# AT2: Machine Learning as a Service (Retail)

**Student Name :** Aibarna Singh Basnet

**Student ID :** 24585717

**Deliverables**

* [Prediction Model](https://github.com/aibarna/basnet_aibarna-24585717-ML_AT2/blob/main/notebooks/predictive/basnet_aibarna_24585717_prediction_lgbm.ipynb)
* [Forecast Model](https://github.com/aibarna/basnet_aibarna-24585717-ML_AT2/blob/main/notebooks/forecasting/basnet_aibarna_24585717_forecasting_arima.ipynb)
* [Github Repo](https://github.com/aibarna/basnet_aibarna-24585717-ML_AT2/tree/main)
* Prediction model on Heroku
* Forecasting model on Heroku

### **The Brief:**

You are working for an American retailer that has 10 stores across 3 different states: California (CA), Texas (TX) and Wisconsin (WI).

You are tasked to build 2 different models that will be deployed into production as API:

* a predictive model using a Machine Learning algorithm to accurately predict the sales revenue for a given item in a specific store at a given date.
* a forecasting model using a time-series analysis algorithm that will forecast the total sales revenue across all stores and items for the next 7 days.

**Business Understanding**

**Business Aim:** Develop a Machine Learning (ML) Model for Sales Forecasting and Prediction.

Our primary objective is to establish a robust machine learning model that can predict and forecast sales for each of our retail stores located in California (CA), Texas (TX), and Wisconsin (WI). This predictive model will serve as a valuable tool for inventory management, staff allocation, and overall strategic business planning.

**Model Target**

**Objective**: The primary objective is to develop a Sales Prediction and Forecasting Model.

**Description**: Leveraging historical sales data, seasonality patterns, and other pertinent factors, our goal is to build a machine-learning model that can accurately predict future sales for our retail stores situated in California (CA), Texas (TX), and Wisconsin (WI). The primary purpose of this model is to provide precise sales forecasts, which will be instrumental in aiding inventory management, staff allocation, and overall store management.

**Key Focus Areas**:

1. Utilize historical sales data.
2. Identify and incorporate seasonality patterns.
3. Develop a machine-learning model for sales prediction.
4. Focus on enhancing inventory management, staff allocation, and overall store management through accurate sales forecasts.

By achieving these objectives, the model aims to assist the retail business in making informed decisions and optimizing its operations.Top of Form

**Hypothesis**

In this undertaking, it is crucial to conduct thorough research and gather data to validate our assumptions and refine the model. Our hypotheses will serve as the guiding principles for selecting data and developing the model, shaping our comprehension of anticipated performance**.Top of Form**

**Hypotheses for Store Performance:**

1. **Sales by State**: Stores located in California (CA) are expected to have a higher sales compared to stores in Texas (TX) and Wisconsin (WI).
2. **Total Sales Leader**: The Texas (TX) store is hypothesized to lead in terms of total sales among all stores.
3. **Customer Retention Rate**: It is assumed that the Wisconsin (WI) store maintains the highest customer retention rate.

**Hypotheses for Customer Behaviour:**

1. **Demographic Variations**: Significant variations in customer demographics, such as age and income, are expected to be observed among the three states.
2. **Preference for High-End Products**: Customers in California (CA) are anticipated to exhibit a greater preference for high-end product purchases compared to customers in Texas (TX) and Wisconsin (WI).

**Hypotheses for Product Categories:**

1. **Regional Product Preferences**: Regional preferences are expected to lead to varying performance of specific product categories in different states.
2. **Outdoor and Sporting Goods Sales**: The store in Texas (TX) is expected to have a larger proportion of sales in outdoor and sporting goods compared to other stores.

**Hypotheses for Geographical Factors:**

1. **Urban Store Performance**: Store performance is assumed to be positively influenced by proximity to urban centres, with urban stores in California (CA) expected to outperform rural stores in the same state.

