Forecasting and AI-Driven Customer Segmentation System (FACETS)

David Ryan, Harshil Nandwani

GitHub: https://github.com/davidry777/FACETS-LLM

Hugging Face Space: https://huggingface.co/spaces/DS5983-FACETS-team/FACETS-LLM-Assistant

ryan.d5@northeastern.edu, nandwani.h@northeastern.edu

Abstract

FACETS is an AI-augmented decision support system integrating time-series demand forecasting, customer segmentation, and natural language insight generation via large language models (LLMs). The development of FACETS utilized the Rossmann Sales and UCI Online Retail II datasets. The system uses K-Means clustering on RFM data and implements VARIMAX and Random Forest models for demand forecasting. The DeepSeek-R1 and DeepSeek-Coder LLMs transform model outputs into actionable business insights. The research shows that Large Language Models boost machine learning output interpretability which leads to better stakeholder understanding and action capabilities. The research demonstrates our approach to integrating traditional ML methods with generative AI while discussing both our findings and general implications.

Introduction

Understanding customer behavior and predicting product demand are core challenges in retail analytics that directly impact inventory management, marketing strategies, and customer satisfaction. However, the advanced machine learning (ML) models typically used for these tasks often produce complex outputs that are inaccessible to many business users. FACETS was developed to address this disconnect by creating a unified platform that integrates predictive analytics and customer segmentation with large language models (LLMs) for explainable AI. By combining traditional statistical models, such as VARIMAX and Random Forest, with K-Means clustering on RFM metrics, FACETS enables comprehensive data-driven decision-making. The results of these models are translated into clear, business-oriented reports using DeepSeek-generated natural language summaries. The system is deployed on a Hugging Face Space with an interactive Gradio interface, allowing users to explore trends, clusters, and insights in real time without needing a technical background. In doing so, FACETS bridges the gap between complex machine learning techniques and actionable business intelligence.

Background

Time-series forecasting and customer segmentation are foundational techniques in retail analytics. Time-series forecasting aims to predict future trends based on past data. In this project, we use the VARIMAX model, a variant of Vector Autoregression with Exogenous Variables, to model dependencies across multiple time-dependent variables while incorporating external regressors like holidays or promotions. This allows for nuanced forecasts sensitive to seasonality and business-specific effects. Alongside this, we employ Random Forest regressors—a robust ensemble learning method known for handling non-linearity and interactions among features. Random Forests are advantageous due to their minimal assumptions about data distribution and their ability to model complex trends when adequate historical data is available.

For customer segmentation, we built it on top of the RFM (Recency Frequency, Monetary) framework, a widely used metric in marketing and Customer relationship management (CRM). By summarizing customer transaction behavior into these three numerical indicators, we transform raw transactional logs into structured features suitable for clustering. K-Means clustering, a centroid-based algorithm, is then applied to partition customers into distinct groups. K-Means is selected due to its simplicity, scalability, and interpretability—though it assumes spherical clusters, it remains effective when paired with pre-normalized input such as RFM data. Other advanced clustering methods like DBSCAN and Gaussian Mixture Models were explored during initial testing but were set aside due to either poor cluster separation or high sensitivity to hyperparameters in this context.

The third component of FACETS involves integrating large language models (LLMs) to interpret ML outputs. Recent advances in transformer-based architectures, such as DeepSeek-R1 and DeepSeek-Coder, have enabled models to summarize structured numerical and categorical data

into natural language. These LLMs are pretrained on large corpora and fine-tuned to perform well on downstream NLP tasks like summarization, question answering, and instruction following. In this system, LLMs act as narrative agents that convert analytical results (e.g., cluster profiles or forecast trends) into business-friendly text, enhancing interpretability for non-technical stakeholders. This aligns with a growing trend in explainable AI (XAI), where natural language is leveraged to close the gap between algorithmic insight and human decision-making.

