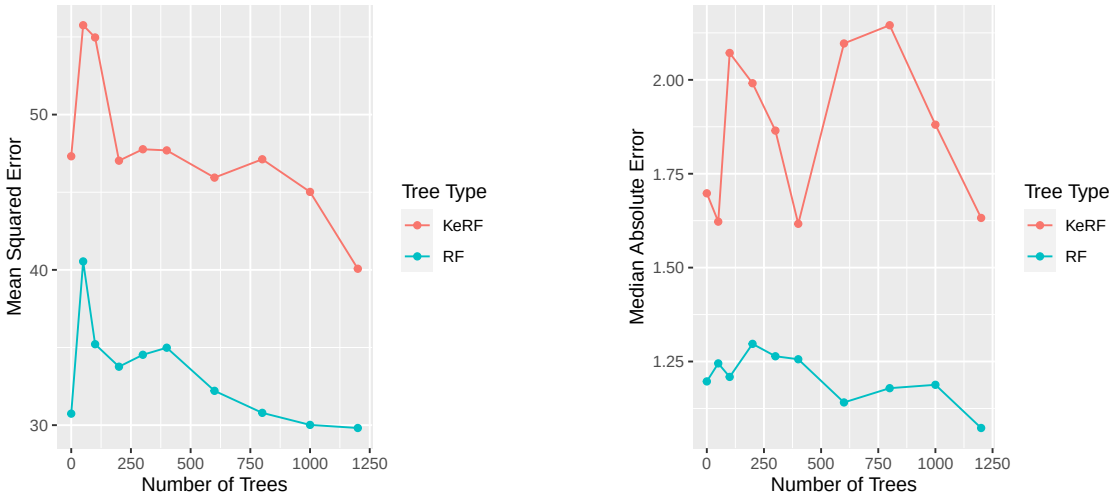
Table 5.4: Random Forest Hyperparameter Optimisation Search Optimal Results

		Default RF	Random Search RF	Grid Search RF
MSE	Mean	30.792	30.016	29.817
	Std	15.137	14.562	14.307
	Time	0:12:55	02:46:10	1:48:11
MedAE	Mean	1.179	1.188	1.073
	Std	0.069	0.0724	0.068
	Time	0:12:59	2:24:06	1:47:01
% Deviance		96.17%	96.62%	96.62%
Max Error		3.3298	2.3168	2.3118
Relative %		0%	0.47%	0.47%

for the Grid Search RF the median absolute error observes an increase in both mean and standard deviation of results - potentially due to the method of subsetting when scoring - using the $n = 15$ cross validation reduces the bias of the results, but still has some margin of error. At a training time of 2:46 minutes, this approach significantly improves upon the default random search if a 0.47% increase or ≈ 1 degree maximum error improvement is worthwhile. However, the grid search RF produces lower mean squared error and median absolute error metrics than both alternate methods whilst requiring nearly 1 hour less training time and producing on par percentage deviance and maximum error.

The so-called best method would depend on the available resources, and the willingness to search the hyperparameter space before training and additionally the frequency of re-training the model. For example, if re-training the model with a new dataset would were to happen daily, a 0.47% increase is not worth the additional searching and training time.

One of the key benefits to using a kernel over direct output from a random forest is to increase the transfer learning. A random forest kernel is a quick and simple solution to transferring the model into interpretable terms when a full deep neural network or equivalent is unnecessary. The kernel method performs equivalently to that of the random forest which it is based upon, see Figure 5.16, except a magnitude worse. For predicting new values accurately, the random forest is preferable, however for succinctly mapping relationships the kernel performs adequately, if not, still significantly above average. The problem the kernel faces is the sheer size of information a random forest contains, especially with 1200 estimators. A kernel is a generalisation of the full random forest, so this behaviour is expected, and is, in some regards, a desired property by which to simplify very complex non-linear



(a) mse (b) MedAE

Figure 5.16: Kernel Random Forest vs Random Forest for Mean Squared Error and Median Absolute Error

relationships.

Visually inspecting a tree with depth $n = 3$ for production line 1 is presented in Figure 5.17; the decision criteria are presented on the top row of a decision node stating the variable, condition and value 'picked SKU total' <= 11451.5. The second row states the regression criteria, here Mean Squared Error and the value of the split. The third row, samples, is how many observations are contained within the node - each level should sum to the total number.

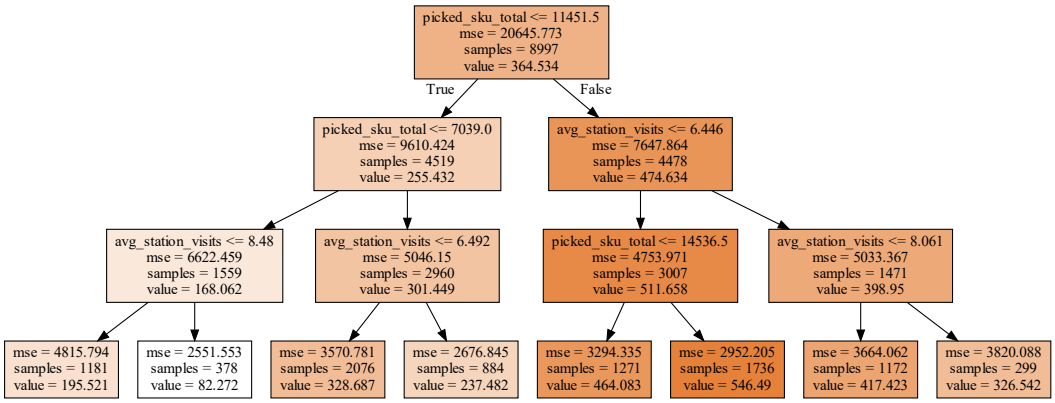


Figure 5.17: Kernel Random Forest nodes pruned to $n = 3$ for Production Line 1

5.3 Time Series Modelling

The first question in this study sought to leverage regression modelling to understand the relationships or indicators of throughput. Most regression modelling ignores that this data is a time series; leveraging this idea leads to applying changepoint analysis. The following experiments were performed in R using the changepoint package [70],

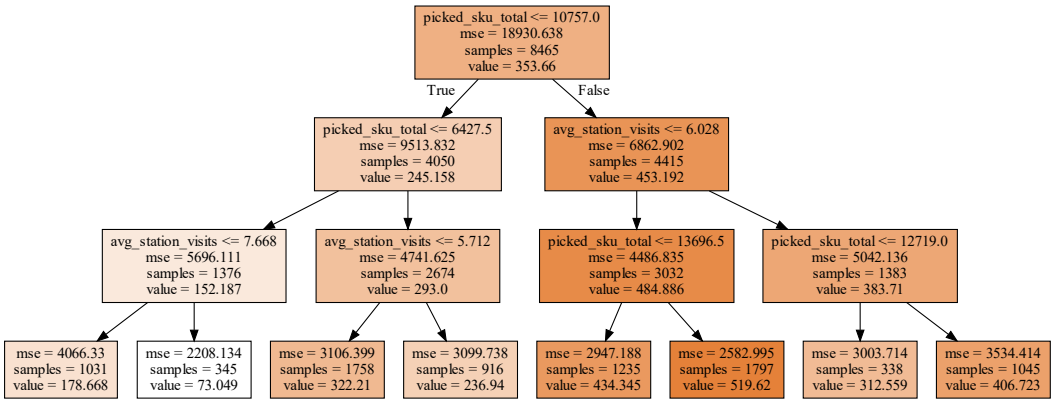


Figure 5.18: Kernel Random Forest nodes pruned to $n = 3$ for Production Line 2

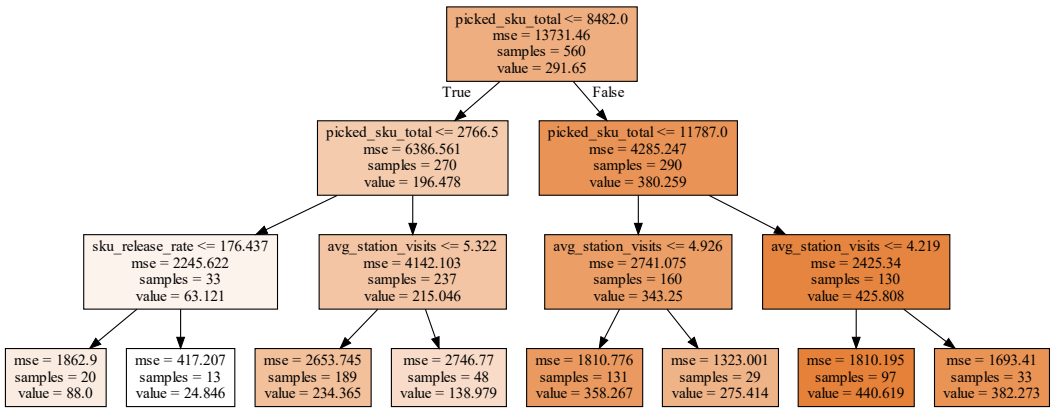


Figure 5.19: Kernel Random Forest nodes pruned to $n = 3$ for Production Line 3

running on a docker instance of RStudio, so to make all results reproducible [3]. The dataset used for the analysis is the MICE imputed dataset to be consistent with the regression models, however there are no imputations of the variables considered within the changepoint analysis.