**Hypotheses for Seasonal Trends:**

1. **Seasonal Sales Variation**: Sales in California (CA) are expected to exhibit more pronounced seasonality compared to those in Texas (TX) and Wisconsin (WI).
2. **Holiday Season Impact**: The holiday season is expected to significantly impact sales in all three states.

Top of Form

**Project Steps**

To initiate our data analysis project for an American retailer with ten outlets across California, Texas, and Wisconsin, we will follow a structured approach:

**1. Data Collection:**

* Gather comprehensive information about each store, including sales data, customer data, inventory data, and any relevant information.
* Ensure data consistency and completeness, covering records for each store in every state.
* Import CSV files from the provided source and upload them to a personal Google Drive for accessibility.

**2. Data Cleansing:**

* Identify and rectify missing values, errors, and inconsistencies within the data.
* Fill in missing information and eliminate erroneous entries to ensure data accuracy.
* Verify that data types are suitable for analysis, such as numeric values represented as numbers and dates as dates.

**3. Data Integration:**

* Combine data from diverse sources and formats into a unified dataset for analysis.

**4. Exploratory Data Analysis (EDA):**

* Utilize basic statistics and visualizations to gain a profound understanding of the data.
* Compute summary statistics, create histograms, scatter plots, and other visualizations to unveil patterns and trends.

**5. Sales Analysis:**

* Evaluate sales data to discern store performance, distinguishing well-performing stores from those needing improvement.
* Key metrics include total sales, average sales per store, and sales growth over time.
* Determine the most popular products or product categories.

**6. Time Series Analysis:**

* Investigate the data's time-series aspects to unearth seasonality, trends, and anomalies in sales or other pertinent metrics.

**7. Deployment:**

* Deploy the predictive model developed for sales forecasting on a platform like Heroku for practical use in the retail business.
* Ensure that the deployed model can provide real-time sales predictions based on the latest data.

**8. Reporting and Visualization:**

* Craft visual reports and dashboards to convey our findings to stakeholders.
* Utilize tools like Tableau, Power BI, and Python libraries such as Matplotlib and Seaborn to facilitate this process.

By following these structured steps, we aim to not only analyse the data comprehensively but also deploy a predictive model that can contribute to the retail business's success by providing accurate sales forecasts for informed decision-making.

Top of Form

**DATA UNDERSTANDING**

This dataset offers an extensive view of the operations of a retail company, encompassing various aspects of its stores, sales, and interactions with customers. The data is meticulously collected from diverse internal sources, ensuring its accuracy and comprehensiveness. It spans across a significant period, enabling historical analysis and the identification of trends.

This dataset holds value for a multitude of analyses and applications, including but not limited to:

1. **Market Targeting and Segmentation:** Understanding and segmenting the market to tailor strategies for specific customer groups.
2. **Managing Inventory and Forecasting Demand:** Efficiently managing stock and predicting product demand to avoid overstock or shortages.
3. **Evaluating Store and Regional Performance:** Assessing the performance of individual stores and regions to make informed decisions.
4. **Enhancing Customer Retention and Loyalty Programs:** Utilizing data to improve customer loyalty and retention strategies.
5. **Refining Marketing Strategies and ROI Assessment:** Optimizing marketing campaigns and measuring their return on investment.
6. Features

**Key Attributes of the Dataset:**

* **Store Details:** Comprehensive information about the ten stores, including their geographic location, size, and store type (e.g., flagship, outlet, etc.).
* **Sales Information:** In-depth retail sales data, covering daily, weekly, and monthly sales figures for individual stores, product categories, and overall revenue.
* **Inventory Insights:** Valuable insights into inventory management at each store, including product turnover rates and the availability of popular items.
* **Customer Profiling:** A glimpse into customer behaviour and preferences, including data on demographics, purchase history, and participation in loyalty programs.
* **Geographical Distribution:** An examination of store locations in California, Texas, and Wisconsin, with a particular focus on regional trends and store performance.
* **Marketing Impact:** An investigation into the influence of marketing campaigns and promotional events on sales and customer engagement.
* **Seasonal Sales Trends:** The identification of patterns in seasonal sales and an analysis of how holidays and special events impact store performance.

