In the rapidly evolving edtech industry, accurately forecasting course sales is essential for effective resource allocation, budgeting, and strategic planning. This project aims to develop a predictive model to forecast course sales over time, leveraging various data sources such as historical sales, course attributes, marketing efforts, and external factors. This literature review surveys relevant research in sales forecasting, machine learning applications in predictive analytics, and specific studies in the edtech domain to build a foundation for this project's methodology and objectives.

**Forecasting Sales in Different Sectors**

One common technique used for predicting future revenues in many industries is time series analysis. Various studies have shown how effective ARIMA models can be at handling temporal dependency issues while dealing with seasonality patterns typically found within sales datasets across various sectors – these include retailing too. Hyndman & Athanasopoulos (2018), for example, illustrate its application when forecasting Australian retail turnover where such factors are deemed critical components towards achieving accurate predictions .

***Methods of Machine Learning for Predictive Analytics***

Machine learning has become popular due to its capability of working with large volumes of data while capturing intricate non-linear relationships among different variables. In this regard, Bohanec et al (2017) showcase how Random Forests and Gradient Boosting Machines – among other algorithms – can predict sales outcomes within the context of the retailing industry based on their study conducted in Slovenia. Traditionally, such statistical techniques tend to have limited feature spaces for making predictions but these models incorporate more explorative attributes which enable them to adapt new data trends faster thereby competing favorably against these methods .

***Ensemble Methods***

Ensemble methods, which combine predictions from multiple models to improve accuracy, have also been explored extensively. The work of Dietterich (2000) outlines how ensemble techniques like Bagging, Boosting, and Stacking can enhance predictive performance by mitigating the limitations of individual models. These approaches are particularly useful in dynamic and volatile markets where single models may fail to capture all underlying patterns.

**Predictive Analytics in Edtech**

***Historical Data and Course Attributes***

In the context of edtech, historical sales data and course attributes are critical for forecasting. A study by De Vries et al. (2019) analyzed historical enrollment data and course characteristics to predict future course demand. Their findings suggest that factors such as course content, instructor reputation, and historical enrollment trends significantly influence future sales.

***Impact of Marketing Efforts***

Marketing efforts are another crucial determinant of course sales. Research by Lewis and Rao (2015) investigated the impact of various marketing strategies on online course enrollments. Their study found that targeted marketing campaigns, including email marketing and social media advertising, positively affect enrollment numbers. By incorporating marketing data into predictive models, companies can better anticipate the outcomes of their marketing efforts and adjust strategies accordingly.

***External Factors***

External factors, such as economic conditions, technological advancements, and regulatory changes, also play a role in forecasting sales. For instance, how macroeconomic indicators like GDP growth and unemployment rates influence the demand for online education. The results indicate that during economic downturns, there is often an increased demand for skill development and online courses, as individuals seek to improve their employability .

**Machine Learning Techniques for Sales Forecasting**

***Regression Models***

Commonly applied in sales projection is linear regression among other types. These models establish connections between dependent variable(s) and multiple independent variable(s) hence making them suitable for predicting various product features based on such data sets.

***Time Series Forecasting Models***

Sales data contains time related patterns that can be captured using advanced time series forecasting models like Long Short-Term Memory (LSTM) networks. In a study conducted by Siami-Namini et al. (2018), these networks were used to predict stock prices showing their effectiveness with sequential information and long term dependencies. Therefore they are perfect for predicting course enrollments where past enrolment history significantly influences future trends .

***Hybrid Models***

To increase prediction accuracy different forecasting methods have been combined into what are known as hybrid models. According to Zhang (2003), a hybrid model that used artificial neural networks (ANN) and ARIMA was developed for time series forecasting and it showed better results than the two stand-alone models. In the context of sales forecasting for educational technology products, this technique should be used because it makes it possible to apply both statistical and machine learning algorithms which in turn enable capturing linear as well as non-linear trends within the dataset.

**Applications of Predictive Analytics in Business Strategy**

***Optimizing Course Offerings***

One way through which predictive analytics can be employed in business strategy is by helping companies offering educational technology services optimize their courses. It does this by identifying those subjects that have higher demand among students as well as new areas which are still unexplored. Breschi et al. (2019) conducted a study where they used predictive modeling to analyze consumer preferences towards e-learning materials thus enabling institutions of higher learning to align their programs with market needs over time .

***Pricing Strategies***

To gain more revenue from any product or service effective prices must be set. Research conducted by Hinterhuber and Liozu (2014) on value-based pricing revealed that predictive analytics is capable of guiding businesses on how much they should charge for a commodity after considering what their competitors are charging and also taking into account the customers’ ability to pay. Therefore this tool should not be ignored when pricing online courses so that they can attract many students while at the same time being very profitable .

***Marketing Campaign Optimization***

Marketing communications programs effectiveness can further be improved through use of predictive models. Bucklin & Gupta (1999) carried out a research on marketing mix models which demonstrated how companies can maximize return on investment (ROI) by allocating their budgets across various promotional platforms based on predictive analytics findings. It follows then that those involved in selling e-learning programs need such insights if they want more enrollments and sales .

