**Data Scientist Assignment: AI-Driven Personalized**

**Marketing Content Generation**

This project leverages Machine Learning (ML) and Generative AI (GenAI) to analyze customer transaction patterns, segment customers, and generate personalized marketing content

**Approach and methodology**

**Feature Engineering and Segmentation**

The process begins with EDA & feature engineering to derive key attributes from transactional data, and users data uncovering patterns in customer behavior. Using clustering algorithms, customers are segmented into distinct groups, with each cluster representing a unique behavioral profile. These profiles enable targeted marketing strategies

**Personalized Content Generation**

Customer profile data, derived features, and cluster insights are used to generate personalized emails, taglines, and summaries through Microsoft Phi3, a large language model (LLM). These outputs align with customer needs and marketing objectives. For visual content, Stable Diffusion generates marketing images based on concise customer summaries.

**Integrated Workflow**

A user-friendly UI consolidates the process, allowing users to input a client\_id and generate personalized marketing content, including emails, taglines, summaries, and images, streamlining content creation for targeted campaigns.

This approach combines feature engineering, clustering, and GenAI to deliver an efficient, end-to-end solution for customer segmentation and personalized marketing.

**Feature engineering decisions**To analyze customer transaction behavior and enable clustering, we created the following feature types:

1. **Static Features**
   * Key metrics: total spending, average transaction value, and transaction frequency.
   * Transaction type analysis (Chip vs. Non-Chip) for client-level insights.
2. **Rolling Transaction Features**
   * Rolling sums, averages, max, and min spending over 3, 6, and 9 months.
   * Growth rates and spending volatility to assess behavior over time.
3. **Merchant Spending Patterns**
   * **Top Categories**: Identifies top 5 spending categories per client.
   * **Category Diversity**: Measures spending variety across unique merchant categories.
   * **Unique Merchants**: Tracks distinct merchants per client.
4. **Transaction Timing Analysis**
   * Average time between transactions and time differences to identify spending regularity.
5. **Spending Patterns and Behaviors**
   * **Consistency**: Spending proportions in recent periods (e.g., last 3, 6, 9 months).
   * **Payment Methods**: Ratios of Chip, Online, and Swipe transactions.
   * **Growth & Changes**: Absolute and percentage changes in spending across periods.
   * **Proportions**: Spending comparisons between recent and earlier periods.

These features provide a comprehensive view of customer transaction habits, enabling effective clustering and personalized insights.

**Clustering approach and justification**

Derived numerical features were analyzed for clustering, with highly correlated features removed to reduce dimensionality. Data was standardized to ensure uniform scale and subjected to PCA to capture maximum variance. The optimal number of clusters was determined using the Elbow Method and Silhouette Score. KMeans++ yielded the best clusters, compared to hierarchical and GMM methods. Clusters were visualized using t-SNE, revealing overlap and indicating scope for additional features to enhance segmentation.

**Cluster Profiling**

1. **Spending Patterns:** Assessed total spending, transaction frequency, and average transaction value.
2. **Financial Metrics:** Evaluated income, debt, and credit scores to gauge financial health.
3. **Transaction Behavior:** Analyzed payment method preferences (chip, online, swipe) for digital adoption insights.
4. **Demographic Indicators:** Factored in age and retirement stage for tailored profiles.
5. **Targeted Strategies:** Named clusters based on patterns to align with marketing and banking products like loans and credit cards.

These are 6 cluster profiles created after examining the data

cluster\_profiles = {

'0': { 'Clustername': 'High Spenders with Diverse Transactions',

'Spending': 'Highest total spending ($69,651), average transaction size ($68), and transaction count (1,065)',

'Transaction Types': 'Predominantly chip-based (~76%), with some online (10%) and swipe transactions (14%)',

'Demographics': 'Moderate income ($38,863 per capita), higher debt ($112,426), and lower credit scores (702)',

'Age': 'Average age is 51'

},

'1': { 'Clustername': 'Moderate Spenders with Consistency',

'Spending': 'Moderate spending ($23,598) and smaller average transactions ($36)',

'Transaction Types': 'Higher proportion of chip transactions (~79%)',

'Demographics': 'Lower income ($18,175 per capita), moderate debt ($55,578), and good credit scores (715)',

'Age': 'Slightly younger demographic (average age 50)'

