# Physician Targeting & Segmentation: A Competitive Intelligence product for Pharmaceutical Retail Drug profits

**Surya Chandra Raju Kurapati**
*Master's in Data Analytics - School of Computing,*
*National College of Ireland, Dublin, Ireland,*
*x23396920@student.ncirl.ie*

**Abstract -** This report addresses the problem of optimizing pharmaceutical sales by identifying and segmenting physicians based on their prescription behaviour in a competitive market. A pharmaceutical company aims to predict physician adoption of a newly launched drug (in this study, a general therapeutic drug) and to segment physicians for targeted marketing. The study utilizes machine learning models, specifically Logistic Regression and Gradient Boosting, for predictive modelling of prescription probability, referred to as headroom analysis throughout the study, and K-Means clustering for physician segmentation. Key findings include a predictive model with 99% ROC AUC, enabling accurate identification of high-potential prescribers, and a four-tiered physician segmentation that highlights distinct characteristics within each group. The analysis identifies competitor drug prescription volume, physician years of experience, trust in pharma, call sample receive rate, and seminar attendance as key factors influencing prescription behaviour. These results provide actionable insights for prioritizing physician targeting, optimizing marketing strategies, and potentially increasing new drug adoption rates and sales.

**Keywords -** Physician, General Therapeutic, Headroom Analysis, Physician Targeting, Segmentation, Logistic Regression, Gradient Boost Classifier, K-Means Clustering, Principal Component Analysis, Pharmaceutical Sales and Marketing, Retail Industry, Healthcare Analytics, Prescription Analysis

## 1. INTRODUCTION

The pharmaceutical industry faces the challenge of effectively marketing new drugs to physicians through Medical Representatives. With numerous healthcare providers and diverse prescription behaviours, it is crucial to identify and target the most receptive physicians. This report details an analysis conducted for a Pharmaceutical company, which seeks to optimize the launch of a new general therapeutic drug. The core objectives are twofold: (1) to predict which doctors are most likely to prescribe the new drug (headroom analysis), and (2) to segment doctors into meaningful tiers to inform targeted marketing strategies. Machine learning techniques are employed to analyse physician profile data and drug prescription records, enabling data-driven decision-making in pharmaceutical sales and marketing. The analysis aims to identify key characteristics that influence prescription behaviour and provide actionable insights for optimizing medical representative call plans and resource allocation.

## 2. LITERATURE REVIEW

Traditional approaches—largely based on demographic data or historical prescribing patterns—are increasingly being replaced by data-driven methods that incorporate behavioural and psychographic insights. According to Deloitte (2020) [2], companies using predictive analytics and advanced segmentation have experienced improved marketing return on investment (ROI) and better physician engagement. McKinsey & Company (2019) [4] further assert that multidimensional segmentation models, which include factors like specialty, patient volume, digital behaviour, and prescribing trends, allow for more accurate and personalized targeting strategies.

Academic literature also supports the adoption of machine learning techniques in this domain.

Unsupervised learning methods such as K-means and hierarchical clustering are commonly used to group physicians based on similar behaviours or characteristics without pre-labelled data (Chen et al., 2021) [1]. Additionally, supervised algorithms like random forests and decision trees have shown promise in predictive modelling for physician responsiveness (Singh & Gaur, 2020) [5]. Despite these technological advancements, ethical considerations remain critical—especially concerning data privacy, transparency, and regulatory compliance, as mandated by laws such as HIPAA (Kostkova et al., 2016) [3]. These studies collectively point to a shift from intuition-driven to evidence-based strategies in physician segmentation and engagement.

## 3. UNDERSTANDING DATASET(S)

The analysis utilizes two primary datasets (Refer to Table 1):

- **doctors_details.json:** An semi-structured dataset containing profile information for 500 doctors, including demographics (age, gender), practice details (specialty, years of experience, location), prescription behaviour metrics, pharma company engagement metrics, and influential patterns such as seminar attendance.
- **drug_transactional_data.csv:** This dataset comprises prescription records for 10 drugs across the 500 doctors, totalling 5000 records. It includes drug details (name, type, manufacturer, cost), prescription outcomes (whether prescribed), and doctor characteristics (mirroring the JSON data).