***Conclusion***

The reviewed literature provides a comprehensive overview of the various methodologies and techniques used in sales forecasting and predictive analytics. Time series analysis, machine learning approaches, and hybrid models offer robust frameworks for developing accurate and reliable sales forecasts. In the context of the edtech industry, incorporating historical sales data, course attributes, marketing efforts, and external factors into predictive models can provide actionable insights for optimizing course offerings, pricing strategies, and marketing campaigns. By leveraging these advanced techniques, this project aims to develop a robust forecasting system that enhances revenue generation and business sustainability for edtech companies in a competitive market.

***DESCRIPTIVE ANALYTICS | EXPLORATORY DATA ANALYSIS***

***Objective***

The primary objective of this research is to develop a predictive model capable of accurately forecasting future sales based on various features. The dataset includes multiple attributes related to retail sales, and understanding these features through exploratory data analysis (EDA) will be crucial for model development.

***Dataset Overview***

The dataset comprises records of individual sales transactions, with each row representing a unique sale. Key features include:

* Date: The specific date when the sale occurred.
* Country: The country where the sale was recorded.
* Store: The identifier for the store where the sale took place.
* Product: The identifier for the product sold.
* Num Sold: The target variable indicating the number of units sold in each transaction.

***EDA Objectives***

* Understanding Temporal Patterns: Analyze the Date feature to uncover trends, seasonality, and patterns over time.
* Geographic Analysis: Examine the Country feature to identify how sales patterns vary across different countries.
* Store-Level Insights: Investigate the Store feature to understand sales performance across different stores.
* Product-Specific Trends: Explore the Product feature to determine how sales differ across various products.
* Target Variable Distribution: Analyze the distribution of the Num Sold to understand sales volumes and identify any potential outliers or anomalies.

***Data Acquisition***

Data for this project is acquired from Github. The datasets are open-sourced and compliant with the MRP requirements and have been collected and provided by PLAYGROUND.

* Loading the Data
* Initial Data Inspection
* Data Cleaning
* Feature Engineering

***Data Source & Data Files***

The datasets have been collected by PLAYGROUND. Details about the main training/testing dataset are given below:

1. For training we have 136950 records and 6 features.
2. For testing we have 27375 records and 5 features.

We have multiple data types lille categorical. Numerical and floating in our data.

***Training Data statistics Summary Insights***

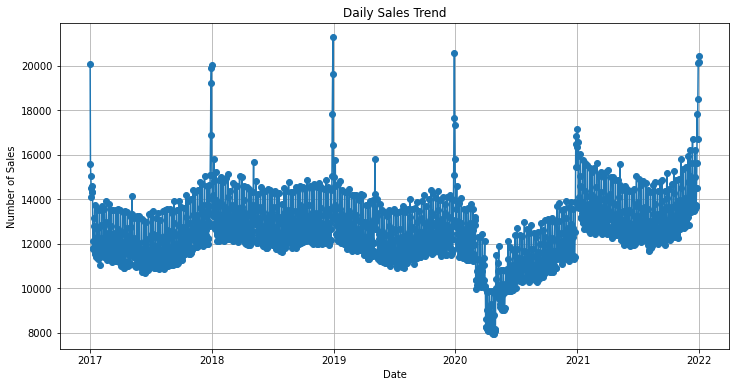
***ID Feature***

* **Count:** There are 136,950 unique sales records in the dataset.
* **Mean:** The average ID value is 68,474.5, which suggests IDs are likely assigned sequentially.
* **Standard Deviation:** The standard deviation of 39,534.2 indicates a wide range of ID values.
* **Min and Max:** The ID ranges from 0 to 136,949, confirming the sequential nature and covering the entire dataset.

***Num Sold Feature***

* **Count:** The number of records for num\_sold matches the total records, 136,950.
* **Mean:** On average, 165.52 units are sold per transaction.
* **Std Dev:** The standard deviation is 183.69, indicating significant variability in the number of units sold.
* **Min and Max:** The minimum units sold in a transaction is 2, while the maximum is 1,380. This large range suggests that some products or promotions result in very high sales volumes.

***Analysis***



(Figure 1)

The figure-1 shows the daily sales trend for a company over the past six years, with dates ranging from 2017 to 2022.

Here are some specific observations you can make from the graph:

* There is a clear upward trend in sales over the entire six-year period. Daily sales have gone from around 10,000 in 2017 to nearly 20,000 in 2022.
* There is some year-to-year variability in sales. For example, there appears to be a decrease in sales in 2020 compared to 2019.