},

'2': { 'Clustername': 'Value-Conscious Users',

'Spending': 'Lower spending ($28,521) and moderate transaction counts (659)',

'Transaction Types': 'Predominantly chip-based, with fewer online and swipe transactions',

'Demographics': 'Lower income ($20,806 per capita) but relatively low debt ($53,509)',

'Age': 'Average age is 51'

},

'3': { 'Clustername': 'Mid-Tier Spenders with Frequent Transactions',

'Spending': 'Mid-level spending ($37,504) and higher transaction counts (915)',

'Transaction Types': 'Predominantly swipe-based (~65%), with fewer chip transactions',

'Demographics': 'Moderate income ($22,370 per capita), moderate debt ($59,842), and good credit scores (720)',

'Age': 'Younger demographic (average age 49)'

},

'4': { 'Clustername': 'High-Volume Spenders',

'Spending': 'High spending ($55,620) with many transactions (1,444)',

'Transaction Types': 'Balanced between chip and online transactions',

'Demographics': 'Moderate income ($24,012 per capita), moderate debt ($54,292), and average credit scores (702)',

'Age': 'Slightly older demographic (average age 54)'

},

'5': { 'Clustername': 'Older, Conservative Spenders',

'Spending': 'Moderate spending ($47,810) with significant chip-based transactions (~81%)',

'Transaction Types': 'Mostly chip and fewer online or swipe transactions',

'Demographics': 'Lower debt-to-income ratio (0.94), strong credit scores (724)',

'Age': 'Oldest group (average age 59)'

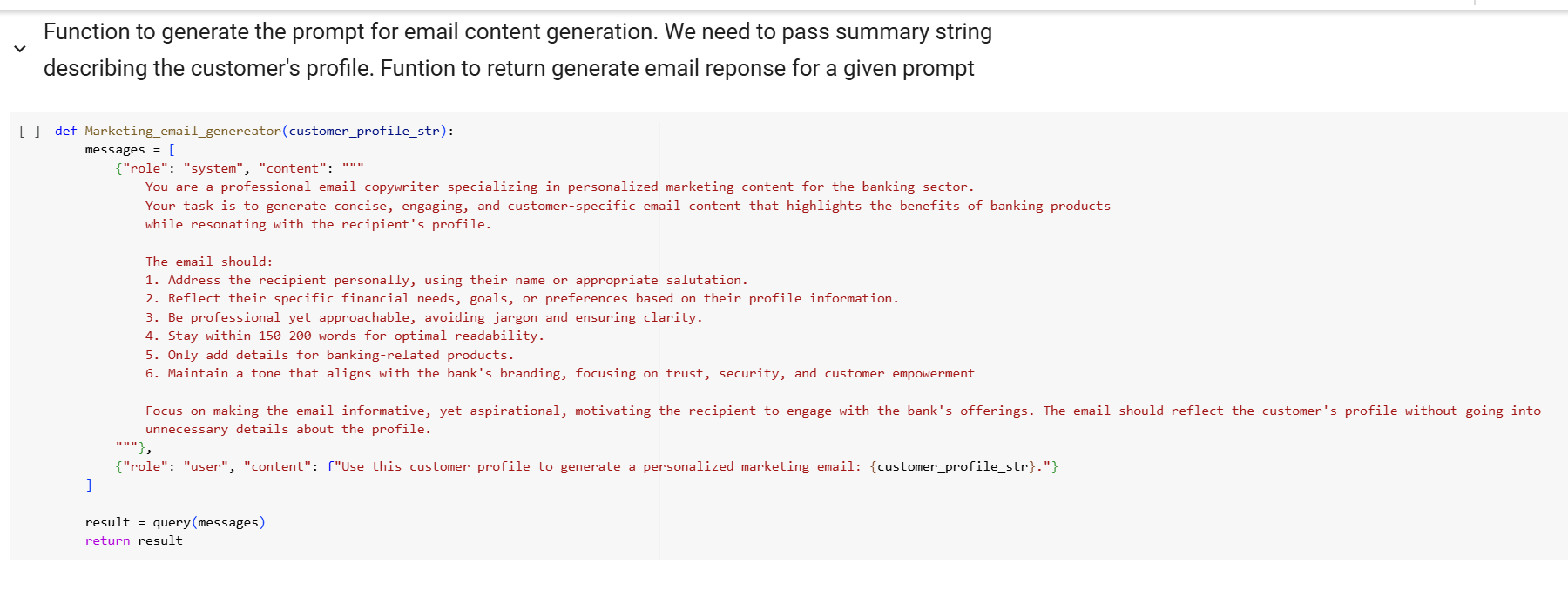
}

}

**Prompt engineering strategy**

To generate personalized content for customers, open-source LLM models were employed, designed specifically for text generation tasks. Crafting effective prompts was critical to achieving high-quality responses. Since the tasks varied across email generation, taglines, customer summaries, and image generation, distinct prompts were developed for each type of content.