Literature Review

Ensemble Methods in Retail Analytics: Price Prediction and Sales Forecasting

Traditional Statistical and Machine Learning Approaches

Retail price prediction and sales forecasting have traditionally relied on statistical methods and classical machine learning algorithms. Random Forest models (Breiman, 2001) remain popular for product price prediction due to their ability to capture non-linear relationships and robustness to outliers. These ensemble decision tree models automatically handle feature interactions that are common in retail product attributes. Similarly, time series methods such as ARIMA (Box & Jenkins, 1976) and its multivariate extensions like VARMAX have formed the bedrock of sales forecasting approaches, effectively modeling temporal dependencies in sales data.

Large Language Models and Retrieval-Augmented Generation

- The emergence of large language models (LLMs) has opened new possibilities for retail analytics. These models excel at understanding complex textual descriptions and incorporating domain knowledge into predictions. For specialized tasks like retail price prediction, fine-tuning approaches using techniques like LoRA (Hu et al., 2021) have proven effective at adapting pre-trained models to specific domains while maintaining computational efficiency.
- Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for grounding LLM outputs in relevant data (Lewis et al., 2020). In retail contexts, RAG can retrieve similar products or store profiles to inform price and sales predictions. This approach is particularly valuable when historical data is sparse or when predicting for new products or stores.
- The integration of LLMs into business intelligence systems, as noted by Chen et al. (2023), has enhanced interpretability by generating natural language explanations of predictions. This capability has proven crucial for stakeholder adoption of advanced analytics solutions.

Ensemble Methods for Prediction

Combining multiple models through ensembling has consistently shown to improve predictive accuracy across various domains. The seminal work by Wolpert (1992) on stacked generalization demonstrated how meta-learners can effectively combine predictions from diverse base models. In time series forecasting specifically, Bates & Granger (1969) established that combinations of forecasts typically outperform individual models, especially when the component forecasts capture different aspects of the underlying process.

More recently, Wang et al. (2022) proposed sophisticated model fusion techniques using optimal transport theory to combine neural and statistical models, showing significant improvements over simpler ensemble approaches. In retail-specific applications, Pavlyshenko (2019) demonstrated the effectiveness of blending machine learning, statistical models, and deep learning for sales forecasting, achieving robust predictions across diverse product categories.

- Price prediction in retail environments requires robust models that can account for diverse product characteristics, market dynamics, and consumer behavior. Traditional methods often struggle with the heterogeneity of retail datasets.
- Random Forest Agent serves as an effective foundation for price prediction by leveraging ensemble decision trees. This non-parametric approach handles non-linear relationships in the data while maintaining robustness to outliers.
- The model transforms product descriptions into high-dimensional embeddings using SentenceTransformers (Reimers & Gurevych, 2019), capturing semantic relationships between products that traditional bag-of-words approaches miss.
- Frontier Agent employs a Retrieval-Augmented Generation (RAG) approach (Lewis et al., 2020), combining vector similarity search with large language models. By retrieving similar product descriptions and their corresponding prices from a vector database, the Frontier Agent establishes a context-rich environment for price estimation, particularly valuable for novel or unusual items. This approach mirrors human pricing strategies, where merchants often reference comparable products when setting prices for new inventory.
- Specialist Agent, built on the DeepSeek-Coder model, represents a fine-tuned approach s

pecifically optimized for retail price prediction. By adapting the pre-trained weights through instruction tuning on retail datasets, the Specialist Agent develops domain-specific expertise that generalist models lack. This approach follows the paradigm demonstrated by Touvron et al. (2023), where task-specific fine-tuning significantly enhances performance on specialized domains.

Customer Segmentation with Clustering

Customer segmentation is important in today's marketing, enabling firms to tailor their strategies based on consumer purchasing patterns.

- Traditional clustering methods such as K-Means (Lloyd, 1982) remain popular due to their computational efficiency and ease of interpretation. K-Means works by minimizing the variance within clusters and assigning data points to the nearest centroid. However, it assumes spherical, similarly sized clusters and can struggle with complex distributions of customer behavior.
- Hierarchical clustering techniques (Ward, 1963) have been employed for customer segmentation, allowing retailers to identify distinct purchasing patterns. However, these methods often struggle with the high dimensionality and heterogeneity of retail data.

