Enhancing Ensemble Algorithm & Calibration Process for Demand Forecasting

PURDUE UNIVERSITY | The Data Mine

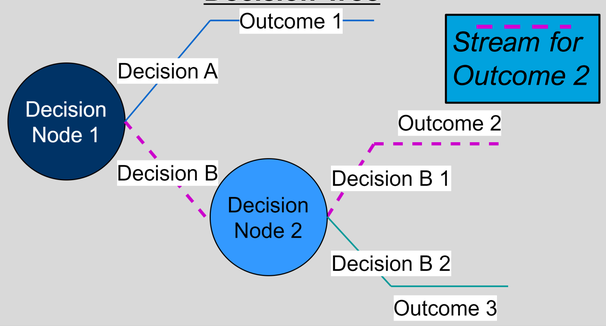
Aman Kumar, Arhan Chaudhary, Kapil Naresh, Rishab Vempati, Sahil Desai,
Shreya Gupta, Siddh Patani, Siddharth Sancheti, Zachary Kirkeby

CISCO

BACKGROUND

Cisco's Demand Planning team **generates diverse forecasting streams** for 10K+ products. A stream represents a **decision tree** that is used to forecast demand for each product by quarter.

Decision Tree



OBJECTIVE

- **Develop** Stream Selection method
- **Enhance** decision tree algorithm
- **Implement** calibration method for selected stream
- **Achieve** higher forecast accuracy with "Best Ball"
- **Optimize** product demand forecasting
- **Improve** performance of ensemble approach

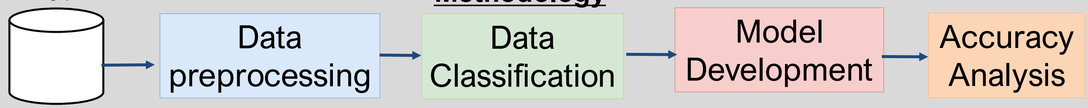
ACKNOWLEDGEMENTS

We would like to thank **Cisco Team**, our Mentor **Ji Qi** and entire **Data Mine Team** for providing support and guidance throughout this project.

1. STREAM SELECTION

PURPOSE: Determine the best performing stream to determine the optimal model for different product types.

Methodology



Bayesian Model

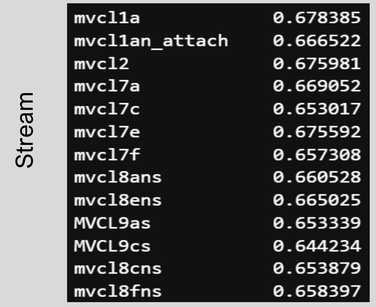
Probabilistic model calculates the likelihood of outcomes based on observed data. Observed data can be harnessed by existing Cisco tree models.

Maximum Likelihood Estimation

Statistical model determines the likelihood of different, independent parameters based on observed data and fits it to a normal curve.

Gaussian Mixture Model

Gaussian Mixture Model clusters PLIDS into three categories and is used to detect inherent trends within the clusters.



Stream	
mvcl1a	0.678385
mvcl1an_attach	0.666522
mvcl2	0.675981
mvcl7a	0.669052
mvcl7c	0.653017
mvcl7e	0.675592
mvcl7f	0.657308
mvcl8ans	0.660528
mvcl8ens	0.665025
MVCL9as	0.653339
MVCL9cs	0.644234
mvcl8cns	0.653879
mvcl8fns	0.658397