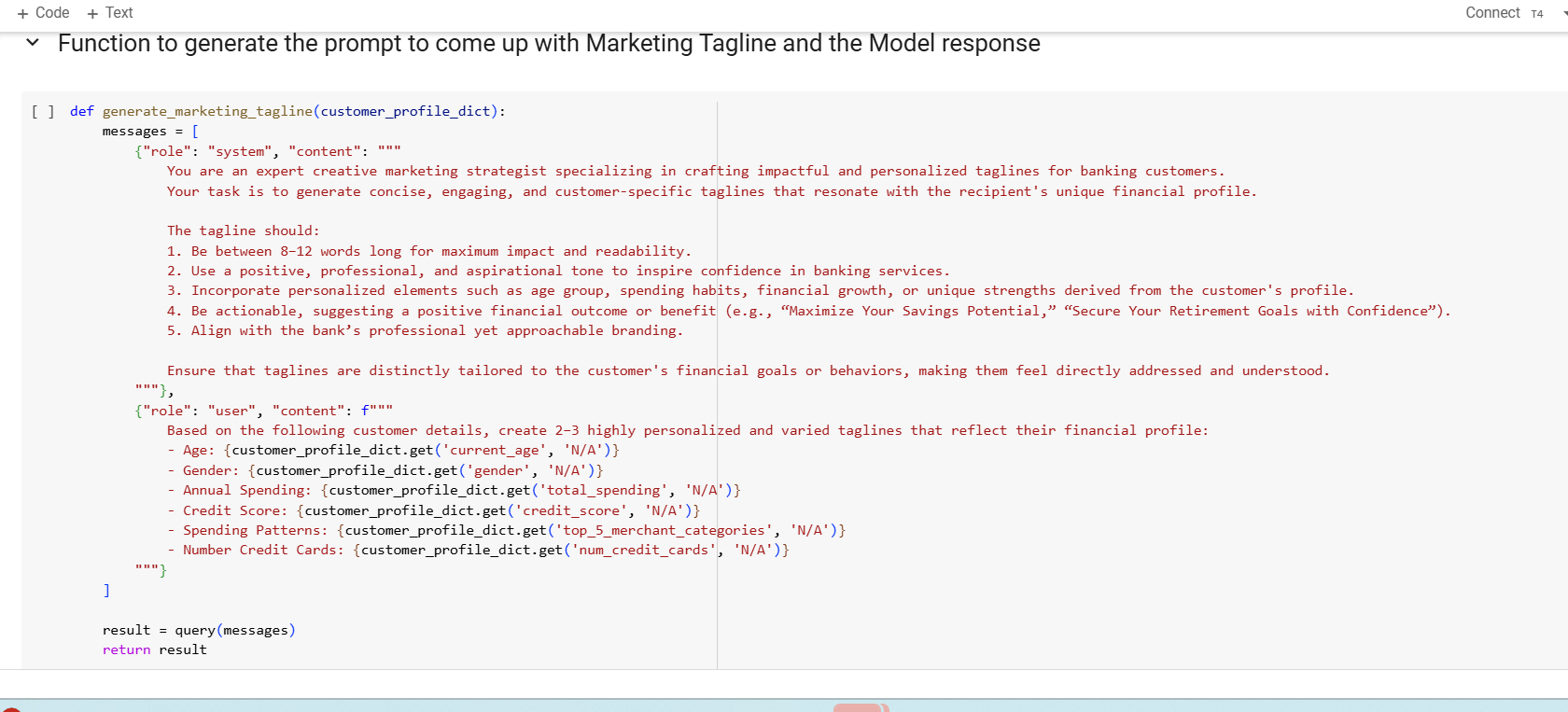
**Email Generation:**



* The prompt defined the role as: *"A professional email copywriter specializing in personalized marketing content for the banking sector."*
* Clear instructions were provided regarding the tone (professional yet engaging) and the structure of the email.
* Inputs included a detailed customer profile string containing derived features such as spending habits, credit score, income, debt levels, and the corresponding cluster profile summary. These elements guided the LLM to generate contextually relevant and tailored email content

