Nottingham University Business School

UNITED KINGDOM · CHINA · MALAYSIA

# Detecting Concept Drift in Machine Learning Models: A Model-Agnostic Approach Using Permutation Importance, SHAP Values, and MCR Bounds

**by**

**Shreyas Kupsad**

**2023-24**

A Dissertation presented in part consideration for the degree of "MSc in Business Analytics"

# Abstract

Machine learning models are at risk of degradation over time due to feature drift and concept drift, which can lead to inaccurate predictions, poor business decisions, and increased maintenance costs. Feature drift refers to changes in the statistical properties of input features, while concept drift involves shifts in the relationship between these features and the target variable. This study focuses on detecting concept drift, a critical issue in dynamic environments where data patterns evolve continuously. The study develops a model-agnostic approach using techniques such as SHAP values, permutation importance, and Model Class Reliance (MCR) bounds to identify and interpret drift across different machine learning models.

The research validated that the changes in SHAP values and feature importance scores can be reliable indicators of concept drift. The methodology proved effective across diverse models, confirming its model-agnostic nature. Furthermore, the study provided insights into the changing relationships between features and predictions, offering practitioners a deeper understanding of how drift impacts model performance.

However, several limitations were identified, including challenges in detecting statistically significant changes in permutation importance, inaccuracies in temporal segmentation, and difficulty in capturing gradual drift. These findings suggest that the methodology, while promising, requires further refinement, particularly in terms of statistical validation and detection of slow-evolving drift. Future work should focus on enhancing these aspects.

Overall, this dissertation contributes to the detection of concept drift, providing a framework that enables proactive intervention to maintain model reliability and accuracy in dynamic environments.

# Acknowledgements

I would like to extend my heartfelt appreciation to those who have been instrumental in the completion of this research.

First and foremost, my deepest gratitude goes to my supervisor, Dr. Gavin Smith, whose unwavering support, and guidance have been invaluable throughout this research journey. His insightful input, patience, and constructive feedback were pivotal in shaping the direction and quality of this study.

I would also like to express my sincere thanks to my family and friends for their continuous encouragement and understanding. Their unwavering support, love, and motivation provided the strength I needed to persevere and successfully complete this dissertation.

Finally, I would like to acknowledge the faculty and staff at The University of Nottingham for providing the resources and academic environment that made this research possible.

**Table of Contents**

1. Introduction ..... 1

2. Literature Review ..... 4

2.1 Reasons for the occurrence of drift in models ..... 4

2.2 Statistical methods for data drift detection ..... 4

2.3 Model based approaches for concept drift detection ..... 6

2.4 Explainable AI (XAI) for drift analysis ..... 8

2.5 Combination of Permutation Importance and XAI techniques ..... 9

3. Methodology ..... 11

3.1 Model Selection and Justification ..... 12

3.2 Introduction to Datasets ..... 13

3.2.1 Electricity Dataset Segmentation ..... 13

3.2.2 Airlines Dataset Segmentation ..... 14

3.3 Measurement of Feature Importance Scores ..... 14

3.4 SHAP Value Analysis ..... 15

3.5 Statistical Analysis of Feature Importance Shifts ..... 15

3.6 Model Class Reliance (MCR) Calculation ..... 16

3.7 Model-Agnostic Comparison ..... 16

4. Implementation and Evaluation ..... 17

4.1 Implementation of Electricity Dataset ..... 17

4.1.1 Data description ..... 17

4.1.2 Cleaning and pre-processing ..... 19

4.1.3 Implementing Logistic Regression ..... 20

4.1.4 Implementing Random Forest ..... 23

4.1.5 Implementing Neural Networks ..... 25

4.2 Implementation of Airline Dataset ..... 27

4.2.1 Data Description ..... 27

4.2.2 Cleaning and Pre-processing ..... 28
4.2.3 Implementing Logistic Regression ..... 29
4.2.4 Implementing Random Forest ..... 33
4.2.5 Implementing Neural Network ..... 36
5. Results and Discussion ..... 40
5.1 Discussion on Detecting Abrupt Concept Drift ..... 40
5.2 Discussion on Detecting Cyclic Concept Drift ..... 42
6. Conclusion ..... 44
6.1 Achievement of Objectives ..... 44
6.2 Limitations of the Approach ..... 45
6.3 Practical Implications ..... 46
6.4 Summary and Future Work ..... 46
7. References ..... 48
8. Appendix ..... v

# 1. Introduction

Feature drift and concept drift are phenomena that cause machine learning models to degrade over time. Understanding these drifts, their consequences, and the importance of early detection is important for maintaining the performance and reliability of models.

Feature drift, also known as data drift or covariate shift, occurs when the statistical properties of the input features change over time. This can happen due to various reasons such as seasonal changes, changes in user behavior, or external events like a pandemic. For example, a model predicting customer purchases might experience feature drift if the age and income distribution of customers changes significantly over time (Hinder, Jakob and Hammer, 2020).

Concept drift or drift in general refers to changes in the relationship between input features and the target variable. This means that the underlying concept the model is trying to predict evolves over time (Brzeziński & Stefanowski, 2014). When only the distribution of input features significantly changes, it is often referred to as feature drift. However, in this dissertation, the focus is on Concept drift.

Concept drift can be further categorized into:

- Sudden Drift: Abrupt changes in the data distribution e.g., A sudden economic downturn can drastically change consumer spending habits, impacting models predicting sales or market trends.
- Gradual Drift: Slow changes over time e.g., Gradual shifts in weather patterns can affect models predicting crop yields or energy consumption.
- Incremental Drift: Continuous but small changes e.g., Customer preferences can evolve gradually over time, affecting models predicting customer churn or product recommendations.

- Recurring Concepts: Patterns that reappear over time e.g., Tourism patterns that repeat annually, such as increased demand for winter clothing or holiday travel (Brzeziński & Stefanowski, 2014).

As drift occurs, any explanations or interpretations of the model based on its original training data become less reliable. Previously inferred relationships learnt by inspecting the models may no longer hold true. Drift often results from changes in hidden contexts that are not explicitly captured in the data. For instance, a change in policy regulations on consumer behavior might alter the underlying patterns, but these shifts aren't directly visible in the input features (Hinder, Jakob and Hammer, 2020).

The primary consequence of data drift is model decay, a reduction in the model's predictive accuracy and reliability. This can lead to:

- Inaccurate Predictions: As the model's assumptions no longer hold, its predictions become less accurate.
- Business Impact: Poor model performance can lead to incorrect business decisions, financial losses, or reduced customer satisfaction.
- Increased Maintenance Costs: More frequent model retraining and updates are required to maintain performance (Hinder, Jakob and Hammer, 2020).

Therefore, early detection of drift is crucial for data scientists and engineers to be proactive rather than reactive. This research aims to address this challenge by developing a model-agnostic approach for detecting concept drift. By leveraging techniques such as permutation importance, SHAP values, and MCR bounds, this study seeks to establish a versatile framework that can identify drift independently of the specific machine learning model being used. In summary, the research aims to answer the following questions:

- Can we create a reliable method to detect concept drift?

- Can we use changes in SHAP values and feature importance scores to determine drift?
- Will this method be independent of the specific machine learning model being used?
- Can we explain the changes in relationships between features and the prediction variable after detecting drift, thereby interpreting the model's workings?

Feature drift and concept drift are common in most of the dynamic environments where data and underlying patterns change over time. Understanding these drifts, their consequences, and the importance of early detection is essential for maintaining the effectiveness of machine learning models. By implementing robust monitoring and detection strategies, organizations can mitigate the adverse effects of drift and ensure their models continue to deliver accurate and reliable predictions.

# 2. Literature Review

The success of machine learning models depends on the consistency of the data they are trained with. However, in the real world, data is not static i.e., it changes over time. This change in the underlying patterns of data, known as feature drift, can greatly affect how the model performs. It can make the previously learned connections between data features and the target outcome outdated and less effective (Gama et al., 2014).

## 2.1 Reasons for the occurrence of drift in models

The concept of drift can be influenced by various factors, such as errors in the data or changes in the importance of features to the underlying model. Hammoodi et al., (2018) mainly discusses these two factors. As real-world data is messy, there can be typos, missing values, or inconsistencies that make their way into the data over time. These errors can mislead the model, causing it to learn faulty patterns. As unexpected erroneous data increases, causing a change in the distribution of input features, the model's performance drifts away from reality. Another factor that the author discusses in the models is changes in the feature importance. As features are the input quantities (or values) used by models to make predictions, their importance can fluctuate over time. For example, a model predicting customer purchases might have price as an important feature. However, during a recession, other features like brand loyalty or discounts might become more important for capturing buying behaviour. If the model doesn't adapt to these changes in feature importance, its predictions become inaccurate (Hammoodi et al., 2018).

## 2.2 Statistical methods for data drift detection

Existing literature highlights several methods to detect feature drift and concept drift in models. One popular method uses statistical tests to spot changes in data patterns (Gama et al., 2014). For example, non-parametric tests like the Kolmogorov-Smirnov (KS) Test compare the cumulative distribution functions (CDFs) of two datasets to find significant differences.

This test is mostly useful in scenarios where the assumption of equal variance and normal distributions cannot be met simultaneously. This allows for a more flexible approach to evaluating differences in distributions (Shao et al., 2014). By examining how the CDFs of two datasets diverge, changes in data distributions can be detected using the KS Test, which can be indicative of concept drift or other shifts in the underlying data generating process.

In addition to the KS Test, Chi-Square test, which is a non-parametric tests as well, can be employed in conjunction with masking (see below) techniques to point out specific feature distributions that are changing over time. The Chi-Square test is a statistical hypothesis testing method used to measure the correlation and dependence between different types of variables (Zhu et al., 2023). By applying the Chi-Square test to compare feature distributions before and after a certain point in time, researchers can identify which features are experiencing significant changes in their distributions, signalling potential drift.

Moreover, the Mann-Whitney U test, again a non-parametric test, is commonly used to compare the distributions of two independent samples. Masking is a technique which involves selectively focusing on specific features while analysing data. This technique when combined with the Mann-Whitney U test can help isolate the effects of feature drift by narrowing down on individual feature distributions (Ditzler & Polikar, 2011). This approach demonstrates a more granular understanding of how each feature contributes to changes in the overall data distribution, enabling data scientists or analysts to find out the exact features that are driving the observed drift.

These statistical tests have several limitations when detecting drift in models. They are very sensitive to the size of the sample, which can lead to false results in both large and small samples. These tests assume certain data characteristics, such as continuous data for KS, similar shapes for

Mann-Whitney, and normal distribution for t-tests. They may also struggle with different types of data, like discrete or skewed data. They are designed for one-variable analysis and don't work well with data that has many features. Furthermore, they are designed for batch processing rather than real-time data analysis and focus on p-values, which may not fully capture the practical significance of the observed drift (Praveen, 2024).

### 2.3 Model based approaches for concept drift detection

Model-based methods for concept drift detection are designed to work specifically with certain types of models and often leverage the unique characteristics of these models to identify when drift has occurred. For example, Yuan et al. (2023) proposes a novel approach for detecting and handling concept drift in random forests by leveraging the structure of individual trees. It introduces "forgetful data structures" that allow decision trees to adapt by selectively forgetting outdated information through node pruning. The authors present an incremental tree-building algorithm that updates trees efficiently as new data arrives, maintaining accuracy without complete retraining. They also explore using changes in tree structure, such as node growth or pruning, as indicators of concept drift. This approach provides a new perspective on managing drift in random forests, developing efficient algorithms for dynamic environments, and demonstrating the potential of tree structure as a drift indicator, thereby contributing significantly to machine learning and data mining.

Baier et al. (2021) proposes a method for detecting concept drift in neural networks by leveraging model uncertainty, which is especially useful when true labels are scarce or costly. The key ideas include using Monte Carlo Dropout to estimate prediction uncertainty and monitoring these estimates with the ADWIN (Adaptive Windowing) algorithm to detect significant changes, signalling potential drift. Adaptive Windowing (ADWIN) is a statistical method used to detect changes in data streams by dynamically adjusting the size of a window of recent observations. It maintains a

window that can grow or shrink in size based on changes detected in the data distribution. When drift is detected, the neural network is retrained using recent data. However, this method relies on model architecture (suitability of dropout for uncertainty estimation) and the sensitivity of ADWIN, which may require careful tuning.

Another paper introduces an ensemble-based approach to address concept drift in SVMs (Klinkenberg and Joachims, 2000). Key ideas include using multiple SVMs trained on different data subsets and employing a sliding window to capture recent data, allowing models to adapt to changes in data distribution. Drift detection is achieved by monitoring the performance of individual SVMs, with significant degradation indicating potential drift. Strengths of this method include improved robustness and accuracy through ensemble use and adaptability via the sliding window. However, it faces challenges in parameter tuning and can be computationally expensive for large datasets.

The above mentioned model-specific techniques based on random forests, neural networks, and SVM ensembles, are inherently tied to the unique characteristics and architectures of these models. For instance, methods leveraging tree structure changes are mainly applicable to decision trees and random forests, while approaches relying on dropout-based uncertainty estimates are specific to neural networks. Similarly, ensemble techniques with sliding windows are tailored to SVMs. These constraints limit the generalizability of such methods across different model types. This highlights the need for a model-agnostic framework that can identify concept drift independently of the underlying model architecture, ensuring broader applicability and consistent performance across diverse machine learning systems.

One such method is Permutation Importance, which measures the contribution of each feature to a model's predictions by randomly shuffling the values of a single feature and observing the resulting decrease in model

performance. Features causing a significant performance drop are considered more important, and changes in feature importance rankings between training and test data can indicate potential drift (Gary et al., 2024).

