Mix Forecasting
2-stage hierarchical Demand Forecasts
Supply Chain Operations

Introduction
Purdue University's Data Mine Team, in collaboration with Cisco, aims to explore a 2-stage hierarchical modeling approach to forecast the demand for its 10,000+ products. The platform forecasts are then disaggregated into the product level (PLU) using Mix prediction approaches ranging from naïve forecasts based on the last observed demand rates under the platform, to more sophisticated approaches. The idea is to improve forecast accuracy at the two levels independently, and hence improve the PLU level forecast accuracy and reduce systematic bias.

We have 4 goals to achieve:
- Develop Mix accuracy metrics to isolate and measure the Mix forecast performance.
- Develop one or more Mix forecasting methods using statistical or multi-variate AI/ML approaches that consider Mix patterns and demand features.
- Evaluate these different Mix forecasting methodologies, including those provided by Cisco and any ensembles, using this metric to identify the best approaches, and to visualize and monitor its performance over time.
- Given platform-level forecasts, measure the final PLU-level accuracy of the 2-stage approach against other single-stage expert forecasting methods.

This comprehensive approach is poised to not only optimize demand forecasting outcomes but also to drive operational excellence and strategic decision making within Cisco's supply chain ecosystem.

Variable Analysis
Handling a dataset with over 1000+ columns requires a systematic approach to effectively reduce dimensionality while retaining meaningful information. One approach involves utilizing techniques such as correlation analysis. Principal Component Analysis (PCA) and random forest feature importance to identify and select the most relevant features.

Random forest:
Random forests (RF) are machine learning algorithms that can provide variable importance measures to rank predictors.

Combine many binary decision trees. Trees are built using several bootstrap samples and randomly selecting a subset of explanatory variables at each node. The results from each tree are then aggregated to give a prediction for each observation.

Highly accurate	Better generalization	Interpretable

PCA: Principal Component Analysis (PCA): a ML based unsupervised learning method to reduce the variables in a dataset by creating linear combinations.
- Reduced our feature size to 70%.
- Correlation analysis remove to very high flow units.
- Enhanced the dataset’s manageability and interpretability.

Regression Model
Linear Regression Overview:
Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Use Case for Our Project:
In our project at Cisco, linear regression can be utilized to model the relationship between demand for products (dependent variable) and various demand indicators such as forward-looking sales predictions and customer growth rates (independent variables). This can help us understand how changes in these indicators affect product demand, aiding in accurate demand forecasting.

Techniques We Are Using:
- Training with Historical Data
- Feature Selection
- Model Evaluation using Error Metrics

Categorical analysis
We categorized the data into four distinct categories: age, lifestyle, or product, seasonality, product lifecycle stage, and sales value.

Data Processing:
- Formulating the data to make it valuable
  - Data Filtering
  - Data Transformation: Scaling, Right/left shift
  - PLU Analysis: Grouping K-means clustering
- Feature selection

Trend Analysis:
Uncovering hidden patterns using key features. Picking out the patterns and relationships within the data to identify notable patterns. Look for upward or downward movement over time.

Moving averages	Exponential Smoothing	Linear Fits

Seasonality Analysis
Post-categorization, it became imperative to visually represent the data to discern significant trends and patterns. Utilizing various visualization techniques, we facilitated a comprehensive understanding of the dataset’s characteristics.

Decomposed time series data into its constituent components: trend, seasonality, and residual (noise)
- Analyze the seasonal patterns by comparing the behavior of the time series across different seasons or time periods
- Adjust forecasts based on the estimated seasonal effects identified during analysis.

Random Forest
Why Random Forest?
A powerful ensemble machine learning algorithm that leverages the collective wisdom of multiple decision trees for accurate forecasting. We selected Random Forest for its robustness, ability to handle various data types, and low feature importance estimation, which is crucial for understanding demand factors in our forecasting project.

Utility in Our Case:
Random Forest is particularly useful for forecasting demand due to its ability to handle complex relationships and interactions between features, providing accurate predictions for Cisco’s diverse product range.

Results
Results signify the achievements of our work. We used two-stage multivariate and univariate KPI, based efficient and well-thought-out hierarchical models.

Dashboard + Images
This Python script creates a web-based dashboard using Dash, a framework for building interactive web applications with Python. The dashboard displays a grid-like component, a graph, and a toggle for scatter points.

The update ‘graph funciona’ serves as a callback to update the graph based on the uploaded CSV file and the scatter points. It reads the contents of the uploaded file, converts it to existing data, generates a new plot using Plotly Express, and updates the scatter points based on the file’s contents.

The update ‘scatter funciona’ allows users to upload CSV files containing data values pairs, visualizes the data with an animation and automatically calculates scatter plots for additional insights. It is designed with a modern dark theme, enhanced readability, aesthetics, facilitating data exploration and analysis.

Conclusion
During our project, we employed several techniques to identify the most influential variables across our dataset.

This preparatory work paved the way for:
- Experimenting with various regression models
- Identifying the optimal K for our analysis
- Focusing on pivotal variables and categories - uncovering key drivers.

Predicting future values of our target variable:
- Deriving Random Forest and XGBoost models
- Capturing complex statistical relationships and interactions using features.

The output of our efforts was showcased on a simple dashboard, a visual narrative that conveys the significant patterns and insights extracted.

Future scope and improvements
In this project’s next phase, we aim to enhance our Mix Forecasting to meet Cisco’s needs by:
- Work on improving Random Forest and XGBoost models
- Optimizing forecast error
- Improving the platform data to sort into more meaningful values to predict sales trends.

The inclusion of a robust model assessment with rigorous testing to ensure reliability. These steps are aimed at giving Cisco a more accurate and acceptable forecasting tool.