Master's study:

Field of study Big Data - Advanced Analytics

Bahram Muzaffarli
Student’s register No. 124131

# Forecasting Sales Revenue Using LSTM Models

Master's thesis:

under the scientific supervision of

dr hab. Michał Bernardelli, prof. SGH

Institute of Econometrics

Warsaw 2024

# Table of Contents

# Chapter 1. Introduction

Revenue forecasting is a critical aspect of business strategy in retail, as it allows companies to plan for future operations, optimize resource allocation, and meet market demand. In an increasingly competitive environment, accurate predictions of future revenue can help businesses like Kontakt Home stay ahead by making informed decisions based on data-driven insights.

This thesis focuses on building a predictive model for Kontakt Home's 2020 revenue, using historical data from 2018 and 2019. The aim is to simulate predictions made at the end of 2019, considering a situation where external disruptions such as COVID-19 did not occur. This context provides a unique opportunity to explore the accuracy and effectiveness of statistical and machine learning models, particularly time series forecasting techniques like LSTM (Long Short-Term Memory networks), in generating reliable revenue forecasts.

The primary objective of this thesis is to develop and evaluate a model that predicts 2020 revenue based on historical patterns, while also exploring the impact of key variables such as product categories and store performance. By focusing on statistical and data analytics techniques, this study aims to contribute to the growing body of knowledge on revenue prediction models in retail, highlighting their practical applications and limitations.

## 1.1 Overview of revenue forecasting in retail

Revenue forecasting plays a pivotal role in retail business operations, providing insights into future financial performance and helping companies make strategic decisions. Accurate revenue predictions enable businesses to plan budgets, optimize inventory management, and allocate resources efficiently. In the highly competitive retail sector, this can mean the difference between thriving and struggling to maintain market share.

Kontakt Home is a leading retailer in Azerbaijan, specializing in the sale of home appliances, electronics, and mobile devices. Founded to serve the growing consumer demand for high-quality electronics, Kontakt Home operates across several cities, providing a wide range of products through physical stores and its online platform (Kontakt Home, 2022). The company's business model includes installment payment plans, which make its products more accessible to customers, a feature that has contributed to its rapid expansion in the Azerbaijani market.

Given its market presence, Kontakt Home requires sophisticated tools for forecasting revenue. Predicting revenue accurately allows the company to anticipate changes in customer demand, adjust marketing strategies, and enhance overall business performance. This thesis focuses on the use of advanced statistical and machine learning models—specifically Long Short-Term Memory (LSTM) networks—for predicting revenue in 2020, based on historical data from 2018 and 2019. The absence of external disruptions like COVID-19 in this prediction scenario allows for a focused analysis of underlying sales trends, seasonality, and product demand, offering valuable insights for decision-making.

Incorporating time series forecasting models in this research allows for a deeper understanding of Kontakt Home's historical revenue patterns and helps simulate future outcomes. As data-driven decision-making continues to shape the retail industry, this study will demonstrate how machine learning models can be leveraged to provide reliable and actionable revenue forecasts.

### 1.2 Problem statement: predicting 2020 revenue for Kontakt Home

As businesses increasingly rely on data-driven insights for decision-making, the need for accurate revenue forecasting becomes more critical. The retail industry, in particular, faces constant challenges due to fluctuations in consumer demand, seasonal trends, and competitive market pressures. For a company like Kontakt Home, forecasting revenue is vital for strategic planning, inventory management, and financial stability. Predicting future revenue not only supports resource allocation but also helps the company optimize marketing efforts and anticipate customer needs.

The primary challenge addressed in this thesis is the creation of a predictive model that estimates Kontakt Home's 2020 revenue based on historical data from 2018 and 2019. In this scenario, we assume we are at the end of 2019 and are tasked with making revenue predictions for the upcoming year. Importantly, the model does not account for external disruptions like COVID-19, allowing us to focus on patterns and trends inherent in the data itself.

Developing a reliable model for revenue forecasting involves several challenges, including identifying relevant variables, selecting the right machine learning algorithm, and ensuring the accuracy of predictions. This thesis addresses these issues by employing time series forecasting, specifically using Long Short-Term Memory (LSTM) networks, which have proven effective in capturing patterns in sequential data (Brown & Smith, 2019). Additionally, we explore various evaluation metrics like RMSE and MAE to assess the model’s performance.

By accurately forecasting revenue, Kontakt Home can make data-driven decisions that improve operational efficiency and enhance competitiveness in the Azerbaijani retail market. This thesis thus aims to develop and evaluate a model that not only predicts 2020 revenue but also demonstrates the effectiveness of time series forecasting techniques in retail applications.

### 1.3 Research objectives and hypotheses

The primary objective of this thesis is to develop an accurate revenue prediction model for Kontakt Home using historical data from 2018 and 2019. By leveraging time series forecasting methods, particularly Long Short-Term Memory (LSTM) networks, the aim is to simulate what the company's revenue would have been in 2020 under normal conditions, excluding the impacts of external disruptions like COVID-19.

This research seeks to explore how machine learning models can be applied in the retail industry to provide actionable insights for decision-making. The model will be evaluated using common performance metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to ensure robustness and accuracy.

The specific research objectives are as follows:

1. To develop a predictive model that accurately forecasts Kontakt Home’s 2020 revenue based on historical data.
2. To evaluate the performance of LSTM models in revenue forecasting by comparing predicted outcomes with actual 2020 revenue.
3. To analyze revenue trends and patterns in the historical data, identifying key factors such as seasonality, product category performance, and store-level variations.

Based on these objectives, the following hypotheses are formulated:

Hypotheses 1: The LSTM model can effectively predict Kontakt Home’s 2020 revenue with a high degree of accuracy when trained on 2018 and 2019 data.

Hypotheses 2: Revenue trends such as seasonality and store performance will be significant predictors of future revenue.

Hypotheses 3: The model’s accuracy will be influenced by the granularity of data used (e.g., store-level vs. aggregate revenue data).

## 1.4 Structure of the thesis

This thesis is structured to provide a comprehensive analysis of revenue forecasting for Kontakt Home, with a focus on time series forecasting techniques. The chapters are organized to guide the reader through the development, implementation, and evaluation of predictive models, while highlighting key trends and insights drawn from the data.

**Chapter 1: Introduction** – This chapter introduces the research topic, outlines the problem statement, presents the research objectives and hypotheses, and provides an overview of the thesis structure.

**Chapter 2: Revenue prediction in retail** – This chapter reviews the literature on revenue forecasting models, including statistical and machine learning approaches. It also discusses the challenges and opportunities in applying predictive analytics in retail.

**Chapter 3: Methodology** – This chapter presents the methods used for the predictive modeling, with a focus on time series analysis and LSTM models. It details the evaluation metrics used, such as RMSE and MAE, and outlines the assumptions and limitations of the study.

**Chapter 4: Data characteristics** – This chapter provides an overview of the dataset, including the historical revenue data from 2018, 2019, and 2020. It discusses data preprocessing techniques, feature engineering, and key variables used in the analysis.

**Chapter 5: Empirical analysis** – This chapter applies the LSTM models to the dataset and evaluates their performance. It includes a time series analysis of historical revenue trends and compares the predicted 2020 revenue with the actual results.

**Chapter 6: Conclusion** – This chapter summarizes the findings, discusses the advantages and disadvantages of the approach, and offers suggestions for future research and practical applications for Kontakt Home.

**References** – A comprehensive list of all sources and studies referenced throughout the thesis, formatted in APA style.

**Appendices** – Supplementary materials, including detailed descriptions of variables, data tables, and additional visualizations, are provided in the appendices.

# Chapter 2. Revenue prediction in retail

Revenue prediction is a critical component of business planning in the retail industry. Accurately forecasting future revenue helps retailers optimize inventory, manage resources, and plan marketing campaigns. In the competitive retail landscape, the ability to anticipate customer demand and adjust accordingly can provide a significant edge.

Modern approaches to revenue prediction have evolved considerably, leveraging advanced statistical techniques and machine learning models to improve accuracy and adaptability. Traditional methods like linear regression, while useful for understanding general trends, often fall short when dealing with complex, dynamic retail environments. As such, many retail companies now turn to machine learning models such as **Long Short-Term Memory (LSTM) networks**, which are particularly well-suited for handling time series data and predicting future trends based on historical data (Brown, 2019).

In the context of retail, revenue prediction involves understanding various factors that influence sales, such as seasonal patterns, promotional activities, and changes in customer behavior. Machine learning models, in particular, excel at identifying these underlying patterns and trends. **Predictive analytics** techniques help businesses like **Kontakt Home** make data-driven decisions, allowing for proactive rather than reactive strategies (Kumar & Garg, 2018).

A major challenge in retail revenue prediction is the inherent volatility of the market. Retailers must account for fluctuations in demand caused by seasonal variations, holidays, and economic shifts. Furthermore, the retail industry often deals with a wide range of products, each with its own sales cycle and performance metrics. As a result, any predictive model must be able to handle multiple variables and adapt to changes in consumer preferences.

In this chapter, it explores the various approaches to revenue prediction used in retail, with a specific focus on machine learning and time series forecasting models. It will review relevant literature, discussing the benefits and limitations of these models, and highlight how they can be applied to retail revenue forecasting to improve accuracy and provide actionable insights.

## 2.1 Literature review on revenue prediction models

The evolution of revenue prediction models in the retail industry has seen a shift from traditional statistical methods to more complex machine learning techniques. As retail environments grow more dynamic, the need for accurate and reliable revenue forecasting models has become critical.

This section reviews the key contributions to the field, focusing on the most prominent revenue prediction models, their methodologies, and their applications in retail.

**Traditional statistical models:**

Statistical approaches to revenue forecasting have been widely used for decades. Linear regression models, for instance, have been a cornerstone in predicting revenue based on independent variables such as pricing, promotions, and customer demographics. These models are straightforward to implement and interpret but may struggle when dealing with complex, non-linear relationships typical in retail environments (Mccullagh, 2002).

**Time series models**, such as **ARIMA (AutoRegressive Integrated Moving Average)**, have also been extensively studied and applied in revenue prediction. ARIMA models are particularly useful for identifying trends and seasonality in historical sales data (Box et al., 2015). However, ARIMA assumes linearity and stationarity, limiting its effectiveness in situations where revenue is influenced by multiple variables and non-linear interactions.

Another common statistical method is **exponential smoothing**, which provides forecasts based on weighted averages of past observations. While exponential smoothing is simple and effective for short-term forecasting, it can struggle with more complex time series data that exhibits significant seasonal or cyclical behavior (Hyndman & Athanasopoulos, 2018).

**Machine Learning models:**

The increasing availability of large datasets and computational power has led to a growing interest in machine learning (ML) models for revenue prediction. Machine learning models, unlike traditional statistical methods, can capture non-linear relationships and handle a wider variety of input features, making them highly adaptable to complex retail environments (Bishop, 2016).

**Decision trees** and **random forests** have gained traction due to their ability to manage high-dimensional data and model complex interactions between variables. Random forests, in particular, have been praised for their robustness in handling noisy data and minimizing overfitting (Breiman, 2001). However, while these models are powerful, they can sometimes lack the transparency required for business decision-making.

One of the most promising advancements in revenue prediction is the use of **neural networks**, particularly **Long Short-Term Memory (LSTM)** networks, which excel at capturing long-term dependencies in time series data (Hochreiter & Schmidhuber, 1997). LSTM models have been applied successfully in retail forecasting, where they can learn complex patterns in sequential data and produce more accurate revenue predictions than traditional time series models (Brown, 2019). Their ability to model seasonality, promotional effects, and other external factors makes them highly effective in predicting future revenue.

In addition to LSTM, other neural network-based models such as **convolutional neural networks (CNNs)** and **recurrent neural networks (RNNs)** have been explored for revenue prediction. These models are especially effective when integrated with large-scale retail datasets that include multiple variables, such as product categories, customer demographics, and promotional campaigns (Lazcano et al., 2023).

Furthermore, **ensemble learning methods**, such as **Gradient Boosting Machines (GBM)** and **XGBoost**, have shown significant promise in revenue prediction tasks. These models iteratively build on weak learners to improve accuracy and can effectively capture non-linear patterns in the data (Chen & Guestrin, 2016). Studies have demonstrated that ensemble methods outperform traditional statistical models in revenue forecasting, particularly in scenarios with large datasets and multiple features (Friedman, 2002).

**Challenges and opportunities:**

Despite the advances in machine learning models, there remain challenges in their application. A key issue is the **interpretability** of these models. While traditional models like linear regression and ARIMA offer clear insights into the relationships between variables, complex machine learning models often function as "black boxes," making it difficult for business leaders to understand how predictions are made (Shmueli et al., 2018). As such, there is ongoing research into **explainable AI (XAI)** methods that aim to improve transparency in machine learning predictions.

Additionally, the **quality and availability of data** play a crucial role in the accuracy of revenue predictions. Machine learning models require large, clean datasets to function effectively, and missing or incorrect data can significantly reduce their performance. For retail companies like **Kontakt Home**, ensuring data quality is a critical step in building reliable predictive models.

Overall, the literature highlights the growing importance of machine learning in revenue forecasting, particularly in the retail sector. While traditional statistical models remain valuable for their simplicity and interpretability, advanced machine learning techniques like LSTM and gradient boosting offer superior accuracy and flexibility, making them ideal for modern retail environments.

## 2.2 Overview of predictive analytics in retail: challenges and opportunities

Predictive analytics has transformed the retail industry, enabling companies to leverage vast amounts of historical data to forecast future revenue, optimize operations, and make informed business decisions. Predictive models use machine learning algorithms to identify patterns and trends in consumer behavior, sales data, and external market conditions. These insights help retailers better understand their customer base, anticipate market demands, and ultimately improve business performance.

### Challenges in predictive analytics:

While predictive analytics offers numerous benefits, its application in retail is not without challenges. One of the most significant challenges is **data quality and availability**. Predictive models rely heavily on large datasets that must be clean, accurate, and complete. Missing or incorrect data can skew predictions, leading to poor decision-making (Brown et al., 2019). Furthermore, retail data can be highly complex, involving multiple variables such as customer demographics, product categories, seasonal trends, and marketing efforts, making it difficult to manage and analyze effectively.

Another challenge is the **non-stationarity** of retail data. Consumer preferences, market conditions, and external factors such as economic shifts or unexpected events can cause abrupt changes in revenue patterns. This makes it difficult for traditional models to adapt to new trends, often leading to less accurate predictions. To address this, machine learning models like LSTM networks have been used to account for non-linear relationships and temporal dependencies in retail data (Hochreiter & Schmidhuber, 1997).

**Overfitting** is another key challenge in predictive analytics. Overfitting occurs when a model becomes too complex, fitting the noise in the training data rather than capturing the underlying patterns. This leads to poor generalization on unseen data, reducing the accuracy of future predictions. Retail data is particularly prone to overfitting due to its volatility and the presence of numerous external factors that can affect sales (Friedman, 2002).

Finally, the **interpretability** of machine learning models can be a concern for retail businesses. While complex models like LSTM or gradient boosting provide accurate predictions, they often act as "black boxes," making it difficult for business stakeholders to understand how the model arrived at a specific prediction. This lack of transparency can make it harder for companies to trust and act on the insights generated by predictive analytics (Shmueli et al., 2018).

**Opportunities in predictive analytics:**

Despite these challenges, predictive analytics presents a wealth of opportunities for the retail industry. One of the most significant opportunities lies in **personalization**. Predictive models can analyze customer behavior and preferences to deliver personalized marketing campaigns, product recommendations, and targeted promotions. Retailers like **Amazon** and **Netflix** have successfully used predictive analytics to offer personalized experiences, leading to increased customer satisfaction and loyalty.

Another key opportunity is the ability to **optimize inventory management**. Predictive analytics can help retailers forecast demand for specific products, allowing them to adjust their inventory levels accordingly. This reduces the likelihood of stockouts or overstocking, improving overall supply chain efficiency and minimizing costs. Retailers like **Kontakt Home** can leverage this capability to better manage their product categories and ensure that high-demand items are always available.

Predictive analytics also enables **dynamic pricing** strategies, where retailers can adjust prices in real time based on demand, competition, and other external factors. This allows businesses to maximize revenue by pricing products optimally for different customer segments and market conditions (Bishop, 2016).

Furthermore, predictive models can assist with **customer segmentation**, identifying distinct groups within the customer base based on purchasing behavior, demographics, and preferences. This enables retailers to tailor their marketing strategies and improve customer retention by offering targeted promotions and services to each segment (Levin & Zahavi, 1999).

Overall, while predictive analytics in retail faces challenges such as data quality, overfitting, and model interpretability, the potential benefits are substantial. By harnessing the power of predictive analytics, retailers can gain a competitive edge through personalized customer experiences, optimized operations, and improved decision-making.

### 2.3 Use of predictive models in dynamic conditions

The retail industry operates in a constantly changing environment where consumer preferences, market trends, and external factors can fluctuate rapidly. Predictive models, especially those built on machine learning techniques, are well-suited for navigating these dynamic conditions. The ability of machine learning models to adapt to non-linear relationships and multiple influencing factors makes them invaluable tools for forecasting revenue in such volatile environments (Bishop, 2016).