The experiments are based upon the changepoints of both mean and variance, using the Pruned Exact Linear Time (PELT) method, with Changepoints for a Range of Penalties (CROPS) as the penalty function for the range $[5, 500]$ [46, 71].

Weekly

The results obtained from the analysis of weekly medians are shown in Figure 5.21, the elbow plot shows the sharp decline of penalty value for $n = 3$ reaching a penalty value of 10 fr $n = 8$ - this is shown in the righthand plot where the segment means are drawn in bold red horizontal lines. Observed is a relatively stable early 2020, leading to an increase in the latter stages of the year and becoming stable again

[3] the changepoint analysis performed is deterministic, so reproducibility is automatic, however the docker instances keep experiments separate from any user which is good practice - making the environment in which they are performed fully controllable

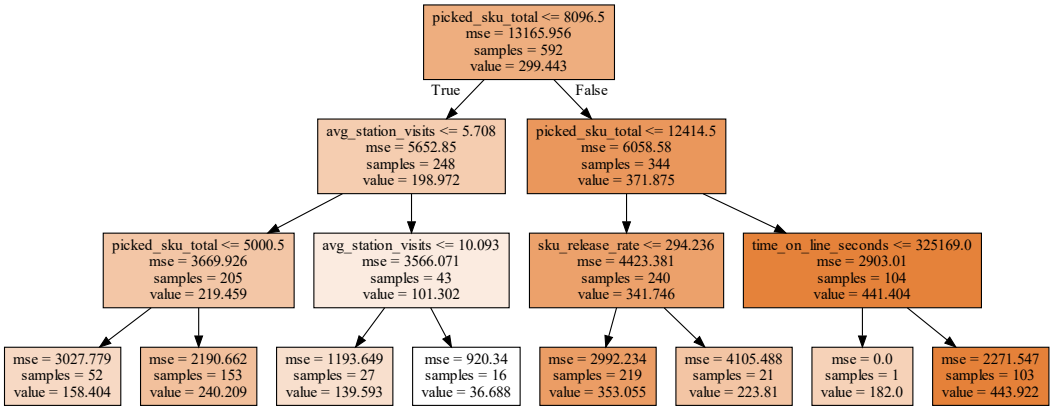
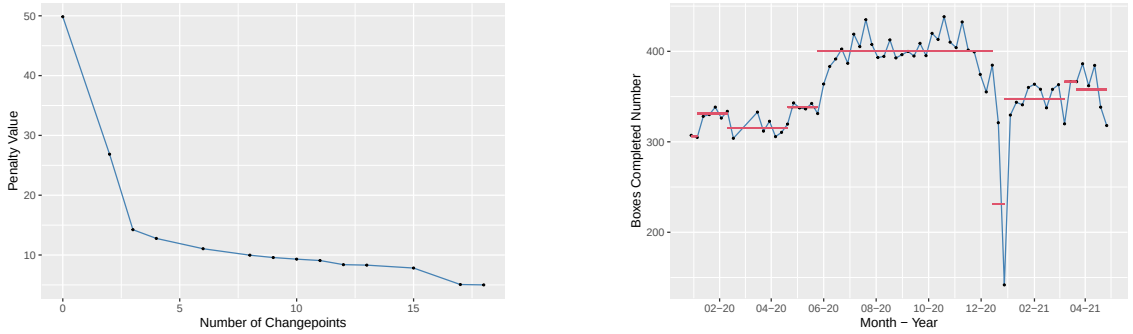


Figure 5.20: Kernel Random Forest nodes pruned to $n = 3$ for Production Line 4



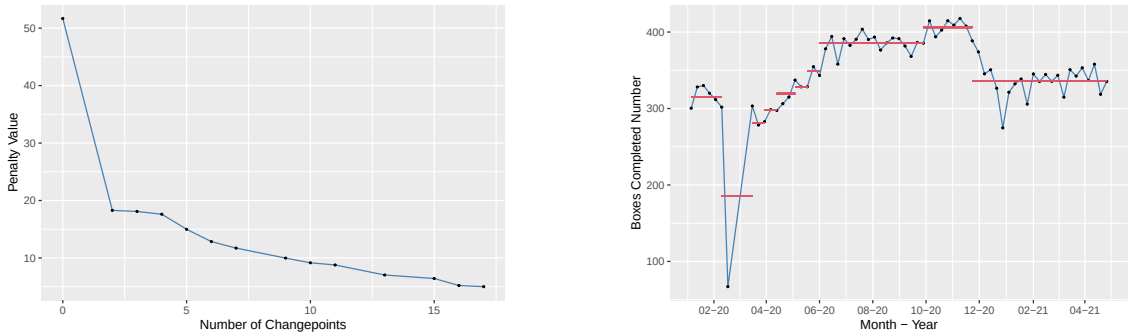
(a) elbow (b) $n = 8$

Figure 5.21: Changepoint Analysis of Weekly Boxes Completed for Production Line 1

at $\bar{x} \approx 400$ and then a slight decline, again, relatively stable with exception to an outlier week - likely due to a production line outage. This step down into 2021 is in line with the opening of the second factory. The same calculations for production line 2 are shown in Figure 5.22, here the elbow plot shows a more considerable single step increase to $n = 2$, but a smaller decline beyond that compared to production line 1; for comparative purposes, $n = 8$ is shown in the righthand plot, and it can be said that this demonstrates roughly similar trends to production line 1. Interestingly, the increasing number of boxes completed is more gradual for line 2 than line 1 and has a more stable peak in December 2020. The results for lines 3 and 4 are shown in Figures 5.23a and 5.23b, respectively; they show relatively little information due to the short running time of the lines - and so will not be considered further in this section as there is insufficient data.

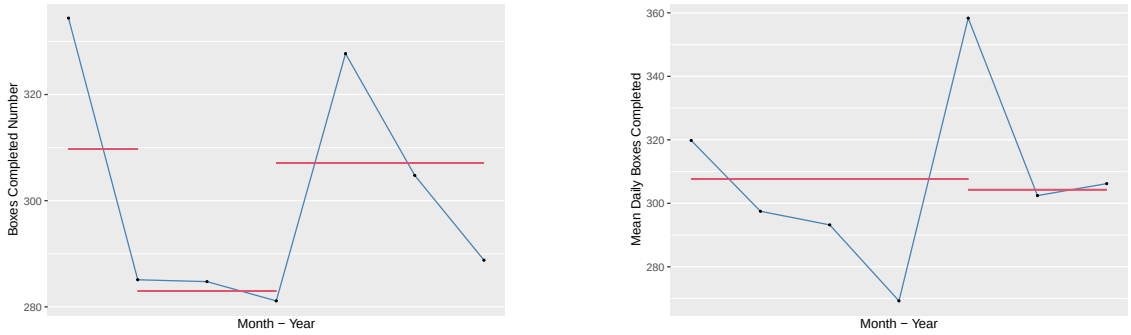
Monthly

The chosen interval can lead to more practical or actionable results, Figures 5.24 presents the average boxes completed for lines 1 and 2. Both the mean and median are presented due to the larger timespan that is being summarised (i.e. 30 days of 24 hours); the median and means can deviate significantly, and as the data is