Permutation importance can be applied across various machine learning models, including deep neural networks, random forests, and support vector machines (Meng et al., 2022). It provides a straightforward and interpretable way to understand the contribution of each feature to the model's predictions. By summing up the permutation importance scores obtained from multiple machine learning models, a final feature importance score can be calculated, aiding in identifying the most critical features in the models (Meng et al., 2022).

### 2.4 Explainable AI (XAI) for drift analysis

In recent years, there has been a growing emphasis on explainable AI (XAI) techniques that not only detect drift but also provide insights into how the relationships between input and output variables change over time. SHAP (SHapley Additive exPlanations) values have emerged as a powerful tool in this context. SHAP values explain how each feature contributes to an individual prediction. Calculated using game theory, they provide insights into how a model utilizes features to arrive at an output (Lundberg, Allen and Lee, 2017).

Three key metrics can be used with SHAP values to compare the trained model data and the latest data:

- **Spread (Variance):** This analyses how the influence of a feature becomes more or less spread out across predictions (higher or lower variance).
- **Central Tendency (Mean):** This examines how the average impact of a feature on predictions shifts significantly.
- **Skewness:** This analyses how the distribution of SHAP values becomes more skewed in one direction compared to the other model.

Another model-agnostic approach gaining attention is the use of Model Class Reliance (MCR) bounds. MCR focuses on the overall performance of the model across different "Rashomon Sets", which represent alternative models with similar predictive power but potentially different emphasis on features (Fisher, Rudin and Dominici, 2019).

Current Variable Importance (VI) measures often don't consider that multiple models can fit the data equally well, each potentially using different variables. This situation, known as the "Rashomon" effect, raises questions about how to accurately describe the importance of each variable and whether different analysts would draw the same conclusions.

To address this, the concept of model class reliance (MCR) was introduced. MCR looks at the range of reliance on a variable across all well-performing models within a certain class, rather than focusing on a single model. This approach aims to reflect the nature of the prediction problem itself rather than the choices made by individual analysts (Smith, Mansilla and Goulding, 2020).

MCR is particularly relevant in contexts where we care about the variables being used, i.e. in criminal recidivism prediction where one should not use protected characteristic such as race. (Smith, Mansilla and Goulding, 2020).

In the context of drift detection, observing a significant decrease in the MCR bound for the latest data compared to the initial data can suggest potential drift.

### 2.5 Combination of Permutation Importance and XAI techniques

In order to better understand the subtleties of concept drift and provide practitioners with more control over when to retrain their models, the combination of model-agnostic methods such as Permutation Importance, SHAP values, and MCR bounds offers a comprehensive framework. These techniques not only detect drift but also explain how the relationships

between input and output variables change over time. For example, a practitioner might observe that both SHAP values and Permutation Importance indicate potential drift while MCR bounds do not. This information allows the practitioner to decide, based on the specific application and tolerance for change, whether to update or retrain the model.

The model-agnostic nature of these methods allows them to be applied to any predictive model, overcoming limitations of traditional model specific and statistical tests, which often require specific data distribution assumptions, struggle with multivariate data, and provide limited interpretability.

As the field of explainable AI evolves, future research may focus on developing more sophisticated techniques for visualizing and interpreting concept drift, as well as automated systems for continuous monitoring and adaptation of machine learning models in production environments.

# 3. Methodology

This chapter explains the methodology undertaken to investigate the detection of concept drift using feature importance scores and SHAP (SHapley Additive exPlanations) values across multiple machine learning models. As outlined in figure 3.1, this methodology aims to provide a robust framework for determining concept drift that is model agnostic. It also aims to address key research questions: whether a reliable method to detect concept drift can be created and whether the detected drift can be explained by analysing changes in the relationships between features and the prediction variable. This framework also aims to provide researchers the flexibility in determining the existence of drift, in cases where only one or two of the metrics show promising shifts e.g. there can be a scenario where the differences in Permutation Importance scores indicate drift but SHAP metrics do not.

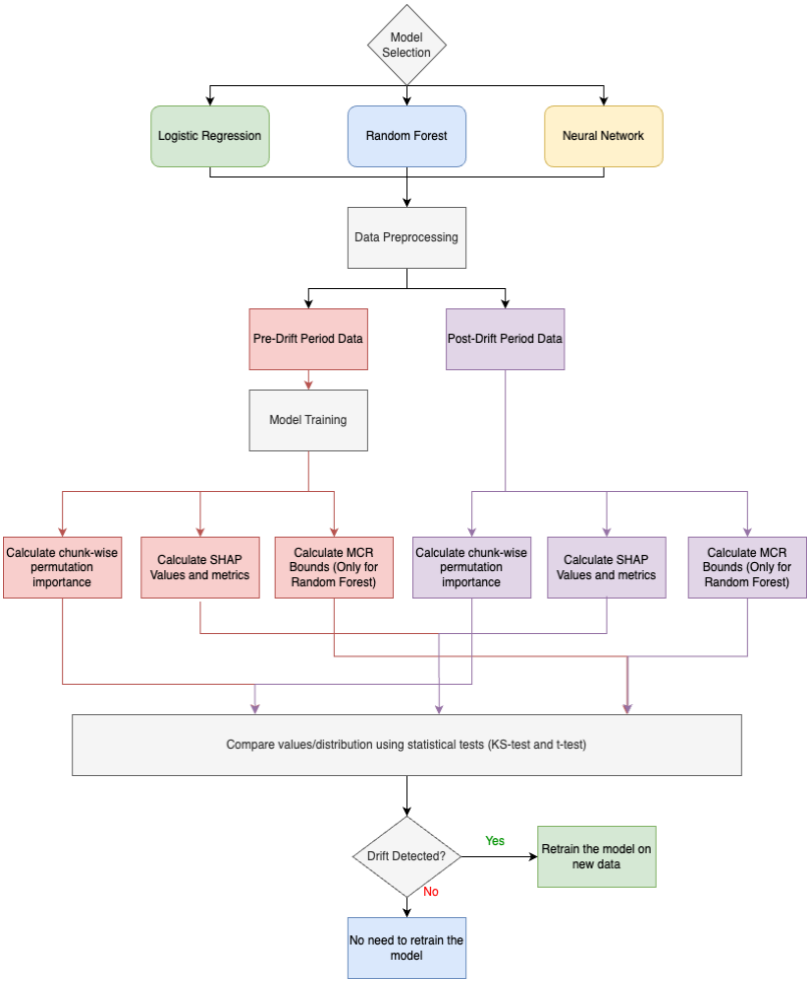


*Figure 3.1: Flowchart for a model agnostic approach in detecting drift*

The approach involves training models on pre-drift data and then applying these models to post-drift data. By analysing changes in feature importance scores and SHAP values across data chunks, this method aims to identify significant shifts in the relationship between features and the prediction variable. These shifts are statistically evaluated using tests like the KS-test and t-test. Therefore, this method attempts to offer a deeper understanding of how and why drift occurs, allowing for more accurate and interpretable insights into model behaviour. It is theoretically superior to the conventional method of merely observing changes in model output, as it provides a bigger perspective how specific features' contributions change over time. This may enhance the detection and explanation of concept drift without the need for immediate model retraining.

### 3.1 Model Selection and Justification

Three distinct machine learning models are selected for this study: Logistic Regression, Random Forest, and Neural Networks (see figure 3.2). These models represent a range of algorithmic approaches, from linear models (Logistic Regression) to ensemble methods (Random Forest) and deep learning techniques (Neural Networks). The selection of these models is motivated by the need to demonstrate that the proposed method for detecting concept drift is independent of the specific machine learning model being used. Therefore, the methodology ensures that the findings

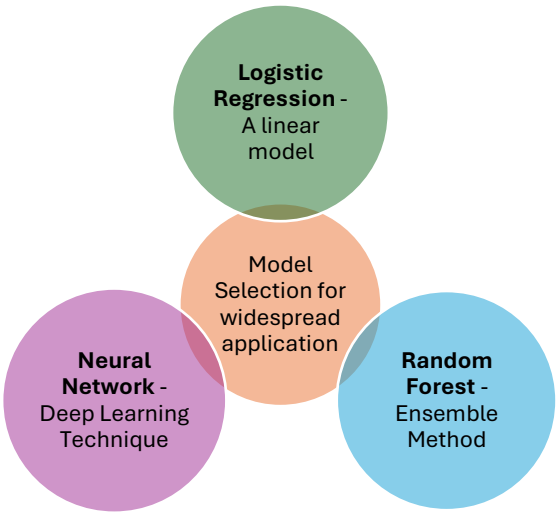


*Figure 3.2: Model selection for detecting concept drift*

are not limited to a particular type of algorithm but can broadly be applicable to different model architectures.

### 3.2 Introduction to Datasets

Two datasets are analysed in this study to evaluate the detection of concept drift. The first dataset is the Electricity dataset, sourced from the MOA dataset repository (MOA, 2017) and originally described by Harries (1999). This dataset captures data from the Australian New South Wales (NSW) Electricity Market, where electricity prices fluctuate based on the dynamic interaction of supply and demand. Prices in this market are set every five minutes, reflecting the constantly changing conditions within the electricity supply network. The Electricity dataset is employed to detect sudden or abrupt drift in the dataset, particularly focusing on the impact of the introduction of the National Electricity Market (NEM).

The second dataset is the Airlines dataset, also sourced from the MOA dataset repository (MOA, 2017). This dataset contains records of individual flights, each described by features such as Airline, Flight Number, Departure Airport (AirportFrom), Arrival Airport (AirportTo, see Appendix for abbreviations of airport names used in the dataset), Day of the Week (DayOfWeek), Departure Time (Time), and Flight Duration (Length). The target variable is binary, indicating whether the flight arrived on time or not. The Airlines dataset is used in an attempt to detect cyclic drift patterns, particularly focusing on differences between weekends and weekdays.

#### 3.2.1 Electricity Dataset Segmentation

Given the objective of identifying sudden or abrupt drift, the model is trained on the first 10 months of data, starting from May 7, 1996. This training period is strategically chosen to include data prior to the introduction of the National Electricity Market (NEM) on May 2, 1997, which falls approximately in the 12th month of the dataset. To observe the impact of this significant market shift, the test dataset is selected to cover a 10-month period, spanning from the 11th month to the 20th month of the

data, thus capturing the period immediately before and after the NEM's introduction. This segmentation allows for a temporal evaluation of the data, providing a clear perspective on the sudden changes in the evaluation metrics that may arise due to this substantial market transition.

### 3.2.2 Airlines Dataset Segmentation

In this study, the model is trained exclusively on weekday data and tested on weekend data to observe whether there are significant changes in the relationship between features and the output variable that could indicate concept drift. This approach is designed to investigate potential shifts that might occur when moving from a weekday operational context to a weekend context. Temporal windows are created within both the weekday and weekend datasets, and the model's permutation importance scores, SHAP metrics, and accuracy scores are compared across these testing period windows. This setup can be considered a semi-synthetic dataset, as the temporal windowing does not follow the natural progression of time, it is not entirely accurate in terms of real-world chronological order. But it is deliberately structured this way to explore the existence of potential drift between weekdays and weekends.

## 3.3 Measurement of Feature Importance Scores

Feature importance scores are measured using permutation importance for each of the three models (as explained in 3.1). Permutation importance involves randomly shuffling the values of each feature in the test data and observing the impact on the model's performance. The greater the decrease in model accuracy after permuting a feature, the more important that feature is considered to be (Meng et al., 2022).

For each model, feature importance scores are calculated chunk wise, using temporal window. The mean permutation importance scores across these periods are then plotted to visualise any noticeable shifts. By comparing the feature importance scores before and after the presumed drift, it is attempted to detect changes that may indicate concept drift.

### 3.4 SHAP Value Analysis

SHAP values are used to interpret the contribution of each feature to the predictions made by the models (Lundberg, Allen and Lee, 2017). For each model, SHAP summary plots are generated for both the pre-drift and post-drift periods. These plots visualise the impact of each feature on the model's predictions, allowing for a comparison of feature contributions before and after the assumed drift.

By examining the SHAP summary plots, it is attempted to identify any shifts in feature contributions that may correspond to concept drift. If significant changes in SHAP values are observed between the two periods, this may indicate that the relationships between the features and the prediction variable might have altered, potentially suggesting concept drift.

In addition to feature importance and SHAP values, the statistical properties of the features themselves are analysed. For each feature, the mean, variance, and skewness are calculated for both the pre-drift and post-drift periods. These statistics are then plotted so that a comparison between the two periods can be made.

Changes in the distribution of a feature can indicate shift in the underlying data distribution, which can be a sign of concept drift. By analysing these distributional changes alongside shifts in feature importance and SHAP values, a more comprehensive understanding of the nature of the drift can be achieved.

### 3.5 Statistical Analysis of Feature Importance Shifts

To quantitatively assess the significance of any observed shifts in feature importance scores or SHAP metrics, statistical tests are applied. The t-test and the Kolmogorov-Smirnov (KS) tests are employed to determine if there are statistically significant differences between the feature importance scores of the pre-drift and post-drift periods. These statistical tests help confirm whether the changes observed in feature importance scores are

due to genuine concept drift or are merely the result of random fluctuations (Shao et al., 2014). A significant result would suggest that the feature importance scores may have shifted, indicating a potential concept drift.

### 3.6 Model Class Reliance (MCR) Calculation

Model Class Reliance (MCR) is calculated specifically for the Random Forest model, as this method is not standard (or complex to achieve) for Logistic Regression and Neural Networks. MCR measures the importance of a specific feature by evaluating the change in model accuracy after permuting the feature's values in the test data. They represent alternative models with similar predictive power but potentially different emphasis on features (Fisher, Rudin and Dominici, 2019).