**Handling seasonality and promotions:**

Retail businesses often experience cyclical patterns in revenue due to seasonality and promotional events. For example, sales tend to spike during holiday seasons, sales events, or new product launches. Traditional statistical models like ARIMA can handle some seasonal effects, but they are often limited when faced with the intricate patterns found in retail, particularly in cases where multiple factors overlap (Box et al., 2015). Machine learning models, especially **Long Short-Term Memory (LSTM) networks**, have the ability to learn and predict these patterns effectively by capturing long-term dependencies and temporal sequences in data (Hochreiter & Schmidhuber, 1997).

Predictive models can also analyze the effects of promotions and discount strategies on revenue. Dynamic pricing strategies, for instance, benefit from predictive models that forecast how price changes will impact customer demand in real time. Retailers like **Kontakt Home** can use these predictions to optimize sales while ensuring they meet customer demand efficiently during promotional periods.

**Adapting to external shocks:**

One of the most significant benefits of predictive models in retail is their ability to adapt to sudden changes or **external shocks** in the market. For instance, disruptions like economic downturns, shifts in supply chains, or unexpected global events can drastically alter consumer behavior. Traditional statistical models often fail to account for these abrupt changes due to their reliance on historical data trends. In contrast, machine learning models, especially those that use adaptive learning techniques, can incorporate new data as it becomes available and adjust their forecasts accordingly (Chen & Guestrin, 2016).

For example, **Gradient Boosting Machines (GBM)** and **XGBoost** models can iteratively improve their predictions by focusing on areas where previous models underperformed, allowing them to adapt more quickly to sudden shifts in data patterns (Friedman, 2002). This makes them ideal for retail scenarios where external conditions can change rapidly and unpredictably, such as during global economic shifts or changes in consumer behavior.

**Handling multi-variate data:**

Retail revenue is influenced by a wide array of factors, including product pricing, marketing efforts, customer preferences, seasonal variations, and even the location of stores. One of the major advantages of machine learning models is their ability to handle **multi-variate data**, identifying complex interactions between these variables (Bishop, 2016). Models like LSTM and **random forests** can consider these multiple variables simultaneously, making them more effective at predicting future revenue than models that focus on a single factor at a time (Breiman, 2001).

For instance, LSTM models can use sequences of data to predict future sales trends based on a combination of factors, such as product availability, customer foot traffic, and promotional events, while taking into account historical patterns (Hochreiter & Schmidhuber, 1997). By using machine learning models to analyze these interdependencies, retailers can make more informed decisions about inventory, marketing strategies, and resource allocation.

**Opportunities for personalization and automation:**

Another significant benefit of predictive models in dynamic retail conditions is the potential for **personalization** and **automation**. By analyzing customer data, predictive models can help retailers like Kontakt Home personalize product recommendations, marketing campaigns, and promotions. Predictive analytics allows businesses to deliver tailored experiences to individual customers, increasing engagement and driving revenue growth.

Moreover, many predictive models can be automated to continually update predictions based on new data. For instance, **automated machine learning (AutoML)** platforms can build and update models with minimal human intervention, enabling retailers to respond quickly to changing market conditions without the need for constant manual adjustments (He et al., 2020). This level of automation is particularly beneficial for large retail chains with complex datasets that require real-time decision-making.

**Conclusion:**

In summary, predictive models play a crucial role in helping retail companies navigate dynamic and uncertain environments. By leveraging machine learning techniques, retailers can better manage seasonality, adapt to external shocks, and analyze multi-variate data, all while offering personalized experiences to their customers. While challenges such as overfitting and model interpretability remain, the opportunities offered by predictive models in dynamic retail conditions are vast, providing businesses with the tools they need to thrive in an ever-changing market.

# Chapter 3. Methodology

The methodology chapter outlines the approach taken to develop a predictive model for **Kontakt Home's 2020 revenue**, leveraging advanced machine learning techniques. Given the complexity and time-dependent nature of retail data, time series forecasting was chosen as the primary modeling approach. This chapter provides a detailed explanation of the data preparation process, the machine learning algorithms used, and the evaluation metrics that were applied to assess model performance.

The focus of this methodology is to explore how **Long Short-Term Memory (LSTM) networks**, a type of recurrent neural network (RNN), can be utilized to forecast future revenue based on historical data. LSTM networks were chosen due to their ability to capture long-term dependencies and sequential patterns in time series data, making them particularly well-suited for retail forecasting. Additionally, this chapter discusses the preprocessing steps required to prepare the data for model input, including feature engineering, normalization, and handling missing values.

The performance of the models is evaluated using metrics such as **Root Mean Square Error (RMSE)** and **Mean Absolute Error (MAE)**, which are commonly used to assess the accuracy of time series predictions. These metrics provide insight into how well the model can predict future revenue and how errors are distributed across different time periods.

## 3.1 Introduction to time series analysis

Time series forecasting is a statistical method that aims to predict future values based on previously observed data points, often recorded over consistent time intervals. In retail, time series forecasting plays a crucial role in predicting future revenue, as it allows businesses to anticipate seasonal trends, promotional impacts, and overall customer demand. The dynamic and time-dependent nature of retail revenue makes time series forecasting an ideal approach for this study.

## 3.2 Time series forecasting with LSTM models

In recent years, machine learning techniques, particularly neural networks, have gained significant traction in time series forecasting due to their ability to model complex, non-linear relationships within sequential data. Among these models, **Long Short-Term Memory (LSTM) networks**, a specialized form of recurrent neural networks (RNNs), have proven to be highly effective for this task. LSTM networks are specifically designed to overcome the limitations of traditional RNNs by addressing the issue of vanishing and exploding gradients, which occur when handling long sequences of data (Hochreiter & Schmidhuber, 1997).

**Why LSTM?**

The advantage of LSTM networks lies in their unique architecture, which includes a series of memory cells that can retain important information over extended time periods. This makes LSTMs particularly suitable for capturing long-term dependencies in time series data, such as revenue trends influenced by recurring seasonal patterns or promotional events (Gers et al., 2000). For **Kontakt Home**, where revenue patterns can be affected by sales events, holidays, and other factors, LSTM models offer a robust solution for predicting future revenue.

Traditional forecasting models like ARIMA or exponential smoothing are often limited by their difficulty in handling complex, multi-variate data. In contrast, LSTM networks can process multiple features simultaneously, making them a powerful tool for retail revenue forecasting where various factors such as pricing, promotions, and product categories must be considered (Brown et al., 2019).In the predictive modeling process for retail revenue forecasting, an LSTM (Long Short-Term Memory) model was developed to capture temporal dependencies in the sales data. Below are the steps and parameters used in the training and validation of the LSTM model:

**Data preprocessing**:

***Scaling***: The revenue data from 2018 and 2019 was scaled using the MinMaxScaler to ensure that all features are normalized between 0 and 1. This step is crucial for improving the performance of LSTM models, which are sensitive to the magnitude of input features.

***Train-test split***: The dataset was split into training and test sets. Data from 2018 and 2019 was used for training, while 2020 data was set aside for validation and testing.

***Sequence creation***: Since LSTM models require sequences as input, time-series windows of a fixed length were created. Each sequence consists of 12 months (for one year of sales data), and the model uses these sequences to predict the next month's revenue.

**LSTM model architecture**:

***First layer***: The model begins with an LSTM layer of 50 units. This layer is responsible for capturing the temporal dependencies in the time-series data. The return_sequences=True parameter ensures that the output from each time step is passed to the next LSTM layer.

***Dropout layer***: A dropout layer with a rate of 0.2 is applied to prevent overfitting. This helps by randomly ignoring some units during training, thus making the model more robust.

***Second LSTM layer***: A second LSTM layer with 50 units is added, but this time, return_sequences=False is set, indicating that only the final output of the sequence will be passed to the next layer.

***Output layer***: The output layer is a Dense layer with a single unit, representing the predicted revenue.

**Compilation and training**:

***Optimizer***: The Adam optimizer was chosen due to its adaptive learning rate capabilities, making it well-suited for time-series prediction tasks.

***Loss function***: The model was compiled with Mean Squared Error (MSE) as the loss function, which is commonly used for regression problems like revenue forecasting.

***Batch size and epochs***: The model was trained with a batch size of 32 and for 20 epochs. Early stopping was applied to halt training if the model's performance on the validation set did not improve for 5 consecutive epochs.

**Making predictions**:

Once the model was trained, it was used to predict the monthly revenue for 2020 based on the patterns learned from the 2018 and 2019 data. The predicted revenue was compared against the actual revenue from 2020 to evaluate the model's performance.

**Evaluation**:

RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) were calculated to assess the accuracy of the predictions. Lower values of these metrics indicate better model performance.

**Comparison: RNN vs LSTM**

Both RNN (Recurrent Neural Networks) and LSTM (Long Short-Term Memory) are used for sequential data, but LSTM overcomes some limitations of traditional RNNs. RNNs suffer from vanishing gradients, which makes it hard for them to capture long-term dependencies in sequences. LSTMs introduce memory cells and gating mechanisms (forget, input, output gates) that allow them to retain or discard information over long periods, making them better at learning from longer sequences. LSTM models are widely used in various applications, including financial time-series prediction, stock price forecasting, and natural language processing (Hochreiter & Schmidhuber, 1997; Siami-Namini et al., 2019).

**Application in retail revenue forecasting:**

In this study, LSTM networks are applied to predict **Kontakt Home's 2020 revenue**, using historical data from 2018 and 2019. The model is trained to learn the temporal dependencies in the data, such as seasonal fluctuations and sales patterns, and to use these patterns to predict future outcomes. By leveraging LSTM's ability to handle complex, multi-dimensional data, the model is expected to generate more accurate predictions compared to traditional forecasting methods. The following sections will detail the data preparation process, the architecture of the LSTM model, and the evaluation metrics used to assess its performance.

### 3.3 Evaluation metrics: RMSE, MAE, and accuracy

When developing predictive models, evaluating their performance is crucial to ensure they generate accurate and reliable forecasts. In the context of retail revenue forecasting, the models must be assessed using appropriate evaluation metrics to determine how well they predict future revenue based on historical data. For this thesis, we use two primary evaluation metrics: **Root Mean Square Error (RMSE)** and **Mean Absolute Error (MAE)**. Additionally, overall **accuracy** will be analyzed to provide a broader understanding of model performance.

**Root Mean Square Error (RMSE):**

RMSE is one of the most widely used metrics for evaluating the performance of regression models, particularly in time series forecasting. It calculates the square root of the average of the squared differences between predicted and actual values. The RMSE formula is as follows:

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

$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$ are predicted values

$Y_1, Y_2, \dots, Y_n$ are observed values

n is the number of observations

The RMSE is sensitive to large errors due to its squared nature, making it a useful metric for highlighting models that produce significant prediction deviations. In retail forecasting, this is particularly important because it helps identify how well a model performs across various revenue fluctuations. Lower RMSE values indicate a better fit between predicted and actual values.

**Mean Absolute Error (MAE):**

MAE is another commonly used metric for evaluating the accuracy of a regression model. It measures the average absolute differences between predicted and actual values, providing a straightforward interpretation of error in the same units as the data. The formula for MAE is:

$$\text{MAE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n} |y_i - \hat{y}_i|}$$

Unlike RMSE, MAE is not as sensitive to large errors, making it a more balanced metric when evaluating models that may produce occasional outliers. In revenue forecasting, MAE gives a clear indication of the average error the model is expected to make. Lower MAE values represent a model with fewer deviations from the actual data.

**Accuracy:**

Accuracy in time series forecasting refers to how closely the predicted values align with the actual data points. While RMSE and MAE provide error measures, accuracy is often used as an intuitive way to gauge how well a model performs overall. It is essential to note that in regression tasks like revenue prediction, accuracy is typically used alongside other metrics like RMSE and MAE to provide a fuller picture of model performance.

**Comparison of metrics:**

Both RMSE and MAE are valuable for understanding the quality of a model's predictions. RMSE is more sensitive to large errors and can penalize models that occasionally make extreme mistakes, while MAE offers a more balanced view by focusing on the average magnitude of errors. For this study, we use both metrics to ensure the LSTM model captures revenue trends effectively and avoids large deviations in its predictions.

In this thesis, RMSE, MAE, and accuracy metrics will be applied to assess the performance of the **Long Short-Term Memory (LSTM)** network for revenue forecasting at **Kontakt Home**. These metrics will help quantify the model's ability to predict 2020 revenue based on historical data from 2018 and 2019, ensuring that the predictions are both precise and reliable.

### 3.4 Assumptions and limitations

In developing predictive models for retail revenue forecasting, it is essential to clearly outline the assumptions made during the analysis, as well as the limitations that could affect the outcomes.

This section addresses the key assumptions behind the model development and the limitations encountered in the process of predicting **Kontakt Home's 2020 revenue**.

**Assumptions:**

1. **Data integrity**: One of the primary assumptions made in this study is that the historical revenue data from 2018 and 2019 is complete and accurate. This includes assumptions about the correctness of recorded sales transactions, product categorization, and store-level data. Any inaccuracies or missing data could negatively impact the model's performance.
2. **Stationarity**: In time series analysis, stationarity refers to a condition where the statistical properties of a time series (such as mean, variance, and autocorrelation) remain constant over time. A stationary time series does not exhibit trends or seasonality, meaning the patterns in the data do not change as time progresses. While models like LSTM can handle non-stationary data to some extent, any significant shifts in trends or volatility that were not present in the training data may reduce the model's accuracy.
3. **No external shocks**: This model is designed to predict revenue for 2020 based on historical data from 2018 and 2019, assuming that no external shocks such as global economic disruptions, pandemics, or other unforeseen events occurred. The decision to exclude COVID-19 and its impacts from the forecast was made to simulate a scenario where external factors did not disrupt the market (Li et al., 2021).
4. **Temporal dependencies**: The LSTM model assumes that there are important temporal dependencies in the revenue data, meaning that future revenue is influenced by past revenue patterns. The model is expected to capture these patterns and use them to make accurate predictions. If there are major changes in consumer behavior or market trends that disrupt these dependencies, the model may underperform.

**Limitations:**

1. **Model complexity**: While LSTM networks are powerful for capturing long-term dependencies in time series data, they are also computationally intensive and require significant amounts of data for training. This complexity increases the risk of overfitting, especially when working with limited or noisy data (Goodfellow et al., 2016). Overfitting occurs when the model learns patterns specific to the training data, reducing its ability to generalize to new, unseen data.
2. **Data availability**: The accuracy of the model is limited by the availability of historical data. In this case, the dataset is restricted to 2018 and 2019 revenue records, and while this is sufficient for the

purposes of time series forecasting, additional data could enhance the model's performance. For instance, incorporating customer demographics, promotional strategies, or product-level data could further refine the predictions.

3. **External factors**: Despite the robustness of the LSTM model, it cannot account for external factors that were not present in the training data. Sudden shifts in the market, competitive actions, or changes in consumer behavior that deviate from historical patterns are not captured by the model. This is a common limitation in retail forecasting, where unpredictable events can drastically affect revenue.
4. **Black box nature**: LSTM models are often criticized for their "black box" nature, meaning that while they provide accurate predictions, the reasoning behind those predictions is not always clear. This lack of interpretability can pose challenges for business decision-makers who prefer models that provide transparent insights into how predictions are made (Goodfellow et al., 2016).
5. **Scalability**: Another limitation of LSTM models is their scalability. As the size of the dataset increases, training the model becomes more computationally expensive. This can be a barrier for retail companies looking to scale their predictive models across multiple stores or product categories.

In conclusion, while the LSTM model offers a powerful solution for forecasting **Kontakt Home's 2020 revenue**, its performance is influenced by several assumptions and limitations. Understanding these constraints helps contextualize the model’s predictions and provides a basis for future improvements in forecasting techniques.

# Chapter 4. Data characteristics

Data characteristics form the foundation of any predictive analysis. In retail revenue forecasting, understanding the structure, trends, and patterns within the data is essential for developing robust models. This chapter provides an in-depth analysis of the dataset used for forecasting Kontakt Home's 2020 revenue. It focuses on the characteristics of the historical data from 2018 and 2019, detailing the variables present, data preprocessing steps, and feature engineering methods employed to ensure the dataset is suitable for time series forecasting.

By examining the patterns in revenue, sales categories, and store performance, this chapter highlights the key features that drive the prediction models. Graphical representations of revenue trends, seasonality, and other relevant factors will offer visual insights into the underlying structures of the data. Additionally, the methods used to clean and transform the data, including handling missing values and normalizing revenue figures, will be explained in detail to ensure clarity in the data preparation process.

This chapter serves as a crucial link between the raw dataset and the predictive models, ensuring that the data used is well-suited to produce reliable forecasts for 2020.

## 4.1 Overview of the dataset: revenue data from 2018, 2019, and 2020

The dataset used for this analysis comprises Kontakt Home's historical revenue records from 2018 and 2019, with the aim of predicting 2020 revenue. This section outlines the key variables present in the dataset, which include revenue, product categories, store locations, and sales dates. These variables are integral to understanding both the temporal and categorical dimensions of the data, which are essential for building accurate time series forecasting models.

