

**Data Storm v6.0**  
Preliminary Round



# Predictive Insights & Performance Solutions for Insurance Agent **Success**





# Table of Content

<b>1. Introduction</b>	4
1.1 Problem Statement	4
1.2 Objective	4
1.3 Relevance	5
<b>2. Data Overview</b>	5
<b>3. Exploratory Data Analysis</b>	7
3.1 Introduction	7
3.2 Univariate Analysis	7
3.3 Bivariate Analysis	8
3.4 Multivariate Analysis	9
3.5 Temporal Analysis	11
3.6 Agents Performance Analysis	13
3.7 Further Analysis	14
<b>4. Data Pre-processing</b>	17
4.1 Initial Inspection and Cleanup	17
4.2 Target Variable Construction	17
4.3 Feature Engineering	17
4.4 Handling Class Imbalance	18
4.5 Preprocessing Pipeline	18
<b>5. Predictive Modelling</b>	19
5.1 Overview	19
5.2 Methodology	19
5.3 Model Development	20
5.4 Feature Importance Analysis	20
5.5 Comparative Model Testing	21
5.6 SMART Action Plans for at-Risk Agents	22
<b>6. Agent Performance Categorization</b>	23
6.1 Objective	23
6.2 Data Aggregation and Feature Engineering	24
6.3 Clustering Approach	24
6.4 Results	25
6.5 Recommended Interventions by Category	25
<b>7. Interactive Dashboard</b>	26
7.1 Dashboard Overview	26
7.2 Key Features and Functionalities	27
<b>8. Conclusion</b>	29

## Table of Figures

Figure 1 Correlation Matrix .....	10
Figure 2 Monthly Sales Stability .....	11
Figure 3 Agent performance trajectories .....	12
Figure 4 Distribution of Month over month Changes .....	14
Figure 5 Net income per policy sold .....	15
Figure 6 ANBP Value Vs Net Income .....	15
Figure 7 Agent Age Vs New Policy Count .....	16
Figure 8 Temporal Trend income Trends by Age Group .....	16
Figure 9 Agent Performance Clustering .....	25
Figure 10 Dashboard Preview - Univariate .....	27
Figure 11 - Dashboard Preview - Bivariate .....	27
Figure 12 Dashboard Preview - Clustering .....	28
Figure 13 Dashboard Preview - Custom intervention plan .....	29

# **1. Introduction**

## **1.1 Problem Statement**

In the ABC Insurance Company, sales agents serve as the primary touchpoint for clients seeking insurance policies. Their performance, especially during the initial few months, plays a pivotal role in defining their long-term contribution and retention within the organization. A recurring challenge faced by the company is the phenomenon of “One Month NILL” agents, those who fail to secure even a single sale in the upcoming month. This has direct implications not only on revenue but also on agent morale, training costs, and customer engagement.

Furthermore, among agents who continue to perform, there exists a wide range of success levels from consistently high performers to those who remain stagnant or show inconsistent trends. The lack of a structured system to monitor, categorize, and support these agents based on their performance limits the company’s ability to nurture its workforce and optimize outcomes. Therefore, there is a clear need for a predictive and analytical framework to both anticipate underperformance and guide agent improvement in a targeted, data-driven manner.

## **1.2 Objective**

The key goals of this project are:

### **1. Early Risk Prediction**

Develop a robust predictive model using historical agent data to identify agents at risk of becoming “One Month NILL”—i.e., agents who are likely to make zero sales in the next month. This prediction will serve as an early warning system, allowing for preemptive intervention before performance drops irreversibly.

### **2. Exploratory Data Insights**

Conduct in-depth exploratory data analysis (EDA) to understand key behavioral, demographic, and performance-related features that differentiate successful agents from underperformers. This includes uncovering hidden patterns in the data, identifying time-based trends, and mapping agent trajectories over their career span.

### **3. Agent Performance Classification**

Implement a categorization system to classify existing agents into performance tiers such as High, Medium, and Low performers. This classification is designed to help the company tailor support mechanisms to the right group of agents and allocate resources more effectively.

### **4. Personalized Improvement Recommendations**

Design a recommendation system that proposes targeted and personalized interventions for at-risk or low-performing agents. These action plans may include SMART (Specific, Measurable, Achievable, Relevant, and Time-bound) strategies such as training sessions, mentorship programs, peer shadowing, or motivational incentives.

## **5. Monitoring & Feedback Loop**

Establish a feedback mechanism that continuously monitors agent progress post-intervention. This will help assess the effectiveness of recommendations and adjust strategies dynamically based on real-time performance changes.

## **1.3 Relevance**

The success of any insurance organization is deeply dependent on the performance of its sales agents, who act as the face of the company for potential and existing clients. Early underperformance by agents can lead to increased turnover, wasted onboarding costs, and missed revenue opportunities. More critically, failing to identify and support struggling agents in a timely manner may result in long-term disengagement, loss of client trust, and decreased operational efficiency.

In today's data-rich environment, relying solely on instinct or retrospective performance reviews is no longer sufficient. By leveraging predictive analytics and data-driven insights, ABC Insurance can proactively manage its workforce, identify risk profiles early, and apply targeted interventions that are more likely to succeed.

Furthermore, categorizing agents based on performance tiers enables better segmentation and resource planning, ensuring that high performers are rewarded and retained, while low performers receive the coaching or tools necessary to improve. Personalized action plans not only boost agent morale and competence but also enhance overall organizational productivity and client satisfaction.

Ultimately, this project is about moving from reactive to proactive performance management, an essential shift for any forward-thinking organization in the insurance sector.

## **2. Data Overview**

The dataset contains 15,308 entries with 23 columns, providing detailed information on insurance agents and their performance metrics over time. Each row in the dataset corresponds to an individual agent's performance record, with various numerical and categorical features capturing agent activities such as proposals, quotations, customer interactions, policy sales, and net income. The columns also track agent-specific details such as agent code, age, join month, first policy sold month, and several performance metrics across different time windows (e.g., last 7 days, 15 days, 21 days). The dataset includes key variables like the count of unique

proposals, quotations, customers, and the number of policies sold, which are critical for understanding and predicting an agent's future performance.

Column Name	Description
row_id	Unique identifier for each row in the dataset.
agent_code	Unique code assigned to each insurance agent.
agent_age	Age of the agent.
agent_join_month	The month when the agent joined the company.
first_policy_sold_month	The month when the agent sold their first policy.
year_month	A timestamp representing the year and month of the record.
unique_proposals_last_7_days	Count of unique proposals made by the agent in the last 7 days.
unique_proposals_last_15_days	Count of unique proposals made by the agent in the last 15 days.
unique_proposals_last_21_days	Count of unique proposals made by the agent in the last 21 days.
unique_proposal	Total count of unique proposals made by the agent.
unique_quotations_last_7_days	Count of unique quotations made by the agent in the last 7 days.
unique_quotations_last_15_days	Count of unique quotations made by the agent in the last 15 days.
unique_quotations_last_21_days	Count of unique quotations made by the agent in the last 21 days.
unique_quotations	Total count of unique quotations made by the agent.
unique_customers_last_7_days	Count of unique customers interacted with by the agent in the last 7 days.
unique_customers_last_15_days	Count of unique customers interacted with by the agent in the last 15 days.
unique_customers_last_21_days	Count of unique customers interacted with by the agent in the last 21 days.

unique_customers	Total count of unique customers interacted with by the agent.
new_policy_count	Number of new policies sold by the agent.
ANBP_value	The agent's ANBP (presumably a performance or commission metric).
net_income	Net income generated by the agent, likely linked to sales or commission.
number_of_policy_holders	Total number of policyholders under the agent's portfolio.
number_of_cash_payment_policies	Number of policies paid by customers through cash.