Therefore, the MCR bounds are calculated for both the pre-drift and post-drift datasets. These bounds are then plotted to identify any noticeable differences. A significant difference in MCR between the two periods would suggest that the model's reliance on certain features has changed, potentially due to concept drift.

### 3.7 Model-Agnostic Comparison

Finally, the results from all three models are compared to determine whether the detected drifts are consistent across different algorithms. By comparing the shifts in feature importance scores, SHAP values, and MCR (where applicable), the methodology aims to assess whether the concept drift is model-independent. If similar patterns of drift are observed across all three models, this would support the hypothesis that the proposed method for detecting concept drift is robust and applicable across different machine learning algorithms.

# 4. Implementation and Evaluation

In this chapter, the descriptive analysis of the datasets, their pre-processing, and the application of the three models under consideration are presented. The evaluation metrics, including accuracy, permutation importance scores, and SHAP values, are plotted for the testing windows to investigate concept drift using the proposed methodology.

Following this, the Airlines dataset is analysed to explore potential cyclic drift patterns between weekdays and weekends. Through these steps, the chapter systematically evaluates the effectiveness of the methodology in detecting and interpreting concept drift across different contexts.

## 4.1 Implementation of Electricity Dataset

The implementation process begins with the Electricity dataset, where the analysis focuses on identifying sudden or abrupt drift.

### 4.1.1 Data description

The Electricity dataset comprises 45,312 instances, each representing a data point collected at 30-minute intervals over a period extending from May 7, 1996, to December 5, 1998. The primary objective of this dataset was to predict how the electricity prices change relative to a moving average of the previous 24 hours. This change is captured by the class label, which indicates whether the electricity price has increased or decreased in comparison to the moving average (MOA, 2017).

The dataset includes the following covariates:

- date: The date on which record was observed.
- day: The serially numbered day on which the record was observed.
- period: The 30-minute interval in the time frame.
- nswprice: The price of electricity in New South Wales.
- nswdemand: The demand for electricity in New South Wales.
- vicprice: The price of electricity in Victoria.

- vicdemand: The demand for electricity in Victoria.
- transfer: The amount of electricity transferred between New South Wales and Victoria.

These covariates provide a detailed overview of the market conditions in both New South Wales and Victoria, as well as the electricity transfer between these states, offering valuable insights into the factors influencing electricity prices. One of the key aspects of this dataset is the period around May 2, 1997, which marks the commencement of the process to introduce a National Electricity Market (NEM) in Australia. This period is particularly significant because, from this date, a trial NEM allowed wholesale electricity trading between the states of New South Wales, Victoria, the Australian Capital Territory, and South Australia (Harries, 1999). This transition had a notable impact on the price and demand for Victoria and Transfer variables. They were constant until the introduction of NEM and began fluctuating since then.

Since the assumed pre-drift period covers the first 10 months, during which these variables were constant, they are excluded from model training as they did not provide meaningful information to the model.

The period, price, demand and transfer variables in the dataset have already been normalized to the interval [0,1], ensuring that the variables are on a comparable scale, which is essential for effective analysis. This normalisation also aids in reducing the Impact of outliers and allows for more straightforward application of machine learning algorithms.

The Electricity dataset serves as a valuable resource for studying concept drift, particularly in the context of the transition to a national market system. The introduction of the NEM and the resultant changes in the electricity market conditions provide a rich ground for analysing how machine learning models can detect such abrupt shifts in data distribution over time.

### 4.1.2 Cleaning and pre-processing

Before applying machine learning models to the Electricity dataset, several essential cleaning and pre-processing steps are undertaken to ensure the data is suitable for analysis.

First, the class variable, which originally indicated whether the electricity price had increased (UP) or decreased (DOWN) relative to a moving average of the last 24 hours, is converted to a binary format. In this conversion, instances labelled as "UP" are assigned a value of 1, and those labelled as "DOWN" are assigned a value of 0. This binary representation of the target variable facilitates straightforward application of classification algorithms in the subsequent modeling steps.

As the price, demand, and transfer-related columns were already normalised to the [0,1] range, no additional scaling or normalisation is required for these features. The normalised values ensured that all features are on a comparable scale, thereby simplifying the training process of the models.

To assess the relationships between the features, a correlation matrix is computed (as shown in the figure 4.1). The analysis reveals that there are no highly correlated features, eliminating the need for any additional feature engineering or dimensionality reduction techniques. With these steps, the dataset is made ready for the implementation of machine learning models to detect concept drift.

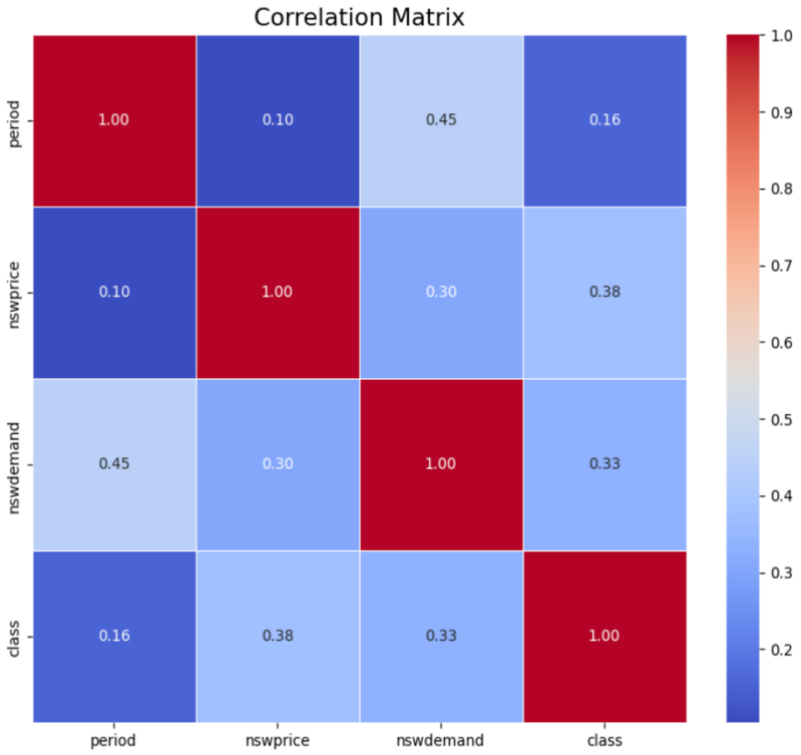


*Figure 4.1: Correlation matrix of features used for training Electricity dataset*

### 4.1.3 Implementing Logistic Regression

Logistic regression is the first model implemented in this study to detect concept drift in the Electricity dataset. As explained in the previous chapter, the dataset is segmented into pre and post drift periods . The pre-drift period, consisting of the first 10 months of data, is used for training to simulate a real-world scenario where a model is trained before any significant changes occur. The drift is anticipated to occur around the 12th month, coinciding with the introduction of the National Electricity Market (NEM). Therefore the testing period data (which covers the period where drift is expected to be observed) would be used for predicting model outputs and evaluating other metrics.

The overall accuracy during the training period was 79%, while the testing period showed a decreased accuracy of 70%, clearly indicating a degradation in overall performance. However, a closer examination of the accuracy over the 10-month testing period (as shown in figure 4.2) reveals a more detailed pattern: there is a downward trend in accuracy from the 11th month to the 20th month, with accuracy dropping as low as 64% in the final temporal window.

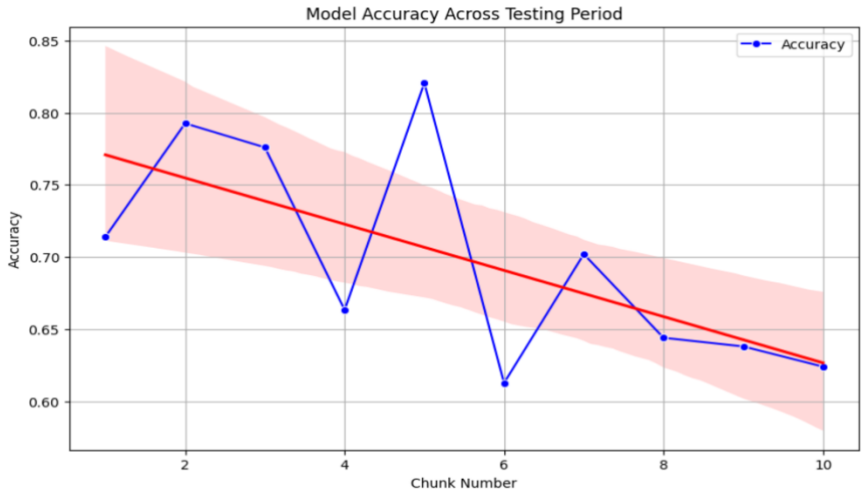


*Figure 4.2: Accuracy scores across different time periods in testing dataset*

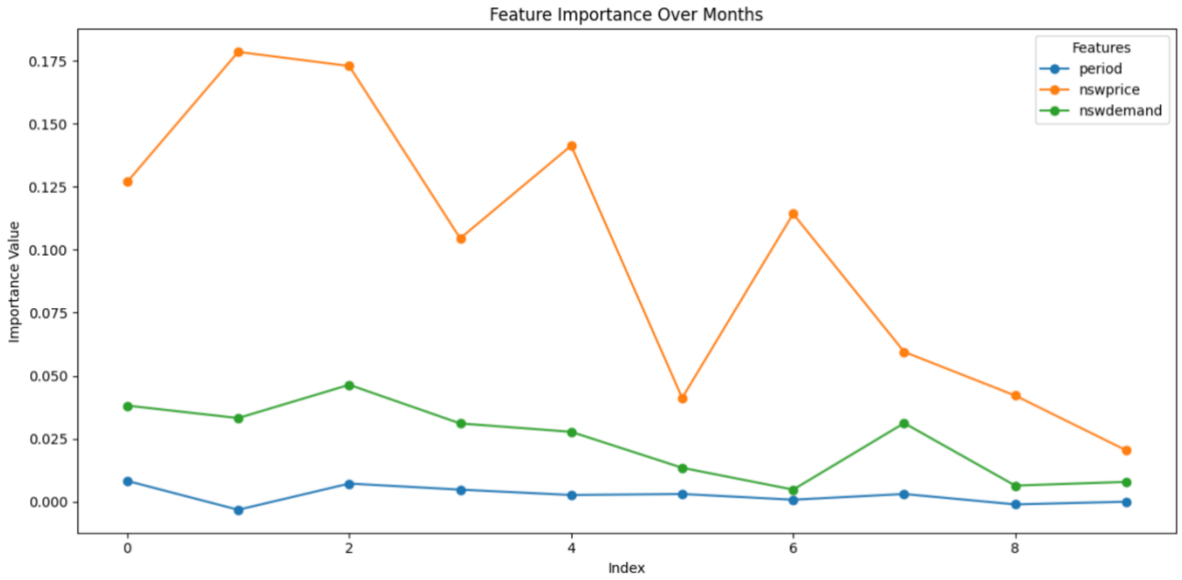


*Figure 4.3: A plot of mean permutation importance scores across testing period*

Following this, permutation importance scores are calculated and plotted for these periods (see figure 4.3). Although the T-test and KS tests did not identify any statistically significant differences, the plotted scores suggest

a notable downward shift in importance values for the NSW Price variable and a slight downward trend for NSW Demand starting from the 12th month, which aligns with the introduction of the National Electricity Market (NEM).

Next, SHAP values are analysed using summary plots over the training and testing periods (see figure 4.4). When statistically compared between the training and testing sets using t-tests and KS tests, the SHAP metrics are found to be significantly different. This distinction is further evident from the SHAP summary plots, as well as from the SHAP metrics plots (see figure 4.5), which include mean, variance, and skewness. These findings suggest that the model's interpretation of the relationships between features and the target variable has probably shifted following the introduction of NEM, hinting towards the potential existence of concept drift.

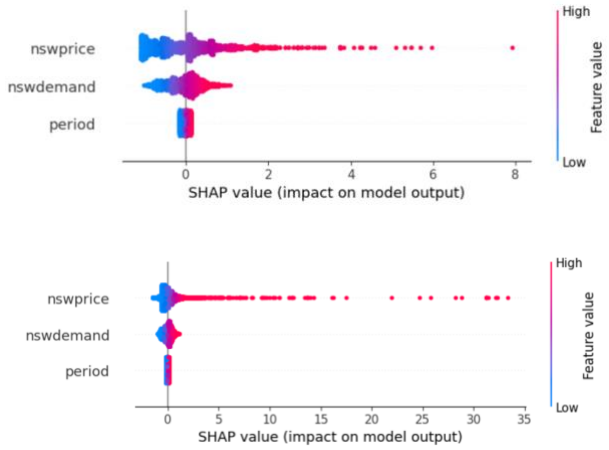


*Figure 4.4: SHAP Summary plots for assumed pre-drift (top) and post-drift (bottom) datasets*

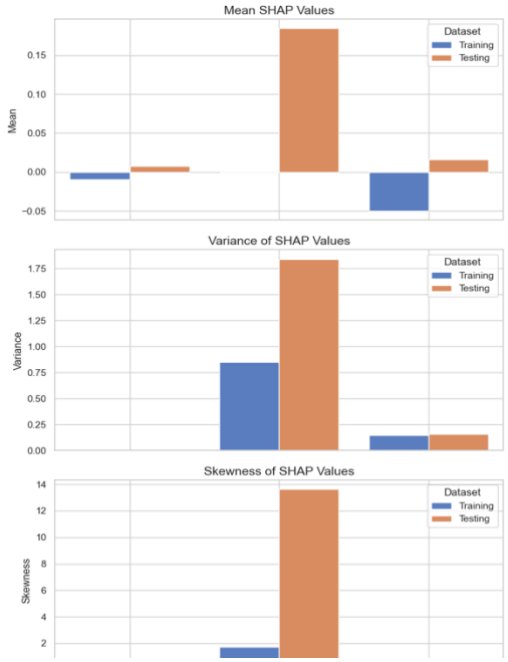


*Figure 4.5: Plot of SHAP metrics for training and testing periods*

### 4.1.4 Implementing Random Forest

Random Forest is the second of the three models implemented in this study to detect concept drift in the Electricity dataset. The assumed pre-drift and post-drift period is chosen as explained in the previous section. The overall accuracy during the training period is 84%, while the testing period showed a reduced accuracy of 68%, clearly indicating a degradation in overall performance. A detailed examination of the accuracy over the 10-month testing period (that includes the period when NEM was introduced) reveals a downward trend (see figure 4.6), with accuracy declining from the 11th month to the 20th month, reaching as low as 64% in the final temporal window.