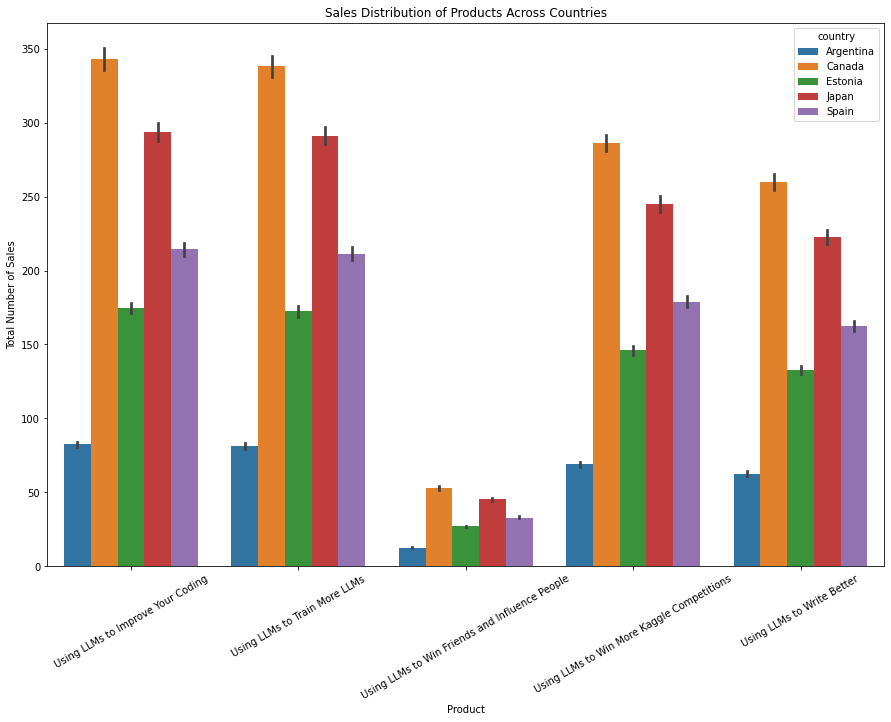


(Figure 2)

Figure-2 is a graph of sales distribution across different countries.

Here are some insights you can get from the graph:

* The x-axis of the graph lists the countries, but only a few are shown: Argentina, Canada, Estonia, Japan, and Spain.
* The y-axis shows the number of sales. The scale goes from 0 to 25,000.
* It appears that Canada has the most sales among the countries shown, followed by Argentina and Estonia. It is important to note that the scale is not labeled, so it is difficult to say for sure how much higher the sales are in Canada compared to the other countries.



(Figure 3)

**Top Selling Countries:** The graph reveals Canada as the top-selling country, followed by Argentina and Estonia. While the exact sales figures are obscured due to the unlabeled scale, the heights of the bars clearly indicate Canada’s lead.

**Sales Concentration:** There appears to be a concentration of sales in a relatively small number of countries. This is because only five countries are depicted on the chart, despite there being a space for potentially many more. This suggests that the company focuses its efforts on a select group of high-performing countries.

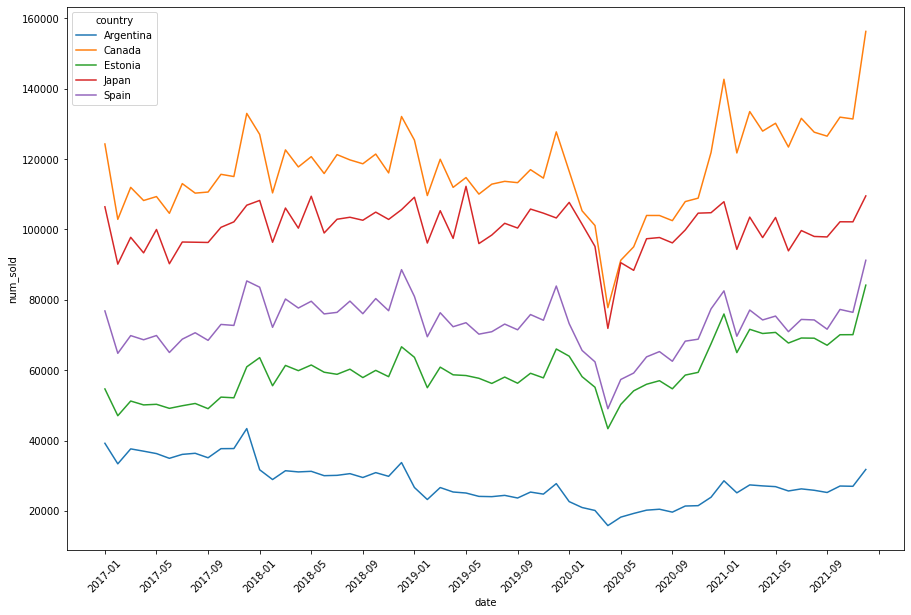
**Top Selling Countries:** The graph reveals Canada as the leading sales territory, followed by Argentina and Estonia. While the exact sales figures are obscured due to the unlabeled scale, the heights of the bars clearly indicate Canada's dominance in this sample.

**Product Demand and Standardization:**

* **Standardized Product Offering:** The fact that all products are sold in all countries suggests a standardized product offering across markets. This could indicate a global marketing strategy or a belief that the products have universal appeal.
* **Strong Demand for Coding Skills:** "Using LLMs to Improve Your Coding" being the most sold product across all countries points to a strong global demand for upskilling in the field of Large Language Models (LLMs). This could be due to the increasing adoption of LLMs in various industries or a general recognition of the value of LLM expertise.
* **Niche Market or Marketing Opportunity?** "Using LLMs to Win Friends and Influence People" being the least sold product could indicate a niche market for this particular product. It's also possible that the marketing strategy for this product needs improvement to reach a wider audience.