Table 1: Understanding the dataset

| Feature Category | Feature Name | Feature Description | Sample Values |
|---|---|---|---|
| Demographics | doctor_id | Unique identifier for the doctor | DOC001, DOC045 |
| Demographics | age | Doctor's age | 45, 58, 66 |
| Demographics | gender | Doctor's gender | Male, Female, Other |
| Demographics | specialty | Doctor's medical specialty | Cardiologist, Gen Physician |
| Practice Details | years_experience | Number of years doctor has been practicing | 5, 18, 30 |
| Practice Details | location | Location type where doctor practices | Urban, Suburban, Rural |
| Practice Details | hospital_affiliation | Whether the doctor is affiliated with a hospital | Yes, No |
| Practice Details | num_patients_per_day | Average number of patients seen daily | 20, 35, 50 |
| Prescription Behaviour Metrics | prescription_volume | Total number of prescriptions written by the doctor | 150, 320, 78 |
| Prescription Behaviour Metrics | prescribed | Whether the drug was prescribed in the record | Yes, No (or 1, 0) |
| Engagement Metrics | attended_drug_seminar | Doctor's attendance at pharma-sponsored seminars | Yes, No |
| Engagement Metrics | received_sample | Doctor received drug samples | Yes, No |
| Influential Metrics | drug_price_sensitivity | Sensitivity of doctor to prescribe drug based on the drug price (higher = more sensitive) | 1, 3, 5 |
| Influential Metrics | trust_in_pharma | Trust level of doctor in pharmaceutical companies | 1, 4, 5 |
| Drug Details | drug_name | Name of the prescribed drug (generic or branded) | Aspirin, Lipitor |
| Drug Details | brand_name | Commercial brand name of the drug | Lipitor, Plavix |
| Drug Details | generic_drug_name | Generic chemical name of the drug | Atorvastatin, Clopidogrel |
| Drug Details | manufacturer | Pharmaceutical company manufacturing the drug | Pfizer, Eli Lilly, Bayer |
| Drug Details | drug_type | Category/type of the drug (e.g., Antidepressant) | Analgesic, Antihypertensive |
| Drug Details | average_cost_eur | Average cost of the drug in Euros (€) | 8.2, 18.5, 28 |
| Drug Details | efficacy_rating | Clinical effectiveness of the drug (higher = better) | 3, 4, 5 |
| Drug Details | side_effect_score | Side effect severity score (higher = more side effects) | 0.2, 0.4, 0.7 |
| Pharma Marketing Influence | marketing_push | Intensity of marketing push by pharma company | Low, Medium, High |

## 4. METHODOLOGY

### *A. Data Preprocessing*

Before applying classification and clustering algorithms, the dataset underwent preprocessing to ensure data integrity and suitability for analysis. Initial inspections were performed using the info() method on both the drug transaction data and the doctor details data. This confirmed that there were no missing values in either dataset. The data was also checked for duplicate records. No exact duplicate records were found in either the physician or drug data. Finally, the numerical variables in both datasets were identified, and their statistical distributions were examined using the describe() method to understand their ranges and central tendencies. This analysis revealed that doctors in the dataset have between 1 and 39 years of experience, with the majority having around 20 years of practice. On average, these doctors see approximately 30 patients per day, maintain a prescription volume of 175, and predominantly prescribe drugs with an efficacy rating above 3 and minimal side effects. This comprehensive assessment provided confidence in the dataset's quality and its suitability for addressing the use-case and problem statement.

### *B. Exploratory Data Analysis (EDA)*

To gain deeper insights into the dataset, a thorough exploratory data analysis (EDA) was performed to understand both the distribution of individual variables and the relationships between them. Key variables including age, specialty, trust in pharmaceutical companies, marketing push, price sensitivity, and receipt of samples were analysed in relation to the target variable, prescribed. In addition, a correlation heatmap was generated to explore the relationships between continuous variables such as years of experience, number of patient visits per day, prescriber prescription volume, and drug price sensitivity.