Top of Form

**DATASET**

[Training Data](https://drive.google.com/file/d/1-0x5Vfri1i-OL3ek2GnhZGye-4eqbqQA/view?usp=drive_link)

[Evaluation Data](https://drive.google.com/file/d/1-2PahhxOmBOCnVNTpgaNGkirjlJz9wYH/view?usp=drive_link)

[Calendar data](https://drive.google.com/file/d/1-6cH8c0tKTFu8EzMJyVfdhxrny6rdgrM/view?usp=drive_link)

[Events](https://drive.google.com/file/d/1_RmDGfRTMkqF4OO9NibNoRhbEjc0OZW4/view?usp=drive_link)

[Items price per week](https://drive.google.com/file/d/1--W-RjAnypyvbwUCsSZVldrA2Ja2jtDA/view?usp=drive_link)

**Modelling and Evaluation**

In the development of predictive and forecasting models for the American retailer, the modelling and evaluation phase within the Cross-Industry Standard Process for Data Mining (CRISP-DM) is a pivotal stage. This section provides an overview of the strategies, algorithms, and performance metrics employed to construct and appraise these models.

Top of Form

**Predictive Model for Sales Revenue Prediction of a given item in a specific store at a given date.**

**1. Data Preprocessing:**

**Data Loading and Transformation:**

* The data was loaded from multiple CSV files containing sales, calendar, calendar events, and item prices data.
* Events data was grouped by date and restructured into a new format for easier integration with the calendar data.
* Data was merged based on date to create a consolidated dataset that includes sales, calendar, and pricing information.
* Missing values in the 'sell\_price' column were filled using the mean sell price for each store, item, and day.
* Data types were downcasted to optimize memory usage.

**2. Feature Engineering:**

**Creating Time-Based Features:**

* Extracted various time-based features from the 'date' column, including day, month, and year.
* Created a 'day' column by splitting the 'd' column, which represents the day of sales data.
* Created additional features related to sales data, events, and time-based characteristics.

**Lag and Rolling Features:**

* Generated lag features (1, 7, and 14 days) to capture the historical behavior of sales.
* Computed rolling mean features (7, 14, and 28 days) to capture moving averages of sales.
* Calculated quarter-based statistics for lag features, including mean, standard deviation, minimum, and maximum.

**Additional Aggregate Features:**

* Calculated mean revenue for various groupings, such as 'id', 'store\_id', and combined 'id' and 'store\_id'.

**3. Model Training and Performance:**

**Model Selection:**

* The LightGBM Regressor model was chosen for sales revenue prediction. LightGBM is a gradient boosting framework that works well for regression tasks and can handle large datasets efficiently.

**Data Splitting:**

* The data was split into training and testing sets based on the 'day' column. Sales data before a certain day (e.g., day 1200) was used for training, and sales data from that day onward was used for testing.

**Training and Prediction:**

* The model was trained using the training data, with features such as 'item\_id', 'store\_id', 'day'.
* The trained model was used to make predictions on the test data.

**Performance Evaluation:**

* The Root Mean Squared Error (RMSE) was used as the performance metric for the predictive model.
* RMSE was calculated to assess the accuracy of the model's revenue predictions.

**4. Model Performance:**

The predictive model achieved an RMSE of approximately 34.22. This indicates the average squared difference between the predicted and actual revenue values is 34.22. The lower the RMSE, the better the model's performance. The top 10 important features contributing to the predictions were identified and displayed in a horizontal bar chart.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

**Forecasting Model for total sales revenue across all stores and items for the next 7 days**

**1. Data Preprocessing:**

* The data consists of sales information, calendar events, and item prices, which have been loaded from CSV files.
* Calendar events have been grouped and transformed into a more structured format.
* Date-related features like weekday, month, year, and 'wday' (weekday as an integer) have been extracted from the calendar data.
* Null values in calendar events have been filled with 'No\_event'.
* Additional sales data with zero sales for the test period has been added.
* Data types have been downcasted to save memory.
* Sales data has been melted into a columnar format for analysis.
* Data from different sources, such as sales, calendar, and prices, have been merged.