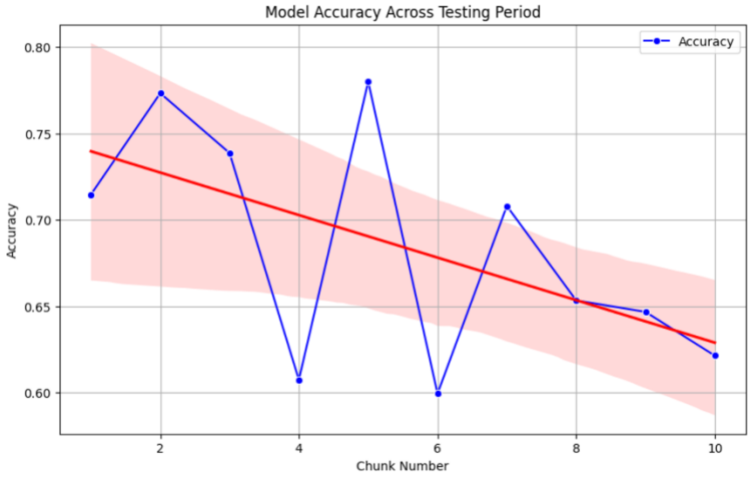


*Figure 4.6: Plot of accuracy scores across testing period*

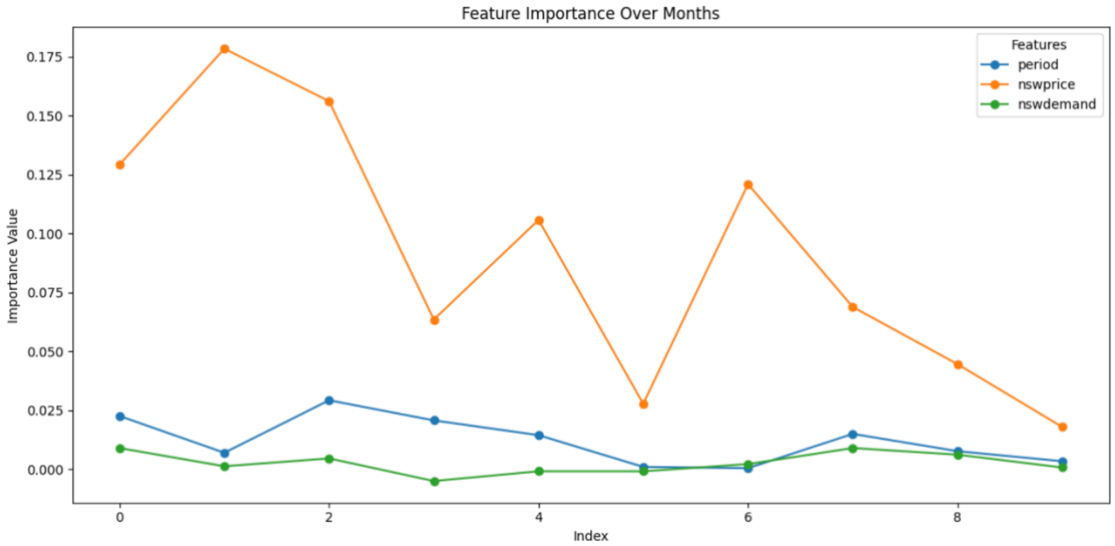


*Figure 4.7: Plot of permutation importance scores across windows in testing period*

Permutation importance scores are then calculated and plotted for these periods as shown in figure 4.7. While the T-test and KS tests did not detect any statistically significant differences, the plotted scores indicate a noticeable downward shift in the importance values for the NSW Price variable starting from the 12th month, which coincides with the introduction of the National Electricity Market (NEM).

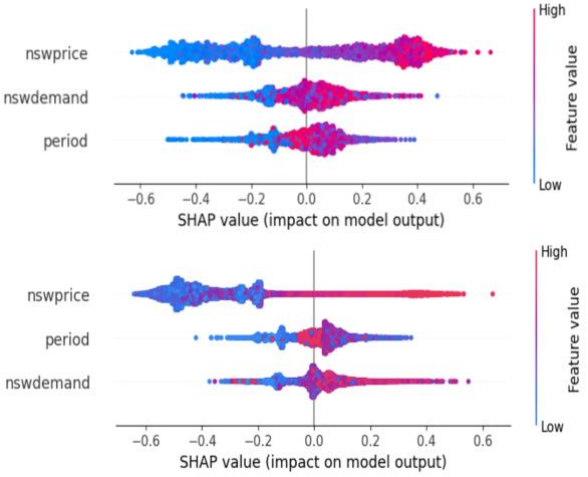


*Figure 4.8: SHAP Summary plots for training (pre-drift, on top) and testing (post-drift, bottom) datasets*

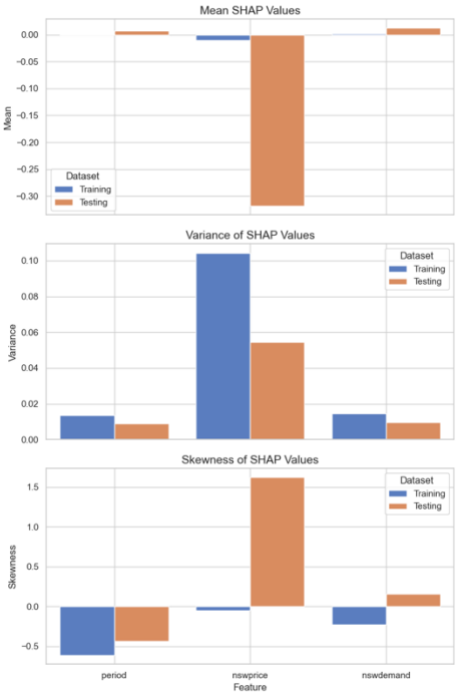


*Figure 4.9: SHAP metrics for training and testing datasets*

Next, SHAP values were analysed using summary plots for both the training and testing periods (see figure 4.8). The summary plots reveal interesting shifts in the patterns between the training and testing datasets. When SHAP metrics were compared statistically between the training and testing sets using t-tests and KS tests, the differences emerged as statistically significant. This is further demonstrated by the SHAP summary plots and SHAP metrics plots (mean, variance, and skewness), highlighting the shifts in feature importance and indicating the presence of concept drift following the introduction of NEM (see figure 4.9).

Additionally, Model Class Reliance (MCR) bounds are calculated for both the training and testing periods (see figure 4.10). These bounds, MCR+ and MCR-, show shifts in their values when comparing the training and testing sets. However, while the shifts in MCR bounds indicate changes in model reliance on certain features, these differences were not statistically significant (when compared using T-test and KS-test).

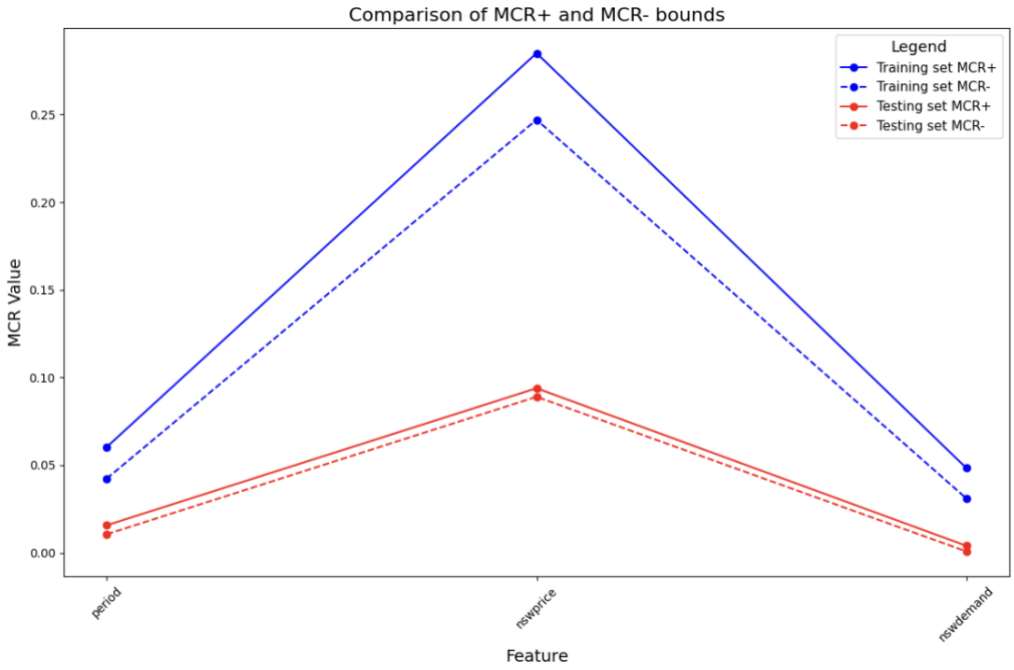


*Figure 4.10: MCR Bounds for training and testing datasets*

### 4.1.5 Implementing Neural Networks

The Neural Network is the third model implemented in this study to detect concept drift in the Electricity dataset. The data segmentation follows the same approach as previously described, with the model being trained on the first 10 months of data, which is assumed to be the pre-drift period. As the National Electricity Market (NEM) was introduced in the 12th month, the testing period spans from the 11th to the 20th month, following the training period to emulate a real-world temporal problem.

The overall accuracy during the training period was 82%, which dropped to 66% for the testing dataset, indicating a decline in model performance.

When examining the accuracy scores across the 10-month testing window, a clear decreasing trend is observed as shown in the figure 4.11.

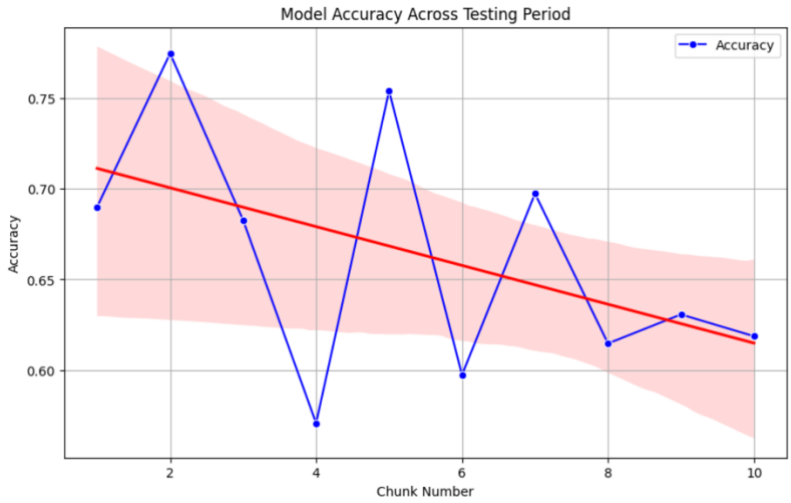


*Figure 4.11: Accuracy scores across testing period windows*

Permutation importance scores are then calculated for these windows, revealing a general decrease in importance scores for the NSW Price variable (see figure 4.12). However, statistical tests such as the t-test and KS-test did not identify this shift as significant.

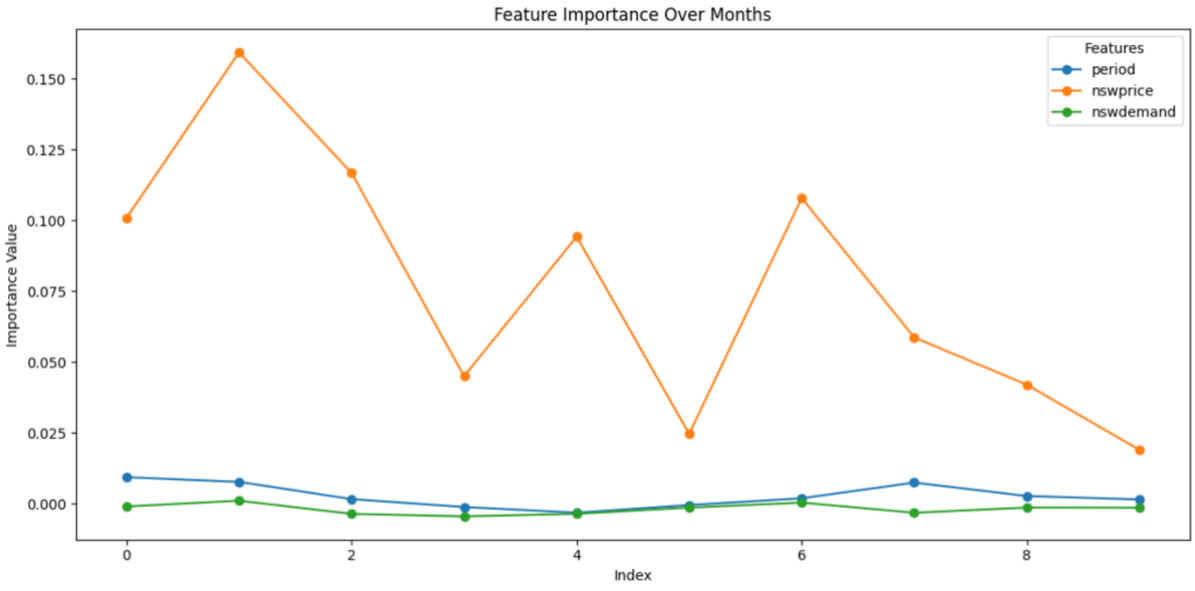


*Figure 4.12: Permutation Importance scores across the testing period*

Next, SHAP summary plots were generated for both the training and testing datasets, which showed noticeable shifts in the shapes and values of the plots (as shown in the figure 4.13.) When SHAP metrics such as mean, variance, and skewness were plotted (see figure 4.14), these shifts became

more evident, indicating changing feature contributions. Furthermore, the KS-test confirmed that there were statistically significant differences between the SHAP metrics of the training and testing sets, supporting the presence of concept drift.