The comprehensive variable analysis revealed distinct patterns influencing physician prescription behaviour. Fig 1, showcasing prescription rates by specialty, highlights that prescribing behaviour is specialty-dependent, with General Physicians and Cardiologists exhibiting slightly higher prescription rates. Fig 2 illustrates that price sensitivity exerts only a minor influence, with prescribing doctors showing marginally higher sensitivity. Fig 3 emphasizes trust in pharmaceutical companies as a significant driver, where higher levels of trust directly correlate with increased prescription likelihood. Furthermore, Fig 4 clearly demonstrates that a strong marketing push substantially boosts prescription rates, positioning marketing efforts as a critical strategic lever. Collectively, these findings suggest that trust in pharma and marketing intensity are the most influential factors, while price sensitivity, specialty, and age play secondary roles.

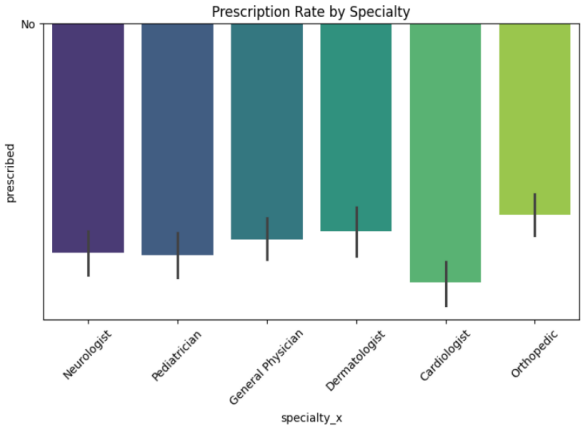


Fig. 1. Prescription Rate by Specialty

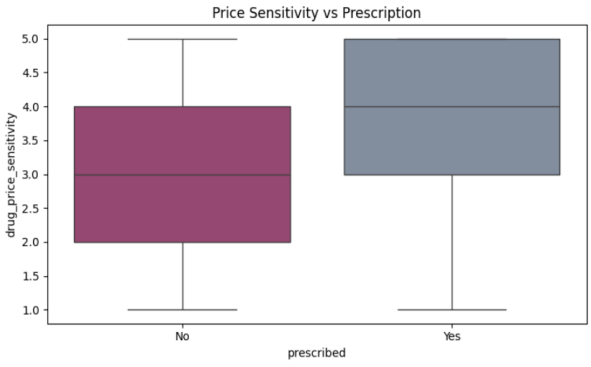


Fig. 2. Sensitivity vs Prescription

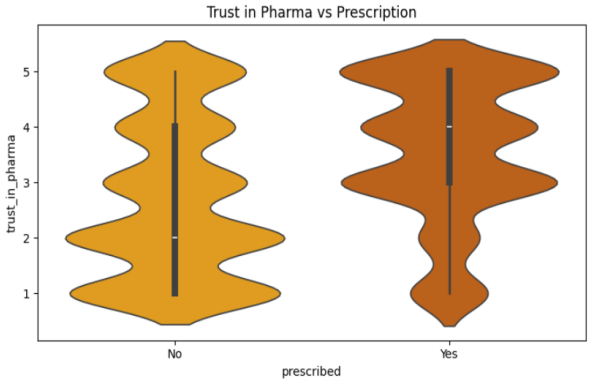


Fig. 3. Trust in Pharma vs Prescription

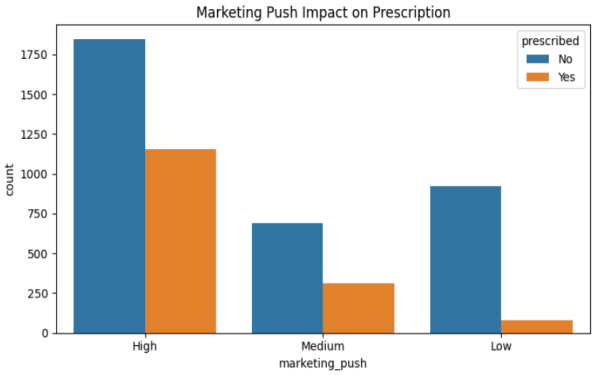


Fig. 4. Marketing Push Distribution

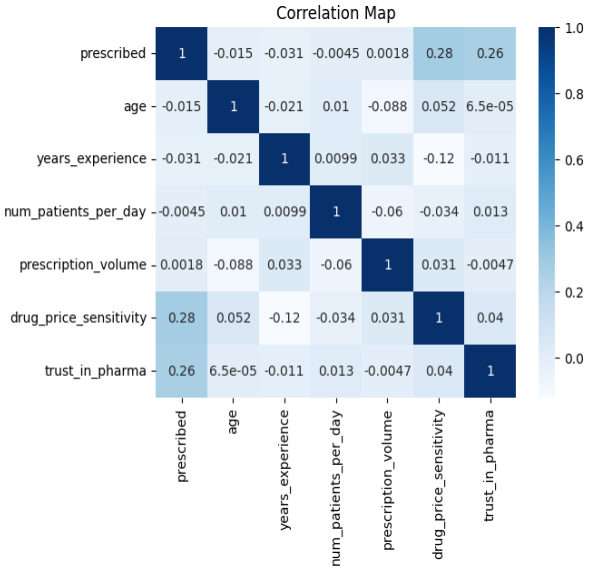


Fig. 5. Correlation Map

Additionally, the correlation analysis (Fig 5) reinforces these findings by identifying two primary numerical factors influencing prescriptions: drug price sensitivity (correlation coefficient of 0.28) and trust in pharma (correlation coefficient of 0.26), both showing the strongest positive associations with prescribing behaviour. Conversely, variables such as years of experience (-0.031) and patient volume (-0.0045) exhibit weak negative correlations, indicating they are poor predictors of prescription likelihood. Importantly, the absence of high correlation values among predictors (all correlations are less than 0.2) confirms that there is no issue of multicollinearity, ensuring the reliability of subsequent modelling efforts.