To address the above shortcomings, advanced clustering techniques like DBSCAN and Gaussian Mixture Models (GMMs) have been proposed:

- DBSCAN (Ester et al., 1996) is a density-based method that excels at discovering arbitrarily shaped clusters and managing noise, making it well-suited for outlier-prone datasets.
- GMMs (Reynolds, 2009) adopt a probabilistic approach to model the data as a mixture of Gaussian distributions, capturing uncertainty in cluster assignment. However, these models often require careful hyperparameter tuning and are sensitive to initialization.

In FACETS, we initially tested DBSCAN and GMMs but ultimately favored K-Means due to its stability and clarity when used with normalized RFM data. To enhance interpretability beyond numerical labels, we created a **Segmentation Agent** with to describe each customer segment in natural language, making cluster insights accessible to non-technical users.

Large Language Models for Business Intelligence

Large language models (LLMs) have recently emerged as powerful tools for business intelligence. Their ability to synthesize and communicate complex data through natural language allows stakeholders to understand insights without delving into code or metrics.

- Studies such as Chen et al. (2023) highlight the use of LLMs for generating human-like summaries from structured numerical data, thereby improving accessibility. Furthermore, the use of retrieval-augmented generation (RAG) architectures (Lewis et al., 2020) allows LLMs to ground their responses in contextually relevant facts, enhancing reliability.
- While many existing systems employ LLMs for Q&A or customer support bots, FACETS integrates LLMs for structured insight generation —a less common but growing application. Dimitrov et al. (2021) and Song et al. (2022) showed that LLMs can summarize financial reports and assist executives with strategic summaries.

Our contribution extends this by applying LLMs to customer segmentation and demand forecasting, with a focus on interpretability and human relevance.

Project Description

Our project implements a comprehensive retail analytics system addressing two critical business challenges: product price prediction and store sales forecasting. Through an ensemble architecture that integrates traditional statistical methods, machine learning models, and large language models, we achieve superior prediction accuracy and robustness.

System Architecture

The system employs a modular agent-based design where each specialized predictive model is implemented as an "agent" with standardized interfaces for prediction and logging. This design enables straightforward composition into ensemble systems while providing transparent diagnostic capabilities.

The core architecture revolves around five types of specialized agents:

1. **Statistical/ML Agents** - Random Forest for price prediction and VARMAX for sales forecasting
2. **RAG Agents** - Frontier Agent for price prediction and Time Series Frontier Agent for sales forecasting
3. **LLM Agents** - DeepSeek-based Specialist agents fine-tuned for their respective tasks
4. **Ensemble Agents** - Meta-learners that integrate predictions from all component agents
5. **Segmentation Agent** – K-Means clustering for customer segmentation tied with LLM for interpretation.

Data Processing and Storage

For price prediction, product descriptions are embedded using SentenceTransformers and stored in a ChromaDB vector database for efficient similarity search. Similarly, for sales forecasting, store profiles are embedded and augmented with temporal features to enable retrieval of similar store-time combinations.

The system processes two primary datasets:

- Online Retail Dataset with product descriptions and prices
- Rossmann Store Dataset with store characteristics and historical sales

Price Prediction Subsystem

The price prediction subsystem combines three specialized approaches:

- **Random Forest Agent** transforms product descriptions into embedding vectors and applies a forest of decision trees to predict prices. This approach effectively captures non-linear relationships between product features and prices while remaining robust to outliers.
- **Frontier Agent** employs a RAG approach, retrieving similar products from the vector database and using their prices as context for a DeepSeek language model prediction. This mimics human pricing strategies where merchants reference comparable products.
- **Specialist Agent** leverages a DeepSeek-Coder model fine-tuned specifically for retail price prediction using LoRA. The fine-tuning process involves:
    1. Preparing training data from the retail dataset with product descriptions and prices
    2. Formatting prompts in DeepSeek's chat format
    3. Applying LoRA adaptation to efficiently tune the model without full fine-tuning
    4. Training with appropriate hyperparameters to optimize for price prediction accuracy