(a) elbow (b) $n = 8$

Figure 5.22: Changepoint Analysis of Weekly Boxes Completed for Production Line 2

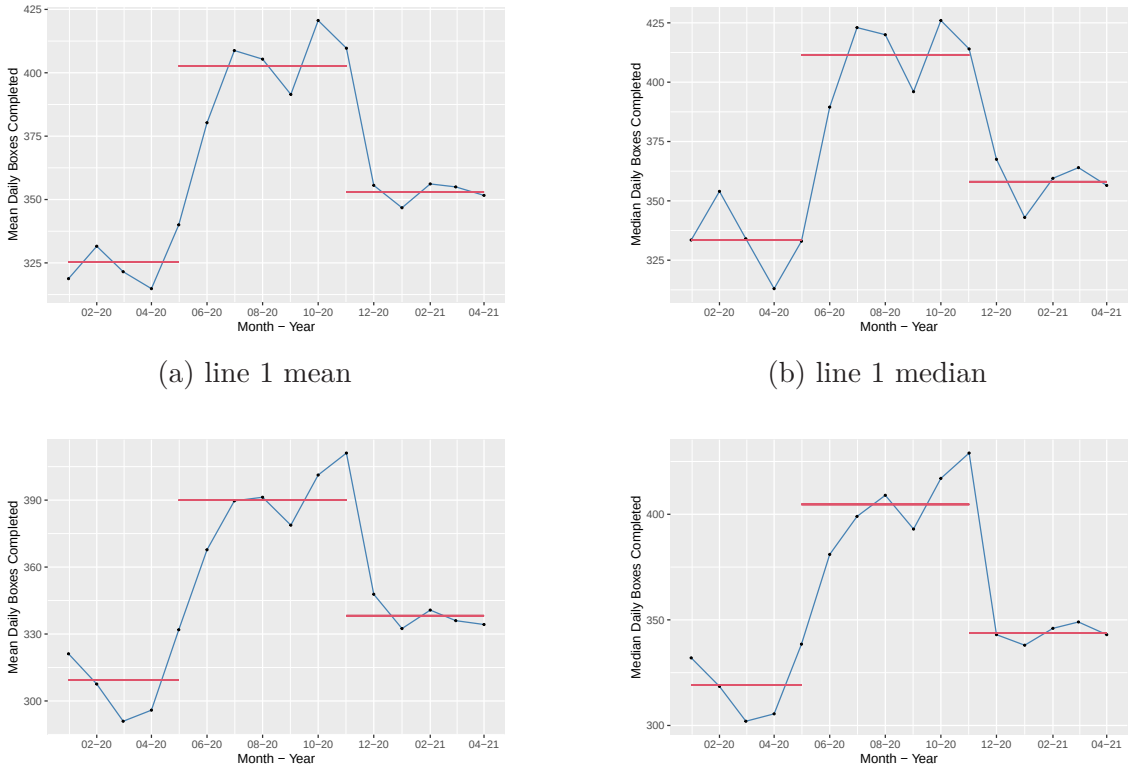


(a) line 3, $n = 2$ (b) line 4, $n = 1$

Figure 5.23: Changepoint Analysis of Weekly Boxes Completed for Production Line 3 and 4

known to have outliers using robust metrics is always desired. The mean for line 1 is shown in Figure 5.24a, the three-tiered patterns mentioned previously is relatively straightforward, starting at $\bar{x} = 325$ for the first four months, increasing to $\bar{x} = 410$ for five months, then decreasing in late 2020 and 2021 has $\bar{x} = 353$. This is similar to the median shown in Figure 5.24b, except the points are clearer; this is surprising, it would appear the median leads to higher magnitudes - this is likely the behaviour of outliers. The median shows that $\bar{x} = 330, 414, 355$, which would suggest that on average ≈ 4 more boxes are produced per hour if using the median than the mean. This would indicate the presence of low magnitude values, which influence the mean. The median approach further emphasises the ramp up and ramp down periods surrounding the peak.

A similar pattern can be observed for line 2; Figure 5.24c shows the three distinct phases with $\bar{x} = 310$ in early 2020, increasing to $\bar{x} = 390$ for mid to late 2020, decreasing to $\bar{x} = 349$ in 2021. Interestingly, this demonstrates that line 2 produces between $\approx 6 - 20$ fewer boxes per hour than line 1. However, it does follow the same trend, as this is driven largely by overall demand, and workload is split relatively evenly. The median for line 2 contains the expected relationship as observed for line 1, with marginally increased x values in the same locations.



(a) line 1 mean

(b) line 1 median

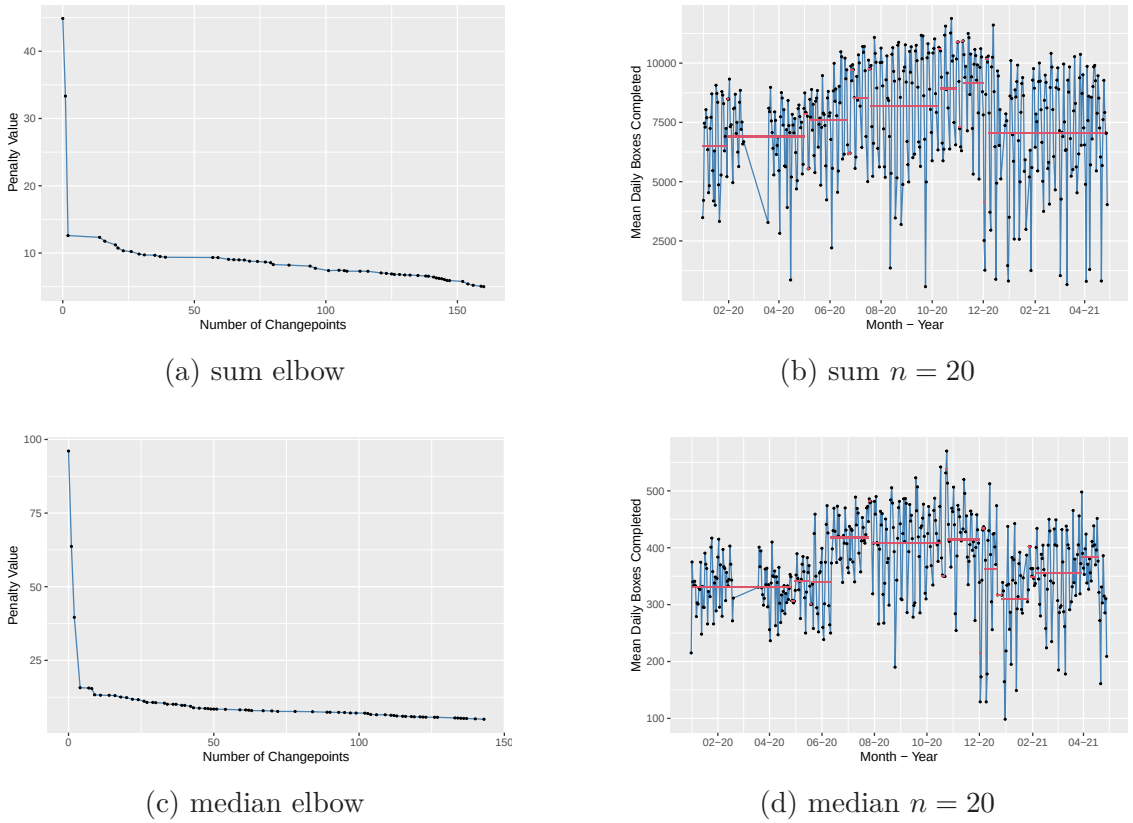
(c) line 2 mean

(d) line 2 median

Figure 5.24: Changepoint Analysis of Monthly Median Boxes Completed for Production lines 1 and 2

Daily

If a smaller time interval is used, the results observed appear to be more sporadic because they are more sensitive to problems with the production lines. Figure 5.25 presents the elbow plot for the daily sum of all boxes produced (5.25a) and the changepoint plot where $n = 20$ (5.25b). The penalty value decreases quickly then flattens after $n = 15$. The changepoint plot uses $n = 20$ as a good balance of interpretability whilst measuring segment changes. The plot demonstrates relatively constant increasing production from early 2020 until early 2021, then a sharp drop to a constant $\sum x \approx 7320$. There are days with significantly lower production; these could be Wednesdays which are known to have low production targets, outliers or production problems - further contextual analysis would be required. The median daily boxes completed is shown in Figures 5.25d, 5.25c which demonstrates a similar trend to that of the sum, except a dip of 21 days is observed at the end of 2020 - this difference between the sum and median could be due to most days observing low production except for specific shifts which were able to produce more. The three phase pattern observed in the monthly and weekly averages is not as evident for the production line 1 daily metrics. Furthermore, the same plots are shown for production line 2 in Figure 5.26, which demonstrates similar summaries to production line 1. A notable difference is that the increase to the peak is more gradual for line 2.