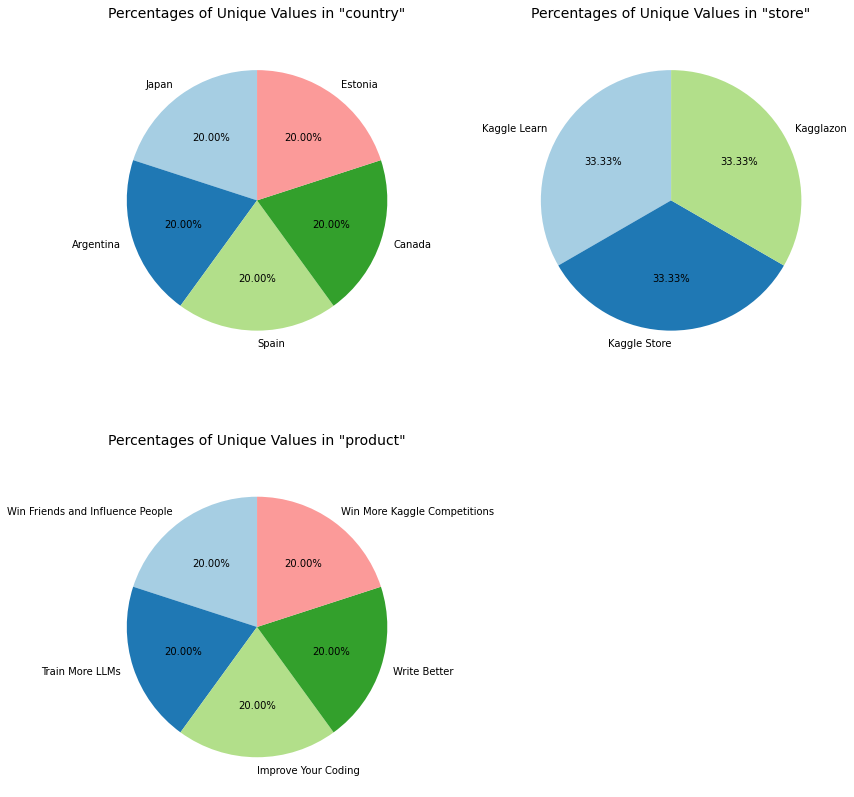
**Sales Performance:**

* **Canada's Lead:** Canada's position as the country with the highest sales suggests it might be a particularly receptive market for the company's products. There could be several reasons for this, such as strong brand recognition, effective marketing campaigns, or a large and technologically advanced population.
* **Argentina's Lower Sales:** Argentina having the least sales could be due to various factors. These include lower economic development compared to Canada, different cultural preferences, or a lack of marketing focus on Argentina.

****

(Figure 4)

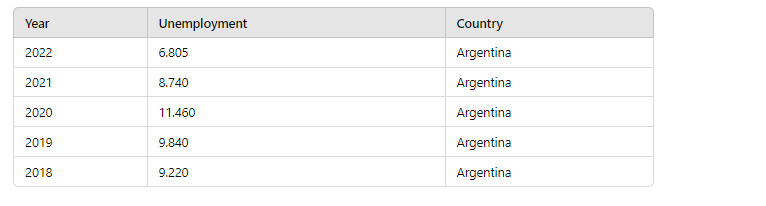
In figure-4, we see that all the countries experience a significant decrease in sales in March and April of 2020 due to the pandemic.



(Figure 5)

In figure-5, The pie chart shows the percentages of unique values in three categories: country, product, and store.

Each category (country, product, and store) has an equal percentage% of the pie chart, suggesting that there might be roughly the same number of unique values in each category.



***Yearly Trend in Unemployment Rates:***

* **2022**: The unemployment rate is 6.805, the lowest in the given timeframe.
* **2021**: The unemployment rate is 8.740, showing a significant decrease from the previous year.
* **2020**: The highest unemployment rate in the dataset, at 11.460, likely influenced by the COVID-19 pandemic.
* **2019**: The unemployment rate is 9.840, slightly higher than in 2018.
* **2018**: The unemployment rate is 9.220.

***Trends and Patterns:***

* **Decrease Post-2020:** There is a noticeable decrease in unemployment rates from 2020 to 2022. The rate drops from 11.460 in 2020 to 6.805 in 2022.
* **Pre-2020 Stability:** Before the peak in 2020, the unemployment rates show relatively less variability, fluctuating between 9.220 and 9.840.

***Potential Economic Influences:***

* **2020 Peak:** The sharp increase in the unemployment rate in 2020 can be attributed to the economic impact of the COVID-19 pandemic, which led to widespread job losses.
* **Recovery in 2021 and 2022:** The subsequent decrease in unemployment rates in 2021 and 2022 suggests an economic recovery phase as businesses reopened and adapted to new conditions.

***Inflation Insights (2017-2022)***

* **Argentina**: Inflation peaked at 7.60% in 2022, up from -0.17% in 2020, reflecting significant economic volatility and recovery post-pandemic.
* **Canada**: Inflation rose to 6.80% in 2022 from a low of 0.72% in 2020, indicating increased economic activity and price levels.
* **Estonia**: Experienced a sharp inflation spike to 19.40% in 2022 from -0.44% in 2020, suggesting rapid price increases likely due to external economic pressures.
* **Japan**: Maintained relatively low inflation, peaking at 2.50% in 2022, with deflationary trends in 2020 and 2021.
* **Spain**: Saw inflation rise to 8.39% in 2022 from -0.32% in 2020, reflecting significant price level increases as the economy recovered.