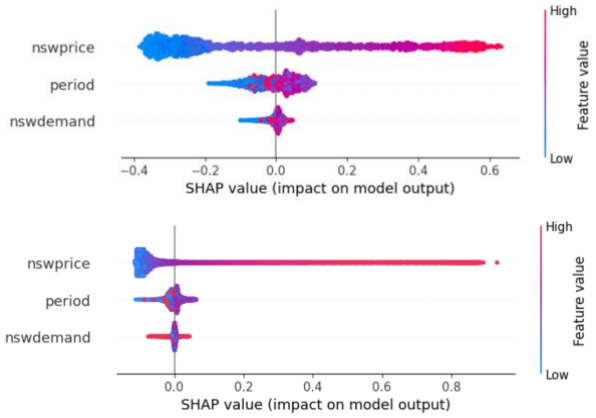


*Figure 4.13: SHAP Summary plots for training and testing sets*

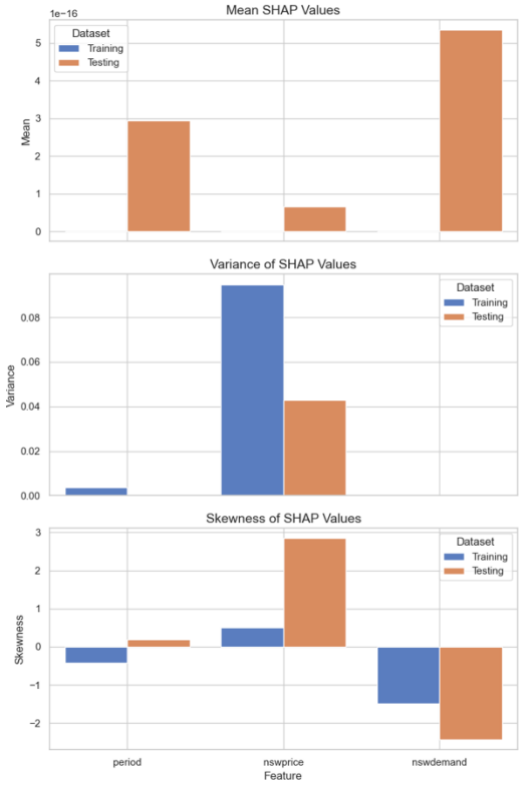


*Figure 4.14: SHAP metrics for training and testing sets*

## 4.2 Implementation of Airline Dataset

For next stage of implementation process, Airline dataset is utilised where the analysis focuses on identifying cyclic drift.

### 4.2.1 Data Description

The Airlines dataset contains a total of 539,383 records, each representing a unique flight. A preliminary inspection of the dataset revealed that it is well-maintained with no missing values across all columns. This ensures that the entire dataset can be utilised without the need for imputation or exclusion of records, preserving the integrity of the analysis.

An interesting aspect of the dataset is the DayOfWeek feature, which allows the data to be naturally segmented into weekday and weekend flights. Given that travel patterns often differ between weekdays and weekends, it is reasonable to hypothesize that the weekend data may exhibit drift from the weekday data. This potential for cyclic drift, where patterns change based on the day of the week, makes the dataset particularly suitable for examining and validating drift detection methods.

### 4.2.2 Cleaning and Pre-processing

The Interquartile Range (IQR) method is used to identify outliers within the dataset. They are not detected in the features except for Length (representing length of the flight). Although outliers are detected in this variable, they are perceived as legitimate and reflective of actual variations in flight durations. These outliers are likely attributable to factors such as extended flights, layovers, or unusual flight routes, which naturally occur

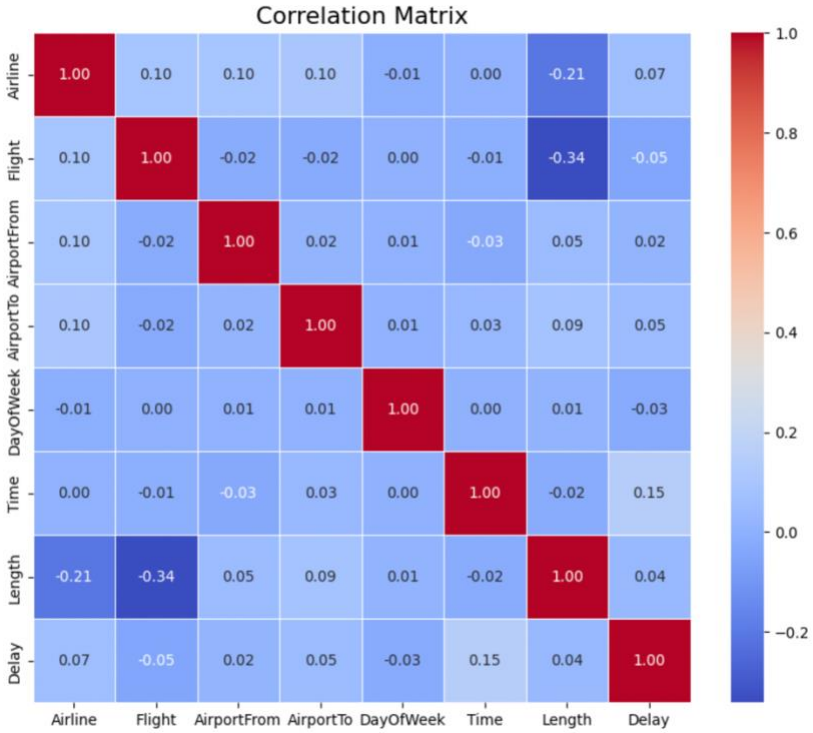


*Figure 4.15: Correlation matrix for the features of Airline dataset*

in air travel. Consequently, these outliers are retained as part of the normal data, ensuring that the analysis accurately captures the full range of flight durations.

To further prepare the data for analysis, standard scaling is applied to the Length and Time features. Since both variables are continuous and of the float type, standard scaling was necessary to normalise their values. This step ensures that the features are on a comparable scale, which is particularly important for models sensitive to the magnitude of input data.

In addition, label encoding is performed on categorical features, including Airlines, From Airport, and To Airport. Label encoding converts these categorical variables into numerical values, making them suitable for use in machine learning models. This transformation allows the models to interpret and utilise these features effectively.

To assess the relationships between the various features, a correlation matrix is plotted (see figure 4.15). The analysis did not reveal any significant correlations between the features, indicating that the variables in the dataset are relatively independent of each other. This lack of multicollinearity may also suggest that each feature contributes unique information to the dataset, which is beneficial for subsequent modeling and analysis tasks.

Day of the week is omitted from the feature set for model analysis as it provides unseen information to the models that can cause the models to behave abnormally.

### 4.2.3 Implementing Logistic Regression

Logistic Regression is the first of the three models implemented to detect concept drift in the Airlines dataset. This linear model is selected due to its simplicity and effectiveness in binary classification problems. The model is trained on the weekday data, under the assumption that this subset represents the pre-drift period. Although the dataset is split into weekday

and weekend subsets, which is not temporally correct, this segmentation allows the creation of a semi-synthetic dataset as outlined in the methodology. This setup is purposefully designed to observe potential drift on weekends. The Logistic Regression model is implemented using default parameters, providing a baseline for further comparisons with more complex models.

The Logistic Regression model is trained on the weekday dataset, achieving an overall accuracy of 58%. And when the same model is applied to the weekend dataset, the overall accuracy remained at 58% as well. However, when analysing the trendlines for accuracy scores over the segmented windows within the weekday and weekend datasets, a notable difference emerged (see figure 4.16). The trendline for the weekend dataset exhibited a steeper decline compared to the more consistent accuracy scores observed in the weekday dataset.

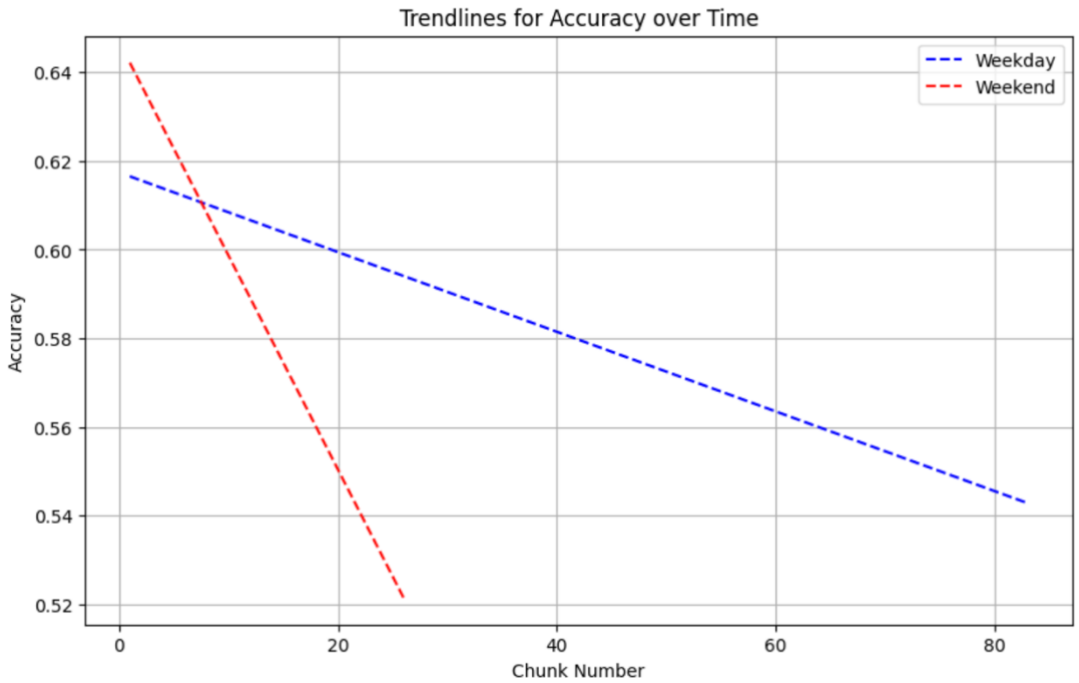


*Figure 4.16: Trendlines for accuracy over time for weekday and weekend datasets*

Therefore, for investigation of potential drift, permutation importance scores are computed for both weekday and weekend data across temporal windows of 5,000 records each. By analysing these scores, differences in feature importance between the weekday and weekend datasets are observed. Specifically, the Flight, Time and Length features showed noticeable shift in their mean permutation importance scores when comparing weekday to weekend data (as shown in the figure 4.17).

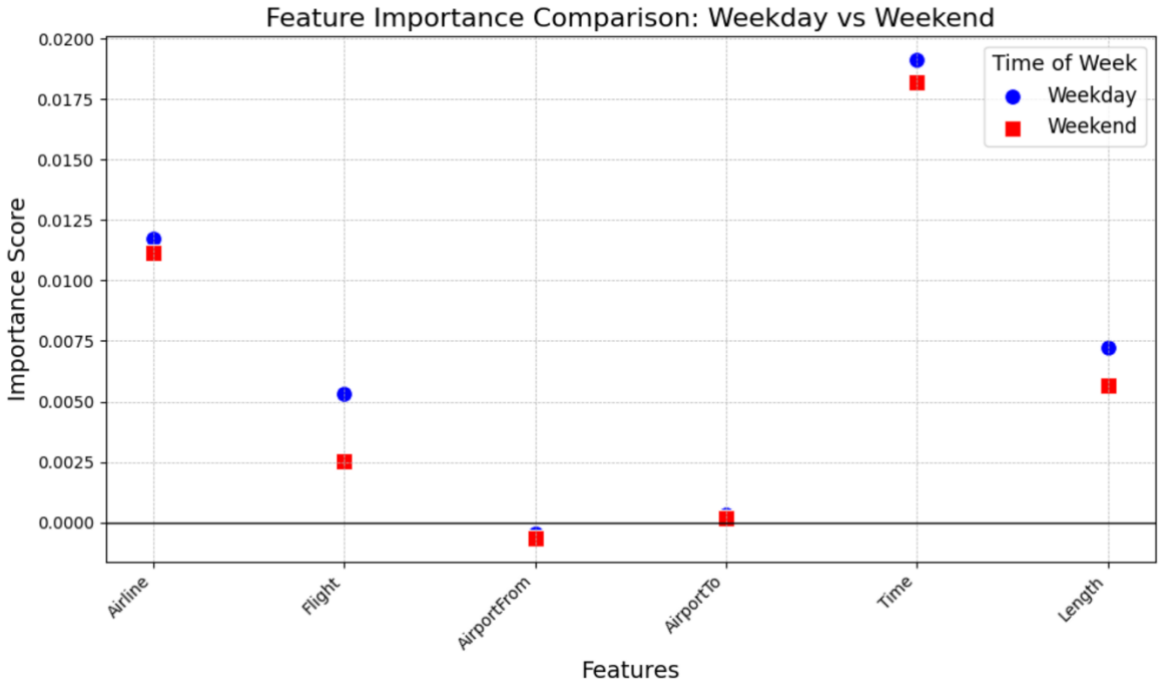


*Figure 4.17: Mean Permutation Importance scores for Weekday and Weekend datasets measured for Logistic Regression*

To quantify these differences, statistical tests are performed. But both the t-test and the Kolmogorov-Smirnov (KS) test could not confirm that there were significant differences in the permutation importance scores between the weekday and weekend datasets for these features.