(a) sum elbow

(b) sum $n = 20$

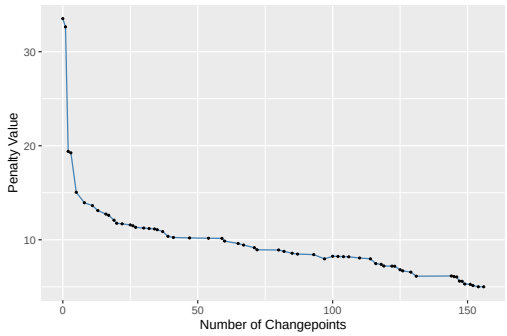
(c) median elbow

(d) median $n = 20$

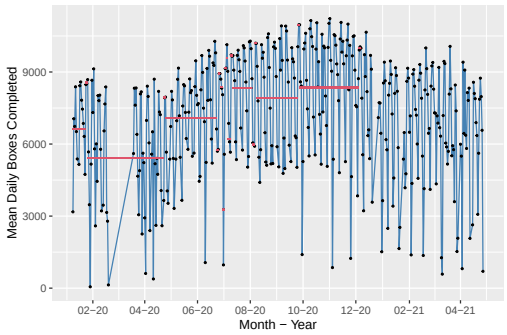
Figure 5.25: Changepoint Analysis of Daily Sum and Median Boxes Completed for Production line 1

Menu Week

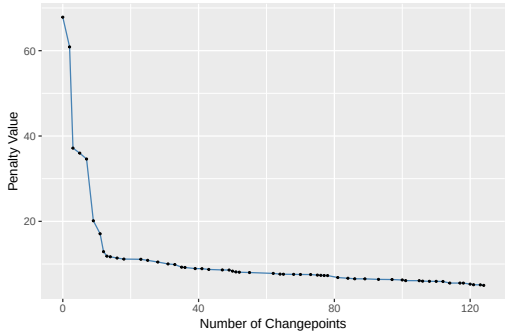
Aggregating by menu weeks is most helpful for the UK Recipe Box Company as it is their primary calendar concept for tracking supply and delivery targets. As a menu week straddles two Gregorian calendar weeks, they are similar to the previous subsection, but translate more directly to a specific menu setup for assessing supply chain performance. Figure 5.27 shows the elbow plot for the penalty value on the left, demonstrating that four changepoints are located for a penalty value of 10, the usual desirable penalty value for other changepoint plots presented. However, a penalty value of ≈ 7 is used to produce 12 points of change as it is found to produce a more accurate reflection of the observed points. This method of visual inspection is by no means a deterministic process. Deviations from the weekly aggregation observed in Figure 5.21 should be minimal, due to both covering the same total number of days - just in different orders. However, the changepoint plot in Figure 5.27b appears to fluctuate more in a polynomial fashion. Furthermore, the first 15 menu weeks appear to have very low variance in mean daily boxes completed and see significantly increased variance around the peak of menu week 420 with $\mu = 410, \sigma = 30$. Comparing this to the same number of changepoints for production line 2, Figure 5.28 - the y scales can be misleading - where the peak is distinctly increased; however, the variance is much increased. Additionally, there is no point at which there is low variance with a stable mean. The elbow plot for this observes a rapid decrease to 13 at $n = 2$, and where $n = 12$, the penalty value is 7.



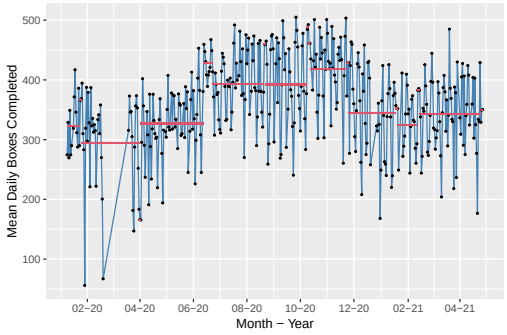
(a) sum elbow



(b) sum $n = 20$

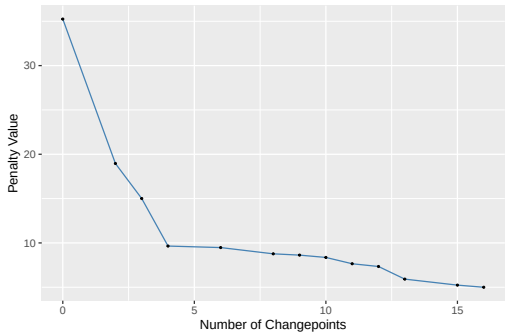


(c) median elbow

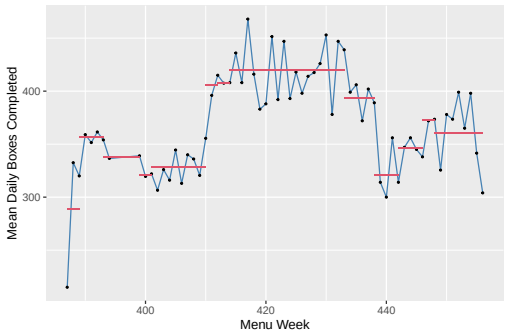


(d) median $n = 18$

Figure 5.26: Changepoint Analysis of Daily Sum and Median Boxes Completed for Production line 2

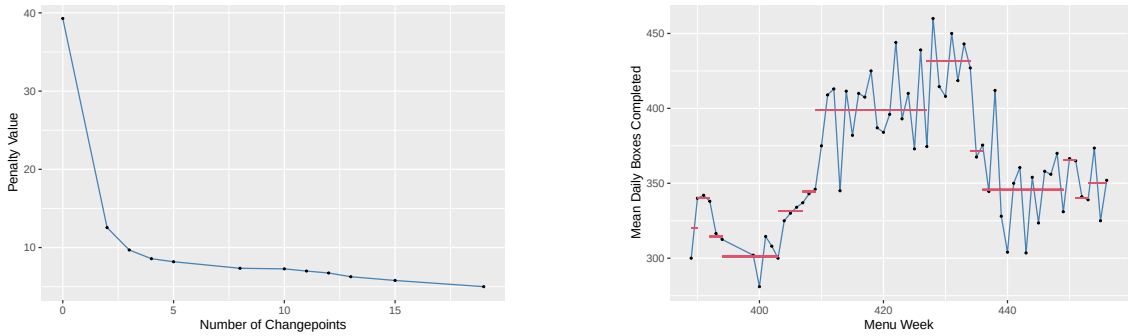


(a) elbow



(b) $n = 12$

Figure 5.27: Changepoint Analysis Mean Boxes Completed vs Menu Weeks for Production Line 1