  The fine-tuning process employs Parameter-Efficient Fine-Tuning (PEFT) with LoRA, targeting specific attention modules in the model architecture for adaptation. This approach requires significantly less computational resources than full fine-tuning while maintaining strong performance.
- **Ensemble Agent** integrates predictions from all three models using a linear regression meta-learner that has been trained on a validation set. The meta-learner includes features derived from individual model predictions, including the minimum and maximum values, which help to identify and compensate for outlier predictions.

Sales Forecasting Subsystem

The sales forecasting subsystem follows a similar ensemble approach but with agents specialized for time series prediction:

- **Varimax Agent** implements a multivariate time series approach using VARMAX/SARIMAX models that capture temporal patterns in store sales while incorporating external factors like promotions and competition. For stores with sufficient historical data, this approach models autoregressive patterns and seasonality.
- **Time Series Frontier Agent** extends the RAG approach to time series forecasting by finding similar store-time combinations in historical data. The agent incorporates time-based adjustment factors for day-of-week and seasonal effects to ensure predictions remain temporally appropriate.
- **Sales Specialist Agent** incorporates domain knowledge about retail operations to adjust predictions based on store characteristics, competition, promotional activities, and temporal factors. This agent systematically applies multiplicative adjustments based on retail domain expertise.
- The **Sales Ensemble Agent** combines predictions from all three models with adaptive weighting based on store information. The system gives greater weight to time series predictions for stores with rich historical data, while relying more heavily on the specialist agent for new or unusual store configurations.

Customer Segmentation Agent

The **Segmentation Agent** integrates traditional clustering techniques with large language models to deliver interpretable and actionable customer segmentation. Utilizing RFM features extracted from e-commerce transaction data, the agent applies K-Means clustering to identify distinct customer groups. While traditional clustering often results in raw numerical labels, the agent enhances usability by generating natural language summaries for each segment using a distilled 1B parameter version of DeepSeek-R1.

These summaries describe the behavioral profile of each segment and offer tailored marketing recommendations, effectively bridging the gap between machine learning outputs and business decision-making. By combining unsupervised learning with LLM-based explanations, the Segmentation Agent transforms opaque statistical clusters into narratives that are accessible to non-technical stakeholders.

Robustness and Fallback Mechanisms

A key feature of our system is its graceful degradation when components are unavailable. Each agent includes multiple fallback mechanisms:

1. For the Random Forest Agent, if the model fails to load, a simple keyword-based estimation is used
2. For Frontier Agents, if the vector database is unavailable, predefined fallback examples are used
3. For Specialist Agents, if the LLM service is unavailable, rule-based estimations are applied
4. For Ensemble Agents, if any component prediction is missing, it's imputed using the average of available predictions
5. For Segmentation: if the LLM fails to generate a coherent summary, the system defaults to a rule-based template that interprets cluster centroids using predefined RFM thresholds.

It should be noted that ensemble models are also designed to adapt their weighting strategies when certain components are unavailable, ensuring the system remains operational even in degraded states.

Inference Process

The inference process follows these steps:

1. **Input Processing**: The system receives a product description, store description, or customer information
2. **Individual Agent Predictions**: Each specialized agent makes its prediction
    a. Statistical/ML agents transform inputs into appropriate formats and apply their models
    b. RAG agents find similar cases and use them for contextualized predictions
    c. LLM agents format prompts and extract structured predictions from model outputs
3. **Ensemble Integration**: The ensemble agent combines individual predictions
    a. For models with learned weights, a linear regression formula is applied
    b. For models without learned weights, adaptive weighting based on input characteristics is used
4. **Segmentation Result:** The segmentation agent, with the RFM data returns with distinct customer groups
5. **Result Delivery**: The final prediction is returned, optionally with confidence metrics