Fig 1: Bayesian Stream Selections

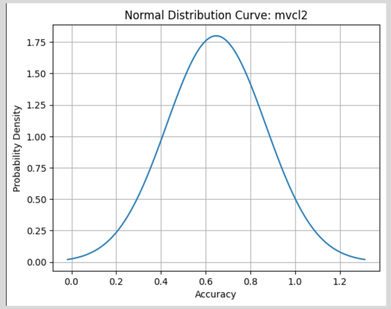


Fig 2: MLE Normal Curve

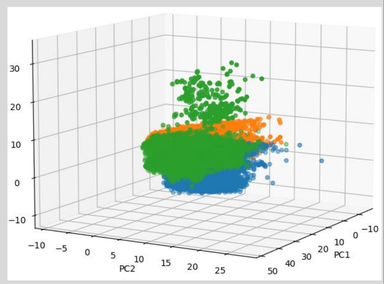
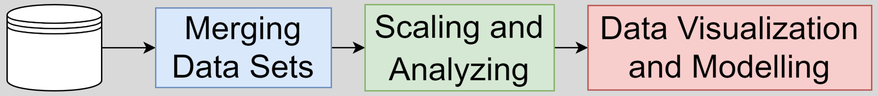


Fig 3: Clustering using GMM

2. FORECAST CALIBRATION

PURPOSE: Finding outliers in the Ensemble models, to better calibrate and refine the algorithm.

Methodology



Isolation Forest Algorithm (IFA)

Computational efficiency with large datasets and robustness to asymmetrical datasets

Standard Deviation

Simpler implementation, but tedious, as it requires iterating through dataset.

Support Vector Machine Model

High-performing classifier for complex datasets, requiring tuning and computation.

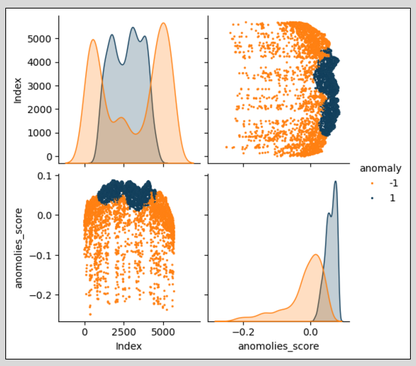


Fig 1: IFA Out & Inliers Relation

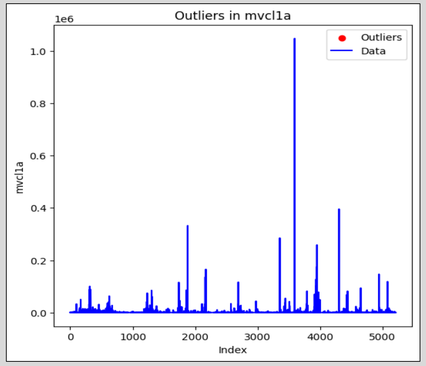


Fig 2: SD Outlier Detection

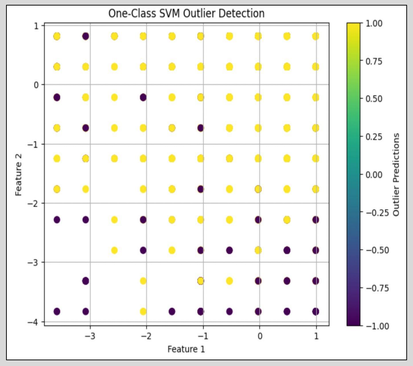


Fig 3: SVM Outlier Detection

RESULTS: IFA *mean accuracy* ↑ by 3% (sample size of 50)

FUTURE GOALS

STREAM SELECTION

We aim to refine accuracy by modifying the algorithms to incorporate quarterly changes in the dataset. We'll also explore integrating Bayesian posterior data to Cisco's tree models to compare decision tree outcomes.

FORECAST CALIBRATION

We will focus on calibrating and re-enhancing the developed stream selection algorithms with our tested Isolation Forest and Support Vector Machine Models to improve their accuracy, robustness and eliminate outliers.

CONCLUSION

- We were able to **leverage** maximum likelihood estimation to generate a normal distribution with each PLID's historical data to **predict** its **most accurate stream** for our own metric.
- We **enhanced the accuracy** using Isolation Forest, though the results are based on smaller data size and **pioneered** Support Vector Machine Model to find outlier more effectively.
- Throughout, we've developed skills in **data visualization**, **collaborated** effectively on coding, and **expanded knowledge** of Cisco's forecasting algorithms.