(a) elbow (b) $n = 12$

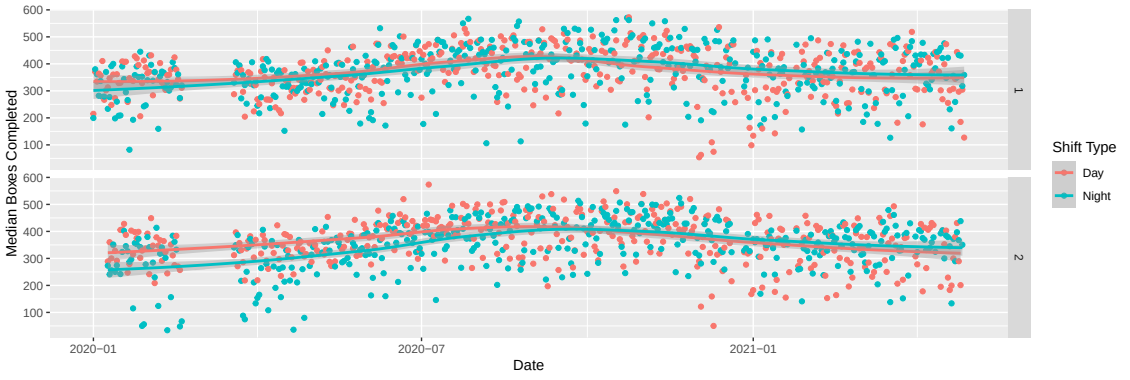
Figure 5.28: Changepoint Analysis Mean Boxes Completed vs Menu Weeks for Production Line 2

Other Categorical Items

Investigating lessor categorical relationships such as shift type or shift colour can allow the observation of the performance of each category relative to the natural fluctuation of demand.

There are two categories within shift type, that is day or night, Figure 5.29 presents the boxes completed and run time for the day, presented in red, and night, presented in blue, across the duration of the case study for both production lines 1 and 2. Slightly more boxes are produced during the day shift for both lines; however, this transitions to the night shift over the course of the observations. This likely reflects a change in algorithms at the leading to a more balanced operating schedule across day and night. This is reflected in the mean run times, which observe similar trends where the day shift has an increased mean run time than the night shift for both lines. This difference decreases with time, leading to a difference of < 1 minutes for line 1 and < 3 minutes for line 2. Furthermore, except for some significant outliers, whilst the mean steadily decreases, the variance of the run time remains relatively constant. Notably, neither production line observes constant outliers, which reinforces the rejection of the assumption that they are not introduced completely at random. Production line 2 observes the most outliers in early 2020, whereas production line 1 observes most around late 2020 and early 2021. These figures have a clear polynomial aspect, peaking near September 2020 and decreasing after that.

When interpreting the median boxes completed vs the fourteen shift sequences, it must be assumed that these are not expected to be equal. For example, sequence 14 is always the resetting shift where ingredients are changed out in preparation for a new menu week. Figure 5.30 demonstrates the median boxes completed per shift hour and the median run time per shift hour. This follows the same polynomial shape trend as the prior. Sequence 6 is the clear maxima across the study with most other shift sequences producing $25 - 100$ fewer boxes per hour, line 14 is significantly below for most of 2020, but in late 2020, sequence 13 decreases to become the lowest producing. This suggests an algorithmic switch which designates the final two shifts resetting periods, compared to just the final shift of a menu week. However, in the most recent weeks in the dataset, sequence 14 has produced an equal number of boxes, on average, with the maxima. The context of this trend is unknown, however as this metric is a median, not a mean (influenced heavily by extremes) or raw data, it would suggest a significant further change in how these later shifts in the week are



(a) boxes completed

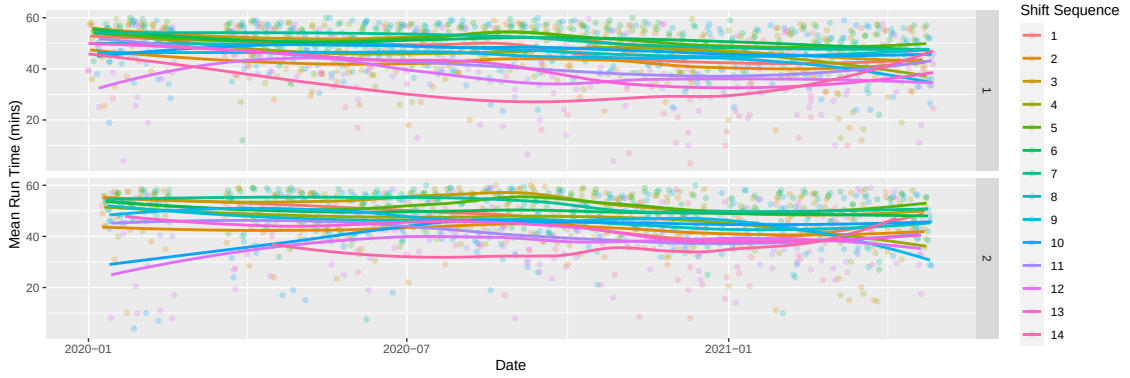
(b) run time

Figure 5.29: Changepoint Analysis Median Boxes Completed and Median Run Time vs Shift Type for Production Lines 1 and 2

utilised. Similar patterns exist for production line 1, sequences 10 and 11 see the lowest median boxes by a large margin in early 2020 but trend upwards to lie within the IQR of all shift sequences, with shift sequence 10 dropping to become the lowest in mid-2021 and shift sequence 11 being third lowest. This trend is odd; the shift in the utilisation rate could not be explained without motivation behind the decisions - it is again likely due to algorithmic changes. For both production lines, the median run time plots show that for most shift sequences (1-9), the run time is relatively constant, except for some highly unstable sequences which see significant changes throughout (and not necessarily lining up with the observed changes in median boxes completed per shift hour).

The shift sequence and shift type can be combined into shift colour (or in practice, shift colour is broken into shift sequence and type. See Figure 5.6). Figure 5.31 presents the median boxes completed and median run time for each shift colour {red, purple, green, yellow}. Both plots demonstrate the tight correlation that the night and day shifts have. The green and yellow day shifts are tightly correlated, as are the red and purple night shifts. Interestingly, the observed peak of the median boxes completed occurs at differing locations for the red and purple to the green and yellow. Additionally, the purple shift deviates from the red in 2021. These plots would highlight any skill differences between shifts; however, there is no discernible difference that would indicate this. The differences are likely due to algorithmic differences that favour specific times of the day or night.

(a) boxes completed



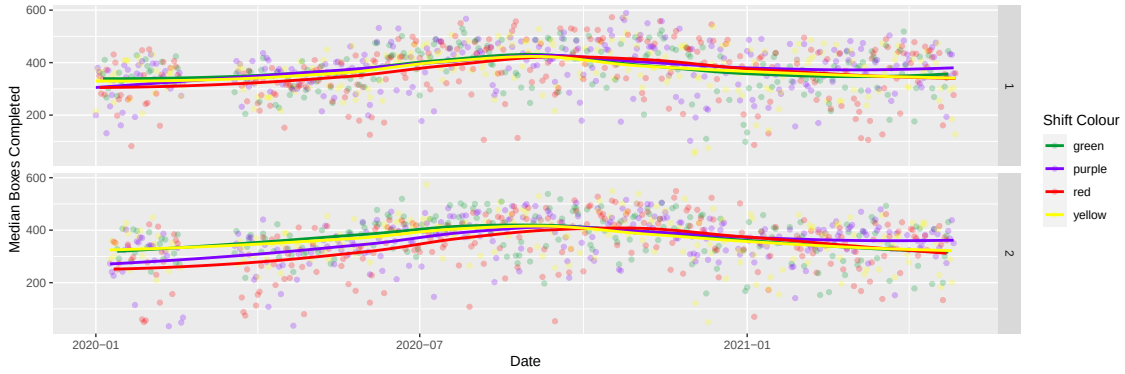
(b) run time

Figure 5.30: Changepoint Analysis Median Boxes Completed and Median Run Time vs Shift Sequence for Production Lines 1 and 2

5.4 Summary

Combining the gained understanding from the regression model and changepoint analysis there are some key points that need to be emphasised.

- Generally, the average station visits being below 4.15 leads to higher throughput. The relationship between average station visits and throughput is similar to a bell curve, the sweet spot occurs between 3.5 and 4.15. Contextually, below this leads to long waiting times at few stations, above this leads to long waiting times across multiple stations.
- A higher total picked number of SKUs always leads to increased throughput, however there is a likely a limit to this. Analysis of pre and post super-SKU would likely show this clearly.
- The regression models are very similar for all production lines, with some nuanced differences. The changepoint models show that this is not the case, and lines 1 and 2 are very similar and 3 and 4 are similar. The two groups are not alike currently, but are trending towards being alike.
- The problem is likely very translatable to new datasets with more rich data, but new hyperparameter searching is probably required for optimality.