**2. Feature Engineering:**

* The primary feature created is 'revenue', which is calculated by multiplying 'sold' and 'sell\_price' columns.
* Columns with more than 50% missing values have been dropped.
* Label encoding has been applied to the 'd' column.

**3. Time Series Analysis Model:**

* The data was initially processed and tested for stationarity using the Augmented Dickey-Fuller Test.
* The data was differenced to make it stationary by calculating the difference between consecutive revenue values.
* An automated ARIMA model was chosen for time series forecasting. The 'auto\_arima\_model' function was used to find the best parameters for the SARIMA (Seasonal ARIMA) model.
* The data was split into a training and validation set, with 80% of the data used for training.
* The best ARIMA model was fitted to the training data.
* The model was saved using pickle for later use.

**4. Model Performance:**

* The performance of the ARIMA model was evaluated on the validation set using RMSE (Root Mean Square Error).
* The validation RMSE was calculated to be 19587.45, indicating the average error in revenue prediction on the validation data.

**5. Forecasting:**

* The trained ARIMA model was used to make sales revenue forecasts for the next 7 days.

A screenshot of a computer program

Description automatically generated

A screen shot of a computer program

Description automatically generated

A screen shot of a computer

Description automatically generated

A computer code on a black background

Description automatically generated



A computer code on a black background

Description automatically generated

**A screenshot of a computer program

Description automatically generated**

**API Structure and Instructions for Running Predictions**

This project offers an accessible API with a structured set of endpoints to provide valuable sales predictions and insights. Here's an overview of the available API endpoints, their functions, and instructions for running predictions:

1. **'/' (GET):**
   * **Description:** This endpoint serves as the project's root and provides a brief overview of the project's objectives.
   * **Expected Output Format:** It includes information about the project's objectives, a list of available endpoints, expected input parameters, and a link to the project's GitHub repository.
   * **How to Access:** Simply make a GET request to the root URL of the API.
2. **'/health/' (GET):**
   * **Description:** This endpoint is designed for health checking and returns a status code 200 along with a welcome message of your choice.
   * **Expected Output Format:** A simple welcome message indicating that the API is operational.
   * **How to Access:** Send a GET request to the '/health/' endpoint.
3. **'/sales/national/' (GET):**
   * **Description:** This endpoint provides a sales volume forecast for the next 7 days based on an input date.
   * **Expected Input Parameters:** Include the desired input date as a parameter in the request.
   * **Expected Output Format:** The output will include the predicted sales volume for the specified date and the following 7 days.
   * **How to Access:** Make a GET request to the '/sales/national/' endpoint with the input date as a parameter.
4. **'/sales/stores/items/' (GET):**
   * **Description:** This endpoint offers the predicted sales volume for an input item, store, and date.
   * **Expected Input Parameters:** Include the item, store, and date as parameters in the request.
   * **Expected Output Format:** The response will contain the predicted sales volume for the specified item, store, and date.
   * **How to Access:** Send a GET request to the '/sales/stores/items/' endpoint with the item, store, and date as parameters.

**Instructions for Running Predictions:**

1. Access the API using the provided endpoint URLs.
2. To obtain national sales forecasts for the next 7 days, make a GET request to '/sales/national/' and include the desired input date as a parameter.
3. For predicted sales volume specific to an item, store, and date, send a GET request to '/sales/stores/items/' with the item, store, and date as parameters.
4. Ensure that your requests follow the specified input parameters format, and the API will respond with the sales predictions accordingly.

Top of Form