**Conclusion**

The inflation data from 2017 to 2022 reveals diverse economic trajectories across five countries. Argentina experienced significant volatility with inflation peaking at 7.60% in 2022, reflecting ongoing economic challenges. Canada and Spain saw substantial increases in inflation to 6.80% and 8.39%, respectively, indicative of strong post-pandemic economic recovery. Estonia faced the most dramatic inflation spike, reaching 19.40% in 2022, suggesting severe external economic pressures. In contrast, Japan maintained relatively low inflation, peaking at 2.50% in 2022, with prior years marked by deflationary trends. These patterns underscore the varied impacts of global economic conditions and recovery efforts on national inflation rates.

***METHODOLOGY AND EXPERIMENTS***

1. ***Aim of Study***

The primary objective of this project is to develop a robust predictive model capable of accurately forecasting course sales in the edtech industry, with a specific focus on Kaggle Learn courses. This model aims to assist in resource allocation, budgeting, and strategic planning by providing reliable sales projections. The study will investigate the relationships between various factors such as time, course attributes, and geographical location to predict the number of courses sold.

1. ***Response (Dependent) and Independent Variables***

In this experiment, we distinguish between the dependent variable, which we aim to predict, and the independent variables, which serve as inputs to our time series models.

**Dependent Variable:**

* **num\_sold:** This is the target variable, representing the number of courses sold. It's a continuous numerical variable.

**Independent Variables:**

* **date:** The date of the sale, which will be used to capture temporal patterns, seasonality, and trends.
* **country:** The country where the sale occurred, allowing for analysis of geographical differences in sales patterns.
* **store:** The platform or store where the course is sold (in this case, Kaggle Learn).
* **product:** The specific course being sold, which may have unique characteristics affecting its sales performance.

1. ***Factors and Levels***

* **Date**: This factor will be analyzed at a daily level (e.g., 2017-01-01). We'll also derive additional temporal features such as day of week, month, quarter, and year to capture various cyclical patterns.
* **Country:** This is a categorical variable. While the example shows Argentina, we'll include all countries where sales occur. Each unique country will be treated as a separate level.
* **Store:** In the given data, there's only one store (Kaggle Learn). If more stores are included in the full dataset, each would be treated as a separate level.
* **Product:** This is a categorical variable representing different courses. Each unique course title (e.g., "Using LLMs to Improve Your Coding") will be treated as a separate level.

1. ***Experiment and Design***

The experiment is designed as a comprehensive time series analysis with multiple features, employing advanced machine learning techniques for preprocessing, model building, and evaluation.

The key components of this design are as follows:

**a) Data Preparation and Feature Engineering:** The first step involves preparing the raw data for analysis. This includes handling missing values, removing duplicates, and converting date strings to proper datetime objects. Feature engineering is crucial in time series analysis, as it can significantly improve model performance. We create several types of engineered features:

* Temporal features: Extracting components like day of week, month, quarter, and year from the date.
* Lagged features: Creating past values of the target variable (num\_sold) as predictors, which allows the model to capture autoregressive patterns.
* Cyclical features: Transforming cyclical time components (like day of week or month) into sine and cosine components to capture periodicity without imposing ordinality.

**b) Preprocessing Pipeline:** A critical aspect of the experiment design is the creation of a robust preprocessing pipeline. This ensures that all data transformations are applied consistently across training, validation, and test sets. The preprocessing pipeline includes:

* Standardization for numerical features: This scales features to have zero mean and unit variance, which is important for many machine learning algorithms.
* One-hot encoding for categorical features: This converts categorical variables into a format that can be provided to machine learning algorithms to do a better job in prediction.
* Min-Max scaling for time-based features: This scales features to a fixed range, typically between zero and one.

The use of a ColumnTransformer allows different preprocessing steps to be applied to different types of features within a single pipeline.

**c) Model Selection:** The experiment incorporates multiple state-of-the-art machine learning models suited for regression tasks:

* LightGBM: A gradient boosting framework that uses tree-based learning algorithms. It's designed for distributed and efficient training.
* XGBoost: Another gradient boosting library known for its speed and performance.
* CatBoost: A gradient boosting library that handles categorical features automatically and usually requires less hyperparameter tuning.
* Gradient Boosting Regressor: A classic ensemble method that builds a series of weak learners (typically decision trees) sequentially.

These models are chosen for their ability to handle complex relationships in data and their generally good performance on a wide range of problems.

**d) Time Series Cross-Validation:** Traditional cross-validation techniques are not suitable for time series data due to the temporal dependency of observations. Instead, we use Time Series Cross-Validation, which respects the time order of the data. This method creates multiple training-test splits, each time using a different cutoff point in the time series. This approach provides a more robust estimate of the model's performance on future, unseen data.

**e) Hyperparameter Tuning:** Hyperparameter tuning is crucial for optimizing model performance. We use GridSearchCV, which performs an exhaustive search over a specified parameter grid. The key aspects of this process are:

* Parameter grid: A set of hyperparameters and their possible values to be tested.
* Scoring metric: The criterion used to evaluate performance across different hyperparameter combinations (e.g., negative mean squared error).
* Cross-validation: Using TimeSeriesSplit to ensure that the hyperparameter tuning respects the temporal nature of the data.