In addition to permutation importance analysis, SHAP (SHapley Additive exPlanations) values are calculated to gain deeper insights into feature contributions during both periods. SHAP summary plots are generated for

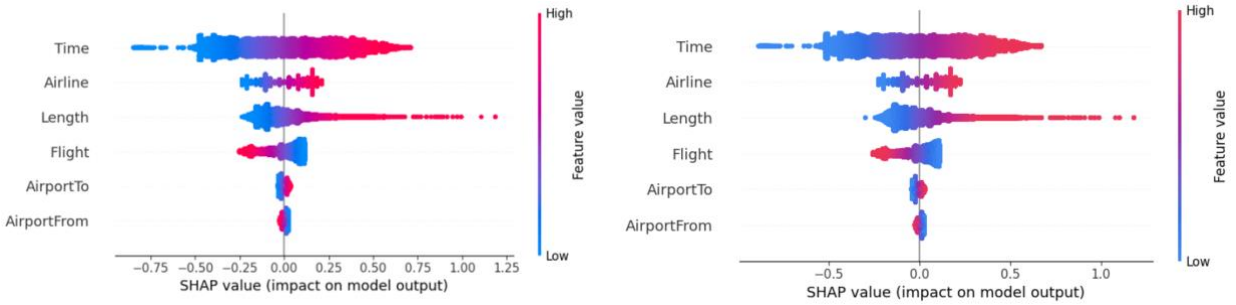


*Figure 4.18: SHAP summary plots for Weekday (left) and Weekend (right) datasets measured for Logistic Regression*

the weekday and weekend datasets (see figure 4.18). The overall shapes of the SHAP value distributions are similar for corresponding features.

To systematically analyse these changes, a plot of SHAP metrics - mean, variance, and skewness is created for both the weekday and weekend datasets. Plot in figure 4.19 clearly illustrates a shift in the mean SHAP values for some features, fortifying the observations from the permutation importance analysis. Statistical tests like T-test and KS test applied to these

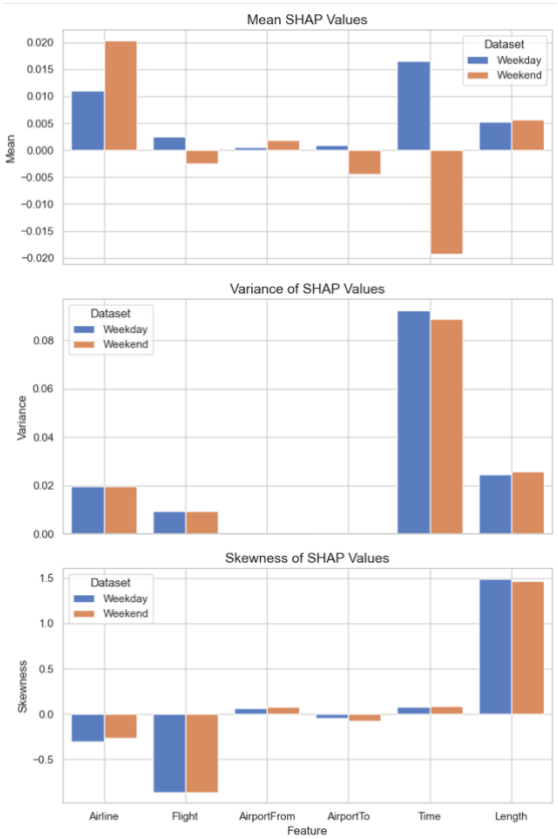


*Figure 4.19: A plot of SHAP metrics for Weekday and Weekend datasets measured for Logistic Regression*

SHAP metrics further confirmed the significant differences between the two sets for Airline, Airport From and Airport To features.

### 4.2.4 Implementing Random Forest

The Random Forest model is the second machine learning algorithm implemented to detect concept drift in the Airlines dataset. Known for its robustness and ability to handle complex interactions between features, Random Forest is trained on the weekday data, which is assumed to represent the pre-drift period. The model is initialised with default

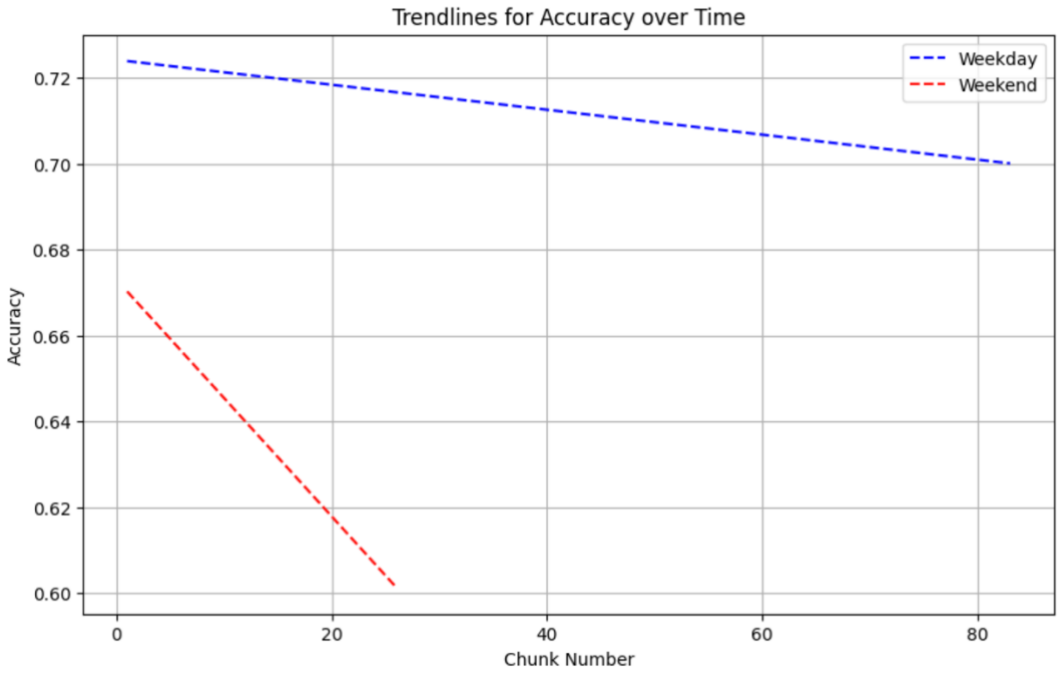


*Figure 4.20: Trendlines for accuracy over time for weekday and weekend datasets*

parameters to establish a baseline performance similar to the logistic regression model.

Upon training, the Random Forest model achieved an accuracy of 66% on the weekday dataset. However, when tested on the weekend dataset, the overall accuracy decreased slightly to 64%. Analyzing the trendlines of accuracy scores for both weekday and weekend data segments revealed a noticeable shift (as shown in figure 4.20), with the weekend datasets exhibiting the lowest accuracy scores. This suggests that the model's

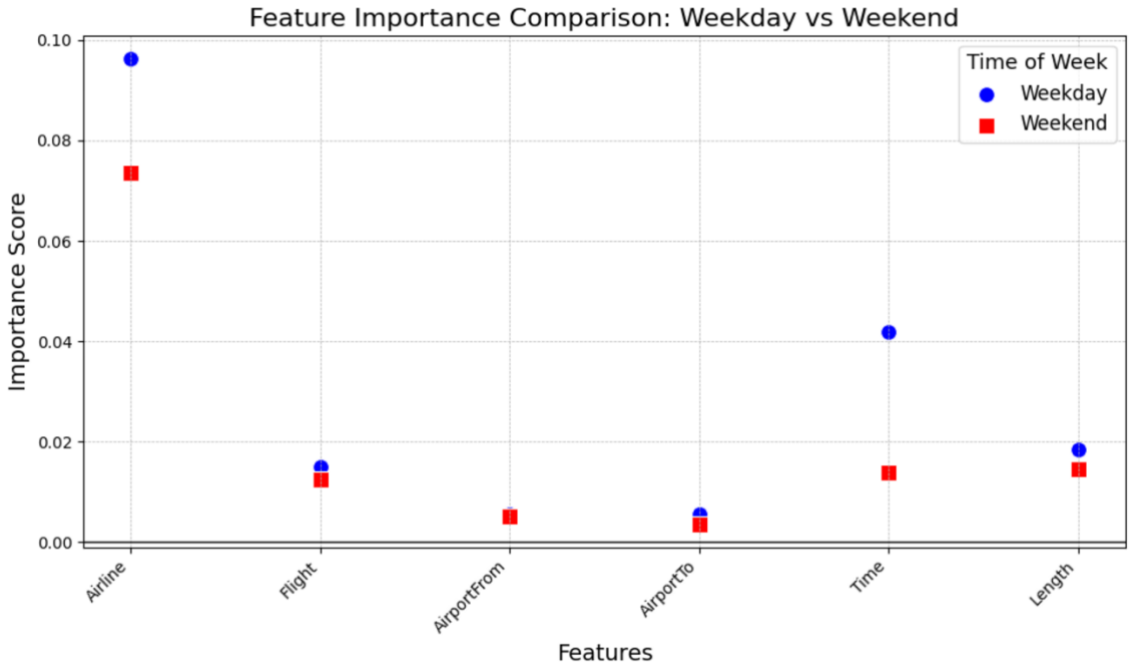


*Figure 4.21: Mean Permutation Importance scores for Weekend and Weekday datasets measured for Random Forest model*

performance varies between weekdays and weekends, potentially indicating the presence of concept drift.

Next, permutation importance scores were calculated for both weekday and weekend datasets (see figure 4.21). Interestingly, statistical tests (T-test and KS-test) identified the importance scores for the Airline and Time features as significantly different between the two datasets. This indicates that these features contribute differently to the model's predictions depending on whether the data is from a weekday or a weekend.

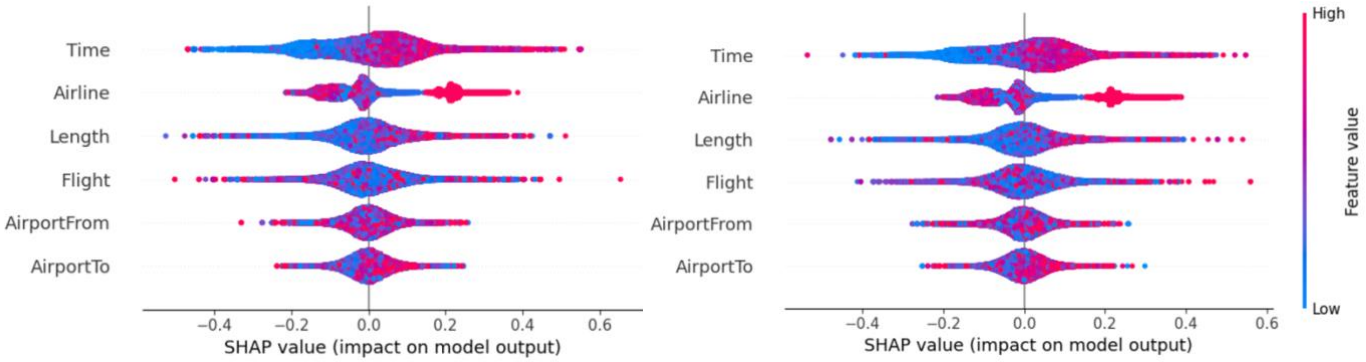


*Figure 4.22: SHAP Summary plots for weekday and weekend datasets*

SHAP summary plots are also generated as shown in figure 4.22, and while they appeared nearly identical at first glance, further analysis revealed differences. When SHAP metrics—mean, variance, and skewness—were

calculated and plotted (see figure 4.23) separately for weekday and weekend datasets, the Airline feature emerged as a significant differentiator. These differences were statistically confirmed using both T-tests and KS-tests.


*Figure 4.23: A plot of SHAP metrics for Weekday and Weekend datasets measured for Random Forest*

Finally, Model Class Reliance (MCR) bounds are calculated and plotted for both datasets (see figure 4.24). Although slight shifts in the MCR bounds are observed, statistical tests did not reveal any major differences between weekday and weekend datasets. This suggests that while certain features like Airline and Time exhibit significant changes in their importance, the overall model reliance remains relatively stable across these temporal segments.

