**MIS 5470 | Insurance Customer Segmentation Analysis |**

**By Alex Ordonez & Vishaka Sharma**

****

# **Final Project Report**

# ***1. Introduction:***

We aim to identify distinct groups of insurance customers based on their characteristics and behaviors. By segmenting the customer base, we uncover valuable insights that inform business strategy, enhance customer engagement, and support decision-making in marketing, service, and product development.

***2. Business Objective & Problem Statement:***

We are developing an **interactive web application** using Python (**Streamlit**) that enables insurance companies to visually explore and segment their clients using the optimal clustering model. The tool will help businesses:

* Dynamically group customers based on real-time data.
* Visualize clustering results and patterns through insightful plots.
* Make informed decisions on pricing and personalized policy offerings.

The project addresses the challenge of lacking granular understanding of the customer base, leading to inefficient marketing and generic offerings.

**Objectives:**

* Identify homogeneous customer segments.
* Improve targeted marketing and sales efforts.
* Enhance customer relationship management.
* Inform pricing and product development strategies.

**Problem Statement:** Organizations struggle with understanding their diverse customer base. This limits their ability to personalize interactions, tailor offerings, and optimize resource allocation across marketing and service channels.

***3. Dataset Overview:***

The dataset consists of demographic and behavioral information about insurance customers, including:

* Time\_to\_Conversion: Duration until policy purchase.
* Conversion\_Status: Indicator of successful conversion.
* Customer Demographics: Age, marital status, etc.
* Insurance Credit Score: A key financial variable used for segmentation.

***4. Data Cleaning & Preprocessing:***

* **Handling Missing Values**: Median imputation applied to numeric features.
* **One-Hot Encoding**: Converted categorical variables to numeric format.
* **Outlier Handling**: Used IQR-based filtering to remove extreme values.
* **Feature Scaling**: Standardized features to have zero mean and unit variance using StandardScaler.
* **Redundancy Removal**: Dropped highly correlated features (Time\_to\_Conversion vs. Conversion\_Status).

***5. Exploratory Data Analysis (EDA):***

* **Spearman Correlation Matrix**: Highlighted relationships between features.
* **Distributions**:
  + *Age*: Centered around the working-age population.
  + *Premium Amount*: Right-skewed with high-value outliers.
  + *Claims Frequency*: Most customers file 0–1 claims.
  + *Credit Score*: Normal distribution with useful segmentation potential.
  + *Website Visits*: Indicates digital engagement.
  + *Total Discounts*: Bimodal distribution reveals pricing sensitivity.
* **Visual Analyses**:
  + Conversion rates by region and policy type.
  + Credit score by conversion status.
  + Boxplots before/after outlier removal.

***6. Clustering Approaches:***

We implemented and compared multiple clustering algorithms:

#### **6.1 KMeans Clustering**

* Tested K = 2 to 9.
* Selected optimal K by maximizing Silhouette Score.
* Optimal K = value with highest score
* Silhouette Score: 0.443

#### **6.2 Gaussian Mixture Models (GMM)**

* Explored K = 1 to 9.
* Selected best K using Silhouette Score.
* Silhouette Score: 0.443

#### **6.3 Hierarchical Clustering**

* Used Ward linkage.
* Evaluated cluster sizes using silhouette method.
* Best Performer with Silhouette Score: 0.517

#### **6.4 DBSCAN**

* Applied PCA to retain 80% variance.
* Used k-distance + KneeLocator to find optimal eps.
* Silhouette Score: 0.425

***7. Additional Application Development (Python Stream lit) for multiple datasets:***

We developed an application that helps the company to automate this analysis, through to the phases:

1. **Data Cleaning and Preprocessing**:

* Missing Values Handling, Categorical Values processing and Data Scaling.

1. **Exploratory Data Analysis:**

* Giving the chance to the user to create their own plots (Histograms, scatter plots, box plots) with their data.

1. **Clustering Models:**

* Implementing optimized clustering models, enhancing performance based on silhouette score.

1. **Retrieving Clustered information:**

* The user can download the results in a CSV format.

The application will enhance the performance of the team without necessarily implementing the clustering models from zero, to have their clients clustered. And will also have the chance to determine which features will be the most important for the clustering models to work.

***8. Cluster Profiling & Insights:***

This crucial stage involves analyzing the characteristics of the identified clusters to understand the distinct customer segments. By examining the distribution of various features within each cluster, the project aimed to develop profiles for each group. This involves identifying the key attributes that differentiate the clusters and uncovering meaningful insights about their behaviors, preferences, and potential value.

Findings include:

* High-value urban customers with high premiums and low claims.
* Price-sensitive rural customers with fast conversion.
* Young digital-first customers with high website engagement.

These profiles help drive targeted strategies and align offerings with customer expectations.

***9. Recommendations for Business Strategy:***

* **Targeted Campaigns**: Use segment-specific language, timing, and channels.
* **Personalized Offerings**: Bundle or recommend products based on segment behavior.
* **Customer Service Optimization**: Tailor service protocols to each cluster's expectations.
* **Pricing and Risk Models**: Adjust pricing or underwriting based on segment risk profiles.

***10. Conclusion:***

The clustering project revealed distinct, actionable customer segments using data-driven analysis. The approach combined robust preprocessing, multiple clustering techniques, and business-aligned insights to improve marketing, pricing, and customer engagement strategies.

### ***Key Takeaways***

* This project adopted a robust data-driven approach to segment insurance customers using various clustering techniques.
* Extensive efforts were made in data cleaning, preprocessing, and exploratory data analysis to ensure model readiness and data quality.
* Multiple clustering algorithms—KMeans, GMM, Hierarchical, and DBSCAN—were implemented and compared using evaluation metrics such as Silhouette Score.
* Beyond just identifying clusters, the project focused on **profiling segments** and extracting **business-relevant insights** from them.
* The ultimate objective was to leverage clustering for better strategic decisions in marketing, customer service, pricing, and product development.

***10. Appendix:***

* Spearman Correlation Matrix
* PCA visualizations of clusters
* Code snippets for KMeans, GMM, Hierarchical, and DBSCAN implementations

### ***Model Comparison Summary***

| **Model** | **Silhouette Score** | **Notes** |
| --- | --- | --- |
| Hierarchical | **0.517** | Best overall separation |
| KMeans | 0.443 | Efficient and stable |
| GMM | 0.443 | Good for overlapping clusters |
| DBSCAN | 0.425 | Needs tuning; good for noise |

Add - how this benefits, solves real world problems