(a) boxes completed

(b) run time

Figure 5.31: Changepoint Analysis Median Boxes Completed and Median Run Time vs Shift Colour for Production Lines 1 and 2

- There are clear trends within the shift type, colour and sequence - which do correlate with the fluctuations of throughput.
- Robust statistics are vital for attaining any results with the current dataset, there are many missing and erroneous values which skew any traditional statistics reducing learning efficiency. Exploration of robust approaches consumed significant time for both developing assumptions and processing data.
- Extreme levels of multi-collinearity present in the dataset leads to limited understanding of influence within the problem

CHAPTER 6

Discussion

The specific discussion and recommendations are the conclusions of this study regarding the main research questions defined in Chapter 1.

6.1 Conclusions

The overall goal is to improve the understanding of, and therefore, to improve the ability to predict throughput. The case study approach reflects the method of data discovery and the handling of dataset imperfections.

Multiple processing approaches were considered, presented and their appropriateness discussed. The regression models built up the required understanding and the motivation for each step. Linear Regression is the core foundation; Support Vector Regression introduced the principles of kernels; Decision Trees introduce explainability; Random Forest improves results and the final model being Kernel Random Forests. Kernel Random Forests are the culmination of all previous models resulting in a highly accurate, efficient and explainable model. The changepoint analysis gives an unparalleled understanding of the change of mean and variance over time for various variables, providing contextual information that can help understand the regression models and providing great insight the UK Recipe Box Company can use for further explanation. There was some change from the original goals stated within the specification, which can largely be attributed to gaining a greater understanding of the dataset; of which a significant amount of time was spent experimenting on the state of the dataset and working on the assumptions required for modelling.

Attempting to quantify the impact of a constantly changing menu requires additional contextual information regarding the specific menu components. To give direct conclusions also requires understanding of how all specific stations are set up. However, in all experiments, an increased number of picked SKUs per hour leads to higher throughput values - there will be a point where this is no longer the case (as the UK Recipe Box already understand with the introduction of the 'Super SKU').

Addressing the differences between lines and factories, the problem is directly translatable between the same factory lines - i.e. production lines 1 and 2 are similar and production lines 3 and 4 are similar. Whilst there exists differences in the setup of the factories (such as differing queue sizes at the stations), and how data is recorded is different, there still exists significant overlap. Nevertheless, not sufficient to make them statistically equivalent.

Firstly, as discussed in the case study, regression models are appropriate to model the relationships surrounding throughput. Linear models were proven to be

inadequate at presenting the model, so non-linear reformulations were presented which provided accurate models that are capable of predicting throughput.

Secondly, the most accurate model presented was Random Forest, with hyperparameter optimisation via grid search. This provided the highest accuracy, both for mean squared error and median absolute error. However, this model came at the cost of significant computation time for hyperparameter searching and model training. A normal random forest, with only 100 estimators was still adequate with only a 0.47% decrease in accuracy. Whilst the Kernel Random Forest is not the most accurate model in the study, it provides value in its easily interpretable results; which was a key deliverable for the UK Recipe Box Company .

Finally, changepoint analysis provides invaluable contextual information which allows greater understanding of the regression results. The three phase pattern observed in all aggregations showed clear changes in mean and variance following a distinct pattern, which would merit further investigation. However, the combination of the changepoint analysis with regression models would require deep industry knowledge.

6.2 Limitations

It should be acknowledged that there may be an over-generalisation of the dataset to reduce correlation partly due to dataset provided. This could be mitigated in future work by collecting more indicator variables that are not closely related.

The dataset is summary data of an hour of run time; it contains multiple values which are 'averages', it is unknown how this average is calculated, and as there are values that are considered outliers we should question whether they are produced via one extra-large outlier or an entire hour of outliers.

The reliability and validity of the model are of utmost importance to make the results worthwhile. As the data is generated from systems used by the UK Recipe Box Company, the reliability of these terms is unknown and the error of these values is unknown.

In addition, the logging of the measurement was performed automatically and remotely, leading to some data loss in a four week period in 2020. The data loss was due to no one at the UK Recipe Box Company noticing it was missing - again questioning the validity of the data.

6.3 Future Work

The current research provides a variety of promising further research directions. One such direction could be to expand the robust elements to investigate how the outliers are treated, including further integrating robust metrics to train decision trees, such as using the Huber loss function. Many other modelling considerations could have been made, such as using stepwise regression, Generalised Linear Models or traditional forecasting, which could yield different results.

An interesting approach would be to combine the results of this problem with a more detailed description of a menu week, whereby instead of the summary statistics of a menu week, the specific recipe and ingredient items are included to investigate if the genuine relationships are simply a number of ingredients or specific

ingredients/recipes that perform poorly or strongly. This could also take a deeper dive looking at specific stations on production lines. Overall using fewer summary statistics in favour of raw data would provide more exciting results and more room for exploration.

An extension of the previous point would be exploring the penalty values further for changepoint analysis; these were mainly by visual inspection or 'somewhere near 10', a more rigorous approach to this would produce more statistically reliable results and individually would provide insight into how the model is coping with overfitting. If raw data was provided, this changepoint analysis would become far more valuable, as it would include a raw approach to performing aggregations.

Investigating alternative hyperparameter searching methods that could provide faster optimal parameters, genetic algorithms or metaheuristics (potentially leveraging some contextual industry knowledge). Originally Particle Swarm Optimisation of hyperparameter optimisation was a goal of the research, but with the increasing size of dataset experimentation it was removed from the scope, initial results proved this approach could give strong results but the experiments showed this to be inconsistent.

A final suggestion for future work is assessing the models capability to predict future values, this was originally within the scope of the project and some experiments were performed in order to assess this, a major roadblock was the relatively limited dataset and unstable parameters due to the demand profile of the company. This led to really poor experiment modelling, $< 50\%$ accuracy, and so would require much more focus if this was an objective (i.e. forecasting). Using entirely unseen periods of data, ideally entire menu weeks where the setup is entirely unknown, would be the best way to measure the forecasting performance but the trend of the data is currently too unstable to provide good results without far more attention.

Bibliography

[1] Rencher, A. C. and Pun, F. C. 'Inflation of R2 in Best Subset Regression'. In: *Technometrics* vol. 22, no. 1 (1980), pp. 49–53.

[2] Chatfield, C. 'Model Uncertainty, Data Mining and Statistical Inference'. In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* vol. 158, no. 3 (1995), p. 419.

[3] Yin, R. *Case study research : design and methods*. Los Angeles, Calif: Sage Publications, 2009.

[4] Ross, D. *Competing through supply chain management : creating market-winning strategies through supply chain partnerships*. New York: Chapman & Hall, 1998.

[5] Simchi-Levi, D. 'Review of:"Competing Through Supply Chain Management: Creating Market-Winning Strategies Through Supply Chain Partnerships"David Frederick Ross Chapman & Hall'. In: *IIE Transactions* vol. 30, no. 8 (Aug. 1998), pp. 762–763.

[6] Levi, D. *Designing and managing the supply chain : concepts, strategies, and case studies*. Boston: Irwin/McGraw-Hill, 2000.

[7] Cooper, M. C. and Ellram, L. M. 'Characteristics of Supply Chain Management and the Implications for Purchasing and Logistics Strategy'. In: *The International Journal of Logistics Management* vol. 4, no. 2 (July 1993), pp. 13–24.

[8] La Londe, B. J. and Masters, J. M. 'Emerging Logistics Strategies'. In: *International Journal of Physical Distribution & Logistics Management* vol. 24, no. 7 (Jan. 1994), pp. 35–47.

[9] Lambert, D. *Fundamentals of logistics management*. Boston: Irwin/McGraw-Hill, 1998.

[10] Lambert, D. and Cooper, M. 'Issues in Supply Chain Management'. In: *Industrial Marketing Management* vol. 29 (Jan. 2000), pp. 65–83.