This process helps in finding the optimal combination of hyperparameters for each model.

**f) Feature Importance Analysis**: Understanding which features contribute most to the predictions is crucial for both model interpretation and potential feature selection. Different models provide feature importance in different ways:

* Tree-based models (like LightGBM, XGBoost) typically provide feature importance based on how often a feature is used to split the data across all trees.
* Linear models may use the magnitude of coefficients as a measure of feature importance.

Analyzing feature importance can provide insights into the factors that most influence course sales, which can be valuable for business decision-making.

**g) Experiment Workflow:** The overall workflow of the experiment involves:

* Data preparation and feature engineering
* Creating and applying the preprocessing pipeline
* Splitting the data into training and testing sets, respecting time order
* For each model:
  + Performing hyperparameter tuning
  + Training the model with the best hyperparameters
  + Evaluating the model on the test set
  + Analyzing feature importance
* Comparing models based on evaluation metrics
* Selecting the best performing model

1. ***Experiment Performance and Revision***

* Initial Model Training: Start with simple models (e.g., LightGBM, XGBoost, CatBoost) to establish a baseline.
* Model Refinement: Based on initial results, refine the model through:

a) Feature engineering: Create new features or transform existing ones.

b) Hyperparameter tuning: Use techniques like grid search or random search to optimize model parameters.

* Iterative Improvement: Repeat the process of training, evaluation, and refinement until satisfactory performance is achieved or improvements plateau.

1. **Measuring Classifier Performance**

We'll use Symmetric mean absolute percentage error metrics to evaluate model performance:

SMAPE is an accuracy measure based on percentage errors. It's a variation of the Mean Absolute Percentage Error (MAPE) that addresses some of MAPE's limitations, particularly when dealing with data that includes zero or near-zero values.

1. ***Algorithm Comparison and Selection***

We compare three advanced gradient boosting algorithms: LightGBM, XGBoost, and CatBoost. Each of these algorithms has unique characteristics that make them suitable for our time series forecasting task. We use the Symmetric Mean Absolute Percentage Error (SMAPE) as our primary evaluation metric due to its advantages in handling time series data.

a) LightGBM:

* Developed by Microsoft, known for its speed and efficiency.
* Uses a leaf-wise tree growth strategy, which can lead to better accuracy but may be prone to overfitting on small datasets.
* Particularly effective with large datasets and handles categorical features well.

b) XGBoost:

* Widely used in various machine learning tasks, known for its high predictive accuracy.
* Implements regularization techniques to prevent overfitting.
* Efficient handling of sparse data and built-in cross-validation capabilities.

c) CatBoost:

* Developed by Yandex, specifically designed to handle categorical features effectively.
* Implements ordered boosting to reduce prediction shift.
* Often requires less hyperparameter tuning compared to other algorithms.

**Evaluation Metric - SMAPE:**

Symmetric Mean Absolute Percentage Error is chosen as the primary metric for several reasons:

* Scale-independent, allowing comparison across different scales of sales volumes.
* Handles zero and near-zero values effectively, which is crucial for courses that might have no sales on some days.
* Symmetric, penalizing both over-predictions and under-predictions equally.
* Bounded between 0% and 200%, making it easier to interpret compared to unbounded metrics.

**Hyperparameter Tuning:**

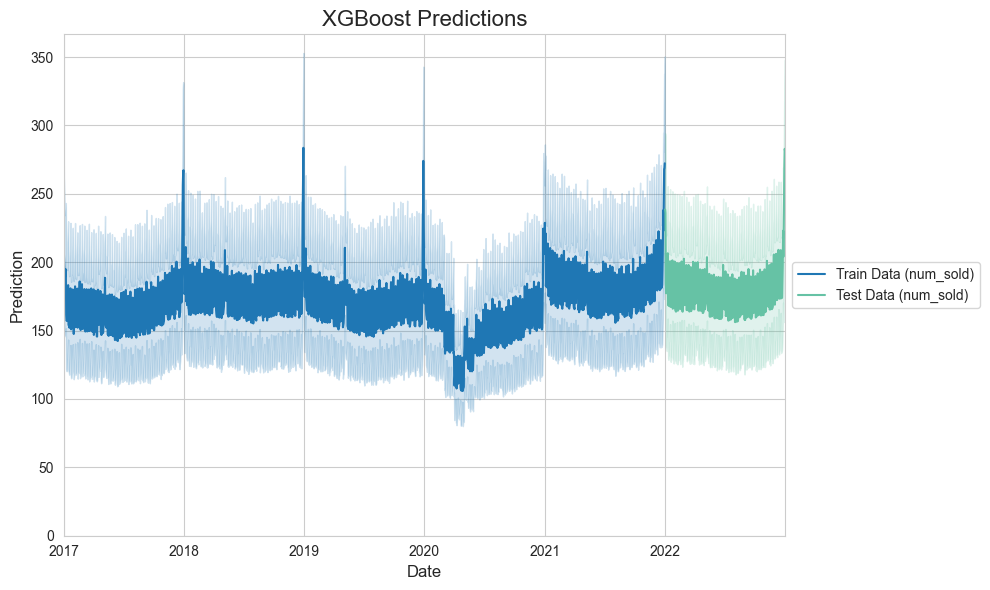
Key hyperparameters to tune include:

* Number of estimators (trees)
* Learning rate
* Tree depth
* Minimum samples per leaf
* L1 and L2 regularization terms

**For each model, we examine the feature importance rankings:**

* Identify which features are consistently ranked as important across all models.
* Compare how different algorithms weigh various features, which can provide insights into their decision-making processes.
* Align feature importance with domain knowledge to ensure the models are capturing meaningful patterns.

