Mix Forecasting

2-stage hierarchical Demand Forecasts

Supply Chain Operations

The Data Mine
Corporate Partners Symposium 2024

Introduction

Purdue University's Data Mine Team, in collaboration with Cisco, aims to exploring 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 (PLID) using Mix prediction approaches ranging from naïve predictions based on the last observed demand ratio under the platform, to more sophisticated approaches. The idea is to improve forecast accuracy at the two levels independently, and hence improve the PLID 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 methods, 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 PLID 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.

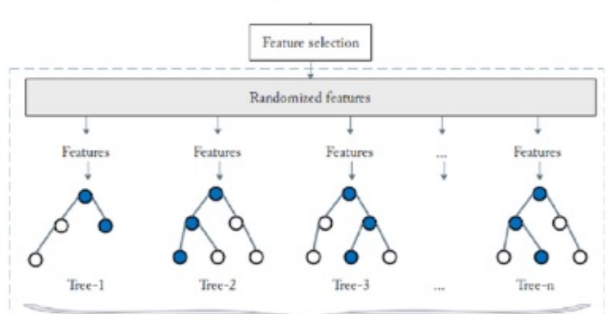
Variable Analysis

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 (RFs) are machine learning algorithms that can provide variable importance measures to rank predictors.

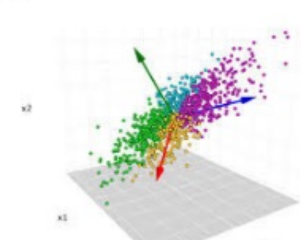
Combine many binary decision trees. Trees are built using several bootstrap samples and randomly selecting a subset of explanatory variables at each node. The results from each tree are then aggregated to give a prediction for each observation.

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/low units.
- Enhanced the dataset's manageability and interpretability.

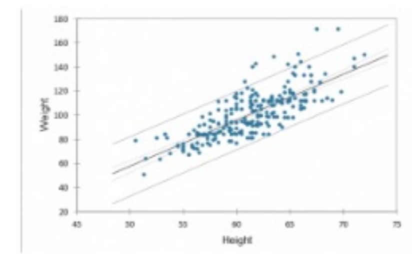


Regression Model

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

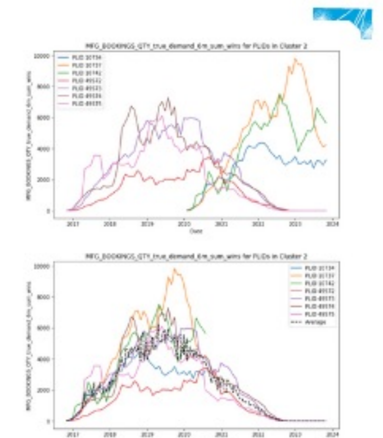


Categorical analysis

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

Data Processing:
Formatting the data to make it valuable

- Data Filtering
- Data Transformation: Scaling, Right/Left shift
- PLID Analysis: Grouping,K-means clustering
- Feature selection

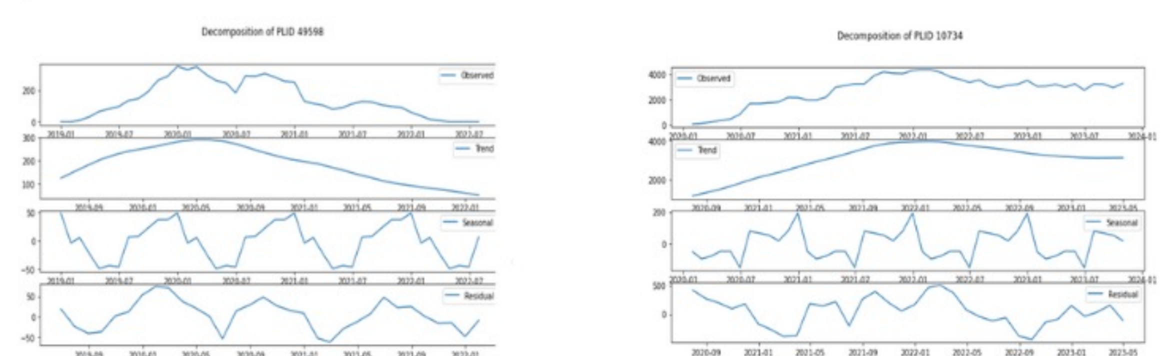


Trend Analysis: Uncovering hidden patterns using key features. Picking out the patterns and relationships within the data to identify notable patterns. Look for Upward or downward movement over time.