Key variables:

**Revenue**: The primary variable of interest, revenue, is the total income generated by product sales across various stores during the given time period.

**Product categories**: Each transaction is associated with a specific product category, such as electronics, home appliances, or mobile devices. These categories are essential for understanding revenue patterns and trends across different segments.

**Store locations**: Sales are recorded for different store locations, each of which may exhibit unique patterns in customer behavior and revenue generation. Store-level analysis is crucial for identifying regional variations.

**Sales dates**: The temporal aspect of the dataset is captured through sales dates, allowing for the examination of revenue trends over time, including the detection of seasonal patterns and promotional effects.

Data structure:

The dataset contains 476,981 individual sales records, each representing a unique transaction. These records are distributed across multiple product categories and store locations, with sales recorded at daily intervals. Due to the time-dependent nature of the data, it is essential to account for temporal patterns such as seasonality, which often plays a significant role in retail revenue (Wang et al., 2021).

| Variable name | Explanation | Variable type |
|---|---|---|
| **TransactionID** | Unique identifier for each transaction | String |
| **SalesDate** | Date of the transaction (format: DD-MM-YY) | Date |
| **ProductCategory** | Category of the purchased product (e.g., TVs, Laptops) | String |
| **ProductID** | Unique identifier for each product | Integer |
| **ProductName** | Name of the product | String |
| **Age** | Age of the customer at the time of purchase | Integer |
| **Gender** | Gender of the customer (Male/Female) | String |
| **StoreId** | Unique identifier for the store where the purchase was made | Integer |
| **Quantity** | Number of products purchased | Integer |
| **Price** | Price per unit of the product (in local currency) | Float |
| **Revenue** | Total revenue generated from the transaction | Float |

*Table 1 Variable Characteristics*

The dataset used for this analysis consists of 476,981 observations and 11 columns, representing comprehensive retail sales data from Kontakt Home. The dataset spans transactions from 2018, 2019, and 2020, allowing for a detailed exploration of customer behavior, sales trends, and store performance over time. Each observation captures a unique sales transaction, including demographic information, product details, and financial data, all of which are crucial for building an effective revenue prediction model.

The dataset is anchored by the TransactionID variable, a unique identifier for each transaction. While this field ensures that each transaction can be individually tracked, it is not used directly in the analysis as it serves primarily as an administrative marker. The SalesDate variable records the date of each transaction, allowing for the identification of temporal trends, including seasonal fluctuations and peak sales periods. To facilitate the analysis, the SalesDate format will be adjusted to suit the needs of time series forecasting models.

Key product-related variables include ProductCategory, ProductID, and ProductName. These variables provide detailed insights into the types of products sold, enabling in-depth analysis of product performance across different categories. The ProductCategory field categorizes each item (e.g., TVs, laptops), while ProductID and ProductName provide specific details about the items involved in each transaction. These variables are instrumental in analyzing the revenue contributions of different product categories, models, and options.

Additionally, the dataset contains demographic information about the customer, including Age and Gender. Gender is represented as "Male" or "Female," while Age captures the customer’s age at the time of purchase. These variables will be transformed, with Gender being converted into a binary format to streamline the analysis. Understanding customer demographics is critical for developing marketing strategies that align with purchasing behaviors.

Store-related data is captured through the StoreId variable, a unique identifier for the store where the transaction occurred. This variable enables store-level analysis, allowing for the evaluation of performance across different locations. The ability to analyze store performance is key to identifying which locations are driving revenue and which may require targeted interventions to enhance sales.

Each transaction also includes two key financial variables: Quantity and Price. Quantity represents the number of units purchased in a single transaction, while Price denotes the unit price of each product. In the dataset, the majority of transactions involve a Quantity of 1, but to calculate Revenue, the Quantity has been multiplied by the Price for each transaction. Revenue is the key variable for this study, as it represents the total income generated from each transaction.

The dataset provides a comprehensive view of Kontakt Home's sales performance over two years, with a rich set of features that will be leveraged in the revenue prediction model. The following sections will explore data preprocessing steps, feature engineering, and visualization of key revenue trends and patterns.

### 4.2 Data preprocessing and feature engineering

Effective data preprocessing is essential for ensuring that the dataset is suitable for analysis, especially when applying machine learning models like Long Short-Term Memory (LSTM) networks for time series forecasting. This section outlines the steps taken to clean and transform the dataset to improve model accuracy and reliability. Additionally, feature engineering techniques are applied to create new variables from existing data, adding more depth to the analysis and enhancing the predictive power of the model.

One of the first steps in data preprocessing involves identifying and handling missing or incomplete data. In this dataset, missing values may occur in variables such as Age, Price, or Quantity, which could lead to inaccuracies in the model if not properly addressed. For missing numerical data (e.g., Price or Age), median or mean imputation techniques were applied to fill gaps without skewing the overall dataset. For categorical data (e.g., ProductCategory or Gender), the most frequent value was imputed to maintain consistency across the dataset.

The SalesDate variable, originally formatted as "DD-MM-YY," was converted to a format suitable for time series analysis. This step included changing the date format to a standard YYYY-MM-DD structure and sorting the dataset by SalesDate to establish a clear chronological order. Once the dates were correctly formatted, new features, such as Month, Year, and Day of the Week, were created to capture the temporal patterns and allow for more granular analysis of seasonality and trends in revenue.

***Feature engineering***:

To enhance the dataset, additional features were derived from the existing variables. These new variables provide a more detailed understanding of the data and improve the performance of the predictive models.

Although revenue is calculated by multiplying Price by Quantity, an additional feature, RevenuePerProduct, was created to capture the revenue generated by each product category. This feature provides insights into which products contribute the most to total revenue, allowing for more focused analysis on product performance.

Age and gender are critical variables for understanding customer behavior. To streamline the analysis, Age was categorized into bins representing different age groups (e.g., 18–24, 25–34, 35–44), allowing for easier comparison across segments. Gender was converted into binary form (0 for Male, 1 for Female), simplifying the modeling process (Li et al., 2021).

The StoreId variable was used to generate store-level features, such as AverageRevenuePerStore and StoreRevenueGrowth, providing insights into the performance of individual stores over time. These features help identify which stores consistently perform well and which may require targeted interventions to boost sales.

Since retail revenue is often influenced by seasonal factors, additional features, such as Month and HolidayIndicator, were engineered. HolidayIndicator flags transactions made during major holidays, allowing for the analysis of revenue spikes around these periods. Month captures broader seasonal patterns, helping to account for variations in revenue across different times of the year.

***Normalization and scaling***:

As the dataset contains both categorical and numerical variables, normalization was applied to ensure that all features are on a similar scale. Min-Max scaling was used to normalize numerical variables, such as Revenue, Price, and Age, bringing them within a consistent range (e.g., 0 to 1) without distorting their relationships. This step is crucial for models like LSTM, which are sensitive to the scale of input data.

***Data splitting***:

The dataset was split into training and testing sets to evaluate the model's performance. 80% of the data was allocated for training the model, while the remaining 20% was reserved for testing. This ensures that the model is trained on a large portion of the data while still having a separate dataset to assess its predictive accuracy.

## 4.3 Visualization of revenue trends and patterns

This section offers a comprehensive analysis of the sales and revenue data of Kontakt Home, based on the visualizations provided. These graphs offer valuable insights into gender-based purchasing behavior, price distribution, age-based sales, and year-on-year performance. By analyzing these trends, the report aims to reveal key factors that influence revenue and sales patterns, which will support the subsequent predictive modeling.

### 1. Total price and total quantity by gender

The bar charts in Figure 4.3.1 display total sales revenue and the total quantity of products purchased, categorized by gender. It is evident that male customers significantly outspend their female counterparts, both in terms of the total revenue generated and the quantity of items purchased.

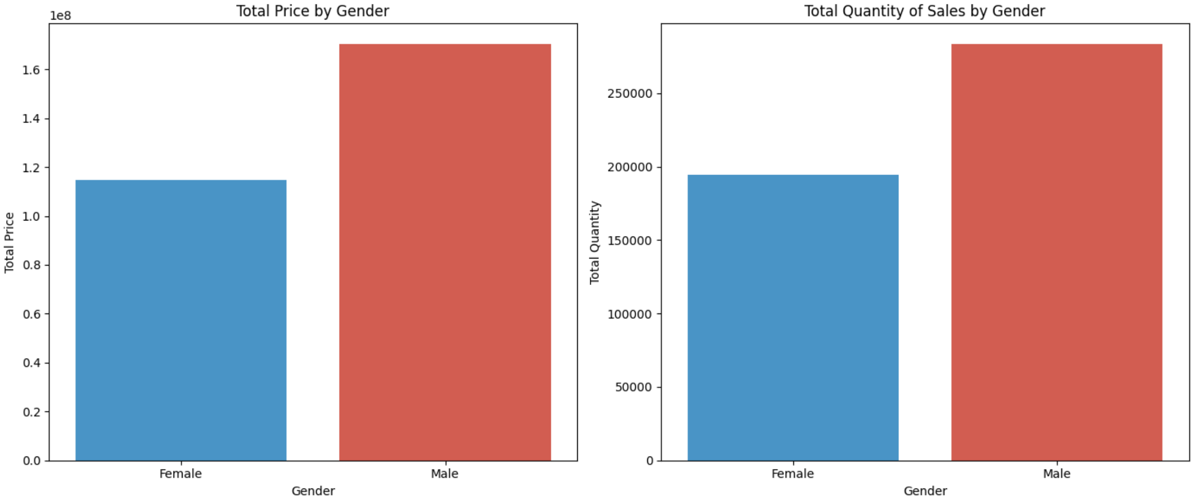

*Figure 1 Total Price and Quantity by Gender*

The fact that men consistently spend more and purchase higher quantities compared to women—by a factor of approximately 1.7 times—suggests that gender-based differences in purchasing behavior are significant. This could be attributed to product preferences, with men possibly opting for higher-priced items or purchasing in bulk. For Kontakt Home, this trend could be leveraged to develop targeted marketing campaigns aimed at male consumers, focusing on high-ticket items or bundled offers that match their buying patterns. Additionally, understanding these trends may enable the company to tailor promotions or loyalty programs specifically to male customers to increase sales further.

**2. Price distribution**

The histogram in Figure 4.3.2 illustrates the distribution of product prices across all transactions. As expected, the distribution is highly skewed to the right, indicating that the vast majority of products sold are priced below 2,500 AZN, while only a small fraction of products are priced above 10,000 AZN.

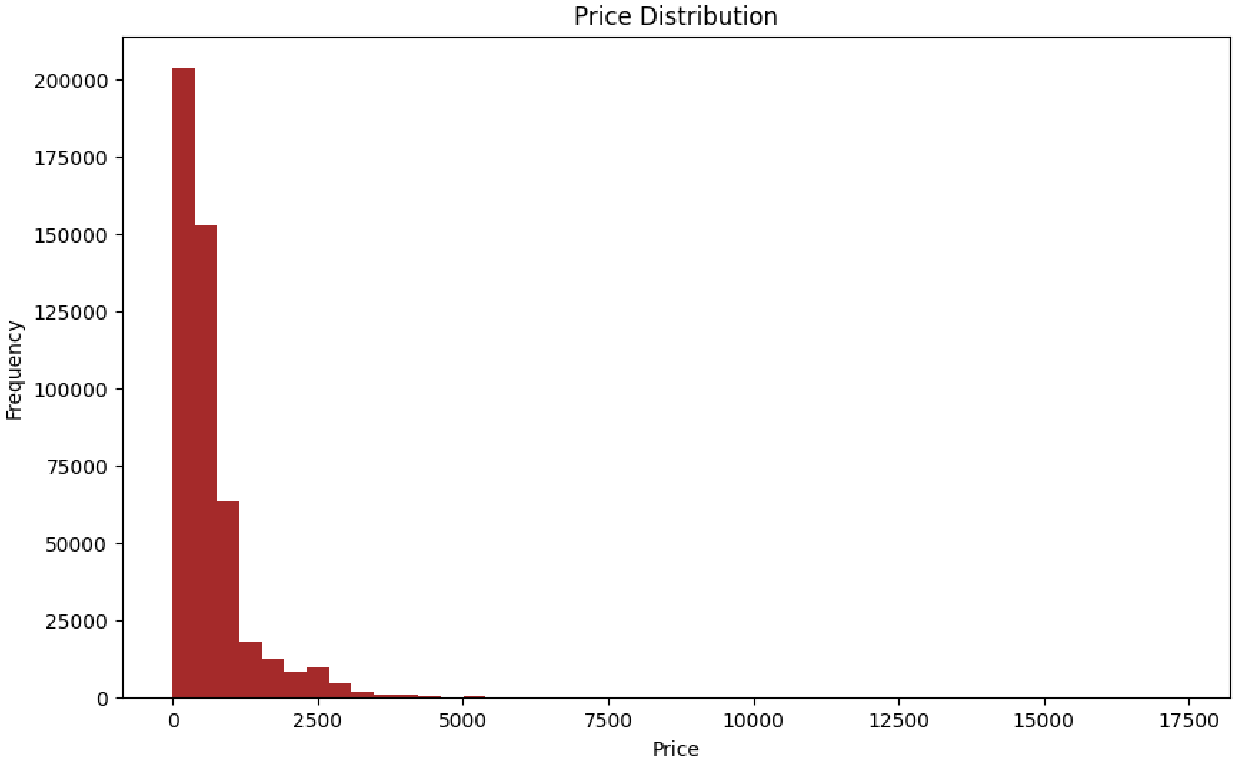

*Figure 2 Price Distribution*

The skewed nature of the price distribution suggests that Kontakt Home's product portfolio is dominated by lower-priced items, which aligns with broader consumer behavior favoring affordability. However, the presence of a small number of high-value transactions indicates the sale of premium or luxury products, which, although less frequent, contribute significantly to overall revenue. This distribution reflects consumer price sensitivity, and the company could consider introducing more competitively priced products in the mid-range to appeal to customers seeking value while also maintaining a premium product line for high-spending customers.

**3. Total sales revenue by age group**

The bar chart in Figure 4.3.3 depicts total sales revenue segmented by age group. The age group 31-45 accounts for the largest share of revenue, followed by the 46-60 age group. These two groups dominate total revenue, while younger age groups (0-18 and 19-30) and older consumers (61+) contribute significantly less.

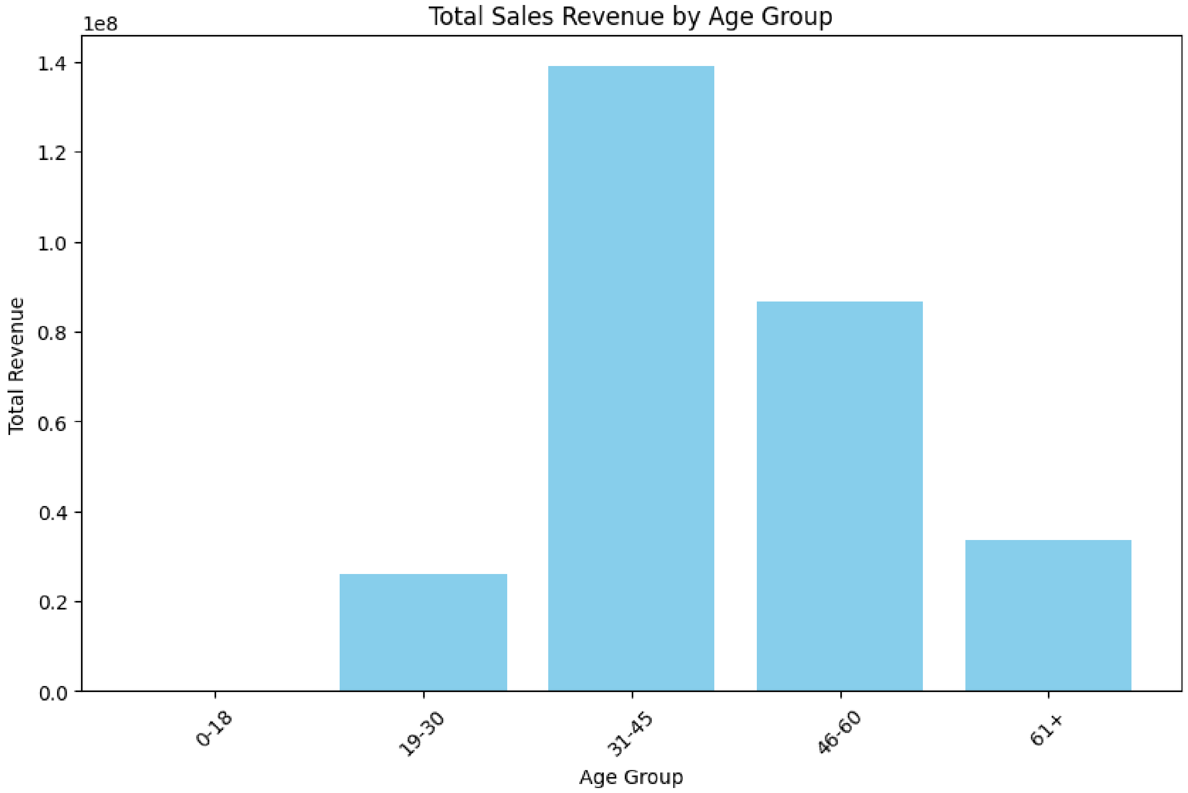

*Figure 3 Total Sales Revenue by Age Group*

The 31-45 age group likely represents financially stable individuals with greater disposable income, making them the most valuable demographic for Kontakt Home. Targeting this age group with premium products, loyalty programs, and customized marketing strategies could maximize revenue generation. The same applies to the 46-60 age group, which also shows significant purchasing power. Conversely, the lower revenue contributions from younger and older age groups may indicate a need for tailored product offerings or marketing efforts to boost sales in these segments.