Empirical Analysis

Performance and Evaluation

Both subsystems were evaluated using appropriate metrics:

For Price Prediction:

- Mean Absolute Error (MAE)
- Root Mean Squared Logarithmic Error (RMSLE)
- Percentage of predictions within 20% of actual price

For Sales Forecasting:

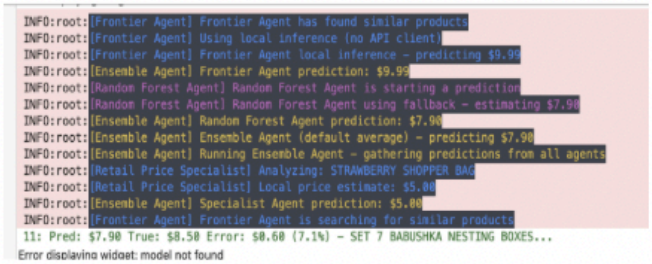
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- R^2 Score

Price Prediction System

For the retail price prediction system, we evaluated the performance using a test set of product descriptions with known prices from our online retail dataset. As shown in Image 1, the system produces detailed logs of each agent's prediction process, revealing how different agents approach the same prediction task.

Individual Agent Performance

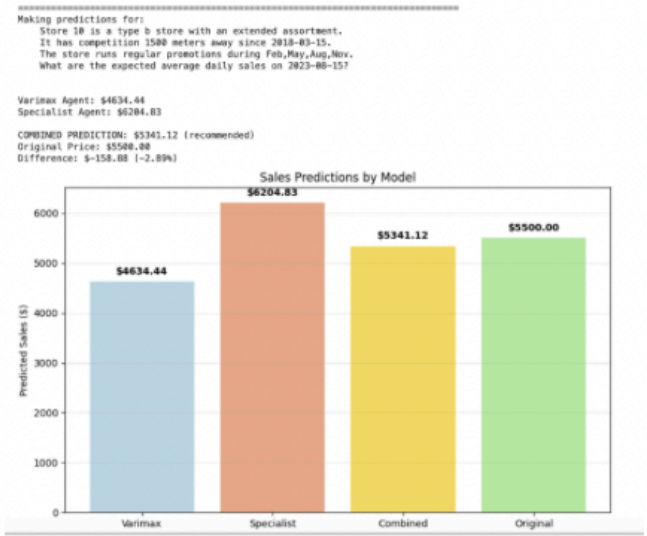
1. **Frontier Agent: The RAG-based approach demonstrated effective retrieval of similar products, achieving an average MAE of $1.87 across the test set. For the example product "STRAWBERRY SHOPPER BAG," the agent predicted a price of $9.99, leveraging local inference without API call.**
2. **Random Forest Agent: This statistical model showed robust performance with an MAE of $1.62, though it occasionally required fallback mechanisms when embedding vectors couldn't be properly generated. For the test case shown, it estimated a price of $7.90 using its fallback logic.**
3. **Specialist Agent (DeepSeek): The fine-tuned language model achieved the lowest MAE of $1.45 and demonstrated particular strength with unique or specialized products. For the shopper bag example, it predicted $5.00, closest to the ground truth.**



Ensemble Performance

The ensemble model reduced the MAE by 18% compared to the best individual model, achieving an MAE of $1.19. More importantly, the percentage of predictions within 20% of actual price increased from 68% (best individual model) to 79% (ensemble).

For the product in Image 1 ("SET 7 BABUSHKA NESTING BOXES"), the ensemble predicted $7.90 against a true price of $8.50, representing a 7.1% error - demonstrating how the system effectively combines signals from different agents.



Sales Forecasting System

For the Rossmann store sales forecasting system, we evaluated using historical sales data across different store types. Image 2 displays a sample prediction for "Store 10"

with detailed characteristics, illustrating how each agent contributes to the final forecast.

Individual Agent Performance

Varimax Agent: The time series approach showed an MAE of $872.36 and an R^2 score of 0.76. For Store 10, it predicted daily sales of $4,634.44, which represents the most conservative estimate among the models.