Moving averages **Exponential Smoothing** **Linear filters**

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.

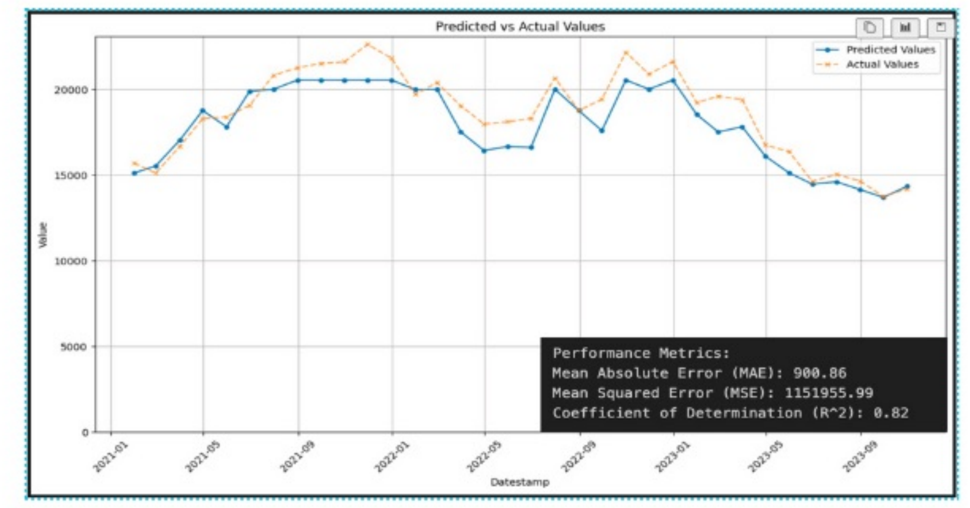
Decomposed time series data into its constituent components: trend, seasonality, and residual (noise).
Analyze the seasonal patterns by comparing the behavior of the time series across different seasons or time periods.
Adjust forecasts based on the estimated seasonal effects identified during analysis.



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 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.

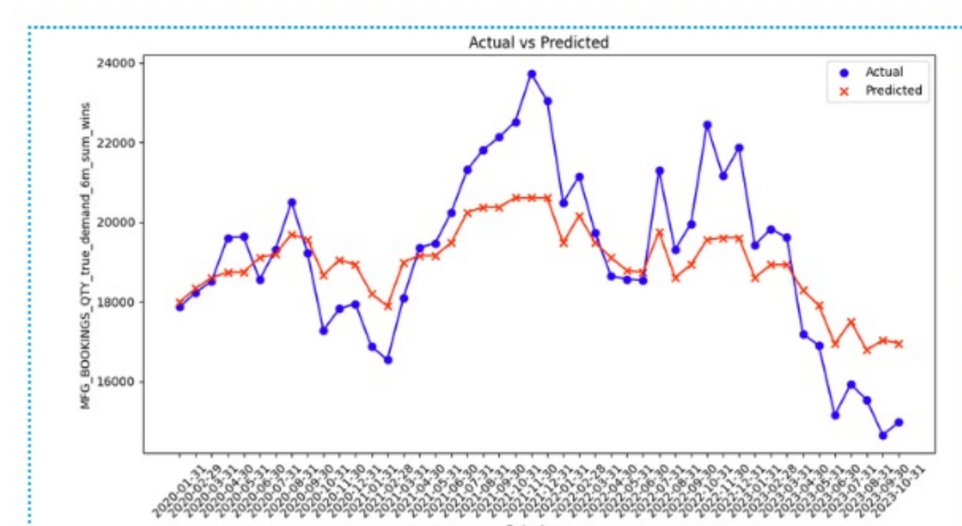


XG Boost

Why XGBoost?
XGBoost is a highly efficient machine learning library known for its speed and effectiveness in predictive modeling.
We chose XGBoost for its exceptional performance, versatility, and ability to handle complex forecasting scenarios. Its regularization techniques and efficient computational speed make it well-suited for improving forecast accuracy and mitigating biases.