### *C. Feature Extraction and Scaling*

Following the exploratory data analysis, feature extraction and scaling were conducted to prepare the dataset for predictive modelling. Key doctor-level attributes such as age, gender, specialty, years of experience, location, hospital affiliation, number of patients per day, prescription volume, marketing push exposure, and receipt of samples were aggregated from transactional records. Categorical variables including gender, specialty, location, hospital affiliation, attended drug seminar, received sample, and marketing push were encoded using one-hot encoding, allowing the model to interpret these non-numeric attributes without imposing any artificial ordering. This transformation enabled the inclusion of important categorical influences on prescription behaviour while preserving the natural structure of the data.

For numerical features such as age, years of experience, number of patients per day, prescription volume, drug price sensitivity, trust in pharma, average cost, efficacy rating, side effect score, sample received ratio, and seminar attended ratio, standardization was performed using StandardScaler. This scaling normalized the features to have a mean of zero and a standard deviation of one, ensuring that variables with larger numeric ranges did not disproportionately influence the model.

The combined application of one-hot encoding for categorical variables and standard scaling for numerical variables created a balanced,

structured input space, optimizing the dataset for downstream algorithms like Logistic Regression and Gradient Boosting.

### *D. Model Development*

#### *Headroom Analysis – Logistic Regression and Gradient Boost Classifier*

To estimate headroom potential and understand physician prescribing behaviour, two algorithms were strategically selected: Logistic Regression and Gradient Boosting Classifier. Each model served a distinct purpose based on the data's characteristics and specific modelling needs.

Logistic Regression was initially chosen because several features during EDA demonstrated a linear or monotonic relationship with the target variable — the probability that a doctor would prescribe. The dataset was split into a 70%-30% train-test split to ensure ample data for both model building and evaluation. Logistic Regression was implemented along with feature selection (using SelectKBest with ANOVA F-statistics) and class balancing to address the skewed target distribution (~31% positive rate).