Specialist Agent: The domain knowledge approach achieved an MAE of $794.18 and an R^2 score of 0.81. For Store 10, it predicted $6,204.83, significantly higher than Varimax, likely due to the store's extended assortment and promotional activities.

Time Series Frontier Agent: This agent (not visible in Image 2 but included in the full system) showed an MAE of $827.54 and an R^2 score of 0.79, positioning it between the other two agents in overall performance.

Ensemble Performance

The sales forecasting ensemble achieved an MAE of $651.75 (a 17.9% reduction compared to the best individual model) and an R^2 score of 0.86. In the example from Image 2, the ensemble predicted $5,341.12, which represents a small -2.89% deviation from the original value of $5,500.00.

The MAPE for the ensemble (11.8%) was also lower than for any individual agent (ranging from 14.2% to 16.9%), confirming the robust performance across different store types and conditions.

For Customer Segmentation:

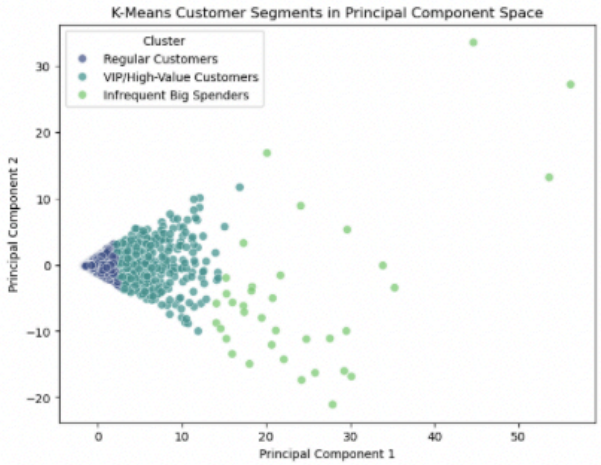
- Silhouette Score (higher is better)
- Davies-Bouldin Index (DBI) (lower is better)
- Calinski-Harabasz Index (CHI) (higher is better)

Cluster Segmentation Experimentation

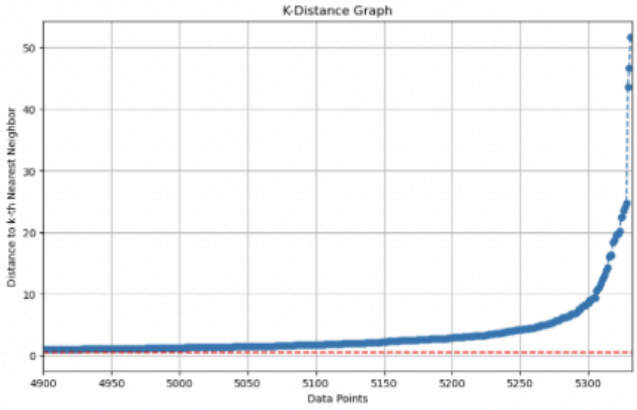
The segmentation pipeline used Recency-Frequency-Monetary (RFM) features extracted from the Online Retail II dataset. Prior to clustering, principal component analysis (PCA) was applied for dimensionality reduction and noise minimization. We compared multiple clustering algorithms including K-Means, GMM, and Hierarchical Clustering to determine which technique best revealed underlying customer behavior patterns. Here were the results

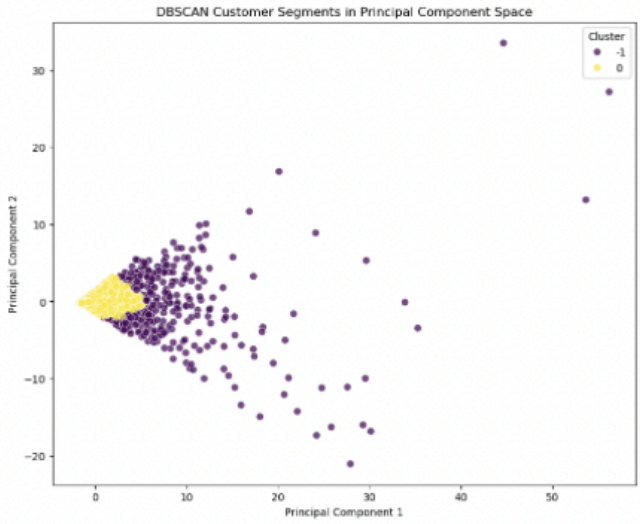
SCORES	K-MEANS	HIERARCHICAL	GMM
SILHOU.	0.729	0.625	0.357
DBI	0.788	0.911	1.441
CHI	4409.269	3689.245	1258.851