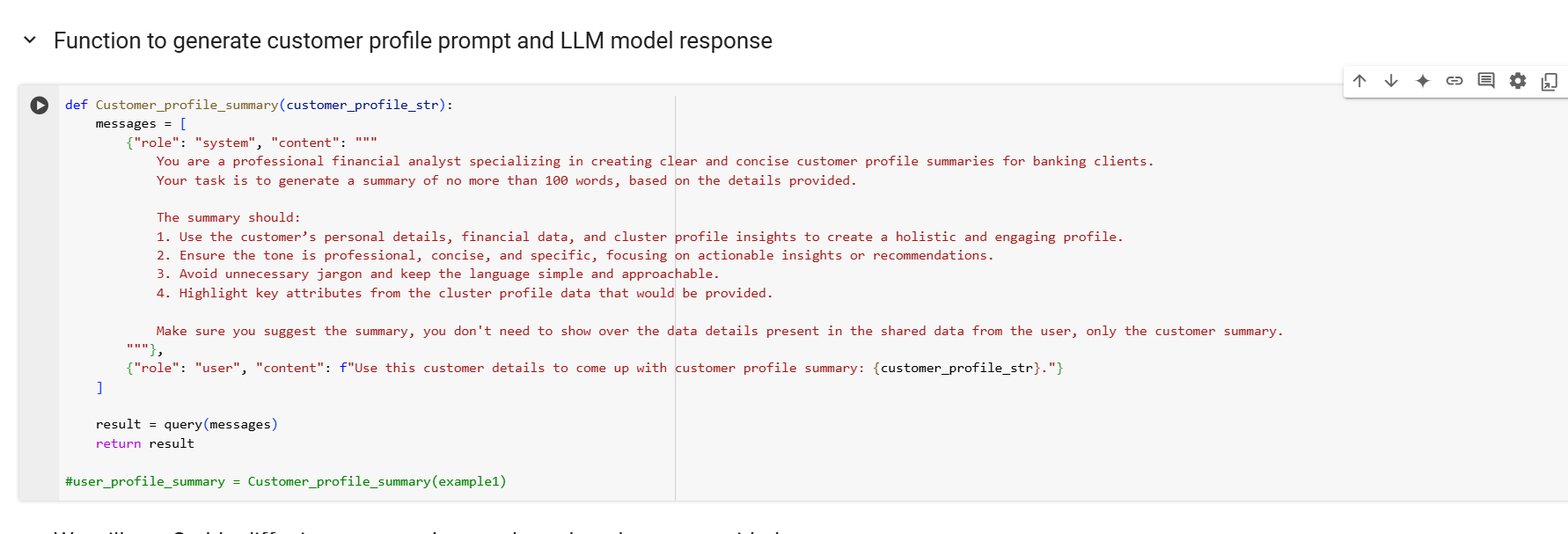
**Tagline Generation:**

* The prompt emphasized the role of: "An expert marketing strategist creating personalized and impactful taglines for banking customers."
* Detailed instructions were given to ensure the tone was concise, aspirational, and customer-focused.
* Inputs included specific financial data, enabling the LLM to craft a tagline that resonated with the customer’s financial profile and goals.