Table 1 - Data Dictionary

This dataset is fundamental for analyzing agent performance, predicting future behavior, and tailoring interventions aimed at improving performance and reducing the risk of agents going "NILL" (not selling anything) in the following month. The features capture both historical and recent performance indicators, which are essential for forecasting and suggesting actionable improvements.

## 3. Exploratory Data Analysis

### 3.1 Introduction

In this section, we will explore and analyze various aspects of the dataset, focusing on agent performance and related features. The primary objective of this analysis is to uncover key insights into agent behaviors, performance, and trends, as well as to identify potential relationships between different features. The features selected for the analysis are related to agent age, performance metrics (like proposals, quotations, and policies sold), and business-related values like ANBP (presumably a business metric) and net income.

### 3.2 Univariate Analysis

Univariate analysis involves examining the distribution and summary statistics of individual variables. The key features of interest for this analysis include agent\_age, unique\_proposal,

unique\_quotations, unique\_customers, ANBP\_value, net\_income, number\_of\_policy\_holders, and new\_policy\_count. Below are the key observations:

### **3.2.1 Distribution of Features**

#### **Agent\_age**

The age distribution of agents ranges from about 20 to 60 years, with the majority of agents concentrated around 40 years. This suggests that the majority of agents are in their prime working age, likely between mid-30s and mid-40s.

#### **Unique proposals and quotations**

The distributions for both unique proposals and unique quotations are roughly normally distributed. This indicates that most agents generate a moderate number of proposals and quotations, with a few exceptional agents generating higher values.

#### **Unique Customers**

Like proposals and quotations, the distribution of unique customers is approximately normal, suggesting most agents interact with a moderate number of unique customers.

#### **ANBP Value**

The ANBP value, representing an agent's business performance, is right-skewed. Most agents have lower ANBP values, but there is a small subset with much higher values, indicating that a few top-performing agents drive a significant portion of the business.

#### **Net income**

Similar to ANBP value, net income is right-skewed, with most agents earning a lower income and a small group of agents earning significantly higher amounts. This suggests a performance-based income distribution.

#### **Number of Policies**

The number of policy holders follows a normal distribution, with most agents managing a moderate number of policyholders.

#### **New Policy Count**

The distribution of new policies sold per agent shows that most agents sell a relatively low number of policies, but there are a few high performers who sell significantly more.

## **3.3 Bivariate Analysis**

Bivariate analysis explores the relationship between pairs of variables.



### 3.3.1 Key observations from bivariate analysis

#### **Agent Age vs. New Policy Count:**

There is no strong visible trend between agent age and the number of new policies sold. This suggests that performance does not correlate strongly with age, and agents across all age groups perform similarly in terms of the number of new policies sold.

#### **ANBP Value vs. Net Income**

There is a clear positive relationship between ANBP value and net income. As expected, higher business values (ANBP) tend to result in higher net income for agents. This indicates that the business value generated by agents is directly tied to their earnings.

#### **New Policy Count by Age Group**

Boxplots reveal that younger and middle-aged agents (20-40 years) tend to have a wider range of performance, with some agents in these groups performing very well, while others perform poorly. This shows that age does not necessarily predict performance in terms of new policy sales, and younger agents can outperform older agents.

## 3.4 Multivariate Analysis

Multivariate analysis examines relationships among multiple variables simultaneously. For this analysis, we used pairplots and correlation heatmaps to understand the relationships among the features.

### 3.4.1 Pair plot visualization

A pairplot was generated to visualize pairwise relationships between `agent_age`, `new_policy_count`, `ANBP_value`, and `net_income`. This helps to identify trends, clusters, and outliers across multiple dimensions.

Key observations from the pairplot:

- There is a positive correlation between `new_policy_count`, `ANBP_value`, and `net_income`. As agents sell more policies, their ANBP value and net income tend to increase.
- Some outliers are visible, especially in `new_policy_count`, where certain agents are outliers in terms of performance.
- The relationship between `agent_age` and the other variables is less clear, further supporting the findings from the bivariate analysis.

### 3.4.2 Correlation Heatmap

A pairplot was generated to visualize pairwise relationships between `agent_age`, `new_policy_count`, `ANBP_value`, and `net_income`. This helps to identify trends, clusters, and outliers across multiple dimensions.

A correlation heatmap was generated to examine the strength of linear relationships between numerical features. Below are the key observations from the heatmap.

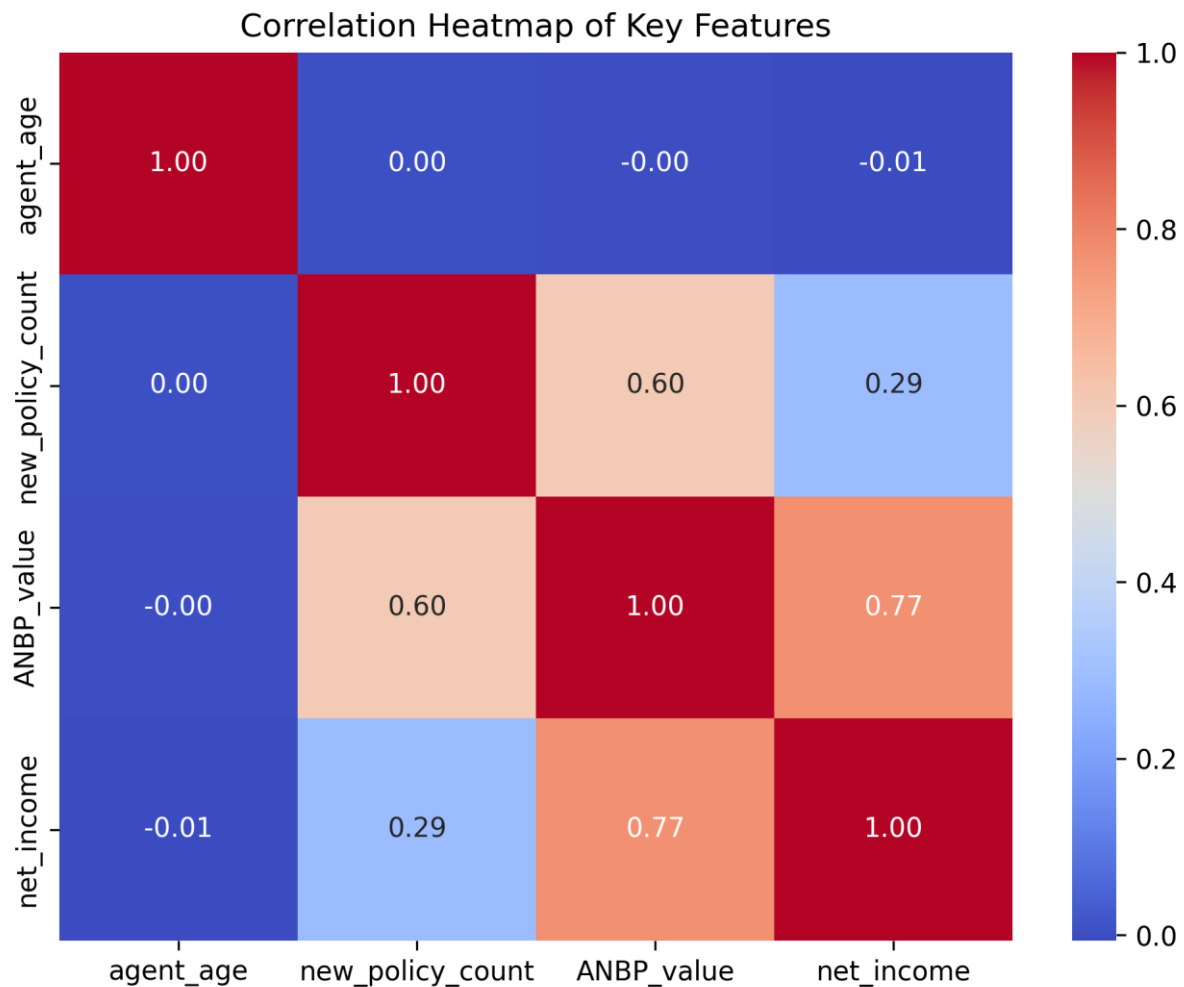


Figure 1 – Correlation Matrix

- **New Policy Count, ANBP Value, and Net Income:** These three features are strongly positively correlated. Agents with higher new policy counts tend to have higher ANBP values and higher net income, as expected from the business model.
- **Agent Age:** There is a weak correlation between agent age and the other features. Age does not appear to significantly influence new policy sales, ANBP value, or net income.