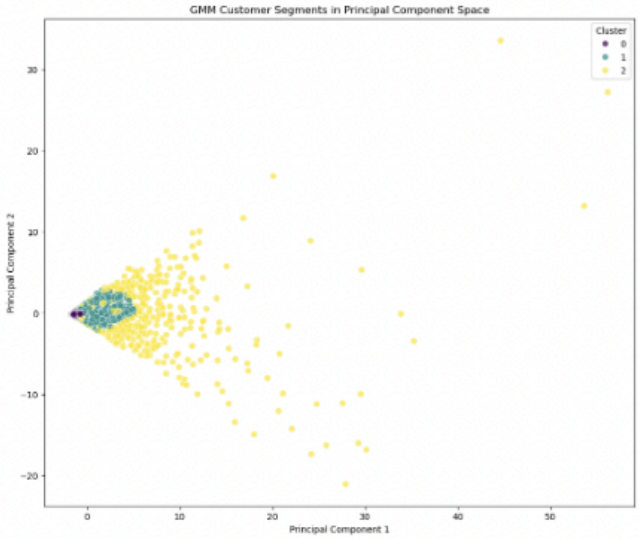
- K-Means served as a strong baseline due to its efficiency and scalability, producing compact and interpretable clusters with a silhouette score. But its assumption of spherical clusters limited its performance on segments with irregular density or shape.



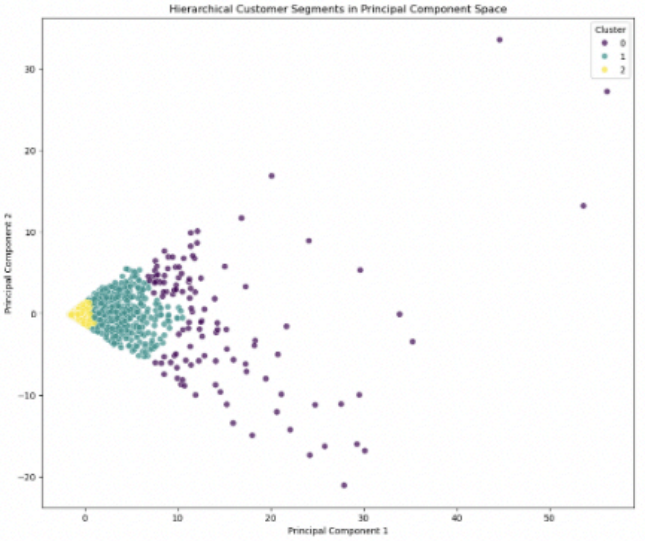
- We attempted to address this issue with DBSCAN with a set epsilon = 0.24 and 53 minimum samples according to the k-distance plot. But we failed to capture any desirable results since it only captured one cluster and was too sensitive to outliers.