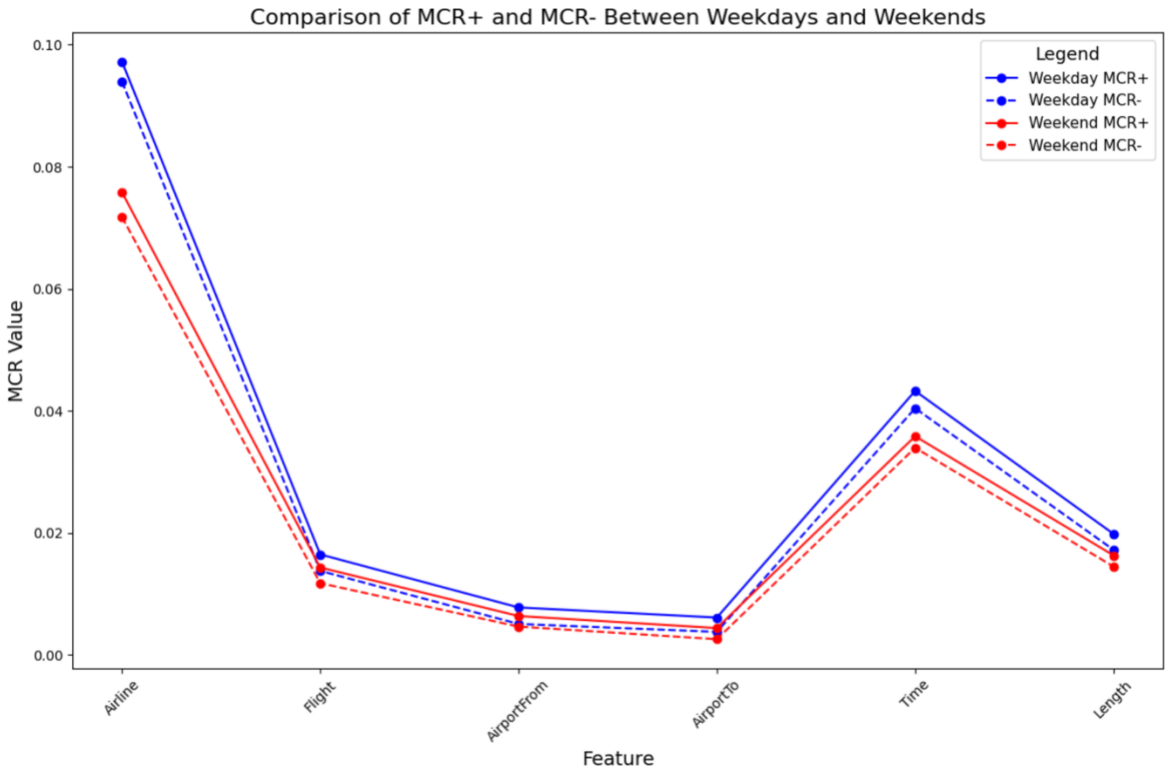


*Figure 4.24: A plot of MCR bounds for Weekday and Weekend datasets*

### 4.2.5 Implementing Neural Network

The third and final model implemented to detect concept drift in the Airlines dataset is a Neural Network. Neural networks are known for their ability to capture complex, non-linear relationships between features, making them a valuable tool in tasks where other models might struggle (Setiyorini and Frieyadie, 2020). The neural network model is trained on the weekday data, assumed to represent the pre-drift period, using a basic architecture suitable for binary classification (two layers with Relu activation functions and an output layer with Sigmoid activation function).

After training the Neural Network model on the weekday dataset, the model achieved an overall accuracy of 59%. When this trained model was tested on the weekend dataset, the overall accuracy remained unchanged at 59%.

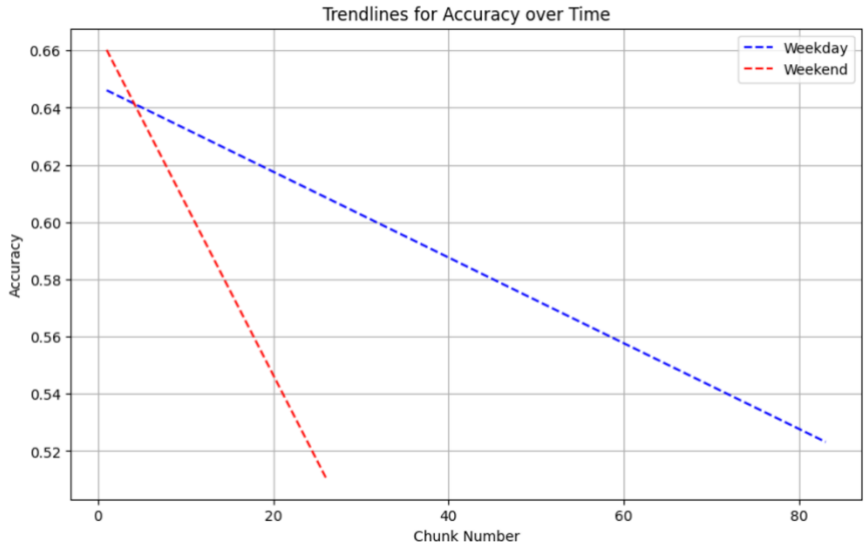


*Figure 4.25: Accuracy scores across subsets of weekday and weekend datasets*

However, when the accuracy plots (see figure 4.25) were examined across different data subsets, a declining trendline was observed for both the weekday and weekend datasets.

To explore potential drift more deeply, permutation importance scores are calculated for both the weekday and weekend datasets, using temporal windows of 5,000 records each. The permutation importance analysis

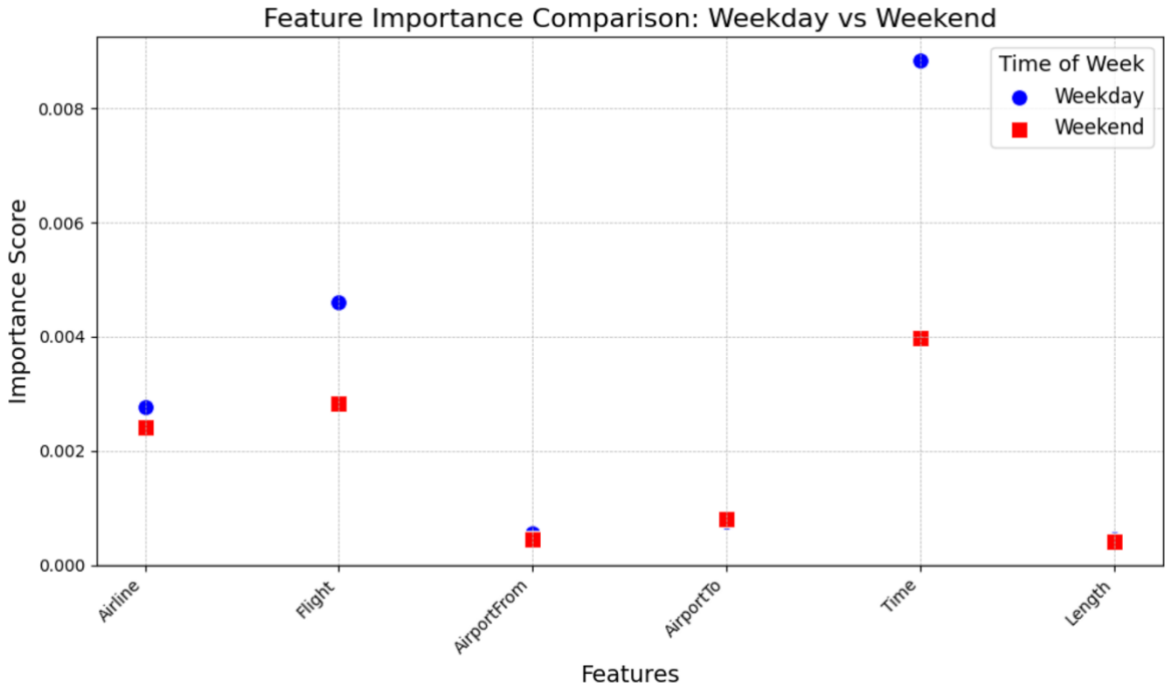


*Figure 4.26: Plot of Permutation Importance scores for Weekday and Weekend datasets measured for Neural Network*

reveals a clear shift in the average importance levels of certain features, specifically Time and Flight, when comparing the weekday to weekend data (see figure 4.26). These shifts are confirmed by statistical tests, including the t-test and Kolmogorov-Smirnov (KS) test, which indicate that there is a statistically significant difference in the permutation importance scores between the two datasets.

Then, SHAP (SHapley Additive exPlanations) values are used to examine the contribution of features during both periods. The SHAP summary plots for the neural network revealed not only shifts in the mean SHAP values but also different shapes in the SHAP value distributions between the weekday and weekend data (see figure 4.27). This suggests that the relationships between the features and the prediction outcome had changed more noticeably for the neural network, likely due to its ability to capture complex interactions.

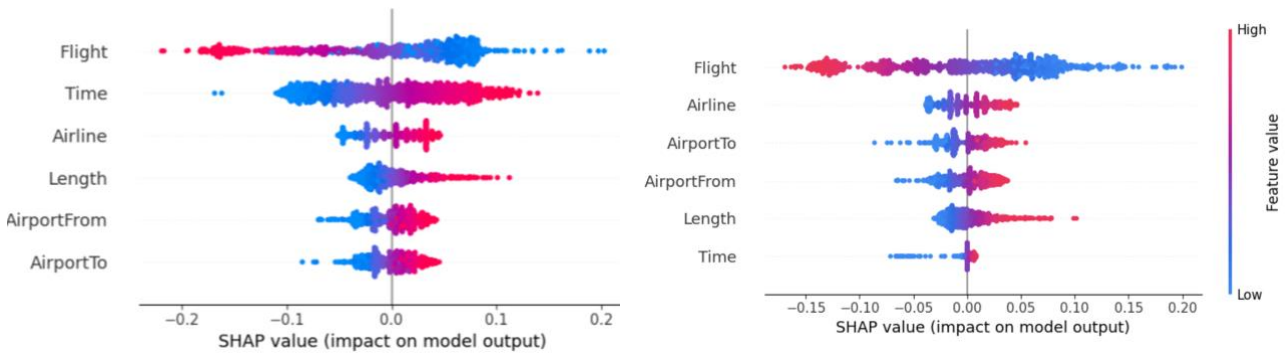


*Figure 4.27: SHAP Summary plots for Weekday (left) and Weekend (right) datasets measured for Neural Network*

To systematically analyse these changes, SHAP metrics - mean, variance, and skewness are plotted for both the weekday and weekend datasets (see figure 4.28). The plots showed clear shifts in the mean, variance, and skewness of SHAP values for the Flight, Time, and Airline features, reinforcing the evidence of concept drift. Statistical tests (using KS Test) confirmed that these shifts are statistically significant, indicating a

difference in how the neural network utilized these features between the two periods.


*Figure 4.28: A plot of SHAP metrics for Weekday and Weekend datasets measured for Neural Network*

# 5. Results and Discussion

In this chapter, the results and evaluations from the implementation of the proposed methodology, as outlined in the previous chapter, will be discussed in detail. The chapter is divided into two sections. The first section focuses on the application of the proposed methodology to the Electricity dataset, where the presence of sudden/abrupt concept drift and its implications are analysed. The second section examines the implementation of the methodology on the Airlines dataset, with particular attention to detecting potential drift between weekday and weekend data (in an attempt to detect cyclic drift). Each section will explore the outcomes of the implemented models, providing insights into the effectiveness of the approach and its ability to identify and interpret concept drift across different datasets.

## 5.1 Discussion on Detecting Abrupt Concept Drift

The implementation of the proposed methodology on the Electricity dataset is aimed to detect sudden/abrupt concept drift associated with the introduction of the National Electricity Market (NEM). The dataset was divided into pre-drift and post-drift periods, with the first 10 months used for training and the subsequent 10 months, covering the period around the introduction of NEM, used for testing.

The results across all three models used in this study consistently showed a degradation in performance when the models trained on the pre-drift period were tested on the post-drift period. Logistic Regression, for instance, demonstrated a drop in accuracy from 79% during training to 70% during testing. Similarly, the Random Forest model's accuracy declined from 84% to 68%, and the Neural Network model saw a decrease from 82% to 66%.

Beyond accuracy, as per the proposed approach, an investigation into the models' performance was conducted through the analysis of permutation importance scores and SHAP values. Permutation importance scores across

all models indicated a downward shift in the importance of the NSW Price variable starting from the 12th month, which aligns with the introduction of NEM. This trend was particularly evident in Random Forest and Neural Network models. Although T-tests and KS-tests did not always confirm these shifts as statistically significant, the visual plots provided some evidence of change.

SHAP summary plots offered additional insights, revealing shifts in the shapes and distributions of SHAP values between the training and testing periods. These shifts were particularly notable in the NSW Price and Demand features, which were critical variables in the models. The SHAP metrics (mean, variance, and skewness) were significantly different between the pre- and post-drift periods, with statistical tests confirming these differences. This was true across all models, indicating that the introduction of NEM had a measurable impact on the relationships between features and the target variable.

Model Class Reliance (MCR) bounds were also calculated to assess the changes in model dependence on specific features. While the bounds showed shifts between the training and testing periods, particularly for the NSW Price feature, these changes were not statistically significant.

In summary, the implementation of the proposed methodology on the Electricity dataset provides practitioners with a flexible approach to determining whether to retrain and reconsider their models. While a decreasing accuracy across the testing period may initially suggest the presence of drift, it may not always be a definitive indicator (Gama et al., 2014). By analysing changes in the relationships between input and output variables, as evidenced by the statistically significant differences in SHAP values and distributions, researchers gain a deeper understanding of the underlying shifts. As these shifts passed (in terms of being significantly different) the statistical tests in all the models used for this study, it can be said this analysis was model independent. This also allows practitioners to

make informed decisions on whether these changes warrant model retraining or further investigation, rather than relying solely on accuracy metrics.

