Machine Learning Engineer Capstone Project:
Kaggle Competition: M5 Forecasting – Accuracy
Estimating unit sales of Walmart's retail goods across various locations

Domain Background:

M5 focused on a retail sales forecasting application with the objective to produce the most accurate point forecasts for 42,840 time series that represent the hierarchical unit sales of the largest retail company in the world, Walmart.

There are concerns in the previous M4 forecasting competition, like range of benchmarks and data sets to avoid overfitting. Also, introducing high-frequency data, such as hourly, daily, and weekly data, to investigate how multiple seasonal patterns and irregularly spaced observations could be properly handled, as well as how data collected from sensors could be optimally used, which is the trend now with the spread of IoT usage in the retail industry.

The M5 competition tried to address these concerns and suggestions by introducing the following innovative features:

- a large data set (42,840 time series).

- focused on accurately predicting the daily unit sales of retail stores across various locations and product categories.

- the data set involved daily data which requires accounting for multiple seasonal patterns, special days, and holidays.

- the data set included exogenous/explanatory variables, such as product prices, promotions, and special events.

Problem Statement:

Time series forecasting of daily sales for the next 28 days by using hierarchical sales data from Walmart, the world's largest company by revenue. This estimation certainly helps Walmart to increase their revenues.

Datasets and Inputs:

Walmart's dataset covers ten stores in three US states (California, Texas, and Wisconsin) and includes item level, department, product, categories, and store details for five years starting from 29th Jan 2011 to 24th April 2016. The data comprises three thousand individual products from three categories and seven departments, sold in ten stores across these three states.

Walmart's dataset provided in the M5 Competition consists of the following three files:

File 1: "calendar.csv" Contains information about the dates the products are sold.

File 2: "sell_prices.csv" Contains information about the price of the products sold per store and date.

File 3: "sales_train.csv" Contains the historical daily unit sales data per product and store.

Solution Statement:

The problem is a time-series data problem and can be solved using Classical Machine Learning techniques to estimate the unit sales to a particular day by using historical sales data. So my solution will be forecasting algorithm using XGBoost and lightGBM.

Benchmark Model:

I will make predictions using moving average (SARIMA) and the very naive last 28 days as benchmark model to see how machine learning can improve the accuracy of forecasting over wildly used methods.

Evaluation Metrics:

The custom performance metric is chosen for this problem which is Root Mean Squared Scaled Error (RMSSE). It is a variant of the well-known Mean Absolute Scaled Error (MASE). Then by using the **Weighted RMSSE** (**WRMSSE**),

Project Design:

On AWS Sage Maker I will use Jupyter Lab because it is seamlessly integrated with github to load libraries for exploratory data analysis including pandas and numpy, then loading data using pandas read_csv.

After that exploratory data analysis is needed to know more about the data before creating the model, then after having good understanding of the data in hand, I will create data frame for training containing all the features required by the model, then I will train two models (XGBoost and LightGBM) to determine the best performing one, after that visualizing results is necessary to understand how our models performed, and then I will select the best model according the evaluation metrics.