[11] Zhong, R. Y. et al. 'A big data approach for logistics trajectory discovery from RFID-enabled production data'. In: *International Journal of Production Economics* vol. 165 (2015), pp. 260–272.

[12] Tan, K. H. et al. 'Harvesting big data to enhance supply chain innovation capabilities: An analytic infrastructure based on deduction graph'. In: *International Journal of Production Economics* vol. 165 (2015), pp. 223–233.

[13] Shukla, N. and Kiridena, S. 'A fuzzy rough sets-based multi-agent analytics framework for dynamic supply chain configuration'. In: *International Journal of Production Research* vol. 54, no. 23 (2016), pp. 6984–6996.

[14] Dutta, D. and Bose, I. 'Managing a big data project: the case of ramco cements limited'. In: *International Journal of Production Economics* vol. 165 (2015), pp. 293–306.

[15] Singh, A. et al. 'Cloud computing technology: Reducing carbon footprint in beef supply chain'. In: *International Journal of Production Economics* vol. 164 (2015), pp. 462–471.

[16] Waller, M. A. and Fawcett, S. E. *Click here for a data scientist: Big data, predictive analytics, and theory development in the era of a maker movement supply chain*. 2013.

[17] Waller, M. A. and Fawcett, S. E. *Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management*. 2013.

[18] Govindan, K. et al. 'Big data analytics and application for logistics and supply chain management'. In: *Transportation Research Part E: Logistics and Transportation Review* vol. 114 (2018), pp. 343–349.

[19] Li, J. et al. 'Throughput analysis of production systems: recent advances and future topics'. In: *International Journal of Production Research* vol. 47, no. 14 (May 2009), pp. 3823–3851.

[20] Koenigsberg, E. 'Production Lines and Internal Storage–A Review'. In: *Management Science* vol. 5, no. 4 (July 1959), pp. 410–433.

[21] Buxey, G. M., Slack, N. D. and Wild, R. 'Production Flow Line System Design–A Review'. In: *A I I E Transactions* vol. 5, no. 1 (1973), pp. 37–48.

[22] Buzacott, J. A. and Hanifin, L. E. 'Models of automatic transfer lines with inventory banks a review and comparison'. In: *A I I E Transactions* vol. 10, no. 2 (1978), pp. 197–207.

[23] Dallery, Y. and Gershwin, S. B. 'Manufacturing flow line systems: a review of models and analytical results'. In: *Queueing Systems* vol. 12, no. 1-2 (Mar. 1992), pp. 3–94.

[24] Govil, M. K. and Fu, M. 'Queueing theory in manufacturing: A survey'. In: *Journal of Manufacturing Systems* vol. 18 (1999), pp. 214–240.

[25] Papadopoulos, H. and Heavey, C. 'Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines'. In: *European Journal of Operational Research* vol. 92, no. 1 (1996), pp. 1–27.

[26] Li, J., Blumenfeld, D. and Alden, J. 'Comparisons of two-machine line models in throughput analysis'. In: *International Journal of Production Research* vol. 44 (May 2006), pp. 1375–1398.

[27] Freedman, D. *Statistical models : theory and practice*. Cambridge New York: Cambridge University Press, 2009.

[28] Efron, B. et al. 'Least angle regression'. In: *The Annals of Statistics* vol. 32, no. 2 (Apr. 2004).

[29] Theil, H. 'A Rank-Invariant Method of Linear and Polynomial Regression Analysis'. In: *Advanced Studies in Theoretical and Applied Econometrics*. Springer Netherlands, 1992, pp. 345–381.

[30] Breiman, L. *Classification and regression trees*. New York: Chapman & Hall, 1993.

[31] Quinlan, J. R. 'Induction of decision trees'. In: *Machine Learning* vol. 1, no. 1 (Mar. 1986), pp. 81–106.

[32] Breiman, L. 'Bagging predictors'. In: *Machine Learning* vol. 24, no. 2 (Aug. 1996), pp. 123–140.

[33] Friedman, J. H. 'Stochastic gradient boosting'. In: *Computational Statistics & Data Analysis* vol. 38, no. 4 (Feb. 2002), pp. 367–378.

[34] Breiman, L. In: *Machine Learning* vol. 45, no. 1 (2001), pp. 5–32.

[35] Barros, R. C. et al. 'A Survey of Evolutionary Algorithms for Decision-Tree Induction'. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* vol. 42, no. 3 (May 2012), pp. 291–312.

[36] Papageorgiou, G., Bouboulis, P. and Theodoridis, S. 'Robust Nonlinear Regression: A Greedy Approach Employing Kernels With Application to Image Denoising'. In: *IEEE Transactions on Signal Processing* vol. 65, no. 16 (Aug. 2017), pp. 4309–4323.

[37] Huber, P. J. 'The 1972 Wald Lecture Robust Statistics: A Review'. In: *The Annals of Mathematical Statistics* vol. 43, no. 4 (Aug. 1972), pp. 1041–1067.

[38] Rousseeuw, P. *Robust regression and outlier detection*. Hoboken, NJ: Wiley-Interscience, 2003.

[39] Maronna, R. *Robust statistics : theory and methods*. Chichester, England: J. Wiley, 2006.

[40] Huber, P. *Robust statistics*. Hoboken, N.J: Wiley, 2009.

[41] Rousseeuw, P. J. 'Least Median of Squares Regression'. In: *Journal of the American Statistical Association* vol. 79, no. 388 (1984), pp. 871–880.

[42] Huber, P. J. 'John W. Tukey's Contributions to Robust Statistics'. In: *The Annals of Statistics* vol. 30, no. 6 (2002), pp. 1640–1648.

[43] Tukey, J. W. 'A survey of sampling from contaminated distributions'. In: *Contributions to Probability and Statistics* (1960).

[44] Scott, A. J. and Knott, M. 'A Cluster Analysis Method for Grouping Means in the Analysis of Variance'. In: *Biometrics* vol. 30, no. 3 (Sept. 1974), p. 507.

[45] Auger, I. E. and Lawrence, C. E. 'Algorithms for the optimal identification of segment neighborhoods'. In: *Bulletin of Mathematical Biology* vol. 51, no. 1 (Jan. 1989), pp. 39–54.

[46] Killick, R., Fearnhead, P. and Eckley, I. A. 'Optimal Detection of Changepoints With a Linear Computational Cost'. In: *Journal of the American Statistical Association* vol. 107, no. 500 (Oct. 2012), pp. 1590–1598.

[47] Kohavi, R. and Wolpert, D. H. 'Bias Plus Variance Decomposition for Zero-One Loss Functions'. In: *MACHINE LEARNING: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL*. Morgan Kaufmann Publishers, 1996, pp. 275–283.

[48] Luxburg, U. von and Schoelkopf, B. 'Statistical Learning Theory: Models, Concepts, and Results'. In: (2008). arXiv: 0810.4752 [stat.ML].

[49] Pawitan, Y. *In all likelihood : statistical modelling and inference using likelihood*. eng. Oxford ; New York: Clarendon Press : Oxford University Press, 2001.

[50] Zimek, A. and Schubert, E. 'Outlier Detection'. In: *Encyclopedia of Database Systems*. Springer New York, 2017, pp. 1–5.

[51] Buuren, S. van. *Flexible imputation of missing data*. Boca Raton, FL: CRC Press, 2018.

[52] Buuren, S. V. et al. 'Fully conditional specification in multivariate imputation'. In: *Journal of Statistical Computation and Simulation* vol. 76, no. 12 (Dec. 2006), pp. 1049–1064.

[53] Buuren, S. van. 'Multiple imputation of discrete and continuous data by fully conditional specification'. In: *Statistical Methods in Medical Research* vol. 16, no. 3 (June 2007), pp. 219–242.

[54] Rubin, D. B. and Schenker, N. 'Efficiently Simulating the Coverage Properties of Interval Estimates'. In: *Applied Statistics* vol. 35, no. 2 (1986), p. 159.

[55] Hopke, P. K., Liu, C. and Rubin, D. B. 'Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic'. In: *Biometrics* vol. 57, no. 1 (Aug. 2001), pp. 22–33.