### 3.4.3 Metric Summary

Metric	Value
Average New Policies per Agent	20.27
Average ANBP Value	1,025,338
Average Net Income	228,041
Average Agent Age	40.59

- On average, agents sell about 20 new policies per month.
- The average ANBP value is just over 1 million.
- The average net income is about 228,000.
- The average agent age is around 41 years.

## 3.5 Temporal Analysis

Temporal analysis provides insights into trends and patterns over time across the agent network, focusing particularly on monthly new policy counts and individual agent sales trajectories.

### 3.5.1 Monthly Sales Stability

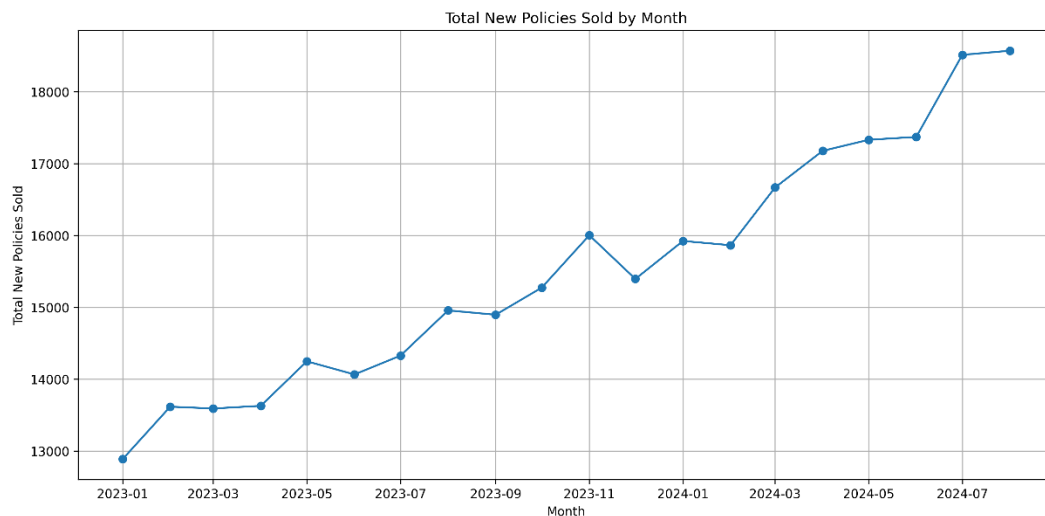


Figure 2 - Monthly Sales Stability

- **Key Observation:** Monthly averages and medians of new policy counts per agent remain stable, typically around 20–21 policies per agent.
- **Agent Base:** The total number of agents is gradually increasing over time, which contributes to an overall upward trend in total policy sales.
- **Anomaly Detection:** Using a z-score method, no statistically significant outliers were detected in monthly sales data (i.e., no months exceeded  $\pm 2$  z-score). This indicates that:
  - The sales process is stable and resilient.

- There are no abrupt shocks or seasonal disruptions affecting performance at scale.

### 3.5.2 Agent level performance trajectories

To understand how individual agent evolve over time in terms of productivity, we examined monthly new policy counts per agent, uncovering key behavioral patterns.

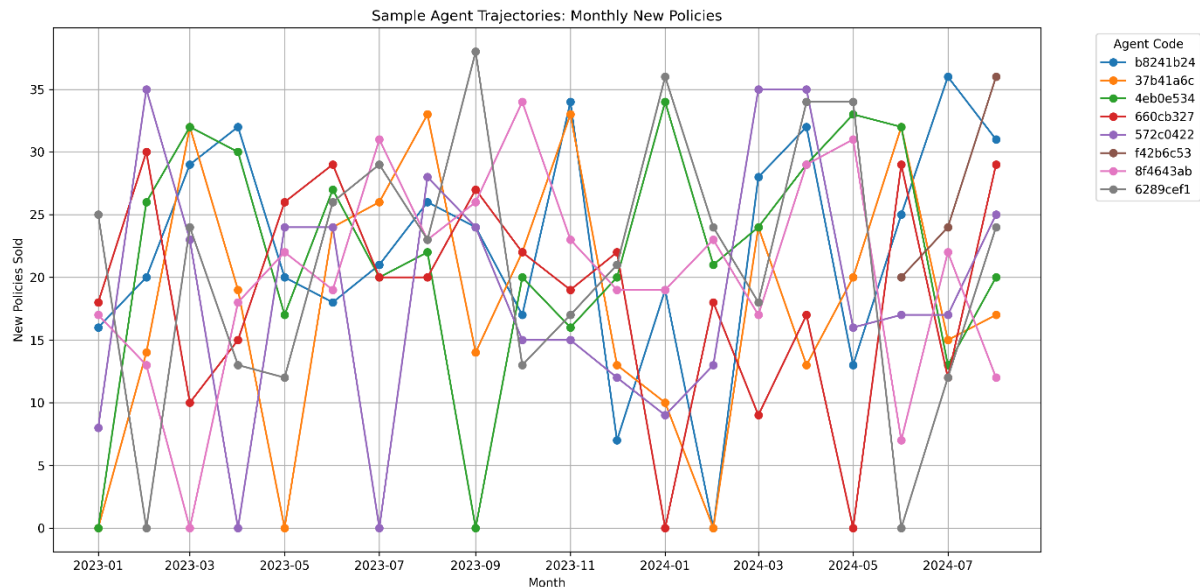


Figure 3 Agent performance trajectories

### Variability in Performance

- **Stable Performers:** Some agents maintain consistent sales levels month-to-month, showcasing reliability and process mastery.
- **Trending Agents:**
  - Some agents exhibit upward trends, indicating improving performance.
  - Others show declining trends, which may signal burnout, disengagement, or external challenges.
- **Volatile Agents:** A subset displays high variability, marked by sharp spikes or drops in monthly sales. These may result from:
  - Special sales campaigns
  - Personal or external circumstances
  - Varying client bases or seasonal influences

### Quantifying Trajectories

- **3-Month Rolling Averages:**
  - Smooth out fluctuations
  - Reveal underlying trends—sustained growth or deterioration
- **Month-over-Month Change:**
  - Most agents show minor regular changes.

- Large swings in a few cases highlight potential instability or exceptional events.

**Sales Stability:** The overall sales process is robust, with minimal volatility at the organizational level.

**Agent Diversity:** Individual agent journeys are heterogeneous—highlighting a need for personalized support or incentive structures.

**Actionable Insight:**

- Track declining performers early and intervene with training or support.
- Recognize consistent high performers for reward and retention.
- Investigate volatility to uncover and replicate high-impact practices or to mitigate underlying issues.