Mathematically, Logistic Regression models the probability pp that a given instance belongs to the positive class using the logistic (sigmoid) function:

$$p = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn)}}$$

The model estimates coefficients that maximize the likelihood of observed outcomes. Logistic Regression was particularly suitable because several features — such as age, years of experience, specialty, gender, and samples received ratio — showed strong, predictable trends with prescribing likelihood. These relationships reinforced the choice of Logistic Regression, leveraging its interpretability and robustness for linearly separable patterns.

However, early data exploration also revealed complex, non-linear interactions among features like drug price sensitivity, trust in pharma, and seminar attendance, which could not be effectively captured through Logistic Regression. This motivated the use of Gradient Boosting Classifier as a complementary modelling approach. Initial Gradient Boosting results showed overfitting (100% training accuracy), prompting a hyperparameter tuning exercise. The number of estimators was optimized to 20, with constraints on maximum depth and minimum leaf nodes to achieve a stable and generalizable model.

Mathematically, Gradient Boosting builds an ensemble of weak learners sequentially, correcting previous errors by fitting a new learner hm(x)hm(x) to the negative gradient of the loss function:

$$\mathrm{Fm(x)} = \mathrm{Fm} - 1\mathrm{(x)} + \gamma \mathrm{mhm(x)}$$

Feature importance analysis (Fig 6) from Gradient Boosting identified drug price sensitivity, trust in pharma, and seminar attendance as the most influential predictors. These features exhibited non-linear and interactive effects on prescribing behaviour — for example, variations in trust levels and price sensitivity impacted prescribing decisions in ways that simple linear models could not explain.

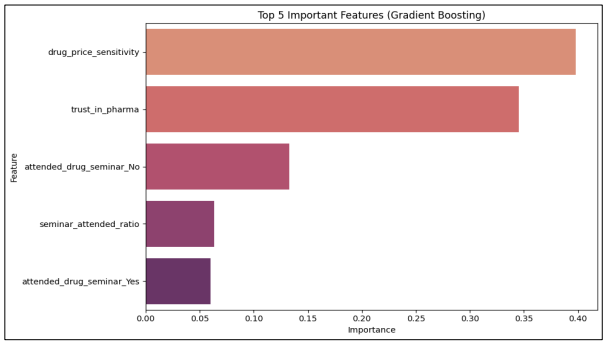


Fig. 6. Feature Importance (Gradient Boost)

In conclusion, while both Logistic Regression and Gradient Boosting produced valuable results for the headroom analysis, Gradient Boosting was ultimately preferred due to its superior ability to handle class imbalance and model complex, non-linear interactions between attributes.

***Segmentation Analysis – K-Means Clustering***

To segment physicians into distinct tiers based on prescriber demographics, prescribing behaviour and engagement characteristics, a clustering approach using K-Means was employed. The dataset included both numerical features (e.g., age, years of experience, number of patients per day, prescription volume, drug price sensitivity, trust in pharma) and categorical features (e.g., gender, specialty, location, hospital affiliation, seminar attendance, sample reception). These features were pre-processed appropriately — numerical attributes standardized and categorical attributes one-hot encoded — to ensure equal weighting during clustering.

To determine the optimal number of clusters, silhouette analysis (Fig 7) was conducted across a range of 2 to 5 clusters. The silhouette score peaked at k=4 clusters (score = 0.079), indicating the best balance between intra-cluster cohesion and inter-cluster separation. Based on this, K-Means clustering was finalized with 4 clusters.


Fig. 7. Silhouette Analysis

A Principal Component Analysis (PCA) was performed (Fig 8) for 2D visualization, clearly illustrating separation among the clusters with minimal overlap, thus reinforcing the choice of four clusters. The PCA scatter plot indicated that the first principal component (PC1) primarily distinguished doctors based on prescribing behaviour, while the second component (PC2) captured differences related to practice patterns and engagement levels.