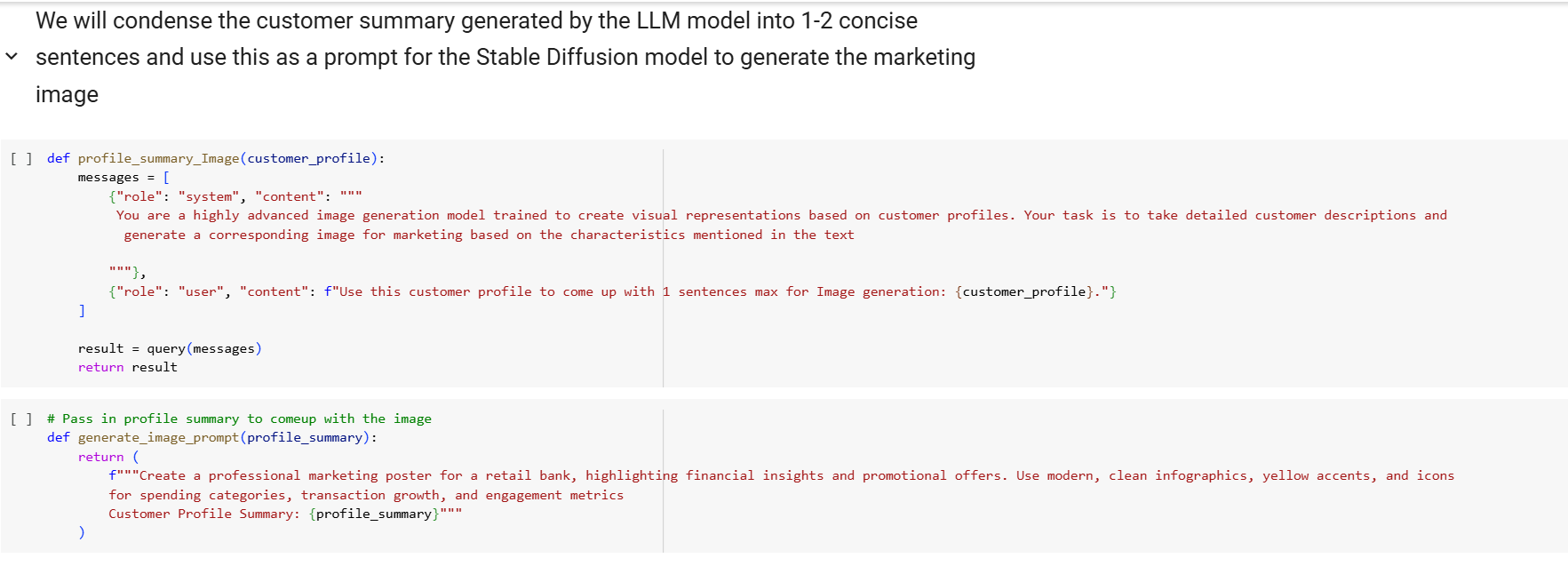
**Customer Profile Summary:**

* The prompt was designed to generate a summary in a clear and engaging tone. It provided explicit instructions for structure and language to ensure consistency.
* Inputs featured detailed customer attributes, enabling the LLM to produce a comprehensive yet concise profile overview.



**Image Generation :**

* The Stable Diffusion model from Hugging Face was used for this task.
* The prompt included the customer profile summary and explicitly outlined the role of the model: "Generate a visually compelling image that aligns with the customer’s summarized profile and preferences."
* This ensured that the generated image visually represented the customer’s traits and financial persona effectively.



**LLM APP UI**

**System architecture diagram**

Email Prompt

Tagline Prompt

Clustering Model

EDA

Transaction Data

Feature generation

Customer Summary

Top Model Selection

Users Data

Feature Selection

Image Generation nn

**Assumptions, limitations and potential improvements**

Given the four-day timeframe for the assignment, certain areas could have been explored more deeply to enhance both clustering accuracy and content generation quality:

1. **Feature Engineering**:  
   Additional derived features could have been created with more time and collaboration with the business team. These features might capture nuanced customer behaviors and improve clustering outcomes.
2. **Clustering Approach**:  
   Current clustering shows overlap among customer groups, indicating that soft clustering techniques like Gaussian Mixture Models (GMM) could be more effective. With better parameter tuning and feature refinement, GMM might provide more accurate segmentations.
3. **Content Generation**:

* Incorporating advanced techniques like few-shot prompting could refine the model’s ability to generate personalized emails, taglines, and summaries.
* Fine-tuning open-source models with domain-specific examples would enhance the contextual relevance and quality of generated text.

1. **Image Generation**:  
   Stable Diffusion results were not fully aligned with customer profiles. Fine-tuning the model or leveraging more advanced image generation tools, such as DALL-E, could yield better results. Hardware constraints limited testing with larger models, which could be revisited with upgraded resources.
2. **Model Exploration**:  
   Exploring other open-source models like LLAMA or Falcon could offer improvements in both text and image quality, particularly with proper parameter adjustments or fine-tuning.