### **3.6 Agents Performance Analysis**

To better understand individual agent behavior over time, we conducted a focused trend analysis using rolling averages, month-over-month changes, and decline identification. This analysis provides valuable insights into consistency, improvement potential, and early warning signals for underperformance.

#### **Rolling 3-Month Averages**

- Purpose: To smooth short-term fluctuations and capture medium-term trends in policy sales.
- Method: Calculated a 3-month rolling average (`rolling_3m_avg`) of `new_policy_count` for each agent.
- Insight:
  - Helps distinguish between temporary dips and consistent declines.
  - Highlights agents showing gradual improvement or deterioration.

#### **Month-over-Month (MoM) Change**

- Calculation: Computed the percentage change in new policy count for each agent from one month to the next.
- Observation:
  - Distribution of MoM changes is mostly centered near 0 but with a long negative tail.
  - Histogram shows many agents have small fluctuations, but a noticeable number experience steep monthly declines.



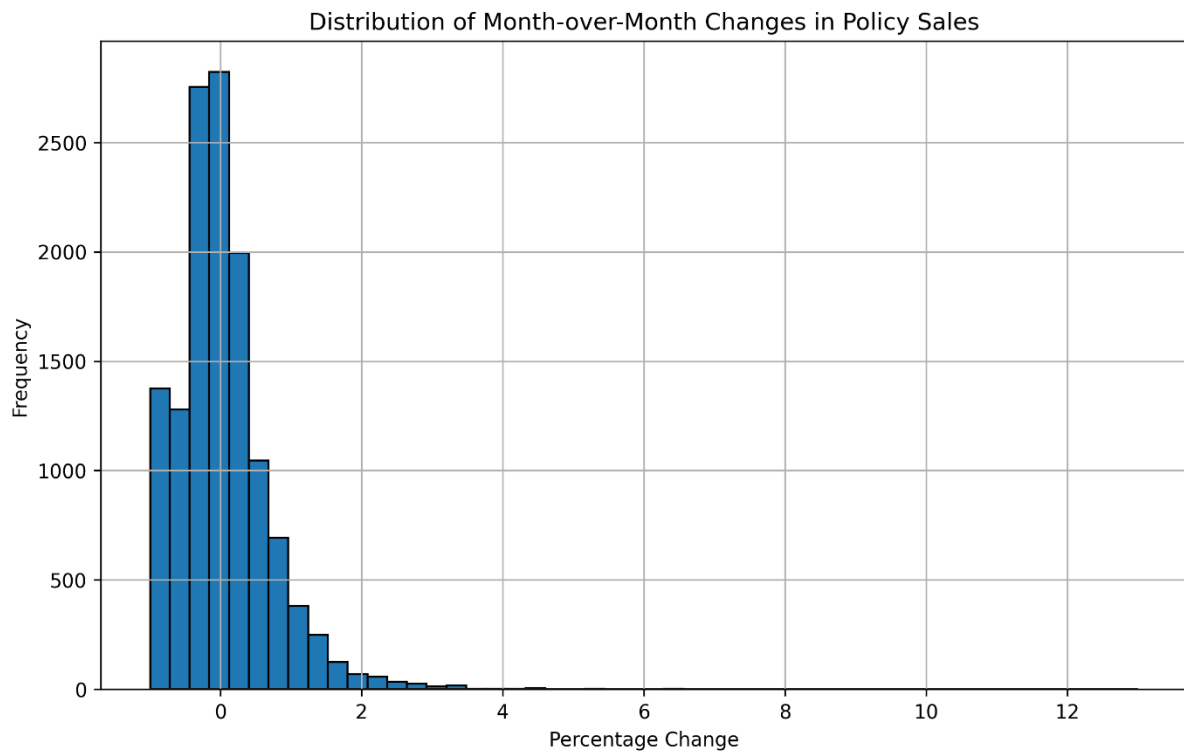


Figure 4 Distribution of Month over month Changes

### Identifying Declining Agents

- Definition: Agents are flagged as "declining" if, in the latest month:
  - Their new\_policy\_count is below the median for that month.
  - Their mom\_change is negative (i.e., they sold fewer policies than the month before).
- Results:
  - 303 agents were identified as showing signs of decline.
  - Their performance summary:

## 3.7 Further Analysis

### Distribution of Net Income per Policy Sold

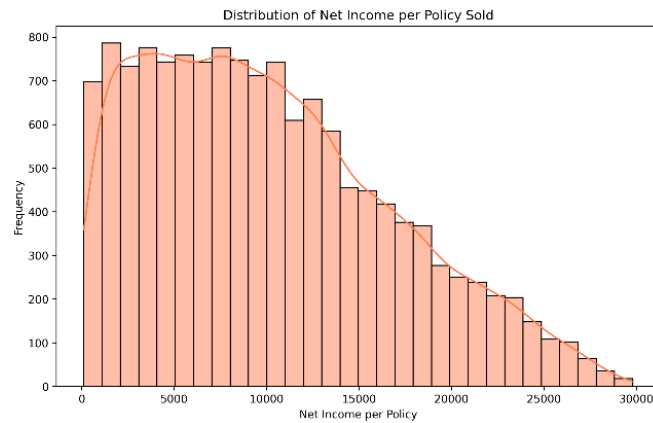


Figure 5 Net income per policy sold

A histogram was plotted to examine the distribution of net income per policy sold. The distribution is right-skewed, suggesting significant variability in how much net income is earned for each policy. This indicates that while most agents earn a moderate income for each policy sold, some agents earn much more, likely due to higher-value policies or bonuses.

#### ANBP Value vs. New Policy Count

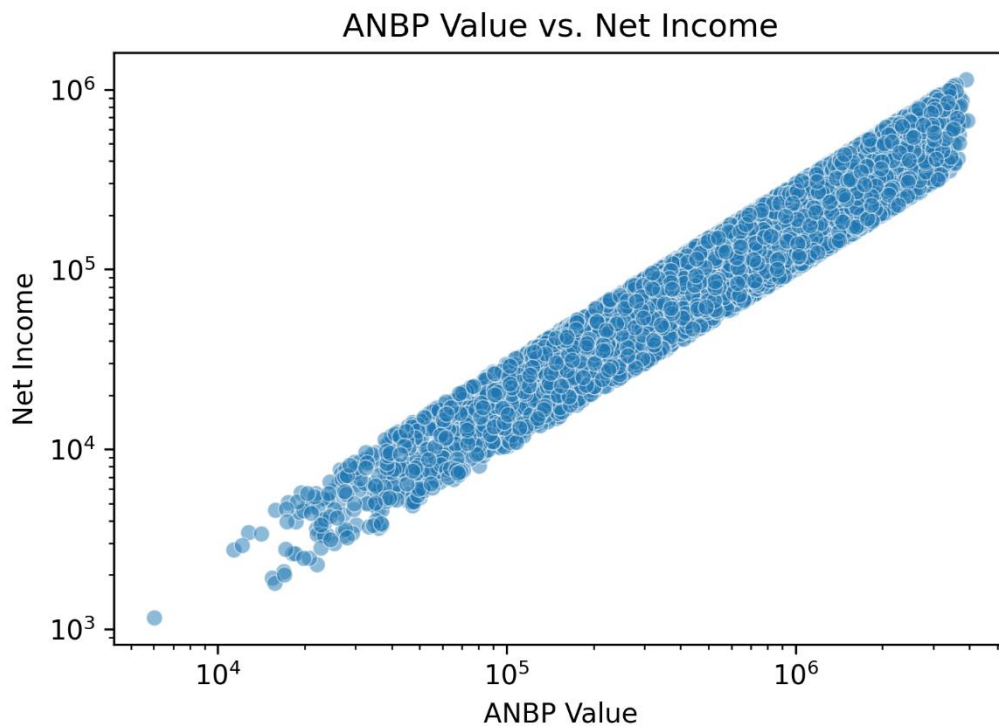


Figure 6- ANBP Value Vs Net Income

A scatter plot was generated to explore whether handling higher ANBP values correlates with selling more policies. While there is a visible positive trend, it is not perfectly linear, indicating that agents who handle higher-value business do not always sell more policies.