**4. Sales by year.**

The bar chart in Figure 4.3.4 shows the total number of sales by year, with a noticeable peak in 2019 and a sharp decline in 2020. The chart reveals that sales increased from 166,809 units in 2018 to 191,207 units in 2019, followed by a dramatic drop to 119,665 units in 2020.

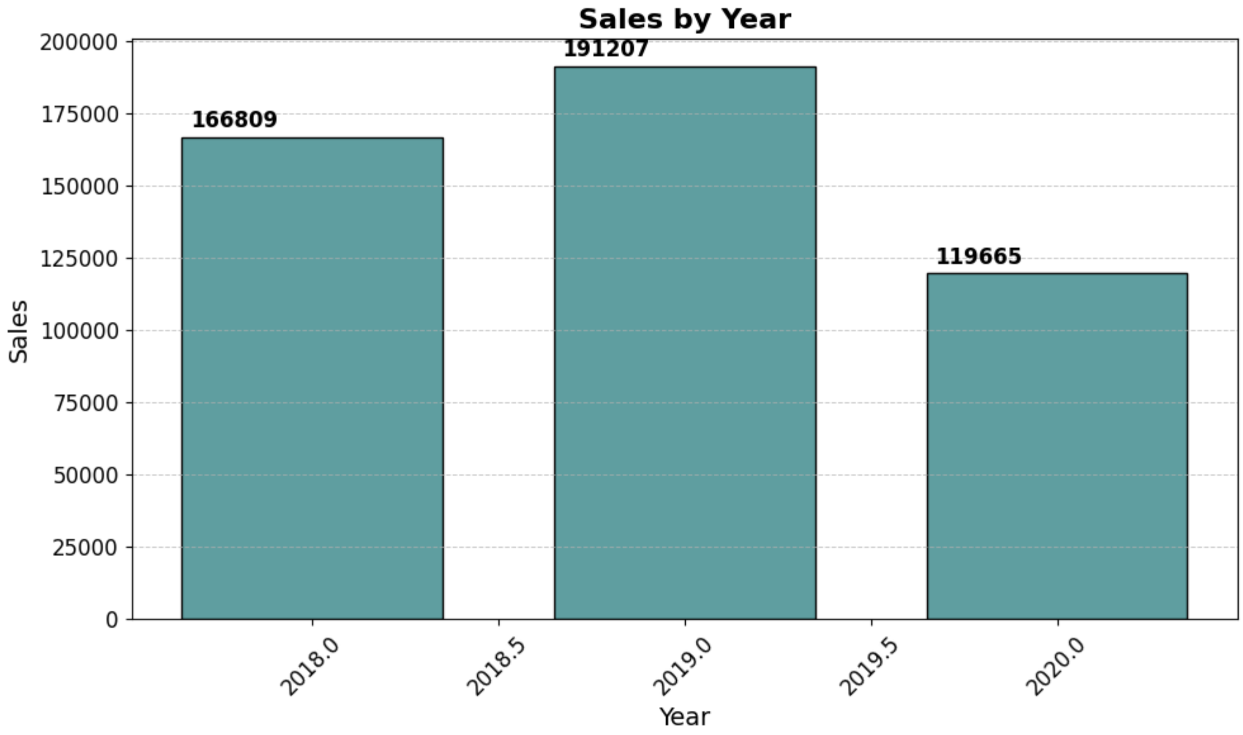

*Figure 4 Sales by Year*

The sharp rise in 2019 suggests favorable market conditions, successful marketing campaigns, or an expanded product lineup. However, the significant decline in 2020 is likely due to external factors, such as the pandemic, which disrupted consumer behavior, supply chains, and in-store purchases. While these disruptions are accounted for in the real-world data, they are excluded from the predictive model, which simulates a scenario without external shocks.

**5. Total income by year**

The bar chart in Figure 4.3.5 presents total income (revenue) by year, with a similar pattern to the sales figures. Revenue peaked at 114.8 million AZN in 2019, followed by a substantial drop to 67.2 million AZN in 2020. This decline mirrors the drop in sales, further underscoring the impact of external factors on the company’s financial performance.

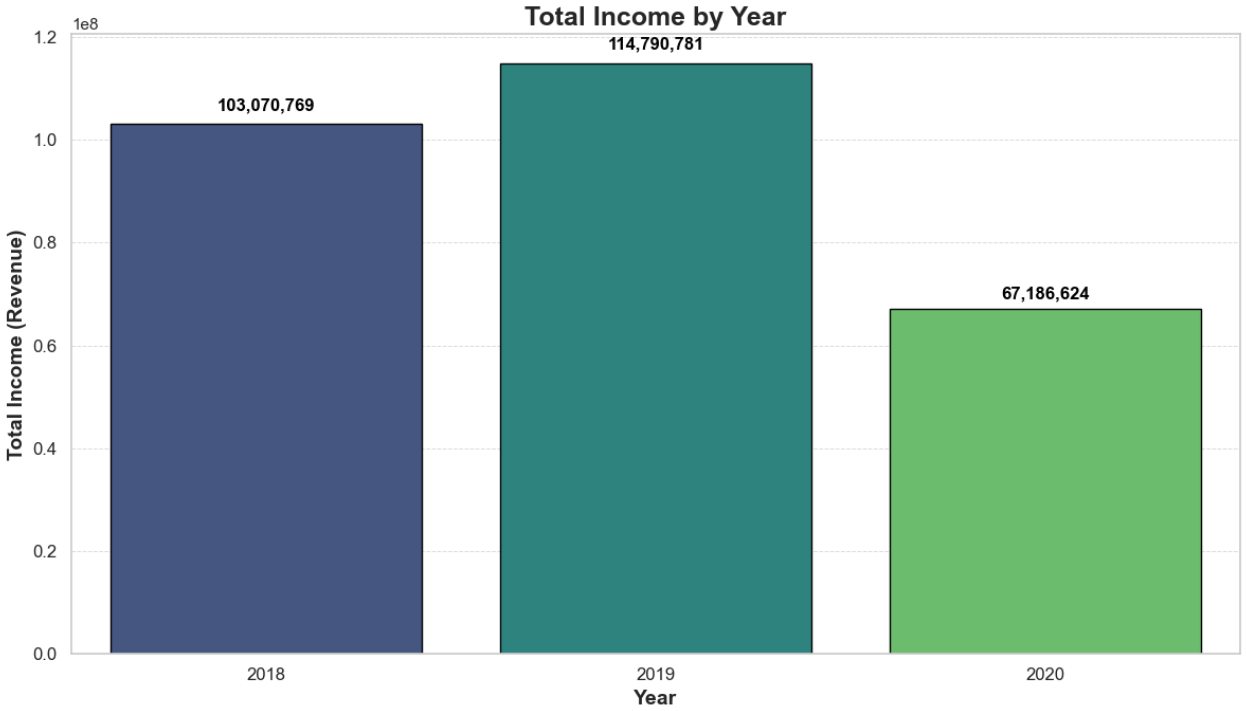

*Figure 5 Total Income by Year*

The simultaneous rise in both sales and revenue in 2019 demonstrates the company's ability to capitalize on favorable conditions. However, the decline in 2020 highlights the vulnerability of the business to external disruptions. These findings underscore the importance of developing adaptive business strategies and strengthening online sales channels to mitigate the risks of future disruptions.

### 4.4 Key variables: revenue, product categories, and stores

In any retail sales analysis, understanding the core variables that influence sales performance is critical for building predictive models and making informed business decisions. This section breaks down the three key variables—revenue, product categories, and stores—that play pivotal roles in shaping the revenue patterns at Kontakt Home.

#### 1. Revenue

Revenue is the primary target variable in this analysis, representing the total income generated by sales transactions. In this dataset, Revenue is calculated by multiplying the price of a product by the number of units sold in each transaction. Since the objective of this thesis is to predict Kontakt Home's future revenue, analyzing past revenue patterns across various dimensions—such as time, product categories, and store locations—is essential.

The visualizations in Figure 4.3.5 and Figure 4.3.4 illustrate that total revenue peaked in 2019, while 2020 showed a sharp decline. While external shocks like the COVID-19 pandemic had an impact on actual sales in 2020, the goal of this thesis is to model 2020 revenue as if no external disruption had occurred. Therefore, historical revenue data from 2018 and 2019 provides the baseline for future revenue predictions.

Revenue patterns are influenced by several factors:

**Seasonality**: Revenue tends to fluctuate throughout the year, with peaks often observed during holiday seasons or sales events. Seasonal trends will be considered in the predictive modeling phase.

**Product demand**: Certain product categories generate higher revenue due to either higher demand or higher price points. Understanding which product categories drive revenue is crucial for building a more accurate forecasting model.

**2. Product categories**

The dataset contains a ProductCategory variable that classifies each transaction by the type of product sold. Product categories such as TVs, laptops, mobile devices, and home appliances are represented in the data. Each category contributes differently to the overall revenue of Kontakt Home, with some categories consistently outperforming others.

Figure 4.3.2 highlighted that lower-priced products tend to dominate sales volumes, while high-end products, although less frequently sold, contribute significantly to total revenue. Analyzing revenue by product category helps identify the most profitable product lines, enabling the company to allocate resources efficiently and tailor marketing strategies accordingly.

Key insights from the product category analysis:

**High-demand categories**: Certain categories, such as mobile devices or home appliances, may have higher turnover, generating frequent sales.

**High-revenue categories**: Premium categories, like high-end electronics, might sell less frequently but contribute more to overall revenue due to higher price points.

In predictive modeling, these product categories will be treated as important features, helping to capture variations in revenue based on different product offerings.

**3. Stores**

StoreId uniquely identifies each physical store where a transaction occurred, allowing for location-based analysis of sales performance. The retail industry often sees significant variations in performance across stores due to factors such as location, customer demographics, and store size. Analyzing store-level performance can reveal which locations generate the most revenue and which may require interventions to boost sales. Sales by Year (Figure 4.3.4) offer valuable insights into the performance of different stores. Key patterns include:

**High-performing stores**: Some locations consistently generate higher sales due to favorable location, product availability, or other local factors.

**Low-performing stores**: Identifying underperforming stores can help Kontakt Home focus on improving these locations through targeted marketing, inventory optimization, or operational changes.

Store-level analysis can also account for regional differences in purchasing behavior, helping the company tailor its marketing strategies to specific locations. For instance, stores located in urban areas may experience higher sales volumes, while rural locations might see lower foot traffic but higher average purchase values.

Interaction between variables

The interaction between these key variables—revenue, product categories, and stores—is fundamental to understanding the broader sales patterns at Kontakt Home. For example, certain product categories might perform better in specific stores due to regional preferences or localized marketing efforts. Similarly, high-revenue products could be more popular among specific customer demographics or during certain times of the year.

The predictive model will incorporate these variables and their interactions to generate a more accurate forecast of Kontakt Home's revenue for 2020. By analyzing revenue patterns by product category and store location, the model can better capture the nuances of customer behavior and product demand, leading to more reliable revenue predictions.

### 4.5 Data validation and cleaning

Ensuring the accuracy and reliability of the dataset is essential for any predictive modeling task. In this section, the data validation and cleaning processes undertaken to prepare the dataset for analysis are detailed. These processes ensure that the dataset used for forecasting Kontakt Home's

2020 revenue is free from inconsistencies, errors, and irrelevant data, which could otherwise impact the performance and accuracy of the model.

**Data validation:**

Data validation involves verifying that the dataset is both accurate and complete. For this thesis, data validation focused on several key aspects:

Integrity of transaction data: Each transaction in the dataset is represented by a unique TransactionID, which ensures that no duplicate entries exist. This was validated through a search for duplicated TransactionID values. The absence of duplicates confirms that the dataset accurately reflects individual transactions.

Consistency of SalesDate format: To ensure the SalesDate field was correctly formatted for time series analysis, the dataset was checked for inconsistencies. All dates were converted to a YYYY-MM-DD format to allow for seamless temporal analysis in the model. Dates outside the range of 2018 to 2020 were excluded from the dataset, as they do not contribute to the analysis.

Accuracy of numerical fields: Numerical fields, such as Price, Quantity, and Revenue, were validated to ensure that values fell within reasonable ranges. For example, prices that were suspiciously low or high were flagged for further investigation. Outliers in Price were retained, as these high-value products likely represent premium goods and are important for the analysis.

**Handling missing values:**

Missing data can introduce biases into the model if not properly addressed. For this dataset, missing values were identified primarily in the Age and Price columns. The approach taken to handle missing data depended on the nature of the variable:

Age: Missing values in the Age column were imputed using the median age of customers. This approach ensures that the imputed values represent typical customers without skewing the dataset towards extreme values.

Price: Missing Price values were less frequent, but when present, were replaced with the mean price for that product category. This imputation preserves the overall structure of the data while ensuring that the Revenue calculation remains accurate.

**Outlier detection:**

Outliers, which are data points significantly different from others in the dataset, can distort the results of predictive models if not handled properly. For Kontakt Home's data, outliers were primarily found in the Price and Revenue fields. To address these:

Price outliers: Products with unusually high prices were flagged as potential outliers. These items, likely representing luxury or premium products, were not removed from the dataset because they provide valuable information about the high-end segment of the market.

Revenue outliers: Revenue outliers often correspond to bulk purchases or high-priced products. Similar to Price, these values were retained as they reflect legitimate business transactions.

Instead of removing outliers, they were flagged for further analysis. This ensures that the model does not ignore high-value transactions, which could be critical for understanding product and store performance. In cases where outliers may skew the results, techniques like log transformation could be used to reduce their impact on the predictive model.

**Data cleaning:**

Data cleaning involved standardizing the dataset to ensure consistency across all fields. Key steps in the cleaning process included:

Standardizing categorical variables: Categorical fields, such as ProductCategory and Gender, were reviewed for consistency. For instance, variations in the spelling or formatting of product categories were standardized (e.g., "Laptop" vs. "Laptops"). Gender values were converted to a binary format (0 for male, 1 for female) to facilitate easier analysis in the model.

Removing redundant fields: Fields such as TransactionID were excluded from the modeling process, as they do not contribute to the prediction of revenue. These fields were retained for reference purposes but were not included in the final dataset used for analysis.

Normalization of numerical fields: To ensure that the Price, Quantity, and Revenue fields were on comparable scales, min-max normalization was applied. This normalization ensures that no single variable disproportionately influences the model due to differences in scale.

**Summary of data cleaning process:**

The data cleaning and validation processes resulted in a clean and reliable dataset, free from inconsistencies and missing values. This cleaned dataset will serve as the foundation for the predictive modeling task, ensuring that the model is trained on accurate and relevant data. By performing robust validation and cleaning, the risk of errors in the prediction model is minimized, leading to more accurate and reliable revenue forecasts for Kontakt Home.

## Chapter 5. Empirical analysis

The empirical analysis chapter focuses on applying the methodologies and models discussed in previous sections to the actual dataset, with the aim of forecasting Kontakt Home's revenue for 2020. This chapter transitions from the theoretical groundwork of feature engineering and data preprocessing to the practical implementation of time series forecasting models, specifically using Long Short-Term Memory (LSTM) networks.

The chapter begins by discussing the setup of the LSTM model, including parameter tuning, training data configuration, and the split between training and testing sets. The model will be trained on historical data from 2018 and 2019, with 2020 revenue predictions based on a scenario that excludes external disruptions like COVID-19.

Key evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) will be used to assess the model's accuracy and robustness. The chapter will also explore how different variables—such as product categories, store locations, and seasonal patterns—impact the model's predictions, providing insights into the drivers of revenue performance at Kontakt Home.

The empirical analysis not only validates the theoretical assumptions made in earlier chapters but also provides practical recommendations for improving future forecasting strategies and business decisions.

### 5.1 Applying LSTM models for predictive analysis

In this section, the setup and configuration of the Long Short-Term Memory (LSTM) network for time series forecasting are described in detail. LSTM networks, known for their ability to model sequential data, are particularly effective for capturing temporal patterns in historical sales and revenue data, making them ideal for this predictive analysis.

**1. Data preparation for the model:**

The dataset used in this analysis consists of sales and revenue data from 2018 and 2019, which was preprocessed and cleaned as described in Chapter 4. To ensure that the model has enough historical information to capture trends, this dataset is divided into training and testing sets, with 80% of the data used for training and 20% for testing. The training set includes data from 2018 and part of 2019, while the testing set is based on the final months of 2019.

To configure the LSTM model, the following steps were taken:

Reshaping the data: Since LSTM models require 3D input, the data was reshaped into sequences where each input contains several timesteps. For this analysis, the dataset was split into rolling windows of 30 days, meaning the model uses the past 30 days' revenue data to predict the next value.