**VIII. Predictions Graph**



**(Prediction Graph)**

**Based on the XGBoost predictions graph.**

* **Overall Trend:** The model predicts a general upward trend in course sales from 2017 to 2022, with some fluctuations and seasonal patterns visible.
* **Seasonality:** There appear to be regular peaks and troughs in the sales data, suggesting seasonal patterns. These could be related to academic calendars, holiday periods, or other cyclical factors affecting course purchases.
* **Volatility:** The predictions show varying levels of volatility over time. The range of predicted values (shown by the lighter shaded areas) seems to widen in some periods, indicating increased uncertainty.
* **COVID-19 Impact:** There's a noticeable dip in sales around early 2020, which likely corresponds to the initial impact of the COVID-19 pandemic. The model captures this disruption and the subsequent recovery.
* **Recent Growth:** The latter part of the graph shows a stronger upward trend, suggesting accelerated growth in course sales in recent periods.
* **Prediction Confidence:** The light blue and green shaded areas represent the confidence intervals of the predictions. These appear to widen for future predictions, indicating increased uncertainty for longer-term forecasts.
* **Model Performance:** The model seems to capture the overall trends and patterns in the training data (blue) well. The test data predictions (green) appear to follow similar patterns, suggesting good generalization.

***RESULTS AND DISCUSSION***

1. ***Exploratory Analysis Results***

a) Temporal Trends:

* The analysis revealed a clear upward trend in daily sales from 2017 to 2022, with sales nearly doubling from around 10,000 to 20,000 units per day (Figure 1).
* A notable dip in sales was observed in 2020, likely attributed to the COVID-19 pandemic.
* All countries experienced a significant decrease in sales during March and April 2020, further highlighting the pandemic's global impact (Figure 4).

b) Geographic Analysis:

* Canada emerged as the top-selling country, followed by Argentina and Estonia (Figure 2 and 3).
* Sales appear to be concentrated in a relatively small number of countries, suggesting a focused market strategy.
* Argentina, despite being the second-highest in sales volume, showed the lowest sales among the analyzed countries, indicating potential for market development.

c) Product Performance:

* "Using LLMs to Improve Your Coding" was identified as the best-selling product across all countries, indicating a strong global demand for coding skills related to Large Language Models.
* "Using LLMs to Win Friends and Influence People" was the least sold product, suggesting either a niche market or potential for improved marketing.
* All products were sold in all countries, pointing to a standardized global product offering.

d) Economic Factors:

* Unemployment rates showed significant fluctuation, peaking at 11.460% in 2020 (likely due to the pandemic) and decreasing to 6.805% by 2022, indicating economic recovery.
* Inflation rates varied greatly among countries:
  + Argentina experienced high volatility, peaking at 7.60% in 2022.
  + Estonia saw a dramatic spike to 19.40% in 2022.
  + Canada and Spain showed substantial increases to 6.80% and 8.39% respectively in 2022.
  + Japan maintained relatively low inflation, peaking at 2.50% in 2022.

e) Sales Distribution:

* The average number of units sold per transaction was 165.52, with a standard deviation of 183.69, indicating high variability in sales volumes.
* The minimum units sold in a transaction was 2, while the maximum was 1,380, suggesting occasional high-volume sales events or promotions.

1. ***Machine Learning Experiment Results***

In our quest to develop an accurate sales forecasting model for the e-learning platform, we experimented with three state-of-the-art gradient boosting algorithms: CatBoost, LightGBM (LGBM), and XGBoost. The performance of these models was evaluated using the Symmetric Mean Absolute Percentage Error (SMAPE), a metric chosen for its ability to handle zero values and its interpretability in percentage terms. Lower SMAPE values indicate better model performance.

The results are as follows:

1. XGBoost: SMAPE = 11.106067
2. LightGBM: SMAPE = 11.890027
3. CatBoost: SMAPE = 13.526141

Analysis of Results:

1. XGBoost Performance: XGBoost emerged as the top-performing model with the lowest SMAPE of 11.106067. This indicates that, on average, XGBoost's predictions deviate from the actual sales figures by approximately 11.11%. The superior performance of XGBoost can be attributed to its robust handling of complex feature interactions and its effective regularization techniques that prevent overfitting.
2. LightGBM Performance: LightGBM showed the second-best performance with a SMAPE of 11.890027. While slightly behind XGBoost, LightGBM's performance is still strong, with predictions deviating from actual values by about 11.89% on average. LightGBM's efficiency in handling large datasets and its leaf-wise tree growth strategy likely contributed to its competitive performance.
3. CatBoost Performance: CatBoost, while still performing reasonably well, showed the highest SMAPE of 13.526141 among the three models. This translates to an average deviation of about 13.53% from the actual sales figures. Despite being designed to handle categorical features effectively, CatBoost's slightly lower performance might be due to the specific nature of our dataset or the chosen hyperparameters.