## Agent Age and New Policy Count Impact on Net Income

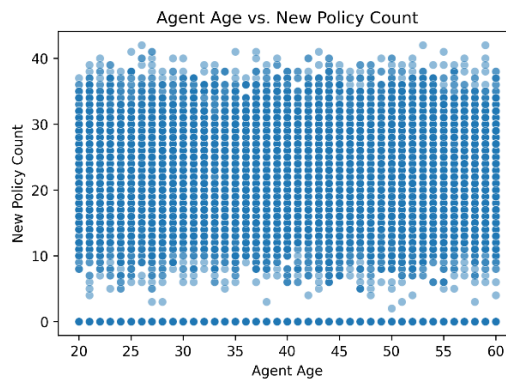


Figure 7 - Agent Age Vs New Policy Count

A bubble plot was created to examine how agent age and new policy count together influence net income. Larger bubbles represent agents with higher policy counts. The plot shows that older agents with higher new policy counts tend to have higher net incomes, but this is not always the case.

## Temporal Net Income Trends Segmented by Age Group

A line plot was used to display temporal trends in net income for different age groups. It highlights that younger agents (20-30 years) show a more volatile income pattern, while older agents (40-50 years) have more consistent monthly performance. This suggests that older agents may have a more stable customer base and better client relationships.

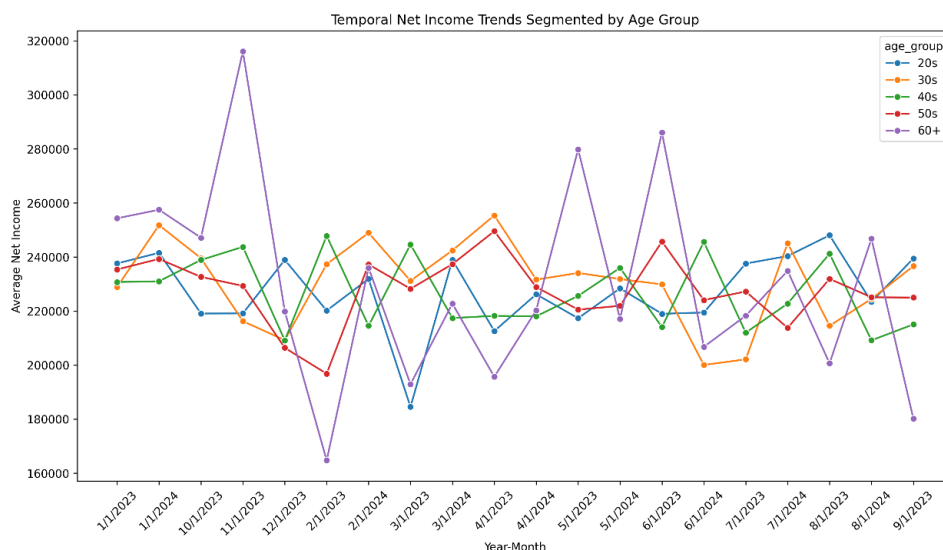


Figure 8 Temporal Trend income Trends by Age Group

The exploratory data analysis (EDA) provided several key insights into the performance of insurance agents. The following points summarize the key findings:

- **Performance Distribution:** Many features like `ANBP_value`, `net_income`, and `new_policy_count` are right-skewed, with a small subset of high performers driving significant business.
- **Agent Age:** Agent age does not strongly correlate with performance metrics like `new_policy_count`, `ANBP_value`, or `net_income`. This suggests that agents across different age groups can perform similarly.
- **Bivariate and Multivariate Relationships:** There are strong positive correlations between performance metrics, especially between `new_policy_count`, `ANBP_value`, and `net_income`. Additionally, age has little effect on these variables.

This report sets the foundation for further in-depth analysis and modeling, including performance prediction or agent classification based on various factors.

## 4. Data Pre-processing

### 4.1 Initial Inspection and Cleanup

The dataset was initially examined for missing values and duplicates. Fortunately, no such issues were identified, allowing the data to be used as-is without requiring additional cleaning steps. All relevant date columns namely `agent_join_month`, `first_policy_sold_month`, and `year_month` were converted to datetime format. This enabled the extraction of time-based features and chronological sorting, which was crucial for the subsequent analysis and modeling.

### 4.2 Target Variable Construction

The goal of the modeling task was to predict whether an agent would stop selling policies in the following month. To facilitate this, a binary target variable `target_column` was created. This column was labeled 1 if the `new_policy_count` in the next month was zero, and 0 otherwise. Because the prediction depends on knowing the next month's activity, the final month of data for each agent was removed, as it lacked the necessary future context to determine the target value.

### 4.3 Feature Engineering

A comprehensive set of features was generated to enrich the model's ability to detect patterns in agent behavior:

- **Time-based features** included the number of months since the agent joined and since they made their first sale. The calendar month and quarter were also extracted from `year_month` to capture seasonal trends.

- **Activity efficiency ratios** were calculated, such as the proposal-to-quotation ratio, customer-to-proposal ratio, and policy-to-customer ratio, which measure how effectively agents move customers through the sales funnel.
- **Short-term trend indicators** like the 7-to-15-day ratio of proposals, quotations, and customer engagements were engineered to detect early signs of declining or increasing activity.
- **Performance metrics** such as `income_per_policy` and `anbp_per_policy` (Annualized New Business Premium per policy) were computed to assess individual agent productivity.
- **Agent tenure categories** were created by binning `months_since_joining` into intervals such as '0–3 months', '3–6 months', etc., helping to capture experience-level differences in agent performance.

## 4.4 Handling Class Imbalance

A significant class imbalance was identified in the target variable, with only about 10% of observations indicating agent inactivity. To address this, a hybrid resampling strategy was employed. First, SMOTE (Synthetic Minority Oversampling Technique) was used to oversample the minority class to reach 50% of the majority class size. Following that, `RandomUnderSampler` was applied to bring down the majority class to 80% of the new, oversampled data. This two-step approach ensured a balanced representation of both active and inactive agents during model training, improving the model's ability to generalize.

## 4.5 Preprocessing Pipeline

To ensure consistency in data transformation, a preprocessing pipeline was built using `scikit-learn`'s `ColumnTransformer`. Numerical features were processed through a pipeline involving median imputation and standard scaling. Categorical features were handled using mode imputation followed by one-hot encoding, with `handle_unknown='ignore'` to safely manage unseen categories in new data. This modular pipeline made the preprocessing steps reproducible and robust, and it seamlessly integrated with downstream modeling processes.



## 5. Predictive Modelling

### 5.1 Overview

This section presents the development, evaluation, and optimization of a machine learning model aimed at predicting insurance agents who are likely to make zero policy sales in the upcoming month. Trained on historical sales and agent performance data, the model incorporates temporal trends, agent activity indicators, and business performance metrics.

The final solution, based on a LightGBM classifier, effectively handled significant class imbalance (9:1 ratio) using resampling techniques and algorithm-level adjustments. LightGBM outperformed other tested models, including XGBoost, Random Forest, and Logistic Regression, and is now recommended for deployment with provisions for continuous performance monitoring and retraining.