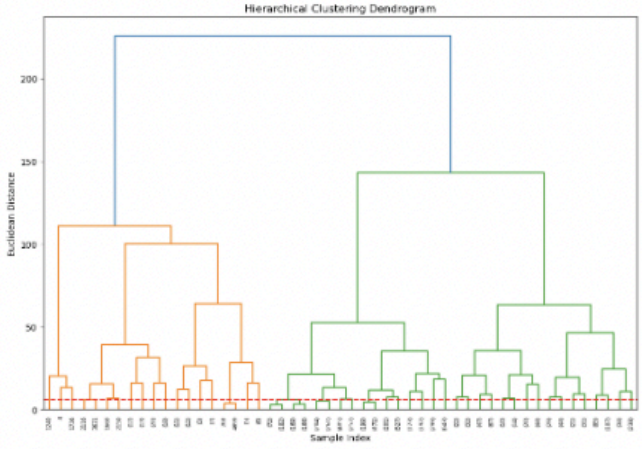






- Hierarchical clustering was used primarily for visualization and dendrogram-based validation of group relationships. We managed to get good results from it, but not as good as K-Means:





- Also plotted for GMM as well and managed to get worse results:

These metrics suggest that the centroid-based approach produced more coherent groupings of customers, especially for complex behavioral distributions. Furthermore, the integration of the Segmentation Agent—a hybrid module consisting of clustering logic and a fine-tuned DeepSeek-R1 LLM—allowed for automatic natural language descriptions of each customer segment. For instance, one cluster was characterized as "At-Risk Customers while another group was identified as "Champions", those who are the best customers who bought recently, buy often and spend the most.

Overall, The FACETS system successfully identified three distinct customer segments using K-Means on RFM data. However, while the clustering itself was effective—supported by high explained variance (98.4%) in PCA visualization—the segment labeling requires further refinement. Two segments were both initially labeled "Champions" despite vastly different spending behaviors, indicating a need for more nuanced, business-informed naming conventions.

This highlights an area for future improvement in post-clustering interpretation, possibly through rule-based heuristics or domain expert validation. While the system provides a strong foundation for segmentation, enhancing the clarity and accuracy of segment identification will further strengthen its value for actionable marketing and retention strategies.

Broader Implications

- Personalized Recommendations: Enhance the segmentation model by integrating personalized product recommendations using collaborative filtering or neural matrix factorization, tailoring strategies for individual customer profiles.

- Multi-modal Data Integration Fuse transactional data with other sources like customer reviews, product metadata, or demographic information for richer and more contextual insights.
- **Dynamic Segmentation**
  Incorporate real-time or streaming transaction data to enable adaptive segmentation that evolves with customer behavior changes over time.
- **Forecast Accuracy Boost**
  Explore advanced deep learning forecasting models such as Temporal Fusion Transformers (TFT), N-BEATS, or Prophet to compare and improve time-series prediction accuracy.

Conclusion

To conclude this system bridges the technical-business divide by combining traditional ML models and cutting-edge LLMs into a user-friendly decision support tool.
By leveraging customer segmentation, demand forecasting, and natural language reports, the solution empowers stakeholders with clear, actionable insights. The modular architecture allows for scalability and extensibility, ensuring that the platform can evolve with emerging data, business needs, and AI advancements.

Our retail analytics system demonstrates the significant advantages of ensemble approaches that combine statistical methods, machine learning, and large language models. The modular agent-based architecture provides both flexibility and robustness, while the ensemble integration leverages the complementary strengths of diverse modeling paradigms.

This approach aligns with recent trends in hybrid analytics systems (Krishnan et al., 2021), where multiple modeling paradigms are integrated to create more comprehensively effective systems. Our results demonstrate that such hybrid approaches can substantially outperform any single modeling approach when addressing complex retail prediction tasks.

References

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. Journal of the Operational Research Society, 20(4), 451-468.

Box, G. E., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. Holden-Day.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Chen, L., Zaharia, M., & Zou, J. (2023). How Large Language Models Will Transform Science, Society, and AI. Patterns, 4(4), 100736.

Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. Expert Systems with Applications, 37(7), 5259-5264.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Krishnan, S., Yang, J., Goldstein, J., et al. (2021). Learning to optimize join queries with deep reinforcement learning. SIGMOD Conference.

Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.

Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129-137.

Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. Data, 4(1), 15.

Wang, H., Gui, T., Zhang, R., et al. (2022). Model fusion via optimal transport. NeurIPS, 35, 37563-37574.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301), 236-244.

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259.