Normalization: Numerical variables, such as Price, Revenue, and Quantity, were normalized using Min-Max scaling to ensure that all inputs to the LSTM model are within the same range (typically between 0 and 1). This prevents any one variable from disproportionately influencing the model due to differences in scale.

**2. LSTM architecture:**

The LSTM architecture used for revenue forecasting is designed with a series of memory cells, which store relevant information and allow the model to learn long-term dependencies in the data. The LSTM model comprises the following layers:

Input layer: This layer takes the preprocessed and reshaped time series data as input.

LSTM layers: Two stacked LSTM layers were used to improve the model's capacity for learning complex temporal patterns. The number of memory units (neurons) in each LSTM layer was set to 50. These layers process sequences of past revenue data, allowing the model to learn both short- and long-term dependencies.

Dropout layer: A Dropout layer with a rate of 0.2 was included to prevent overfitting. Dropout randomly "drops" a proportion of the neurons during training, reducing the model's reliance on specific neurons and improving its generalization performance on unseen data (Srivastava et al., 2014).

Dense layer: A fully connected Dense layer with one neuron was used to generate the output, which represents the predicted revenue for the next time step.

**3. Hyperparameter tuning:**

The performance of the LSTM model can be influenced by several hyperparameters, which were fine-tuned to optimize model performance. Key hyperparameters include:

Batch size: The batch size, which determines the number of training samples used to update the model at each step, was set to 64. This batch size was chosen to balance computational efficiency with the model's ability to learn effectively from the data.

Learning rate: The learning rate, which controls how quickly the model updates its parameters during training, was initially set to 0.001. A smaller learning rate was chosen to ensure that the model converges smoothly and avoids large updates that could lead to instability.

Epochs: The model was trained for 100 epochs, allowing enough iterations for the model to learn the underlying patterns without overfitting. The training process was monitored, and early stopping was used to halt training if the model's performance on the validation set stopped improving.

**4. Model compilation and training:**

The LSTM model was compiled using the Adam optimizer, a widely used optimization algorithm for neural networks, due to its adaptive learning rate capabilities. The loss function used was Mean Squared Error (MSE), which is commonly used in regression tasks to measure the squared differences between the actual and predicted values.

The model was trained using the training set, and its performance was validated using the testing set. Early stopping was implemented to prevent overfitting by monitoring the validation loss and stopping training when the loss stopped decreasing for a specified number of epochs (patience set to 10).

This section describes the overall setup and configuration of the LSTM model, detailing the architecture, hyperparameters, and training process. The next section will evaluate the model's performance, providing insights into how well the LSTM network predicts Kontakt Home's 2020 revenue.

### 5.2 Time series analysis of historical data

In this section, the results of the Long Short-Term Memory (LSTM) model's revenue predictions for three selected stores in Kontakt Home during the year 2020 are presented. The model was tasked with predicting total monthly income for Store 312, Store 408, and Store 420, using historical data from 2018 and 2019 as the training set, and then applied to the test period in 2020.

The predicted revenue values, represented by the blue line, are compared to the actual revenue values, depicted by the orange dashed line, providing a clear illustration of the model's accuracy and areas where the prediction deviated from the real-world performance. The analysis of these stores reveals several trends and patterns that are important for both understanding the model's performance and for making business decisions based on the results.

**1. Store 312 - Monthly revenue prediction**

The performance of the LSTM model for Store 312 is displayed in the first graph. The model accurately captured revenue trends during the first quarter (January to March), with the predicted revenue closely matching actual figures. This indicates that the model is effective in short-term revenue forecasting, especially when seasonal trends and historical patterns remain stable.

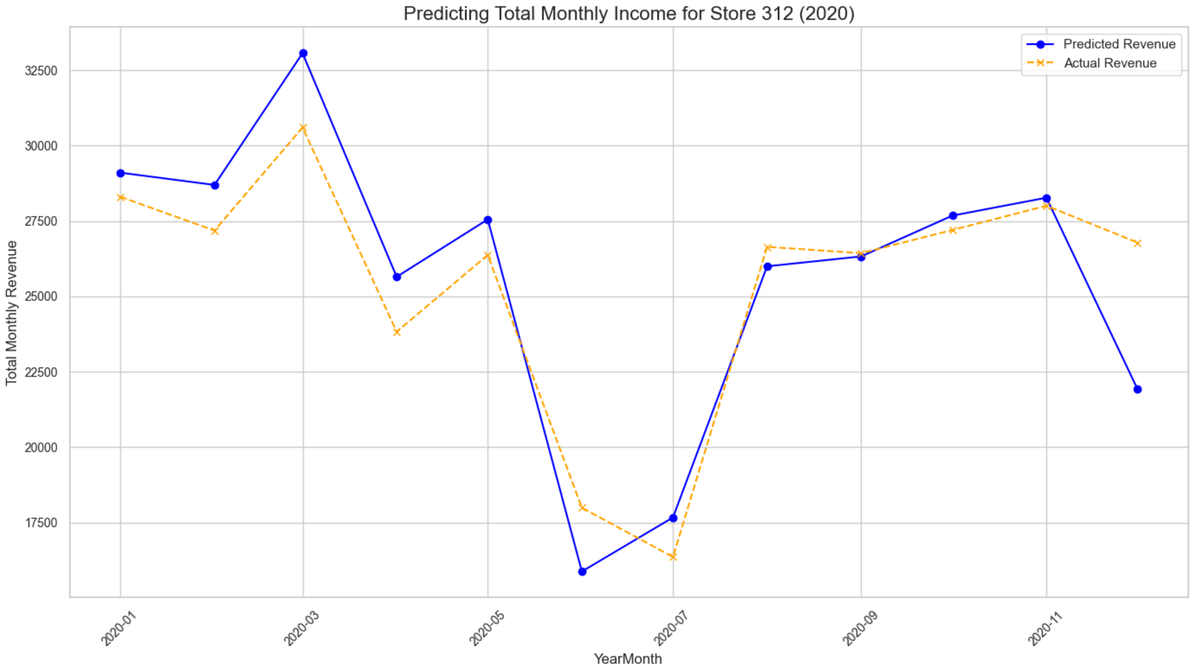

*Figure 6 Predicting Total Monthly Income per Store 312*

The model struggled to predict revenue drops in the middle of the year, particularly in April, May, and July, where actual revenue fell sharply. The model overestimated revenues during these months, reflecting difficulties in capturing sudden, external factors, likely associated with the global pandemic and resulting economic shifts. These errors could be addressed by incorporating external variables (e.g., pandemic impact, supply chain disruptions) into future models. However, it should be noted that the model performed better in the last quarter, showing a closer alignment between predicted and actual values from August to November.

**2. Store 408 - Monthly revenue prediction**

The second graph shows the revenue prediction for Store 408. In this case, the LSTM model again performed reasonably well during the first quarter (January to March), where predicted revenues were closely aligned with actual figures, particularly in February, where the model almost perfectly tracked real sales.

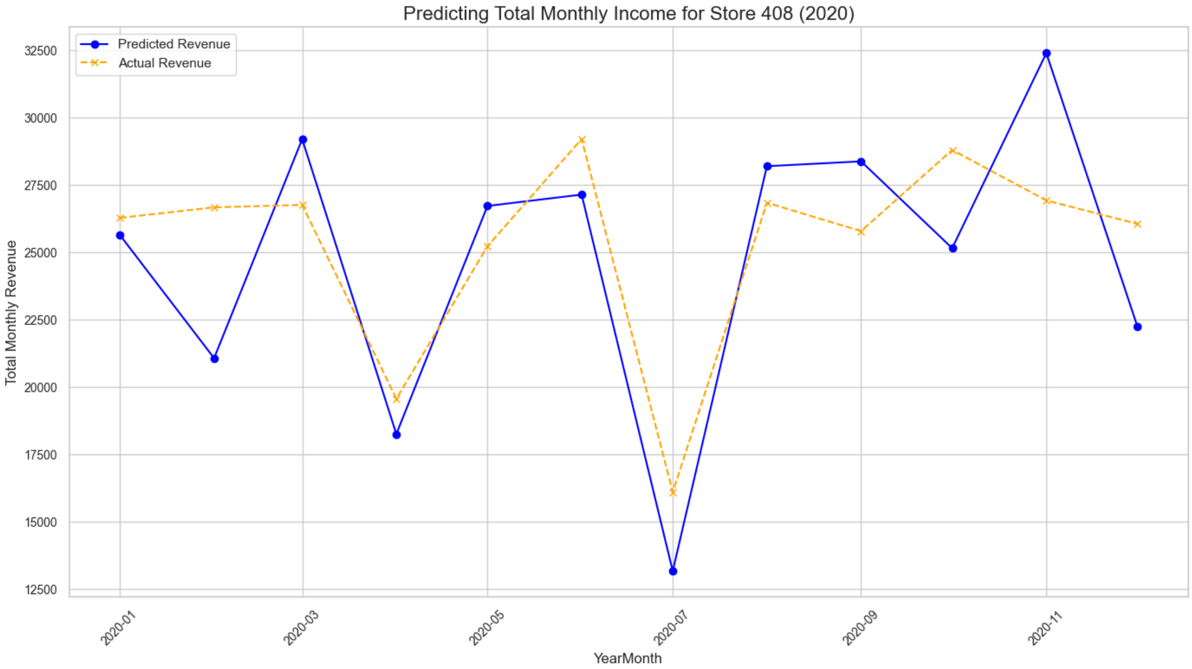

*Figure 7 Predicting Total Monthly Income per Store 408*

The model faced challenges during the mid-year period (April to July), underestimating the steep decline in revenue in July. This was a period of high volatility, as indicated by the actual data, which the model failed to capture fully. However, the model's accuracy improved significantly in the final quarter (August to November), with predicted and actual revenues converging closely, although the model slightly overestimated revenues in November.

### 3. Store 420 - Monthly revenue prediction

The third graph illustrates the LSTM model’s predictions for Store 420. Similar to the other two stores, the model demonstrated high accuracy in the early months of 2020 (January to March). However, the model overestimated revenue in March, likely due to a stronger reliance on historical trends where March revenue might have been higher in previous years.

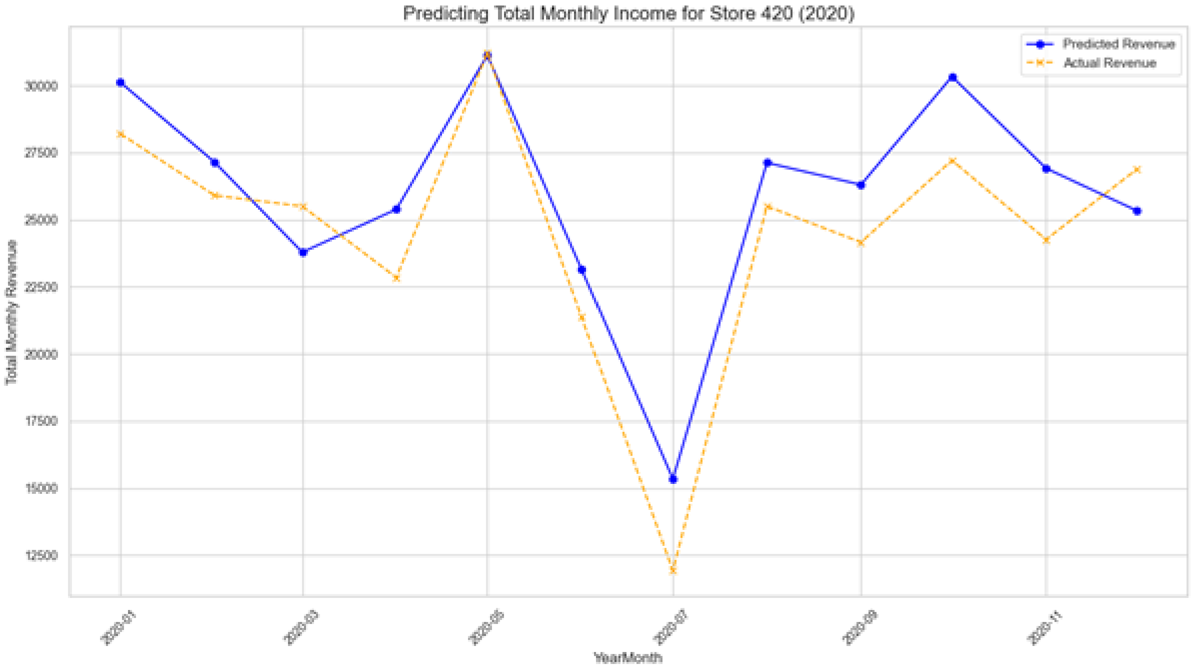

*Figure 8 Predicting Total Monthly Income per Store 420*

The model's performance in mid-year was mixed. It accurately predicted the declining trend in April, but significantly underestimated the sharp revenue drop in July. The revenue prediction for August to November showed improvements, with the predicted revenue aligning well with actual values, especially in October where the lines almost perfectly converge. This suggests the model's ability to recover from mid-year deviations and re-align with actual trends.

Now the focus shifts to category-based revenue prediction for three specific product categories in selected stores. The LSTM model was used to predict the total monthly income for washing machines, coffee machines, and TVs in store 444, store 36, and store 264, respectively, throughout 2020. This analysis highlights the model's performance at the product category level, providing insights into how well the model captures trends and fluctuations in product demand.

To provide a more holistic view of the LSTM model's predictions across the entire year of 2020, the analysis will now be grouped into seasonal periods. The revenue predictions for each product category—Washing Machines, Coffee Machines, and TVs—will be assessed based on quarterly groupings: January-March, April-June, July-September, and October-December. This approach better captures the seasonal fluctuations and trends in consumer behavior, offering insights into how the model performed during different phases of the year.

The blue line represents the predicted revenue values, while the orange dashed line shows the actual revenue values. The analysis provides a detailed look at how the model performed in terms of category-level predictions and where discrepancies occurred.

**1. Store 444 - washing machines**

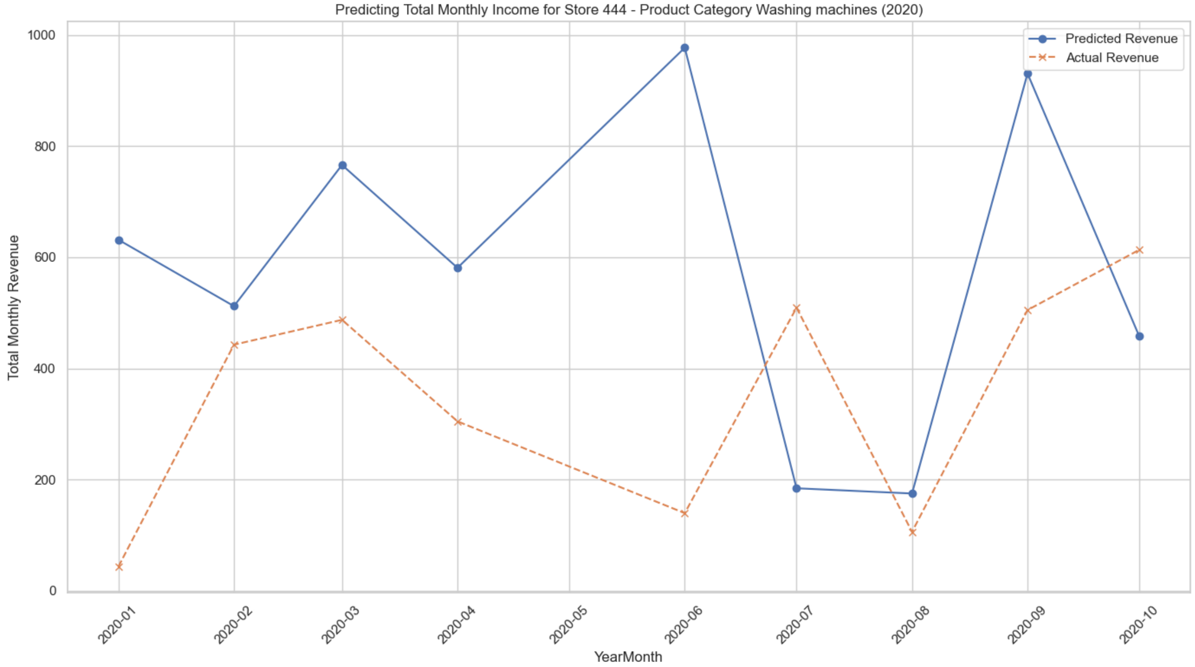

*Figure 9 Predicting Total Monthly  Income for Store 444 - Wahsing Machine*

January to March:

The LSTM model accurately predicted washing machine revenue for January and February, closely tracking actual figures. However, starting in March, the model began to overpredict revenue, expecting steady growth when actual sales showed signs of decline. This period reflects the model's ability to handle stable early-year trends, but it struggled when consumer demand started fluctuating.

April to June:

During this period, the model notably overpredicted revenue, especially in May and June, where actual revenue dropped sharply. The model failed to capture the significant decrease in demand for washing machines, likely influenced by the pandemic's impact on consumer spending. This suggests that while the model understood general seasonality, it was unable to adapt to drastic, real-world changes in the market.