### 5.2 Methodology

#### 5.2.1 Data Preparation

Date fields such as `year_month`, `agent_join_month`, and `first_policy_sold_month` were converted into datetime formats. The target variable was defined as binary, where a value of 1 indicates that an agent had zero sales in the subsequent month. To prevent data leakage, the final observed month for each agent was excluded from the training data.

#### 5.2.2 Feature Engineering

Several categories of features were created:

Temporal Features: Including `months_since_joining` and `months_since_first_sale`.

Activity Ratios: Derived metrics such as `prop_to_quote_ratio` (smoothed using Laplace smoothing) and `sale_conversion_rate`.

Rolling Trends: Calculated 7-day, 14-day, and 21-day moving averages of the `new_policy_count`.

Business Metrics: Included `income_per_policy` and `ANBP_per_policy`.

#### 5.2.3 Handling Class Imbalance

To address the 9:1 class imbalance, a combination of techniques was used:

Resampling: SMOTE was applied to generate 60% additional synthetic minority samples.

Random undersampling reduced the majority class by 70%.

Algorithmic Adjustments:

LightGBM's `scale_pos_weight` was set to 9.0, proportional to the class ratio.

## 5.3 Model Development

### 5.3.1 Algorithm Selection

The LightGBM algorithm was selected as the primary model for its superior speed and accuracy, particularly in dealing with large-scale datasets and class imbalance. Its native support for weighted training and its ability to handle high-dimensional feature spaces made it a fitting choice for this task.

To ensure robustness and justify the model selection, a benchmarking exercise was conducted comparing LightGBM with XGBoost, Random Forest, and Logistic Regression. These models were evaluated using identical data splits and preprocessing pipelines. While each model demonstrated competence to varying degrees, LightGBM consistently outperformed others in both accuracy and computational efficiency.

### 5.3.2 Validation Strategy

The model's performance was evaluated using stratified 5-fold cross-validation to preserve class distribution across folds. This ensured a reliable assessment of generalization performance. The classification threshold was optimized using the F2-score, favoring recall to align with business objectives that prioritize identifying agents at risk of zero policy sales.

### 5.3.4 Hyperparameter Tuning

Extensive tuning was performed to enhance model performance. The finalized hyperparameters for the LightGBM model included 500 estimators, a learning rate of 0.05, a maximum depth of 5, and 31 leaves. The `scale_pos_weight` parameter was set to 9.0 to reflect the 9:1 imbalance in the dataset. These parameters were found to strike an effective balance between bias and variance while optimizing for business-relevant metrics.

## 5.4 Feature Importance Analysis

The model's feature importance rankings revealed that long-term temporal trends and agent tenure were the most influential predictors. Specifically, the 21-day rolling policy trend (`policy_trend_21`) emerged as the top feature, followed closely by `months_since_first_sale` and the 14-day trend metric. Features related to agent efficiency, such as `sale_conversion_rate` and `prop_to_quote_ratio`, also ranked highly, reinforcing the value of behavioral indicators in predicting future inactivity.

Longer-term performance indicators provided more stable signals of potential zero-policy months than short-term changes. Furthermore, the model highlighted the predictive strength of agent tenure and customer engagement levels. These insights can inform targeted interventions and strategic agent support initiatives.

### Top 10 important Features

Rank	Feature
1	months_since_joining
2	months_since_first_sale
3	net_income
4	unique_quotations_last_21_days
5	unique_quotations
6	unique_customers_last_21_days
7	prop_to_quote_ratio
8	policy_trend_7
9	number_of_policy_holders
10	number_of_cash_payment_policies

## 5.5 Comparative Model Testing

XGBoost offered competitive predictive power but was computationally more intensive. Random Forest exhibited poorer handling of imbalanced data. Logistic Regression, constrained by its linear assumptions, underperformed, making it unsuitable for this complex predictive task.

LightGBM was ultimately favored due to its superior capability in handling class imbalance, its rapid training time compared to XGBoost, and its built-in regularization mechanisms that mitigated overfitting. These characteristics made it the most suitable model for real-world deployment where computational efficiency and accuracy are critical.

The final LightGBM model demonstrates a high level of predictive performance and operational efficiency, making it ready for deployment. It is recommended that the model be integrated into the sales platform with a continuous performance monitoring system to detect any degradation over time. Periodic retraining using the latest sales data is also advised to ensure adaptability to shifting patterns in agent behavior. Additionally, the classification threshold may be revisited periodically based on evolving business priorities, particularly the trade-off between minimizing false positives and maximizing recall.

## 5.6 SMART Action Plans for At-Risk Agents

To sustain high performance and reduce agent attrition, it is crucial to identify and proactively support agents who show early signs of underperformance. By using behavioral indicators such as quoting frequency, income trends, conversion efficiency, and recent engagement activity, we categorized at-risk agents into meaningful segments. For each segment, we developed targeted SMART (Specific, Measurable, Achievable, Relevant, Time-bound) action plans combined with personalized interventions. These plans not only aim to uplift agent productivity but also create a culture of continuous learning, accountability, and motivation. The strategies outlined below provide structured support tailored to the unique challenges faced by each type of at-risk agent.

### 1. Low months\_since\_joining + Low unique\_quotations (New & Inactive Agents)

Issue: New agents struggling to generate quotes or engage customers.

SMART Plan:

- Specific: Assign a mentor + weekly sales training.
- Measurable: Target 5+ new quotes/week, tracked via CRM.
- Achievable: Shadow top performers for 2 hrs/week.
- Relevant: Focus on cold-calling/lead-generation techniques.
- Time-bound: Improve within 8 weeks or reassess role fit.

Interventions:

Role-playing sessions on objection handling.

Motivation: Small bonuses for first 10 quotes.

### 2. High months\_since\_first\_sale + Low net\_income (Experienced but Underperforming)

Issue: Veteran agents with declining productivity.

SMART Plan:

- Specific: Upskill on high-value products (e.g., premium policies).
- Measurable: Increase net income by 15% in 3 months.
- Achievable: Attend advanced negotiation workshops.
- Relevant: Analyze their customer segments (e.g., upsell to renewals).
- Time-bound: Biweekly reviews with manager.

Interventions:

Peer mentoring with top earners.

Gamification: Leaderboard for most improved net income.

### **3. Low prop\_to\_quote\_ratio (Poor Conversion from Quotes to Sales)**

Issue: Generating quotes but not closing deals.

SMART Plan:

- Specific: Improve closing techniques (e.g., follow-up scripts).
- Measurable: Boost conversion by 10% in 6 weeks.
- Achievable: Record & review 2 sales calls/week with manager.
- Relevant: Focus on high-intent leads (e.g., repeat quote requesters).
- Time-bound: Weekly feedback sessions.

Interventions:

Bonus for conversions above 20%.

Toolkit: CRM alerts for follow-ups on stale quotes.

### **4. Declining policy\_trend\_7 or unique\_customers\_last\_21\_days (Slumping Engagement)**

Issue: Recent drop in customer acquisition.

SMART Plan:

- Specific: Reactivate past customers via targeted campaigns.
- Measurable: Add 3+ new customers/week.
- Achievable: Use CRM to identify lapsed clients for outreach.
- Relevant: Offer limited-time discounts for referrals.
- Time-bound: Track progress biweekly.

Interventions:

Training: Social selling/webinar strategies.

Motivation: Reward for most reactivated customers.