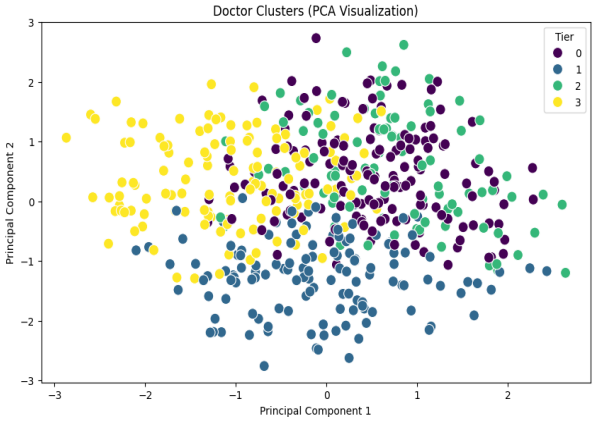


Fig. 8. Principal Component Analysis

## 4. RESULTS

### *A. Headroom Analysis*

Logistic Regression and Gradient Boost Classifier were evaluated (Table 2) on multiple performance metrics, including accuracy, precision, recall, F1-score, and ROC AUC. Gradient Boosting consistently outperformed Logistic Regression across all key metrics, achieving an accuracy of 93% and a perfect ROC AUC of 1.00, compared to 89% accuracy and 0.96 ROC AUC for Logistic Regression. Importantly, both models demonstrated strong recall (85%), ensuring that prescribing physicians were correctly identified. The top features highlighted by Gradient Boosting, such as drug price sensitivity, trust in pharma, and seminar attendance, were consistent with earlier EDA findings, providing further validation. Given its superior ability to handle non-linear relationships, class imbalance, and deliver higher predictive accuracy, Gradient Boosting was selected as the final model for headroom estimation and downstream analysis.

Table 2: Model(s) Evaluation

| Model | Accuracy | Precision | Recall | F1 | ROC AUC |
|---|---|---|---|---|---|
| Gradient Boosting | 0.93 | 1 | 0.85 | 0.91 | 1 |
| Logistic Regression | 0.89 | 0.88 | 0.85 | 0.87 | 0.96 |

### *B. Segmentation Analysis*

The analysis identified four distinct physician segments. The characteristics of each segment are summarized below:

- **Tier 0: Balanced General Practitioners:** This segment comprises middle-aged general physicians with moderate experience and prescription volumes. They exhibit high price sensitivity, primarily operate in rural, individual practices, and tend to receive a good amount of samples but show limited attendance at seminars.
- **Tier 1: Conservative Urban Specialists:** This group is characterized by the lowest prescription volumes and is largely made up of urban-based specialists (primarily dermatologists) affiliated with hospitals. Physicians in this tier demonstrate the lowest price sensitivity, exhibit good trust in the pharmaceutical industry, and tend to receive samples but attend fewer seminars.
- **Tier 2: Young, Trusting Potentials:** This segment consists of younger, urban, hospital-affiliated doctors (primarily paediatricians) with the second-highest prescription volumes and engagement levels. They show the highest price sensitivity and trust in the pharmaceutical industry, and they frequently receive samples and attend seminars.
- **Tier 3: Experienced, High-Volume Specialists:** This tier includes older, more experienced, urban, hospital-affiliated specialists (primarily cardiologists) with the highest prescription volumes and strong existing relationships with pharmaceutical companies. They exhibit the lowest price sensitivity and moderately receive samples and attend seminars.

The final profiling indicated clear patterns based on factors like experience, price sensitivity, trust in pharma, seminar attendance, and specialty, enabling actionable segmentation for targeting.

## 5. Discussion

The results of the headroom analysis provide a powerful tool for identifying high-potential prescribers for the new general therapeutic drug. The Gradient Boosting model's perfect ROC AUC score suggests that the model can accurately predict which doctors are most likely to prescribe the drug. This enables the pharmaceutical company to focus its marketing efforts on the most receptive physicians, maximizing the efficiency of their sales force and marketing budget. The segmentation analysis provides a nuanced understanding of physician behaviour, allowing for the development of targeted marketing strategies.