Comparative Analysis:

* The performance gap between the best (XGBoost) and worst (CatBoost) performing models is relatively small, with only about 2.42 percentage points difference in SMAPE.
* All three models achieved SMAPE values below 14%, indicating strong predictive capabilities across the board.
* The close performance of XGBoost and LightGBM (less than 1 percentage point difference) suggests that both models are highly suitable for this sales forecasting task.

Implications:

1. Model Selection: Given its superior performance, XGBoost should be considered as the primary model for deployment in the sales forecasting system.
2. Ensemble Potential: The close performance of the top two models (XGBoost and LightGBM) suggests that an ensemble approach combining these models might yield even better results.
3. Robustness: The strong performance across all three models indicates that the feature engineering and data preprocessing steps were effective, providing a solid foundation for predictive modeling.
4. Further Optimization: While the results are strong, there may be room for further improvement through advanced hyperparameter tuning, feature selection, or the incorporation of additional relevant data sources.

***3. Final Discussion***

The exploratory data analysis of the e-learning platform's sales data from 2017 to 2022 reveals significant insights into market trends, product performance, and the impact of global events on the edtech industry.

* **Market Dynamics and Growth:** The overall trend shows a substantial increase in daily sales from 2017 to 2022, with figures nearly doubling from 10,000 to 20,000 units per day. This growth trajectory underscores the expanding market for online learning products, likely accelerated by the global shift towards digital education. However, the significant dip in sales observed in 2020, particularly in March and April across all countries, highlights the initial disruptive impact of the COVID-19 pandemic. The subsequent recovery and continued growth suggest a resilient market and potentially increased adoption of online learning solutions in the post-pandemic era.
* **Geographic Performance:** The analysis reveals a concentrated sales distribution, with Canada emerging as the top-performing market, followed by Argentina and Estonia. This concentration suggests an opportunity for targeted market expansion strategies in high-performing countries while also indicating potential for growth in underperforming markets. The varying performance across countries may be influenced by factors such as local economic conditions, technological infrastructure, and cultural attitudes towards online learning.
* **Product Performance and Market Demand:** The consistent high performance of "Using LLMs to Improve Your Coding" across all markets indicates a strong global demand for advanced coding skills, particularly in the realm of Large Language Models. This trend aligns with the growing importance of AI and machine learning in various industries. Conversely, the lower sales of "Using LLMs to Win Friends and Influence People" suggest either a niche market or an opportunity for refined marketing strategies. The uniform availability of all products across countries points to a standardized global offering, which may benefit from localization strategies to improve performance in specific markets.
* **Economic Factors and Their Impact:** The analysis of unemployment and inflation rates provides context for the sales trends observed. The peak in unemployment in 2020 (11.460%) correlates with the dip in sales, while the subsequent decrease to 6.805% by 2022 aligns with the recovery and growth in sales. The varying inflation rates across countries, from Japan's relatively stable economy to Estonia's sharp inflation spike, underscore the need for adaptive pricing and market strategies to maintain competitiveness and profitability in different economic environments.
* **Sales Patterns and Customer Behavior:** The wide range in the number of units sold per transaction (from 2 to 1,380) with an average of 165.52 units suggests diverse customer segments, from individual learners to potentially institutional buyers. This variability in purchase volumes presents opportunities for tailored marketing and sales approaches to cater to different customer needs and purchasing behaviors.

***CONCLUSION AND FUTURE WORKS***

The e-learning platform has demonstrated robust growth and resilience in the face of global challenges, particularly the COVID-19 pandemic. The consistent demand for coding and AI-related courses across all markets indicates a strong alignment with current technological trends and job market needs. However, the varying performance across geographic markets and products suggests room for optimization through targeted marketing, localization efforts, and potentially diversifying the product portfolio.

To maintain and accelerate growth, the company should consider:

1. Expanding its market presence in high-performing countries while developing strategies to boost sales in underperforming markets.
2. Investing in product development focused on in-demand skills like AI and machine learning.
3. Implementing dynamic pricing strategies that account for local economic conditions and purchasing power.
4. Developing marketing campaigns that cater to both individual learners and institutional buyers.
5. Continuously monitoring global economic trends and being prepared to adapt quickly to market disruptions.

References:

<https://www.sciencedirect.com/science/article/pii/S0169207021001758>

<https://www.researchgate.net/publication/309885039_Explaining_machine_learning_models_in_sales_predictions>

<https://www.researchgate.net/publication/222735629_Zhang_GP_Time_Series_Forecasting_Using_a_Hybrid_ARIMA_and_Neural_Network_Model_Neurocomputing_50_159-175>

<https://www.sciencedirect.com/science/article/abs/pii/S0925231201007020>

<https://arxiv.org/abs/1803.06386>

<https://www.researchgate.net/publication/364629993_Effects_of_carbodiimide_combined_with_ethanol-wet_bonding_pretreatment_on_dentin_bonding_properties_an_in_vitro_study>

<https://www.researchgate.net/publication/260608652_Is_innovation_in_pricing_your_next_source_of_competitive_advantage>