[56] Anscombe, F. J. 'Graphs in Statistical Analysis'. In: *The American Statistician* vol. 27, no. 1 (Feb. 1973), pp. 17–21.

[57] Vapnik, V. N. *The Nature of Statistical Learning Theory*. Berlin, Heidelberg: Springer-Verlag, 1995.

[58] Koberstein, A. 'The dual simplex method, techniques for a fast and stable implementation'. In: *Unpublished doctoral thesis, Universität Paderborn, Paderborn, Germany* (2005).

[59] Platt, J. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Tech. rep. MSR-TR-98-14. Microsoft, Apr. 1998.

[60] Chen, P.-H., Fan, R.-E. and Lin, C.-J. 'A Study on SMO-Type Decomposition Methods for Support Vector Machines'. In: *IEEE Transactions on Neural Networks* vol. 17, no. 4 (July 2006), pp. 893–908.

[61] Aslam, J. A., Popa, R. A. and Rivest, R. L. 'On Estimating the Size and Confidence of a Statistical Audit'. In: *Proceedings of the USENIX Workshop on Accurate Electronic Voting Technology*. EVT'07. Boston, MA: USENIX Association, 2007, p. 8.

[62] Scornet, E. 'Random Forests and Kernel Methods'. In: *IEEE Transactions on Information Theory* vol. 62, no. 3 (2016), pp. 1485–1500.

[63] Davies, A. and Ghahramani, Z. *The Random Forest Kernel and other kernels for big data from random partitions*. 2014. arXiv: 1402.4293 [stat.ML].

[64] Cawley, G. C. and Talbot, N. L. 'On Over-Fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation'. In: *J. Mach. Learn. Res.* vol. 11 (Aug. 2010), pp. 2079–2107.

[65] Raju, N. S. et al. 'Methodology Review: Estimation of Population Validity and Cross-Validity, and the Use of Equal Weights in Prediction'. In: *Applied Psychological Measurement* vol. 21, no. 4 (Dec. 1997), pp. 291–305.

[66] Berger, J. *Statistical decision theory and Bayesian analysis*. New York: Springer-Verlag, 1985.

[67] Birgé, L. and Massart, P. 'Minimal Penalties for Gaussian Model Selection'. In: *Probability Theory and Related Fields* vol. 138, no. 1-2 (July 2006), pp. 33–73.

[68] Breusch, T. S. and Pagan, A. R. 'A Simple Test for Heteroscedasticity and Random Coefficient Variation'. In: *Econometrica* vol. 47, no. 5 (Sept. 1979), p. 1287.

[69] Koenker, R. 'A note on studentizing a test for heteroscedasticity'. In: *Journal of Econometrics* vol. 17, no. 1 (Sept. 1981), pp. 107–112.

[70] Killick, R. and Eckley, I. A. 'changepoint: AnRPackage for Changepoint Analysis'. In: *Journal of Statistical Software* vol. 58, no. 3 (2014).

[71] Haynes, K., Eckley, I. A. and Fearnhead, P. *Efficient penalty search for multiple changepoint problems*. 2014. arXiv: 1412.3617 [stat.CO].

Appendices

Specification

Brief

Description
We need to understand exactly what factors in a factory influence throughput, to help us learn which factors we might change to improve factory metrics. To achieve this, a variety of data sources will be available, including detailed picking data from factories to data describing the configuration of the factory. A successful project will combine these data sources to produce an accurate algorithm that successfully predicts what throughput can be achieved for a given setup, and which factors should be changed to improve throughput.
Background
A UK recipe box company is on a mission to become the UK's favourite way to eat dinner! With over 50 weekly recipes on the menu and plans to expand this further, it's a constant challenge to provide sufficient capacity in our factories to allow us to pick so many different recipes each week for so many customers. Sufficient capacity comes from setting our factory up optimally, to ensure that we can process as many recipe boxes as possible every hour. So many things inside a factory influence this throughput, from the layout itself to the efficacy of pickers putting the boxes together.
Deliverables
There are a series of deliverables, some specific to the company, some specific to the Lancaster University MSc process.

1. Contextual Interpretation of throughput modelling
2. Algorithms/Code for All modelling approaches
3. Presentation to interested parties
4. MSc Thesis

Overview

A loose direction of the thesis categorised by months:

June Processing and Basic Linear Regression Models

July Advanced Regression Modelling Techniques

August Experimental Time Series Analysis

The company are mainly interested in quantifying some relationship between throughput and a series of factors using some form of simple regression, however some more interesting models will also be investigated to attempt to gain increased performance - either accuracy or ease of interpretation.
The company may be interested in some of the advanced models but they are not currently the primary outcome.

Weekly Plan

A more detailed version of the loose outline above is shown for each w/c date is shown in Table 1.

Week	Date	Details
1	05/30	Kick-Off Form
2	06/07	Feature Importance & Specification Document
3	06/14	Pre-Processing Experiments
4	06/21	Feature Reduction Experiments
5	06/28	Basic Linear Regression
6	07/05	Non-Linear Regression
7	07/12	Tree's & Kernels
8	07/19	Validation and Comparison Evaluation
9	07/26	Document Check Week & Code Refactoring
10	08/02	Time Series Experiments
11	08/09	Deep Learning Time Series Models
12	08/16	Document Check Week 2 & Code Packaging
13	08/23	Final Poster & Presentation changes
14	08/30	Final Document changes and submission

Table 1: Detailed Weekly Plan

Data Collection Considerations

The data collection is via an automated within-company process which collates performance information regarding the supply line every hour. Most of these are either informative descriptions of shift or summary statistics of performance - this means that any correlation of variables with throughput are only indicators of a problem and the true influence of throughput may be obfuscated. However the paper intends to outline some approaches to remedy throughput issues identified.

The data has been semi-cleaned via industrial supervisor as the cleaning required was extremely contextual to factory-wide related issues such as (1) line-maintenance (2) large-scale shutdown or (3) shift-problems. However some preliminary investigation appears to have introduced some NaN values which require investigation and dealing with on a feature by feature basis: this could mean setting

to 0, setting to mean or row deletion. Some examples of this data are shown in Table 2

Data	Type	Description
shift_sequence	Numeric	A set of workers assigned to the shift. (In effect, categorical)
production_line	Categorical	Production Line data is for
box_release_num	Numeric	Number of boxes released onto the supply line
shortpicked_sku_total	Numeric	Number of pickups missed
ingredient_skus	Numeric	Number of individual "ingredients" [1]
percentage_idle_time	Numeric	Percentage of time there was no box in a station

Table 2: Example of Data Points

Key Research Questions

I consider there to be 3 key research questions

Do regression models show consistent predictor variables for throughput?

This is the key deliverable for the company as it the area of greatest interest, to date, the company have only investigated some simple linear regression models and they believe this area can provide them some real insight into which factors can optimise their throughput [19]. Simple linear regression models such as Ordinary Least Squares or Weighted Least Squares provide a foundation upon which to investigate collinearity or non-linearity. This can be addressed using more advanced regression techniques which is addressed in the second point (Section 6.3). This focuses on time agnostic throughput, and purely looks at the factors that lead to increased throughput.

What is the most accurate model/algorithm/pipeline to predict throughput based on factory setup?

A successful thesis will include a thorough evaluation and comparison of various pipelines - where a pipeline encompasses the processing of data through to the automatic model performance evaluation. The methods of comparison are yet to be determined, the company has stated they are interested in R^2 however this has several problems, which will be discussed in the thesis. The remedy of this is to use many metrics and to utilise some form of k-fold cross over validation performed on split sets.

[1] A SKU represents an ingredient by code, however there are also "supersku's" which are a collection of ingredients merged into a single SKU when there are a set of ingredients always used together.

Can time series analysis predict hourly movement?

This is an experimental investigation into detecting time series trends in the data - it is known that the production lines are busier in specific times of the week in line with their delivery schedule. This section will validate this belief but look to further extend relationships between busier periods and quieter periods and also attempt to discover other previously unknown relationships that could provide the company with further insight in how to optimise their weekly throughput.