Utility in Our Case:
XGBoost serves as a powerful tool for generating accurate demand forecasts for Cisco's product range. It handles large datasets, extracts complex patterns, and minimizes overfitting, aligning perfectly with our project objectives.

Techniques We Are Using:
Training with Historical Data
Cross-Validation
Feature Importance Analysis
Ensemble Learning



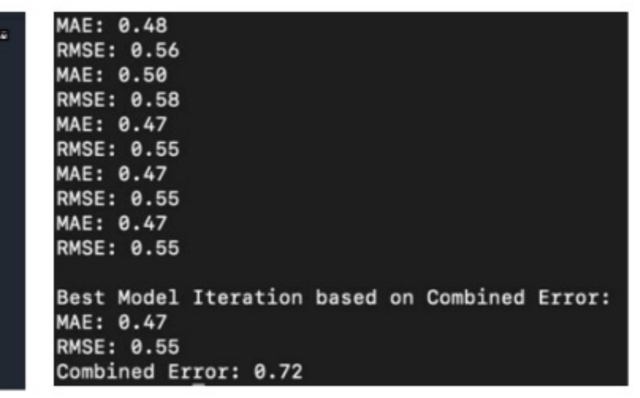
Results

Results signify the achievements of our work. We used 2- stage multivariate and univariate ML based efficient and well though out hierarchical models:

Accuracy Matrix

	Xg Boost	Random Forest	Inhouse
Platform	90	90	90
Platform Mul	90	90	90
PLID	90	90	90
PLID Mul	90	90	90

Dashboard + Images



This Python script creates a web-based dashboard using Dash, a framework for building interactive web applications with Python. The dashboard layout includes a title, a file upload component, a graph, and a toggle for scatter points.

The update_graph function serves as a callback to update the graph based on the uploaded CSV file and the scatter toggle value. It reads the contents of the uploaded file, concatenates it with existing data, generates a line plot using Plotly Express, and updates the scatter points based on the user's selection.

The dashboard allows users to upload CSV files containing date-value pairs, visualize the data using line plots, and optionally display scatter points for additional insights. It's 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 from our dataset.

This preparatory work paved the way for:

- experimenting with various regression models
- identifying the optimal fit for our analysis
- focusing on pivotal variables and categories - uncovering key drivers.

Predicting future values of our target variable:

- delve into Random Forest and XGBoost models
- capture complex nonlinear relationships and interactions among features.

The culmination of our efforts was showcased in a sample dashboard, a visual narrative that conveys the significant patterns and insights discovered.

Future scope and improvements

In this project's next phase, we aim to enhance our Mix Forecasting to meet Cisco's needs by:

- **W**ork on fine-tuning our Random Forest and XGBoost models
- **O**ptimize their settings
- **I**mprove how platform data is sorted into detailed ID levels to predict sales trends

This includes a roadmap with model assessments with rigorous testing to ensure reliability.
These steps are aimed at giving Cisco a more accurate and adaptable forecasting tool.

PURDUE UNIVERSITY | The Data Mine

The Team: Devansh Khandelwal, Piya Saxena, Tanmayi Sharat, Nicholas Lauw, Amrita Konaje, Li Qingwei, Aaryan Shah, Akhil Kota, Aarav Garg, Gaurav Sable.
Corporate Partners: Zijing Zhu, Puneet Gupta, Aishwarya, Ameya Chhatre
The Data Mine staff: Lauren Dalder
Our TAs - Saanidhya Jain, Arvind Gomathy