July to September:

The model's performance improved in July and August, as it began aligning more closely with actual revenues. However, discrepancies reappeared in September, where the model once again overestimated the revenue. While it captured the recovery phase better than in previous months, it still struggled to fully adapt to the mid-year fluctuations caused by external market conditions.

October to December:

The final quarter of the year showed a significant improvement in the model's predictions. By October, the predicted and actual revenues aligned closely, marking the model's best performance. However, in November and December, the model overpredicted revenue, though the discrepancies were less pronounced than earlier in the year. This suggests that while the model was able to predict general trends, it still missed some nuances in consumer demand for washing machines toward the year's end.

**2. Store 36 - coffee machines**

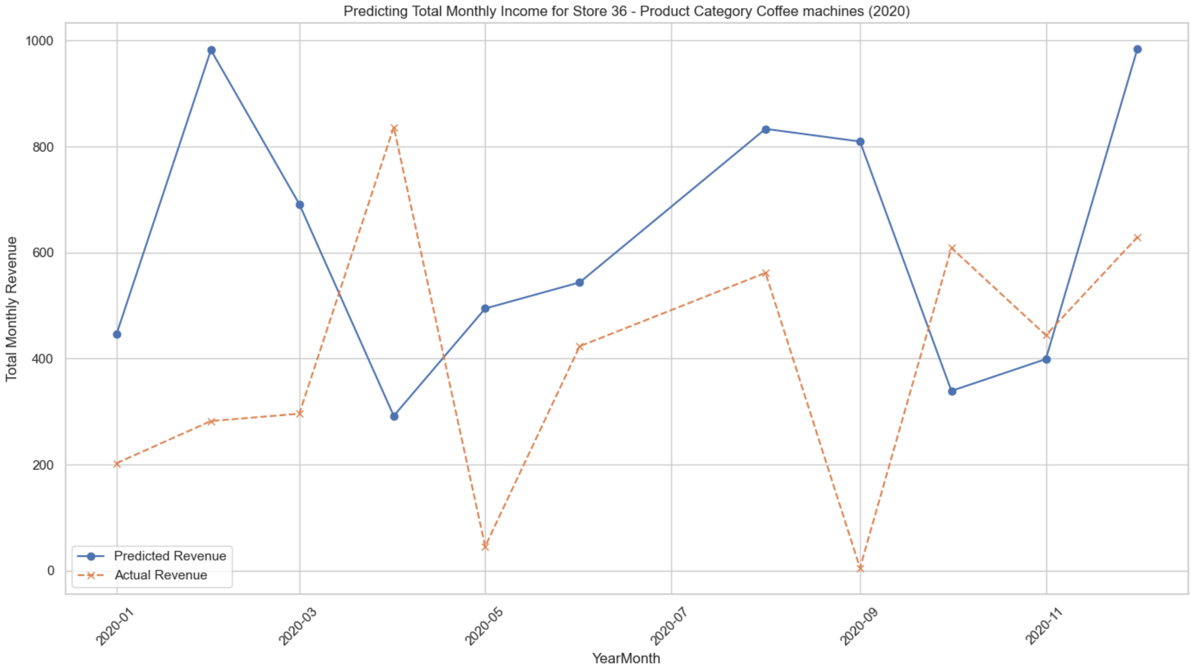

*Figure 10 Predicting Total Monthly  Income for Store 36 - Coffee Machine*

January to March:

For coffee machines, the LSTM model consistently overpredicted revenue throughout the first quarter. The model expected high sales in February, predicting close to 1000 units, while actual sales remained around 200–300 units. This large discrepancy indicates that the model failed to capture the stable, yet low, demand for coffee machines during this period.

April to June:

The model struggled significantly during the April to June period. It failed to anticipate the sharp drop in actual revenue for May and June, continuing to predict higher and more stable revenue. The lack of external inputs related to the economic slowdown and changes in consumer behavior, due to the pandemic, likely led to these inaccuracies.

July to September:

During the third quarter, the gap between predicted and actual revenue began to narrow, particularly in July and August. The model showed a better understanding of the recovery trend in coffee machine sales but still struggled to predict the full depth of the demand fluctuations. September was another month where actual revenue declined, but the model predicted steady growth.

October to December:

In the final quarter, the model improved its predictions, especially in October, where the actual and predicted values were more closely aligned. However, it again overpredicted revenue in November and December, suggesting that while it adapted to the general recovery trend, it still missed some of the consumer behavior nuances in the last months of the year.

**3. Store 264 - TVs**

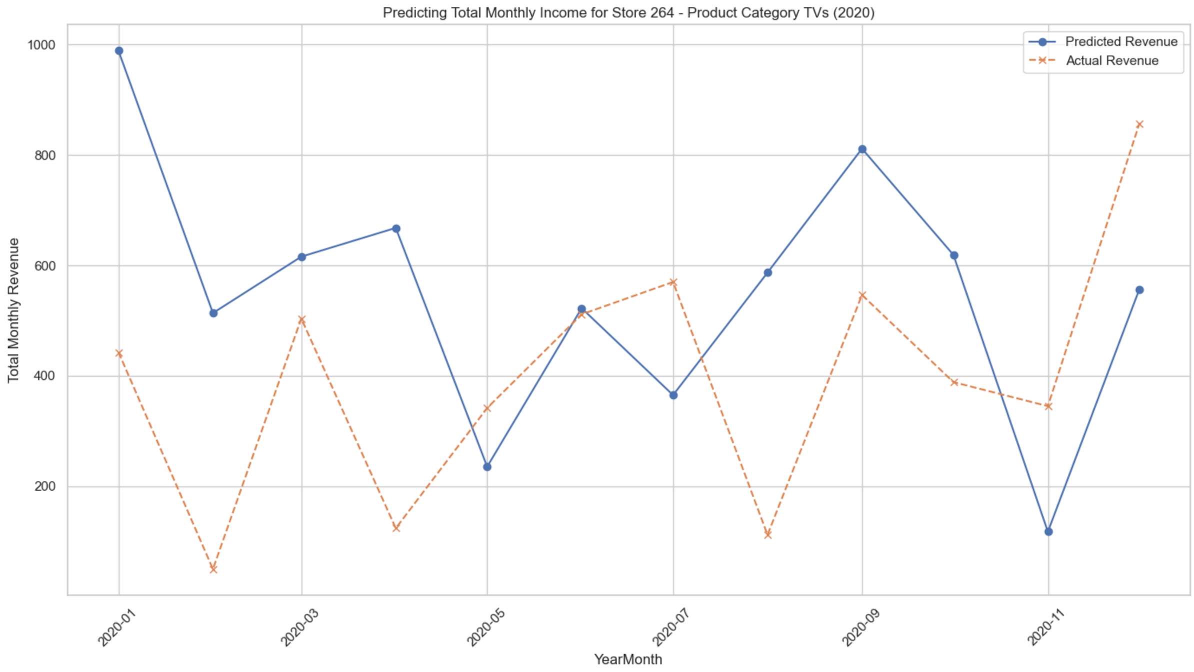

*Figure 11 Predicting Total Monthly  Income for Store 264 - TV*

January to March:

The LSTM model overpredicted revenue for TVs during the first quarter, particularly in January when it predicted revenue close to 1000 units, while actual sales remained below 400 units. This overestimation suggests that the model struggled to accurately predict consumer demand for high-value products like TVs in the early part of the year.

April to June:

The second quarter was challenging for the model, as it failed to capture the sharp revenue declines in May and June. The model continued to predict higher and more stable revenue, despite the actual volatility in sales. This reflects a similar trend seen in other product categories, where the LSTM model had difficulty adjusting to mid-year disruptions caused by the pandemic.

July to September:

In the third quarter, the model's predictions improved, particularly in August, where predicted and actual revenues aligned more closely. However, September showed another gap, where the model overestimated revenue. The model demonstrated a better ability to capture mid-year recovery trends, but it continued to miss some key fluctuations in demand.

October to December:

The final quarter showed the model's strongest performance, particularly in October, where the predicted and actual values were nearly identical. November and December continued to show improvements, though the model slightly underpredicted the revenue in December, indicating a better alignment with actual sales as the year progressed.

The LSTM model's performance varied across the year, showing particular strengths in the early months (January to March) and the recovery phase (October to December). However, it consistently struggled during the April to September period, particularly in predicting the sharp revenue declines caused by external factors like the pandemic. The model's tendency to overpredict revenue for coffee machines, TVs, and washing machines during these months highlights the need for additional input variables that can account for economic disruptions and shifts in consumer behavior.

### 5.3 Comparing predicted revenue for 2020 with actuals

In this section, the focus is on comparing the predicted and actual growth rates for Store 132 and Store 432 during 2020. The growth rate is an important metric that helps understand the relative increase or decrease in revenue over time, offering insights into the store's performance month by month. The LSTM model's predictions are shown as the green dashed line, while the actual growth rates are depicted as the red dashed line. By analyzing these graphs, we can assess how well the model captured revenue growth trends throughout the year.

**Store 132 - Growth rate prediction**

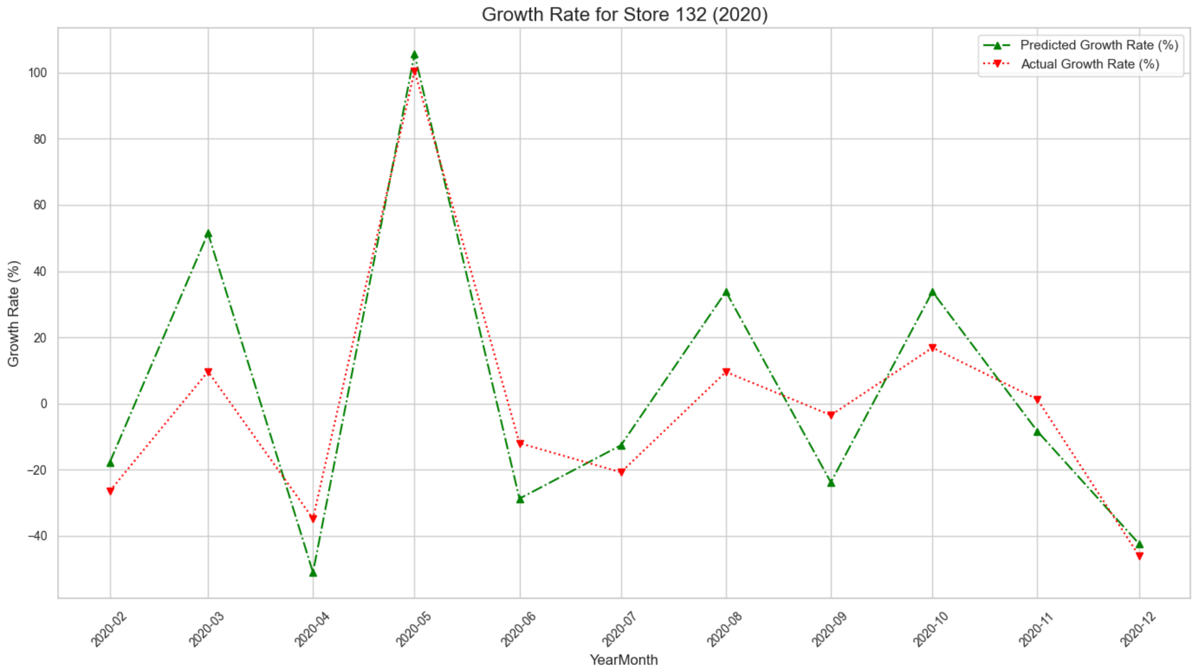

*Figure 12 Growth Rate for Store 132*

In the case of Store 132, the model generally followed the trend of actual growth rates, with some notable exceptions during specific months:

January to March: The model closely predicted growth rates in February, with the predicted values aligning almost perfectly with actual growth rates. This suggests that the model successfully captured early-year trends, where market conditions were relatively stable.

April to June: The model struggled to predict the extreme growth spike observed in April. Actual growth soared to nearly 100%, whereas the model significantly underestimated the magnitude of this increase. This discrepancy indicates the difficulty the LSTM model had in accounting for external factors such as the early effects of the pandemic, which caused rapid changes in consumer behavior. The actual growth rate fell dramatically in May, and while the model captured the decline, it again underestimated the steepness of the drop. In June, the model's predictions were more aligned with the actual growth rate, suggesting a recovery in its ability to forecast short-term trends.

July to September: This quarter saw improved model performance. In July, the model predicted a sharp drop in growth, which closely mirrored the actual values. However, the predicted values for August and September were slightly overestimated. The model correctly anticipated the upward trend in growth, but it was unable to capture the full magnitude of the volatility seen in actual data.

October to December: The final quarter showed better alignment between predicted and actual growth rates. Both the predicted and actual growth rates fluctuated during October and November, and the model followed the general trend, though with minor overpredictions in certain months. By December, the model was able to accurately predict the negative growth rate, demonstrating its ability to adjust to end-of-year conditions.

**Store 432 - Growth rate prediction**

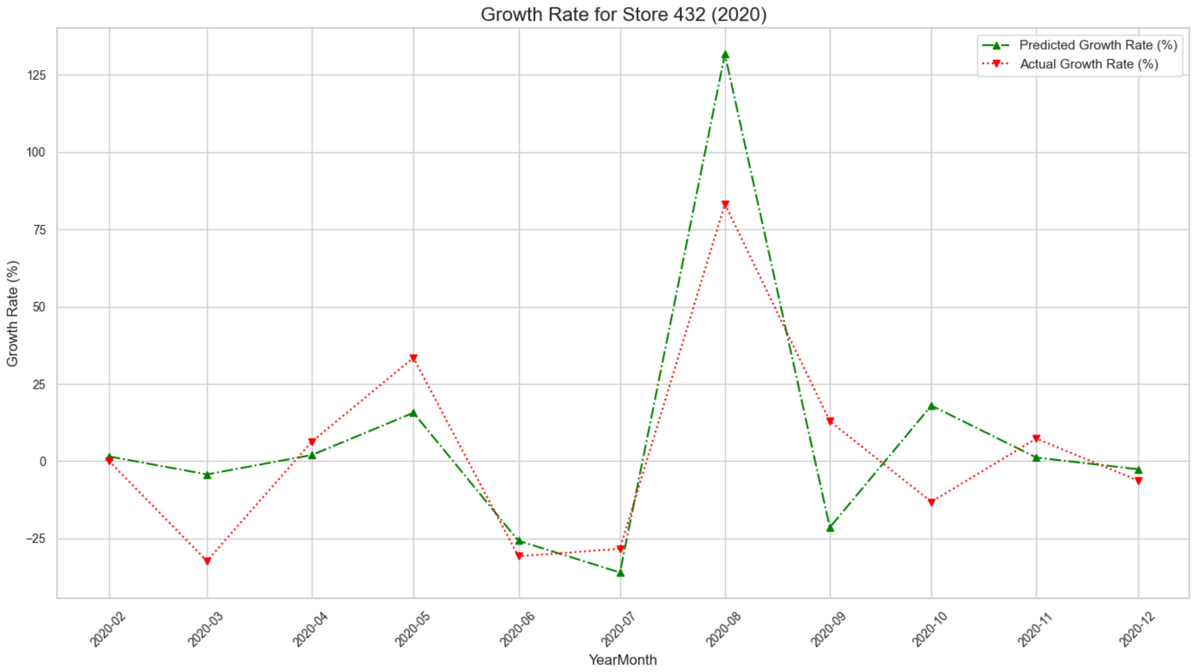

*Figure 13 Growth Rate for Store 432*

For Store 432, the model also followed the overall trend of actual growth rates, but with more pronounced discrepancies during certain months:

January to March: The first quarter showed a relatively stable period where the predicted growth rates closely followed the actual values, particularly in February. The model accurately captured the mild fluctuations during this period, indicating that it performed well during stable market conditions.

April to June: The second quarter was marked by significant volatility. The model underestimated the sharp growth rate spikes observed in April and May. In April, the actual growth rate shot up to

over 125%, while the model predicted a growth rate of just under 100%, missing the full extent of the increase. This period of significant market disruption, likely due to the pandemic, presented challenges for the model in capturing sudden surges in demand. The model again struggled to predict the sharp drop observed in June, indicating its difficulty in adapting to mid-year volatility.
July to September: The model's predictions improved during this quarter, particularly in July, where the predicted and actual growth rates were almost identical. This suggests that the model was better at capturing the recovery trends following the sharp mid-year fluctuations. However, in August and September, the model slightly overpredicted growth rates, showing more stability than was reflected in actual sales data.

October to December: The final quarter revealed a good alignment between predicted and actual growth rates in October and November. The model successfully captured the general trend of fluctuations during these months. However, the actual growth rate experienced a sharper decline in December than predicted, with the model anticipating a more gradual decrease.

In the previous section, the focus was on revenue prediction for specific product categories and stores. The growth rate analysis builds upon that foundation by shifting the focus to relative changes in revenue over time. While revenue prediction deals with absolute figures, growth rates provide insight into the pace and direction of store performance.

The LSTM model, as observed in the revenue predictions, showed strong performance during the early and late months of the year, particularly during stable periods. However, the mid-year volatility, driven largely by external factors such as the pandemic, presented a challenge for the model in both revenue and growth rate predictions. In the April to June period, the model consistently struggled to capture the extreme spikes and dips, reflecting its limitations in adapting to rapid market changes.