For example, based on Table 3, the final prediction results suggest that the medical representative should prioritize DOC011 — an experienced, high-volume prescriber — followed by DOC024, representing young, emerging practitioners with strong prescription potential. Next in the sequence would be DOC022, characterized as an urban prescriber with a more conservative adoption pattern, and finally DOC004, who reflects a balanced general physician with moderate prescribing behaviour.

Table 3: Sample Prediction Results

| doctor_id | prescription_probability | tier | tier_description |
|---|---|---|---|
| DOC0011 | 0.95 | 3 | Experienced and high Volume-Driven specialists: Experienced, Highest prescription volume, urban, hospital affiliated, highly influential |
| DOC0024 | 0.95 | 2 | Young potentials with trust: Young potentials, high prescription volume, high trust in pharma |
| DOC0022 | 0.95 | 1 | Conservative Urban Prescribers: Low prescription volume, urban, hospital affiliated |
| DOC0004 | 0.95 | 0 | Balanced General Physicians: Medium prescription volume, high price sensitivity, run individual practices |

## 6. Conclusion

This report demonstrates the value of machine learning in optimizing pharmaceutical sales and marketing. By accurately predicting physician prescription behaviour and segmenting physicians into meaningful tiers, the analysis provides actionable insights for a Pharmaceutical company. The findings enable the company to prioritize its marketing efforts, develop targeted engagement strategies, and ultimately increase the adoption rate of its new general therapeutic drug. The models provide a data-driven framework for retail drug sales business decisions.

## 7. Recommendations

To maximize the drug's market penetration, it is recommended that medical representatives of the pharmaceutical company focus their in-person visits and structured calls on the top 38% of targeted physicians from Table 4. Engagement strategies should be customized according to the unique characteristics of each physician segment. This targeted approach will optimize resource allocation, address specific prescribing drivers, and is projected to boost prescription rates, helping achieve a sales growth target of 60–70%.

Table 4: Recommendations

| Tier | Tier Description | Total Physicians | Target Physicians | Target Physician Rate |
|---|---|---|---|---|
| Tier3 | Experienced and high Volume-Driven specialists: Experienced, Highest prescription volume, urban, hospital affiliated, highly influential | 126 | 29 | **23%** |
| Tier2 | Young potentials with trust: Young potentials, high prescription volume, high trust in pharma | 94 | 66 | **70%** |
| Tier0 | Balanced General Physicians: Medium prescription volume, high price sensitivity, run individual practices | 151 | 48 | **32%** |
| Tier1 | Conservative Urban Prescribers: Low prescription volume, urban, hospital affiliated | 129 | 46 | **36%** |
| **Grand Total** | | **500** | **189** | **38%** |


## ACKNOWLEDGMENT

The author would like to thank "Dr. Eric Gyamfi" for supporting this research and project.


## REFERENCES


[1] Chen, L., Zhang, Y., & Wang, H. (2021). *Clustering algorithms in healthcare: A review*. Journal of Biomedical Informatics, 113, 103634. https://doi.org/10.1016/j.jbi.2020.103634

[2] Deloitte. (2020). *Data-driven marketing in pharmaceuticals: Harnessing insights for greater ROI*. Deloitte Insights. https://www2.deloitte.com

[3] Kostkova, P., Brewer, H., de Lusignan, S., et al. (2016). *Who owns the data? Open data for healthcare*. Frontiers in Public Health, 4, 7. https://doi.org/10.3389/fpubh.2016.00007

[4] McKinsey & Company. (2019). *Pharmaceutical marketing: The shift toward personalization and analytics*. https://www.mckinsey.com

[5] Singh, A., & Gaur, S. (2020). *Machine learning approaches for healthcare marketing: A predictive modelling perspective*. International Journal of Healthcare Management, 13(4), 327–334. https://doi.org/10.1080/20479700.2020.1717794