## **6. Agent Performance Categorization**

### **6.1 Objective**

The primary objective of this analysis is to systematically categorize insurance agents based on their historical performance metrics. This enables the identification of top performers, agents with growth potential, and those requiring foundational support. By segmenting agents



into High, Medium, and Low performers, targeted development interventions can be planned to enhance productivity, optimize resource allocation, and improve overall organizational efficiency.

## 6.2 Data Aggregation and Feature Engineering

To evaluate agent performance, we computed aggregated metrics on a monthly basis for each agent using the available transactional data. The following key indicators were derived:

- **Average policies per month (avg\_policies):** Reflects an agent's overall productivity.
- **Standard deviation of policies (std\_policies):** Captures fluctuations in performance, which helps assess consistency.
- **Average ANBP value (avg\_anbp):** Indicates the average business premium value brought in by the agent.
- **Average net income (avg\_income):** Reflects the agent's profitability to the organization.
- **Average unique customers (avg\_customers):** Demonstrates the agent's customer reach and client base.
- **Tenure:** Measured in months, it is calculated as the time span between the agent's first and last active months.
- **Stability:** Defined as the inverse of policy standard deviation. Lower variability implies higher stability in monthly performance.

These features were selected to capture both quantity and quality dimensions of an agent's contribution, ensuring a holistic assessment.

## 6.3 Clustering Approach

A K-Means clustering algorithm was used with three clusters to represent distinct performance categories. The clustering was performed on five standardized features: average policies, ANBP value, net income, unique customers, and stability.

Post clustering, clusters were interpreted and labeled as High, Medium, or Low performers based on their average number of policies per month. The labeling criteria were anchored in the overall mean values within each cluster, ensuring that segmentation aligns with organizational definitions of productivity.

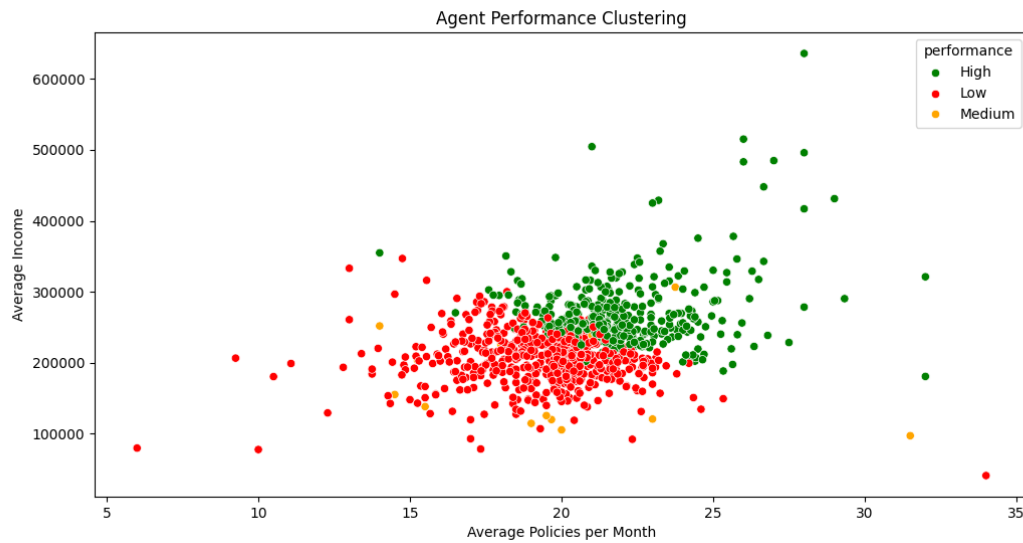


Figure 9 Agent Performance Clustering

## 6.4 Results

The analysis revealed a clear segmentation of agents into three distinct clusters. The resulting distribution of agents is summarized as follows:

Performance Category	Number of Agents
High Performers	350
Medium Performers	14
Low Performers	541

## 6.5 Recommended interventions by category

Based on the profile of each performance segment, the following targeted development strategies are recommended:

### High Performers

Agents in this category demonstrate strong productivity and consistency. They serve as role models and can contribute more strategically to the organization. Recommended interventions include:

- Assigning mentorship roles to support the growth of newer or underperforming agents.
- Enrolling in advanced sales and leadership training to further elevate their performance.

- Allocating high-value or complex clients to leverage their experience.
- Defining stretch goals to maintain engagement and challenge them further.

### **Medium Performers**

These agents have shown moderate productivity and possess potential for growth. Their performance may benefit from consistent skill development and motivational programs. Suggested interventions include:

- Participation in workshops to enhance soft skills, product knowledge, and customer interaction techniques.
- Regular coaching sessions focused on performance feedback and action planning.
- Training programs in proposal writing, negotiation, and relationship management.
- Providing moderate performance-based incentives to encourage upward movement to the high-performing category.

### **Low Performers**

This group requires structured support and closer monitoring. Many agents in this segment may be early in their careers or experiencing challenges in adapting to the sales process. Recommended strategies include:

- Re-engagement through foundational onboarding programs that reinforce product and process knowledge.
- Daily or weekly performance reviews with structured targets and feedback loops.
- Pairing with experienced agents for job shadowing to observe effective sales strategies in practice.
- Setting incremental, short-term goals to rebuild confidence and drive small successes.

The clustering-based agent performance categorization presents a data-driven framework for agent segmentation. By translating complex performance metrics into intuitive performance bands, organizations can drive customized developmental programs and maximize the return on investment in human capital. This framework is not only scalable but also adaptable to include future variables such as customer satisfaction scores or digital engagement rates, making it a sustainable strategy for long-term talent management.

## **7. Interactive Dashboard**

### **7.1 Dashboard Overview**

The Agent Performance Analytics Dashboard is a comprehensive analytical tool developed with streamlit to evaluate, predict, and optimize insurance agent performance. Using machine learning and data visualization techniques, the dashboard provides actionable insights into agent productivity, identifies high-risk performers, and recommends targeted interventions. It

integrates predictive modelling, clustering analysis, and interactive reporting to help managers make data-driven decisions for workforce optimization.

Accessible at: <https://insurance-agents-performance-analyzer.streamlit.app/>

## 7.2 Key Features and Functionalities

### 1. Exploratory Data Analysis

The dashboard opens with a comprehensive exploratory analysis designed to help users understand agent performance patterns through interactive visualizations. It begins with univariate analysis to examine the distribution of individual metrics like agent age, policy sales, and income. Bivariate analysis explores relationships between variables such as policy count and income or tenure and conversion rates, revealing key performance correlations. Multivariate analysis enables deeper insights into how multiple factors interact to influence success. Additionally, temporal analysis tracks trends over time, highlighting seasonal patterns or productivity.