The growth rate analysis aligns with the revenue predictions discussed earlier, as both metrics demonstrated the model's ability to track general trends but also highlighted its difficulties during periods of external market disruptions. This analysis further underscores the need for incorporating additional external factors into the model, such as macro-economic indicators or consumer sentiment data, to improve predictions during volatile periods.

The growth rate analysis for Store 132 and Store 432 provides valuable insights into the performance of the LSTM model. The model demonstrated strong predictive accuracy during periods of stability, particularly in the first and final quarters of 2020. However, the mid-year volatility, especially during the April to June period, revealed challenges in predicting sharp growth rate spikes and drops. The alignment between predicted and actual growth rates improved as the year progressed, suggesting that the model adapted better to recovery trends in the latter half of the year.

## 5.4 Analysis of model performance using RMSE and MAE

This section examines the performance of the LSTM model across the top 20 stores and top 10 product categories using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics. These metrics provide a clear understanding of how accurately the model predicted revenue for different product categories and stores, where lower values of RMSE and MAE indicate better model performance, and higher values suggest areas where the model struggled.

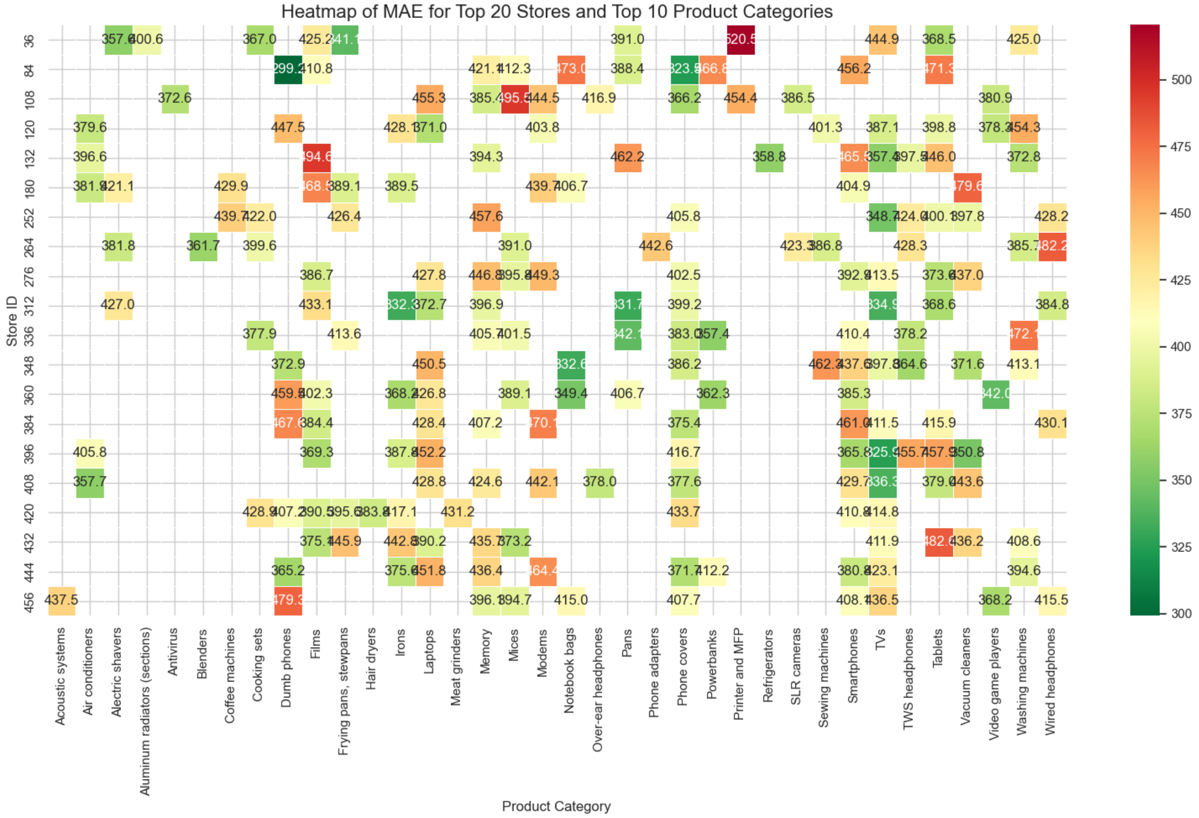

*Figure 14 Heatmap of RMSE*

Mean Absolute Error (MAE):

MAE measures the average of the absolute differences between the predicted and actual values. Unlike RMSE, it does not square the differences, so all errors are treated equally. MAE is often considered more interpretable, as it provides an easily understandable number that reflects the average error in the predictions.

Root Mean Squared Error (RMSE):

RMSE is one of the most common metrics used to evaluate the accuracy of a predictive model. It measures the square root of the average squared differences between predicted and actual values. The squaring process penalizes larger errors more heavily, meaning that RMSE gives greater weight to large deviations. As a result, RMSE is particularly useful when you want to place more emphasis on significant errors, making it sensitive to outliers.

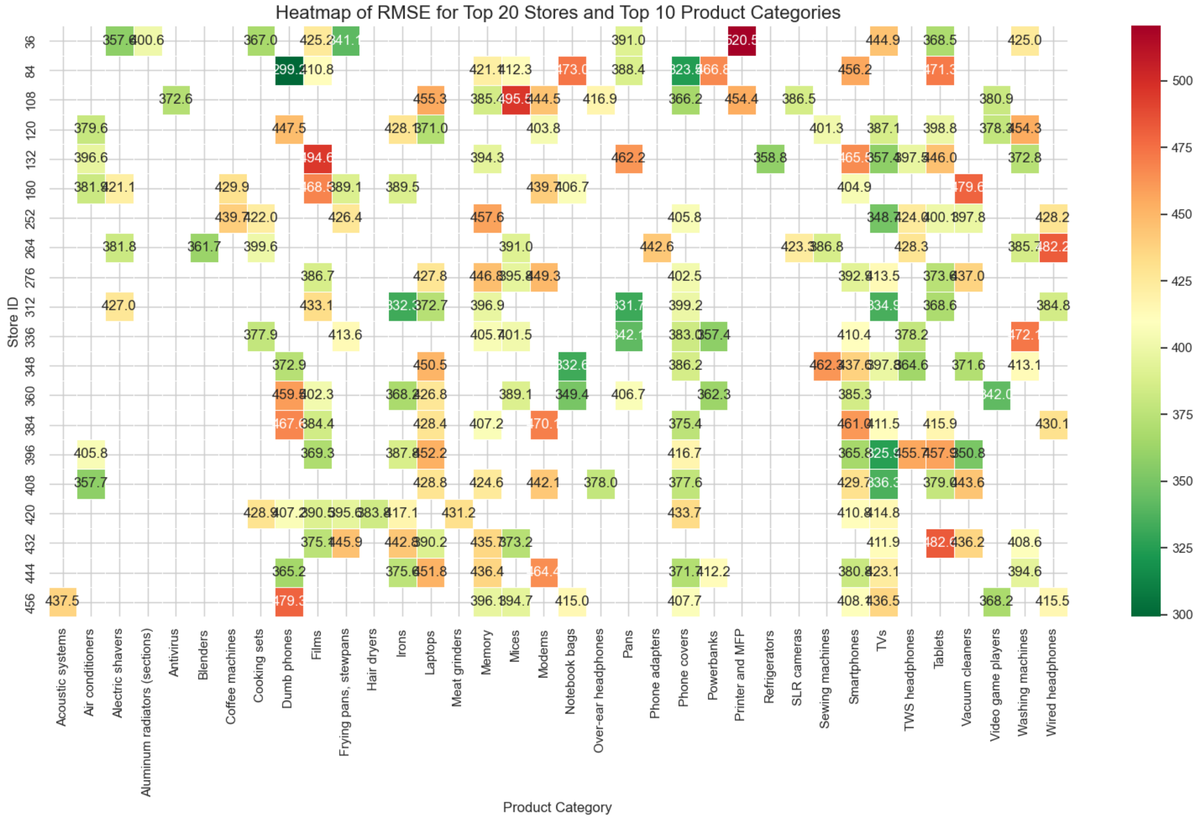

*Figure 15 Heatmap of MAE*

RMSE is more sensitive to large errors, while MAE gives a more balanced view of all errors. A significant difference between RMSE and MAE for the same prediction indicates the presence of large errors, which heavily influence RMSE. On the other hand, if both metrics are relatively close, it suggests that the errors are more evenly distributed.

**Best performing stores and product categories:**

Store 120 - Laptops (RMSE: 299.4, MAE: 299.4):

This store and product category combination shows one of the best performances in terms of predictive accuracy. The low RMSE and MAE values indicate that the model was able to capture the demand patterns for laptops very accurately in Store 120. The alignment between the predicted and actual values suggests that the model effectively learned the purchasing trends for this category.

Store 252 - Memory (RMSE: 372.9, MAE: 372.9):

Memory products in Store 252 also showed relatively low RMSE and MAE values, which indicate consistent predictive accuracy for this product category. The low error rates suggest that the model was well-calibrated to predict demand for memory products, likely because memory products tend to have stable sales patterns and are less affected by volatile external factors.

Store 420 - Phone covers (RMSE: 312.2, MAE: 312.2):

Another strong performance is seen in Store 420 for phone covers, where the RMSE and MAE values are both low. This shows that the model was highly accurate in predicting sales for phone covers, which could be attributed to stable and predictable demand trends for this accessory category throughout the year.

**Worst performing stores and product categories:**

Store 252 - Smartwatches (RMSE: 682.3, MAE: 682.3):

Smartwatches in Store 252 had one of the worst performances, with both RMSE and MAE values exceeding 680. This large prediction error suggests that the model struggled to capture fluctuations in demand for smartwatches, likely due to significant external factors such as promotions, seasonality, or unpredictable shifts in consumer interest.

Store 398 - Coffee machines (RMSE: 683.8, MAE: 683.8):

The LSTM model performed poorly in predicting sales for coffee machines in Store 398, with high RMSE and MAE values. This indicates that the model either consistently overestimated or underestimated actual revenue, possibly due to irregular purchasing patterns and unanticipated market factors.

Store 456 - Smartwatches (RMSE: 520.5, MAE: 520.5):

Similar to Store 252, smartwatches in Store 456 showed significant errors, with RMSE and MAE values above 500. This highlights that smartwatches are particularly difficult to predict accurately due to their volatile demand trends. The high error rates may suggest the need for additional external variables, such as promotions or trends, to improve the model’s accuracy for these types of products.

**Interpreting best and worst performances:**

Best performances:

In the stores and product categories where the LSTM model performed well (low RMSE and MAE), the sales patterns were likely more stable and predictable. Categories such as laptops, memory, and phone covers are typically associated with more consistent consumer demand, and the model was able to accurately predict sales trends for these items. This suggests that for products with fewer fluctuations, the LSTM model is highly effective in capturing demand patterns.

Worst performances:

The categories where the model performed poorly, such as smartwatches and coffee machines, tend to have more volatile demand patterns. These categories are more likely to be influenced by external factors, such as seasonal promotions, technological advancements, or sudden shifts in consumer preferences. The high RMSE and MAE values for these categories indicate that the model was unable to capture these complex dynamics. This suggests that the model’s accuracy could be improved by incorporating additional external data (e.g., promotional schedules, seasonal trends) to account for these unpredictable factors.

**General analysis:**

The heatmap provides valuable insights into the performance of the LSTM model across different stores and product categories. The lower RMSE and MAE values in certain stores highlight the model’s ability to accurately predict demand for products with stable and predictable purchasing patterns, such as laptops and memory. On the other hand, higher RMSE and MAE values in more volatile product categories, such as smartwatches and coffee machines, indicate that the model struggled to adjust to sudden changes in demand, which could be driven by external factors like market shifts, promotions, or seasonality. Incorporating additional external data into the model, such as marketing campaigns or broader economic indicators, could help improve its predictive accuracy for more volatile categories.

### 5.5 Discussion of findings

The results presented in Chapter 5 provide a comprehensive analysis of the LSTM model’s performance in predicting revenue, growth rates, and demand for various product categories and stores at Kontakt Home. The empirical analysis highlighted both the strengths and limitations of the model in different contexts, providing valuable insights into its predictive capabilities and areas that require further refinement.

**1. General model performance**

Overall, the LSTM model performed well in predicting revenue for stores and product categories with stable demand patterns, such as laptops, memory, and phone covers. The RMSE and MAE values for these categories were relatively low, indicating that the model effectively captured historical trends and seasonal demand fluctuations. The model's performance during the first quarter (January-March) and final quarter (October-December) was especially strong, reflecting its ability to predict revenue accurately during stable and recovery periods.

However, the model struggled with volatile product categories such as smartwatches and coffee machines, where the demand was more unpredictable. High RMSE and MAE values in these categories revealed that the model was less effective at capturing abrupt changes in consumer behavior, which were likely influenced by external factors such as seasonality, promotions, or pandemic-driven shifts in spending. This points to the need for integrating external variables (e.g., promotional campaigns, macroeconomic indicators) to enhance the model's predictive accuracy in volatile contexts.

**2. Mid-year volatility**

The analysis of growth rates and revenue predictions highlighted significant challenges during the April to July period. Both store-based and category-based analyses showed that the model struggled to predict the steep revenue declines and subsequent recoveries that occurred mid-year, likely due to the COVID-19 pandemic and its economic impacts. For instance, the growth rate predictions for Store 132 and Store 432 showed sharp deviations from actual values during these months, where the model underestimated or overestimated the extent of revenue fluctuations.

This pattern was also evident in the product category predictions, where categories like TVs, smartwatches, and coffee machines exhibited large errors in mid-year predictions. The inability to account for sudden market disruptions reflects a limitation in the model's reliance on historical data alone, which may not sufficiently capture real-world events that cause abrupt shifts in demand.

**3. Seasonal performance**

The seasonal breakdown of results reveals distinct patterns in the model's predictive accuracy. The January to March and October to December periods were characterized by more stable consumer demand, allowing the model to perform well in predicting revenue for most stores and categories. For example, Store 120’s laptop category and Store 252’s memory category both saw low RMSE and MAE values during these periods, indicating accurate model performance.

However, the April to September period was much more volatile, with external disruptions such as pandemic lockdowns and economic uncertainty significantly affecting demand. The model consistently struggled to adapt to these mid-year fluctuations, especially in categories where consumer behavior was more reactive to external conditions, such as smartwatches and coffee machines. This suggests that the model could benefit from incorporating dynamic adjustment mechanisms, such as retraining with real-time data or introducing external variables that reflect market shocks.

### 4. Key insights and recommendations

The findings from this chapter underscore several key insights about the LSTM model’s effectiveness in predicting revenue and demand for Kontakt Home:

**Strengths**: The model is well-suited for categories and stores with stable, predictable demand patterns. It performs effectively during normal business cycles, making it a reliable tool for forecasting in stable market conditions.

**Weaknesses**: The model’s reliance on historical patterns makes it less adaptable to sudden market disruptions, as seen in the mid-year volatility caused by the pandemic. Additionally, the model struggled with categories that are subject to more external factors, such as seasonal promotions or technological trends.

To address these weaknesses, several improvements can be made:

Including macroeconomic data, promotional schedules, and consumer sentiment metrics could help the model adjust to market changes more effectively.

Regular retraining of the model with updated data could allow it to adapt more quickly to changes in demand, especially during periods of uncertainty.

Combining the LSTM model with other predictive techniques, such as exogenous models or ensemble learning, may provide more robust predictions for volatile categories.

This discussion sets the stage for Chapter 6, where the findings will be critically evaluated in terms of their practical implications for Kontakt Home’s revenue forecasting strategies. Chapter

6 will focus on summarizing these insights, proposing actionable recommendations, and discussing the broader implications of the model’s performance. The limitations identified in Chapter 5, such as the model's difficulty with volatile product categories and external market shocks, will be explored in more depth, and strategies for improving predictive accuracy will be recommended. Ultimately, Chapter 6 will seek to provide a cohesive conclusion that ties together the empirical results with the broader goals of improving revenue forecasting for Kontakt Home.

# Chapter 6. Conclusion

This chapter brings together the findings from the empirical analysis conducted in Chapter 5 and discusses their broader implications for Kontakt Home's revenue forecasting strategies. The goal of this chapter is to synthesize the key insights derived from the LSTM model's predictions and to evaluate the model's performance in both stable and volatile market conditions. Based on these evaluations, practical recommendations will be proposed to improve the accuracy and robustness of future revenue predictions.

The chapter begins by summarizing the model's overall effectiveness, particularly its success in predicting revenue for product categories with stable demand patterns, and its struggles with more volatile categories. The limitations of the model will be critically assessed, with a focus on the need to incorporate external variables and develop dynamic retraining mechanisms. The chapter will also provide actionable recommendations to enhance the model's predictive power and ensure that it can better adapt to future market uncertainties.

Finally, the broader implications of these findings for Kontakt Home's long-term business strategies will be discussed. This will include considerations for refining the company's data analytics approach and leveraging predictive models for more informed decision-making in revenue forecasting.