### 5.2 Discussion on Detecting Cyclic Concept Drift

The proposed methodology was applied to the Airlines dataset to detect cyclic concept drift between weekday and weekend flight data. The dataset's division into weekday and weekend subsets allowed the investigation of whether the semi-synthetic segmentation would reveal shifts in feature importance and model performance, indicative of drift.

Initially, all the three models were trained on the weekday data, which was assumed as the pre-drift period. When tested on the weekend dataset, the model's overall accuracy did not very much for both subsets. It was stable at 58% for Logistic Regression, dropped by 2% for Random Forest and remained 59% for Neural Networks. Therefore, the accuracy alone couldn't provide enough information to the practitioner regarding the existence of presumed drift. However, the accuracy trendline for the weekend data exhibited a noticeable decline compared to the more stable trend observed in the weekday data.

When Permutation Importance scores were calculated across all models, shifts in feature importance were observed, particularly for the Flight, Time, and Length features. However, statistical tests (t-test and Kolmogorov-Smirnov (KS) test) did not consistently confirm these shifts as significant. Only Permutation Importance scores from Random Forest model were confirmed to be statistically different for Time and Airline features.

Then, as per the proposed approach, SHAP values were computed for both weekday and weekend datasets across all models. The SHAP summary plots revealed subtle but important shifts in the distribution of SHAP values, particularly for critical features like Airline, Airport From, and Airport To. While these shifts were visually apparent, statistical tests (T-test and KS-

test) provided more concrete evidence, confirming significant changes in the mean, variance, and skewness of SHAP metrics for these features. The Neural Network model, known for capturing complex non-linear relationships, showed more pronounced shifts in SHAP value distributions, indicating a greater sensitivity to the cyclic nature of the data.

In addition to SHAP analysis, Model Class Reliance (MCR) bounds were calculated to assess the stability of feature importance across the different temporal segments. Although slight shifts were observed in MCR bounds, particularly for the Airline and Time features, these changes were not statistically significant. This suggests that while feature importance varied between weekday and weekend data, the overall model reliance remained relatively stable across these periods.

In summary, the application of the proposed methodology on the Airlines dataset demonstrated that concept drift, particularly cyclic drift between weekdays and weekends, could be detected through a combination of permutation importance, and SHAP value examination. While accuracy alone did not always reflect the presence of drift, the shifts in feature importance and SHAP metrics provided a deeper understanding of the underlying changes in data patterns. This comprehensive approach allows practitioners to identify and interpret drift more effectively, informing decisions on whether model retraining or further investigation is necessary.

# 6. Conclusion

This study aimed at addressing the challenge of concept drift detection, which can lead to model decay, inaccurate predictions, and increased maintenance costs. The primary objective was to develop a model-agnostic approach for detecting concept drift using techniques such as permutation importance, SHAP values, and Model Class Reliance (MCR) bounds. The research attempted to establish a reliable framework that could identify drift independently of the specific machine learning model and explain changes in feature-prediction relationships after drift detection.

## 6.1 Achievement of Objectives

**Utilization of SHAP Values and Feature Importance Scores**: The research validated that changes in SHAP values and feature importance scores are critical indicators of concept drift. Across all models and datasets, shifts in SHAP values and permutation importance scores provided deeper insights into the nature of the drift. These shifts were statistically significant in most cases, particularly for key features like `NSW Price` in the Electricity dataset and `Airline` in the Airlines dataset. This confirms the utility of SHAP values and feature importance as reliable markers for detecting and understanding drift.

**Model-Agnostic Approach**: The study aimed to create a method independent of the specific machine learning model being used. The methodology proved to be model-agnostic, as it was applied successfully across different models - Logistic Regression, Random Forest, and Neural Networks. This versatility makes the framework applicable to a wide range of models and scenarios.

**Interpretation of Changes in Feature-Prediction Relationships**: Another goal was to explain the changes in relationships between features and the prediction variable after detecting drift. By analysing SHAP values and MCR bounds, the research provided some interpretations of how the predictive relationships within the models changed over time. This

interpretability can be important for practitioners, enabling them to understand the implications of drift and make informed decisions about model retraining or adjustments.

### 6.2 Limitations of the Approach

Despite the promising results, there are several limitations to the proposed methodology:

**Statistical Significance of Permutation Importance**: The methodology was tested on datasets with limited temporal windows. Due to this, detecting statistically significant changes in permutation importance scores was challenging. This may be particularly relevant in scenarios where drift is gradual or where data is segmented into smaller temporal intervals. While visual and summary statistics indicated drift, the statistical significance of these changes was not consistently confirmed, suggesting that additional data or refined statistical methods may be required.

**Inaccuracy of Temporal Segmentation in the Airlines Dataset**: Although the Electricity dataset emulated a real-world scenario, the analysis of cyclic drift in the Airlines dataset (using weekday and weekend data) did not represent a fully accurate real-world application. The segmentation of data into weekday and weekend subsets was temporally incorrect because it did not align with actual flow of time. As a result, the analysis may not have captured genuine cyclic patterns, limiting the practical applicability of the findings. Practitioners should take care to ensure accurate temporal segmentation when applying this methodology to real-world datasets.

**Inability to Detect Gradual Drift**: Another limitation of the approach is its potential difficulty in detecting gradual drift. Gradual drift, where changes in data distribution or relationships evolve slowly over time, may not result in significant shifts in permutation importance, SHAP values, or MCR bounds within the short temporal windows used in this study. This

type of drift might require more extended observation periods and finer temporal granularity to be detected, which the current methodology may not fully capture.

### 6.3 Practical Implications

Practitioners can leverage this methodology to gain a comprehensive understanding of concept drift, even in scenarios where accuracy does not significantly drop. By examining shifts in SHAP values, feature importance, and MCR bounds, practitioners can detect subtle forms of drift that might not be immediately evident through traditional metrics. This approach allows for more proactive management of machine learning models, enabling timely interventions such as model retraining or adjustment before drift leads to significant performance degradation.

### 6.4 Summary and Future Work

In conclusion, this study has partially fulfilled its objectives by developing and validating a model-agnostic approach for detecting and interpreting concept drift. The methodology proved effective across different datasets and models, offering a comprehensive framework for early drift detection. The reliability of the model, however, was debatable as there were inconsistencies found with statistical differences in permutation importance scores across models. Additionally the reliance on limited temporal windows, the inaccuracies in temporal segmentation, and the challenge of detecting gradual drift were identified.

Future work should focus on addressing these limitations. Enhancing the statistical validation of drift detection by employing larger datasets with longer and more accurate temporal windows might improve the robustness of the methodology. Additionally, refining the methodology to better detect gradual drift could involve incorporating more sophisticated techniques or machine learning models that are sensitive to slow-evolving patterns. Further, exploring automated methods for accurate temporal segmentation could improve the application of this methodology in real-world scenarios,

particularly in dynamic environments where data patterns change over time. Finally, expanding the methodology to include additional datasets from various domains would help generalize the approach, making it more versatile and applicable to a broader range of industries and applications.

# 7. References

D. Brzezinski and J. Stefanowski, "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm," in IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 1, pp. 81-94, Jan. 2014, doi: 10.1109/TNNLS.2013.2251352.

Hinder, F., Jakob, J. and Hammer, B. (2020). Analysis of Drifting Features. Available at: https://arxiv.org/pdf/2012.00499

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. and Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), pp.1–37. doi:https://doi.org/10.1145/2523813.

Hammoodi, M. S., Stahl, F., & Badii, A. (2018). Real-time feature selection technique with concept drift detection using adaptive micro-clusters for data stream mining. Knowledge-Based Systems, 161, 205-239. https://doi.org/10.1016/j.knosys.2018.08.007.

Shao, J., Ahmadi, Z., & Krämer, S. (2014). Prototype-based learning on concept-drifting data streams. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2623330.2623609.

Zhu, Y., Lü, Z., Chen, X., Li, Y., & Wang, J. (2023). Identification of cashmere and wool based on lbp and glcm texture feature selection. Journal of Engineered Fibers and Fabrics, 18, 155892502211465. https://doi.org/10.1177/15589250221146548.

Praveen, M. V. R., kuchhal, P., & Choudhury, S. (2024). Novel statistical method for data drift detection in satellite telemetry. International Journal of Communication Systems, 37(9). https://doi.org/10.1002/dac.5766.

Ditzler, G. and Polikar, R. (2011). Hellinger distance based drift detection for nonstationary environments. 2011 IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments (CIDUE). https://doi.org/10.1109/cidue.2011.5948491.

Yuan, Z., Sun, Y., & Shasha, D. (2023). Forgetful forests: data structures for machine learning on streaming data under concept drift. Algorithms, 16(6), 278. https://doi.org/10.3390/a16060278.

Baier, L., Schlör, T., Schöffer, J., & Kühl, N. (2021). Detecting concept drift with neural network model uncertainty. https://doi.org/10.48550/arxiv.2107.01873.

Klinkenberg, Ralf & Joachims, Thorsten. (2000). Detecting Concept Drift with Support Vector Machines. Proceedings of ICML.

Gary, S. F., Scheibe, T. D., Rexer, E., Torreira, A. V., Garayburu-Caruso, V. A., Goldman, A. E., … & Stegen, J. C. (2024). Prediction of distributed river sediment respiration rates using community-generated data and machine learning. https://doi.org/10.22541/essoar.171136943.37512936/v1.

Meng, L., Treem, W., Heap, G. A., & Chen, J. (2022). A stacking ensemble machine learning model to predict alpha-1 antitrypsin deficiency-associated liver disease clinical outcomes based on uk biobank data. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-21389-9.

Lundberg, S., Allen, P. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. [online] Available at: https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Fisher, A., Rudin, C. and Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. [online] Available at: https://arxiv.org/pdf/1801.01489.

Smith, G., Mansilla, R. and Goulding, J. (2020). Model Class Reliance for Random Forests. [online] Available at: https://proceedings.neurips.cc/paper_files/paper/2020/file/fd512441a1a791770a6fa573d688bff5-Paper.pdf.

Harries, Michael & Nsw-cse-tr, U & Wales, New. (2003). SPLICE-2 Comparative Evaluation: Electricity Pricing.

(2017) MOA dataset repository. URL http://moa.cms.waikato.ac.nz/datasets/.

Setiyorini, T. and Frieyadie, F. (2020). Comparison of linear regressions and neural networks for forecasting electricity consumption. Jurnal Pilar Nusa Mandiri, 16(2), 135-140. https://doi.org/10.33480/pilar.v16i2.1459.

## 8. Appendix

Below are the links for python scripts:

Detecting abrupt concept drift for Electricity dataset – https://drive.google.com/file/d/1b_a4mFMxj5PpriRMPM_oXL9oOM06-DUE/view?usp=sharing

Detecting cyclic concept drift for Airlines dataset – https://drive.google.com/file/d/1tYW8_jjnVEzEvJ0OAWK1HQ9wtvR5uXLb/view?usp=sharing

Airlines Abbreviations used in the dataset –

Alaska Airlines AS / ASA
American Airlines AA/AAL
Air Canada AC/ACA
Aeromexico AM / AMX
Continental Airlines CO / COA
Delta Airlines DL / DAL
FedEx FX / FDX
Hawaiian Airlines HA / HAL
Northwest Airlines NW / NWA
Polar Air Cargo PO / PAC
Southwest Airlines SW / SWA
United Airlines UA / UAL
United Parcel (UPS) 5X / UPS
Virgin Atlantic VS / VIR
VivaAerobús VB / VIV
WestJet WS / WJ

ATL - Hartsfield-Jackson Atlanta International Airport - Georgia
AUS - Austin-Bergstrom International Airport - Texas
BNA - Nashville International Airport - Tennessee

BOS - Boston Logan International Airport - Massachusetts

BWI - Baltimore-Washington International Thurgood Marshall Airport - Washington

CLT - Charlotte Douglas International Airport - North Carolina

DAL - Dallas Love Field - Texas

DCA - Ronald Reagan Washington National Airport - Arlington, Virginia

DEN - Denver International Airport - Colorado

DFW - Dallas/Fort Worth International Airport - Texas

DTW - Detroit Metropolitan Airport - Michigan

EWR - Newark Liberty International Airport - New Jersey

FLL - Fort Lauderdale–Hollywood International Airport - Florida

HNL - Daniel K. Inouye International Airport - Honolulu, Hawaii

HOU - William P. Hobby Airport - Houston, Texas

IAD - Dulles International Airport - Virginia

IAH - George Bush Intercontinental Airport - Houston, Texas

JFK - John F. Kennedy International Airport - Queens, New York

LAS - McCarran International Airport - Las Vegas, Nevada

LAX - Los Angeles International Airport - California

LGA - LaGuardia Airport - Queens, New York

MCO - Orlando International Airport - Florida

MDW - Chicago Midway International Airport - Illinois

MIA - Miami International Airport - Florida

MSP - Minneapolis–Saint Paul International Airport - Minnesota

MSY - Louis Armstrong New Orleans International Airport - Louisiana

OAK - Oakland International Airport - California

ORD - O'Hare International Airport - Chicago, Illinois

PDX - Portland International Airport - Oregon

PHL - Philadelphia International Airport - Pennsylvania

PHX - Phoenix Sky Harbor International Airport - Arizona

RDU - Raleigh-Durham International Airport - North Carolina

SAN - San Diego International Airport - California

SEA - Seattle–Tacoma International Airport - Washington

SFO - San Francisco International Airport - California

SJC - Norman Y. Mineta San Jose International Airport - California

SLC - Salt Lake City International Airport - Utah

SMF - Sacramento International Airport - California

STL - St. Louis Lambert International Airport - Missouri

TPA - Tampa International Airport - Florida