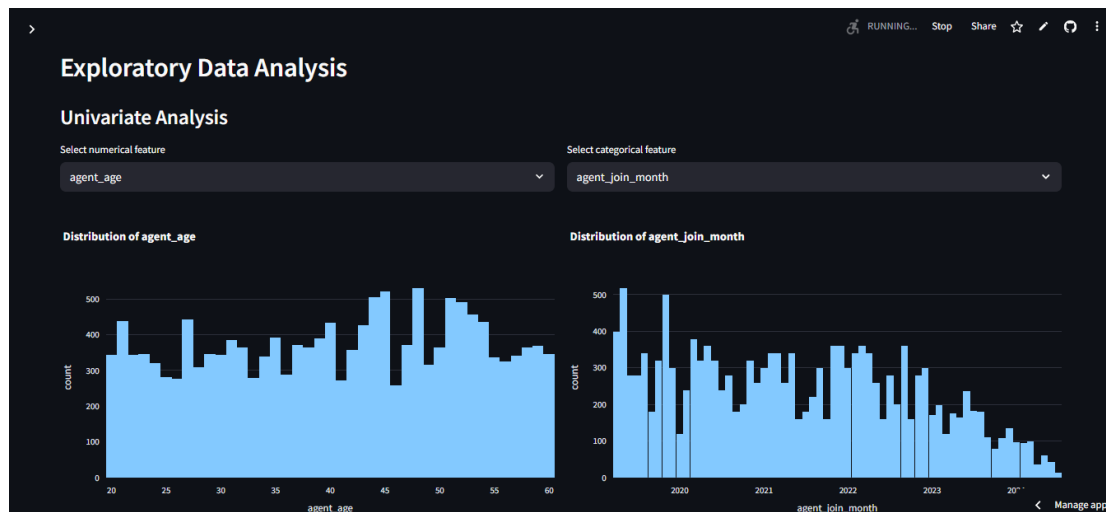


Figure 10 - Dashboard Preview | Univariate

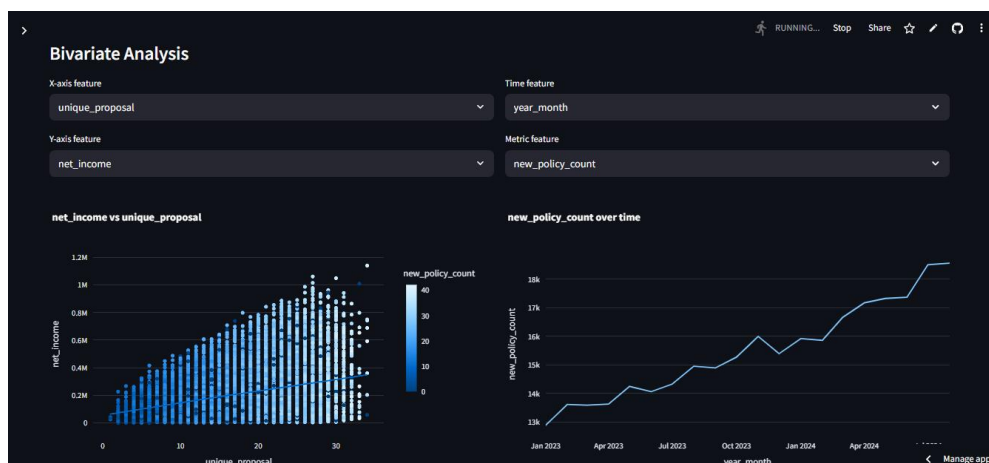


Figure 11 - Dashboard Preview | Bivariate

## 2. Agent Performance Clustering

Agents are categorized into High, Medium, and Low performers using K-means clustering. This segmentation is based on:

Average Policies Sold – Measures sales consistency.

ANBP Value & Net Income – Evaluates revenue contribution.

Customer Engagement – Assesses agent reach and retention.

Stability Index – Quantifies performance consistency over time.

The dashboard provides interactive cluster visualizations, including t-SNE plots for high-dimensional data interpretation, enabling managers to quickly identify top and underperforming agents.



Figure 12 Dashboard Preview | Clustering

## 3. One NILL Agent Prediction

The previously built predictive model has been integrated to forecast agents at risk of becoming "One NILL" those likely to sell zero policies in the upcoming month. Utilizing a LightGBM classifier enhanced with SMOTE and undersampling techniques to address class imbalance, the model delivers reliable results measured through AUC-ROC scores and fine-tuned probability thresholds. It identifies key predictors such as recent sales trends, conversion efficiency, and customer engagement levels, enabling managers to take timely, targeted action. By flagging high-risk agents in advance, the model supports proactive interventions, ultimately reducing agent attrition and strengthening overall retention strategies.

## 4. Intervention Planning & Tracking

The dashboard includes a structured framework for designing, implementing, and monitoring performance interventions.



## Standardized Recommendations

- High Performers – Advanced training, leadership development, and mentorship roles.
- Medium Performers – Targeted coaching, goal-setting, and skill-building programs.
- Low Performers – Intensive performance improvement plans with daily monitoring.

## Custom Intervention Management

- Interactive Form – Allows managers to create tailored intervention plans with timelines and success metrics.
- Status Tracking – Monitors progress (Planned, In Progress, Completed, On Hold).
- Export & Reporting – Facilitates data sharing via CSV exports for further analysis.

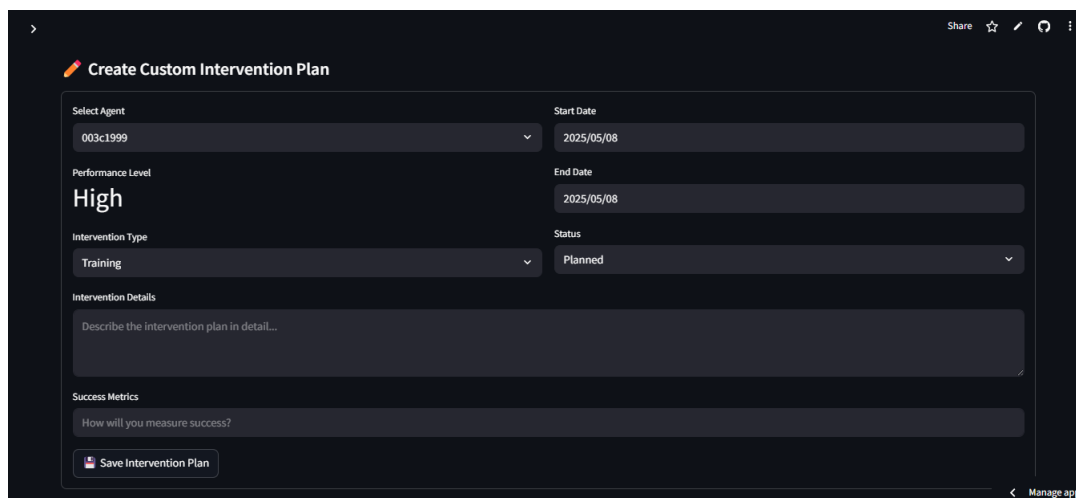
The image shows a Streamlit web application interface for creating a custom intervention plan. The title bar at the top says 'Create Custom Intervention Plan' with a pencil icon. The form is divided into several sections: 'Select Agent' with a dropdown menu showing '003c1999'; 'Performance Level' with a text input showing 'High'; 'Intervention Type' with a dropdown menu showing 'Training'; 'Start Date' and 'End Date' both with date pickers showing '2025/05/08'; 'Status' with a dropdown menu showing 'Planned'; 'Intervention Details' with a large text area containing the placeholder 'Describe the intervention plan in detail...'; and 'Success Metrics' with a text area containing the placeholder 'How will you measure success?'. At the bottom left is a 'Save Intervention Plan' button with a save icon. At the bottom right is a 'Manage app' link with a left arrow icon. The top right corner has 'Share', a star icon, a refresh icon, and a menu icon.

Figure 13 Dashboard Preview - Custom intervention plan

## 8. Conclusion

This report presents a comprehensive data-driven approach to understanding and enhancing the performance of insurance agents. Through meticulous exploratory data analysis, we identified critical patterns and trends in agent behaviour and sales outcomes. Using a combination of predictive modelling techniques and performance clustering, the project successfully highlighted agents at risk of going "NIL" (making no sales in the next month) and categorized agents into meaningful performance groups. These insights were further translated into personalized action plans aimed at improving agent productivity and engagement.

The deployment of an interactive Streamlit dashboard has made these insights easily accessible and actionable for stakeholders, offering visual summaries, predictive results, and targeted recommendations. Altogether, the work not only demonstrates the potential of machine learning in performance prediction but also provides a strategic foundation for data-informed decision-making in sales force management.