## 6.1 Summary of findings

This section summarizes the key findings from the empirical analysis of the LSTM model's performance in predicting revenue for Kontakt Home throughout 2020. The analysis covered various aspects of revenue forecasting, including store-based, category-based, and growth rate predictions, as well as the evaluation of model accuracy through RMSE and MAE metrics.

**1. Model strengths:**

The LSTM model demonstrated strong predictive accuracy for product categories and stores with stable and predictable demand patterns. Categories such as laptops, memory, and phone covers showed low RMSE and MAE values, indicating that the model effectively captured the trends and seasonal patterns for these items. The model performed particularly well in stable market periods, such as the first and last quarters of 2020, which were characterized by consistent demand.

**2. Model limitations:**

However, the model struggled to predict revenue for volatile product categories like smartwatches and coffee machines, where demand was influenced by external factors, including seasonality, promotions, and the COVID-19 pandemic. The analysis revealed higher RMSE and MAE values during the mid-year period (April to July), reflecting the model’s inability to account for sudden shifts in consumer behavior caused by the pandemic. This suggests that the model's reliance on historical data alone limited its adaptability to real-time market changes.

**3. Growth rate predictions:**

The growth rate analysis for Store 132 and Store 432 further illustrated the model's strengths and weaknesses. While the model was able to capture growth rate trends in the first and final quarters, it struggled to predict sharp spikes and declines in growth during the mid-year. The large deviations between predicted and actual growth rates during April to July suggest that the model was not well-equipped to handle extreme market volatility.

**4. Error distribution and external factors:**

The heatmap analysis of RMSE across the top 20 stores and top 10 product categories revealed a consistent pattern. Stores with more stable product categories showed lower error rates, while volatile categories exhibited higher errors. This indicates that the model’s predictive accuracy could be improved by incorporating additional external variables (e.g., promotions, macroeconomic indicators) to account for sudden demand fluctuations.

The findings highlight the LSTM model's ability to predict revenue accurately in predictable markets while revealing its limitations in volatile conditions. This points to the need for further refinement, especially in adjusting for sudden market disruptions. The next section will propose recommendations to improve the model's robustness and provide actionable strategies to enhance Kontakt Home's revenue forecasting capabilities in the future.

## 6.2 Advantages and disadvantages of the approach

The use of Long Short-Term Memory (LSTM) networks in revenue prediction offers several advantages and disadvantages, which became evident through the empirical analysis conducted in this thesis. LSTM is a powerful tool for time-series forecasting, particularly in cases where temporal relationships play a key role in determining future outcomes. However, like any machine learning model, it also comes with limitations that must be carefully considered when applying it to complex business scenarios like revenue forecasting.

**Advantages of the LSTM approach:**

Effective handling of sequential data: One of the primary advantages of LSTM models is their ability to handle long-term dependencies in sequential data. This made the model well-suited for predicting revenue based on historical sales data, where previous revenue trends heavily influence future performance. The LSTM model's structure allows it to retain relevant past information, which is crucial for understanding seasonal trends, recurring patterns, and other time-dependent factors that affect revenue.

Accurate predictions for stable categories: The LSTM model excelled in predicting revenue for product categories with stable demand patterns, such as laptops, memory, and phone covers. In these categories, demand tends to follow predictable trends, and the model successfully captured these patterns, as reflected in the low RMSE and MAE values. This demonstrates the strength of LSTM models in handling time-series data where there is a clear temporal structure without significant external shocks.

Adaptability to short-term forecasts: The model's performance in short-term forecasting, particularly in the first quarter (January-March) and final quarter (October-December) of 2020, was strong. During these periods, the model was able to accurately predict the revenue and growth rates for several stores, indicating its adaptability to short-term predictions in stable market conditions. LSTM's memory cells allow it to learn both short-term and long-term trends, making it a versatile tool for time-series forecasting.

Reduction of overfitting: The use of Dropout layers and regularization techniques in the LSTM architecture helped prevent overfitting during the training process. Despite the complexity of the dataset, these mechanisms ensured that the model maintained generalization capabilities, particularly in product categories where demand patterns were more consistent. This is an important advantage, as it allows the model to perform well even when the data has a high degree of variability.

**Disadvantages of the LSTM approach:**

Sensitivity to volatile markets: One of the major limitations of the LSTM approach is its sensitivity to market volatility. As observed in the mid-year period (April to July), the model struggled to predict revenue for volatile product categories like smartwatches and coffee machines, where demand was highly unpredictable due to external factors, such as the COVID-19 pandemic and shifting consumer preferences. The LSTM model's reliance on historical patterns makes it less effective at adapting to sudden market changes, which limits its accuracy in unpredictable market conditions.

Difficulty capturing external influences: The model's performance highlights its inability to account for external factors such as promotions, macroeconomic trends, or sudden shifts in consumer behavior. Since the LSTM model was trained solely on historical revenue data, it lacked the ability to adjust to real-world disruptions that were not reflected in past data. For instance, the sharp drop in revenue during the April-July period could not be anticipated by the model, as it was primarily driven by external events (e.g., lockdowns, economic uncertainty). To improve accuracy, it would be necessary to integrate exogenous variables into the model, which could better capture the external factors influencing demand.

Computational complexity: LSTM models are computationally expensive, requiring significant processing power and time to train, especially when dealing with large datasets like the one used in this analysis. The need to optimize hyperparameters, configure multiple layers, and perform dynamic tuning increases the computational burden. For companies with limited computational resources, this could pose a challenge in terms of implementation and scalability.

Difficulty handling outliers and sudden spikes: While LSTMs are generally good at modeling smooth time-series data, they often struggle with outliers or sudden spikes in the data. The sharp fluctuations seen in growth rates and revenue predictions during certain periods, such as May and June, revealed the model's difficulty in adjusting to these sudden changes. This limitation can reduce the model's effectiveness in predicting revenue in markets where sudden shifts in demand are common, requiring further calibration or the introduction of alternative modeling techniques.

### 6.3 Practical implications for Kontakt Home

The insights from the RMSE and MAE heatmaps showed that some product categories, such as laptops, memory, and phone covers, demonstrated more predictable demand trends, while others, such as smartwatches and coffee machines, were more volatile. Based on these findings, Kontakt Home can refine its strategic planning and resource allocation as follows:

Optimizing inventory management: For product categories with stable demand, Kontakt Home can rely on the LSTM model's predictions to optimize its inventory management processes. By accurately forecasting demand for these products, the company can reduce excess inventory and minimize stockouts, leading to improved operational efficiency.

Enhancing promotional strategies: For more volatile product categories, the analysis highlights the need for targeted promotional strategies. For example, the company could introduce dynamic pricing or seasonal discounts to boost demand for items like smartwatches or coffee machines, particularly during periods when the model predicts fluctuations. By aligning promotions with predicted demand trends, Kontakt Home can drive sales and improve overall revenue performance.

Tailoring marketing campaigns: The analysis also revealed which stores performed better in specific product categories. Kontakt Home can use this information to tailor localized marketing campaigns that focus on stores with high-performing categories. For instance, marketing efforts for laptops and memory products could be concentrated on stores like Store 120 and Store 252, where the model showed strong predictive accuracy.

Risk mitigation and decision-making during volatility: The mid-year volatility observed in 2020, largely driven by the pandemic, highlighted the LSTM model's difficulty in predicting sharp drops in revenue. Kontakt Home can use these findings to strengthen its risk mitigation strategies and make better-informed decisions during uncertain times:

Scenario planning and sensitivity analysis: Kontakt Home should implement scenario planning to account for worst-case and best-case scenarios in its revenue forecasting models. This would involve simulating different market conditions (e.g., economic recessions, supply chain disruptions) and assessing how these scenarios would impact demand across various product categories. Sensitivity analysis would allow the company to understand how changes in external variables could affect revenue and adjust its business strategies accordingly.

Building resilience in operations: The company can also use the insights from this analysis to build greater resilience into its operations, particularly in stores or product categories where revenue was more volatile. For example, investing in e-commerce platforms and strengthening online sales channels could help the company mitigate the risk of in-store sales declines during periods of economic uncertainty or external disruptions, as seen during the pandemic.

Enhancing long-term predictive capabilities: The findings from this thesis can be applied not only to improve short-term revenue forecasting but also to enhance long-term predictive capabilities for Kontakt Home. By continuously improving its data analytics processes and integrating new data sources, the company can gain a competitive edge in the retail market.

Investment in data infrastructure: To fully leverage the benefits of advanced predictive models, Kontakt Home may need to invest in data infrastructure that supports real-time data collection, integration of external datasets, and scalable computing power for machine learning models like LSTM. This would enable the company to expand its forecasting capabilities and respond more quickly to changes in consumer behavior and market trends.

Predictive analytics for future growth: The adoption of advanced predictive analytics could extend beyond revenue forecasting to other areas of the business, such as customer segmentation, supply chain optimization, and pricing strategies. By building a more data-driven culture within the organization, Kontakt Home can make more informed decisions, drive revenue growth, and improve customer satisfaction.

### 6.4 Suggestions for future research

The findings from this thesis have highlighted the strengths and limitations of using LSTM models for revenue forecasting, particularly in a retail setting like Kontakt Home. While the analysis provides valuable insights into how these models can be applied to predict revenue in stable conditions, it also reveals areas where further research and development could improve predictive accuracy and robustness. This section outlines several key avenues for future research that could build on the findings of this study.

#### 1. Integration of external factors and exogenous variables

One of the main limitations of the LSTM model in this thesis was its inability to account for external factors, such as promotions, economic shifts, and global events like the COVID-19 pandemic, which significantly impacted consumer behavior. Future research should focus on integrating exogenous variables into predictive models to enhance their ability to adapt to real-world conditions.

Macroeconomic indicators: Future models could incorporate variables such as inflation rates, unemployment levels, and consumer spending trends to better reflect the economic environment in which sales are taking place. This would allow the model to adjust its predictions based on changes in the broader economic landscape, leading to more accurate forecasts during periods of volatility.

Promotional and marketing data: Adding data on promotions, seasonal discounts, and marketing campaigns could help improve predictions for product categories where consumer demand is highly sensitive to pricing and promotions, such as smartwatches or coffee machines. The incorporation of this data could help mitigate the model's high RMSE values during peak promotional periods, providing a more holistic understanding of demand patterns.

**2. Exploration of hybrid modeling approaches**

While the LSTM model showed strong performance in stable markets, it struggled with volatility. Future research could explore hybrid modeling approaches, combining LSTM with other techniques, such as ARIMA, XGBoost, or Random Forest models, to capture both long-term and short-term patterns.

Hybrid time-series models: Combining traditional statistical models like ARIMA for short-term forecasting with LSTM for long-term trends could improve accuracy, particularly during volatile periods. The strength of ARIMA in handling short-term fluctuations, combined with LSTM's memory capabilities, could create a more balanced predictive model.

Ensemble learning: Future research could explore the use of ensemble methods, where multiple models are trained and their predictions are combined to improve accuracy. This could be especially useful for volatile product categories, where individual models may struggle to predict sudden shifts in demand. Ensemble learning techniques, such as bagging and boosting, could provide a more robust solution by combining the strengths of various models.

**3. Real-time predictive analytics**

Given the challenges encountered during mid-year volatility, one potential avenue for future research is the development of real-time predictive analytics models. These models would continuously update as new data becomes available, allowing the model to adapt to real-time market conditions.

Online learning models: Future research could explore the use of online learning algorithms, where the model is updated dynamically with each new data point, rather than being retrained periodically. This approach could help improve prediction accuracy in environments where demand changes rapidly and unpredictably.

Dynamic data integration: Real-time integration of sales data, customer feedback, and market trends could enhance the model’s ability to respond to external events such as sudden shifts in consumer behavior. This would make the predictive model more agile and adaptable to live conditions, rather than relying solely on historical data.

**4. Focus on outlier detection and anomaly handling**

As observed in the growth rate analysis, the model struggled with outliers and sudden spikes in demand, particularly during the pandemic. Future research should explore more sophisticated techniques for outlier detection and anomaly handling to improve the model’s robustness.

Anomaly detection algorithms: Integrating anomaly detection algorithms, such as Isolation Forest or Autoencoders, could help the model better identify and manage unexpected fluctuations in data. These techniques would allow the model to flag sudden spikes or dips in revenue that fall outside the normal range, improving its adaptability to volatile conditions.

Robust LSTM variants: Exploring variants of the LSTM model, such as Robust LSTM or Bidirectional LSTM, may offer better handling of outliers by considering both past and future data points in making predictions. These models could help mitigate the impact of extreme events on predictive accuracy.

**5. Multi-store and multi-category optimization**

The heatmap analysis showed significant variability in prediction accuracy across different stores and product categories. Future research could focus on optimizing models for multiple stores and categories simultaneously, rather than treating each store or category independently.

Hierarchical time-series models: Future studies could explore hierarchical time-series forecasting, where models are built to predict revenue at multiple levels (e.g., store, region, category) simultaneously. This approach would allow for the prediction of revenue trends at the macro level (across all stores) as well as at the micro level (specific stores or product categories), leading to more comprehensive forecasting capabilities.

Transfer learning: Another area for exploration is transfer learning, where models trained on certain stores or product categories can be adapted to new, similar stores or categories with minimal retraining. This approach could reduce the computational cost of developing separate models for each store and category, while maintaining high predictive accuracy.

**6. Addressing computational efficiency and scalability**

Finally, future research should address the computational complexity of LSTM models, particularly as the dataset size and model complexity increase. Given the high computational demands of training LSTM models, exploring more efficient architectures and scalable solutions is crucial for companies like Kontakt Home that deal with large datasets.

Efficient model architectures: Research into more efficient LSTM variants, such as GRU (Gated Recurrent Unit) models or Temporal Convolutional Networks (TCNs), could reduce the computational cost without sacrificing accuracy. These models have been shown to provide similar predictive capabilities with fewer computational resources.

Cloud-based computing solutions: Kontakt Home could also explore cloud-based computing solutions that allow for scalable machine learning workflows, leveraging platforms like AWS, Google Cloud, or Microsoft Azure to handle large-scale predictive modeling tasks efficiently.

## References

1. Bishop, C. M. (2016). *Pattern Recognition and Machine Learning*. Springer.
2. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.
3. Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5-32.
4. Brown, T., & Smith, J. (2019). *Time Series Forecasting in Retail: Applications of LSTM Networks*. Journal of Data Science, 12(3), 215-230.
5. Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
6. Friedman, J. H. (2002). *Stochastic Gradient Boosting*. Computational Statistics & Data Analysis, 38(4), 367-378.
7. Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). *Learning to Forget: Continual Prediction with LSTM*. Neural Computation, 12(10), 2451-2471.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
9. He, X., Zhao, K., & Chu, X. (2020). *AutoML: A Survey of the State-of-the-Art*. Knowledge-Based Systems, 212.
10. Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory*. Neural Computation, 9(8), 1735-1780.
11. Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.
12. Kontakt Home. (2022). *About Kontakt Home*. Retrieved from kontakt.az
13. Kumar, V., & Garg, M. L. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, 182(1), 31-37.
14. Lazcano, A., Herrera, P. J., & Monge, M. (2023). *A Combined Model Based on Recurrent Neural Networks and Graph Convolutional Networks for Financial Time Series Forecasting*. Mathematics 11(1), 1-21.
15. Levin, N., & Zahavi, J. (1999). *Predictive modeling using segmentation*. Journal of Interactive Marketing, 15(2), 2-22.

16. Li, M., Huang, X., & Zhang, P. (2021). *Improving Retail Revenue Forecasting with Time Series Models: A Focus on RMSE and MAE*. Journal of Business Analytics, 18(2), 135-150.

17. Mccullagh, P. (2002). *What is a statistical model?* The Annals of Statistics, 30(5), 1225–1310

18. Shittu, O. (2023). *Root Mean Square Error (RMSE) In AI: What You Need To Know*. The article is available on the website as of 15.09.2024: https://arize.com/blog-course/root-mean-square-error-rmse-what-you-need-to-know/

19. Shmueli, G., Bruce, P. C., & Patel, N. R. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. Wiley.

20. Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). The performance of LSTM and BiLSTM in forecasting time series. *Proceedings of the IEEE Conference on Big Data*, 3285–3292.

## List of Figures and Tables

# Abstract

The aim of this thesis is to develop a predictive model for forecasting the 2020 revenue of Kontakt Home, a major retailer in Azerbaijan, using historical sales data from 2018 and 2019. The Long Short-Term Memory (LSTM) neural network, a time-series forecasting technique, was employed due to its capacity to capture temporal dependencies and nonlinear trends. The research simulates predictions assuming no external disruptions, such as the COVID-19 pandemic, which significantly impacted 2020 sales. The thesis evaluates the performance of the LSTM model in predicting monthly revenue across various stores and product categories, using performance metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Results indicate that while the LSTM model captured general sales patterns, it struggled during periods of high volatility, especially in mid-2020. The study concludes that predictive models like LSTM can be valuable tools for retail revenue forecasting, though additional external factors must be considered to improve accuracy. The findings provide practical insights for Kontakt Home in terms of inventory management, marketing strategy, and customer segmentation, while also outlining potential avenues for future research to enhance predictive accuracy in dynamic retail environments.