**Capstone Initial Report – OTC Sales Forecasting**  
**Student Name:** Alex Saura  
**Course:** Professional Certificate in Machine Learning and Artificial Intelligence

**Overview**

This project focused on forecasting weekly over the counter (OTC) pharmaceutical shipments using historical sales data. The primary objective was to develop a model that could predict future box shipments and assist in planning for potential demand surges or disruptions, such as those caused by seasonal illnesses or public health events.

The dataset consisted of weekly OTC shipment volumes extracted from a raw sales CSV file, aggregated by week. The analysis was conducted using a Jupyter Notebook (Capstone\_Initial\_Report.ipynb) and covered exploratory data analysis (EDA), statistical profiling, model development, and forecasting.

**Methodology**

The notebook is structured in logical steps from data cleaning to modeling and forecasting:

1. **Data Aggregation & Preprocessing**  
   Sales data were converted to a weekly format using the transaction date. Missing values were addressed, and time-based features such as lag variables and calendar indicators were added to enrich the dataset.
2. **Exploratory Data Analysis (EDA)**  
   The notebook begins by loading and inspecting the dataset, converting date values into a standardized weekly format. Using visualizations like line plots, the notebook identifies fluctuations in weekly shipments and seasonal patterns. A Shapiro-Wilk test and skewness analysis indicated a non-normal distribution in sales data, with signs of seasonality and sporadic spikes that could be linked to health-related events.
3. **Model Development**  
   Two models were developed using the RandomForestRegressor from Scikit-learn:
   * **Baseline Model**: Utilized only a simple Time Index feature to capture basic temporal progression.
   * **Enhanced Model**: Leveraged lag features (1–3-week lags), a rolling 3-week average, and calendar-based features (week of year and month).
4. **Evaluation Metrics**  
   Both models were evaluated on a test set using R² and Mean Squared Error (MSE), Root Mean Squared Error(RMSE). The enhanced model demonstrated significantly stronger performance.

These results show that the enhanced model substantially reduced error compared to the baseline, though both models yielded negative R² scores, indicating that more sophisticated time-aware models or richer features might be required for optimal performance.

| **Model** |  |  |  |  |  |  |  |  | **R² Score** |  |  |  | **MSE** |  |  |  |  |  |  |  |  | **RMSE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Baseline RF (Time Index) |  |  |  |  |  |  |  |  | -3.01588 |  |  |  | 6001.693929 |  |  |  |  |  |  |  |  | 77.470600 |
| Enhanced RF (Lag + Calendar Features) |  |  |  |  |  |  |  |  | -0.42212 |  |  |  | 2125.343178 |  |  |  |  |  |  |  |  | 46.101444 |

These metrics indicate that incorporating lag and seasonal features provides a more accurate and reliable model for predicting shipment volumes.

1. **Forecasting Future Shipments**  
   A 10-week rolling forecast was generated using the enhanced model. The model used its own predictions as lag inputs for each subsequent step. Forecasted shipments showed a steady trend consistent with recent history and within expected fluctuation ranges.

**Key Insights**

* The enhanced Random Forest model significantly improved predictive performance, with a notable reduction in RMSE of nearly 40% compared to the baseline.
* Forecasted values continued the general trend seen in recent data, supporting the model’s value for operational planning and forecasting.
* Temporal feature engineering—particularly the inclusion of lagged shipment values and calendar-based patterns—proved essential for performance improvements.

**Future Recommendations**

* **Incorporate External Factors**: Add contextual data like holidays, weather, or disease outbreaks for improved seasonal accuracy.
* **Test Alternative Models**: Future iterations could compare additional models (e.g., linear models or others introduced in previous coursework).
* **Segment Product Lines**: Focused modeling on categories such as cold & flu or seasonal allergy products could improve sensitivity to public health trends.
* **Anomaly Detection**: Residual analysis could flag outlier weeks and help identify unusual demand surges.

**Final Notes**

The complete notebook includes clean, structured code and well-labeled visualizations that align with best practices for EDA and machine learning modeling. All findings, metrics, and forecasts are clearly documented, making this project a strong baseline for further time series development and practical business use.