Data Storm 6.0

Predict NILL Agents Report

Team Cognic AI

Data Preprocessing

Data Loading and Initial Inspection

	type	count	nunique	null	mode	least_frequent	mean	min	max
agent_code	object	15308	905	0	NaN	NaN	NaN	NaN	NaN
agent_age	int64	15308	41	0	48.0	46.0	4.058577e+01	20.0	60.0
agent_join_month	datetime64[ns]	15308	64	0	2019-05-01 00:00:00	2024-07-01 00:00:00	NaN	2019-04-01 00:00:00	2024-07-01 00:00:00
first_policy_sold_month	datetime64[ns]	15308	28	0	2024-04-01 00:00:00	2022-02-01 00:00:00	NaN	2022-01-01 00:00:00	2024-04-01 00:00:00
year_month	datetime64[ns]	15308	20	0	2024-08-01 00:00:00	2023-01-01 00:00:00	NaN	2023-01-01 00:00:00	2024-08-01 00:00:00
unique_proposals_last_7_days	int64	15308	4	0	0.0	3	1.491246e+00	0.0	3.0
unique_proposals_last_15_days	int64	15308	7	0	NaN	5	2.991116e+00	0.0	6.0
unique_proposals_last_21_days	int64	15308	21	0	11.0	6	1.004573e+01	0.0	20.0
unique_proposal	int64	15308	34	0	14.0	1	1.751653e+01	1.0	34.0
unique_quotations_last_7_days	int64	15308	5	0	0.0	2	1.989679e+00	0.0	4.0
unique_quotations_last_15_days	int64	15308	7	0	1.0	3	2.996538e+00	0.0	6.0
unique_quotations_last_21_days	int64	15308	10	0	1.0	8	4.469624e+00	0.0	9.0
unique_quotations	int64	15308	32	0	14.0	32	1.392239e+01	1.0	32.0
unique_customers_last_7_days	int64	15308	7	0	3.0	NaN	2.995493e+00	0.0	6.0
unique_customers_last_15_days	int64	15308	11	0	6.0	0	4.997714e+00	0.0	10.0
unique_customers_last_21_days	int64	15308	16	0	9.0	0	7.500523e+00	0.0	15.0
unique_customers	int64	15308	31	0	15.0	31	1.549373e+01	1.0	31.0
new_policy_count	int64	15308	42	0	0.0	2	2.026999e+01	0.0	42.0
ANBP_value	int64	15308	13672	0	0.0	NaN	1.025338e+06	0.0	3933840.0
net_income	int64	15308	15074	0	NaN	NaN	2.280414e+05	1160.0	1140237.0
number_of_policy_holders	int64	15308	73	0	0.0	NaN	3.096832e+01	0.0	116.0
number_of_cash_payment_policies	int64	15308	188	0	0.0	NaN	1.011809e+02	0.0	378.0

The train dataset primarily contains the historical performance data of agents. From 2023-01-01 to 2024-08-01, we have monthly performance records for each agent. Out of a total of 905 agents, 650 agents have complete historical data spanning the entire time range in the training dataset.

The test dataset provides agent performance data for the month immediately following the last available month in the training dataset. Additionally, there are 9 agents (3f4c6d56, 6a0d02ea, d8e13589, 335b4b20, 421aec8b, c37f2ebf, 5444f88c, f9f0169c, b5338672) who have recently joined; therefore, their records are not present in the training dataset. For the remaining agents, their historical data is available in the training dataset.

Date Formatting and Duplicate Handling

To ensure consistency, the date column was converted into a proper datetime format, and the dataset was sorted by agent and in chronological order. Additional date-related features, such as year and month, were extracted to enable potential time-based analyses.

After formatting the dates, we checked for and removed any duplicate records to prevent redundancy. No duplicate dates were found in the dataset.

Furthermore, we applied label encoding to the *agent_code* column to convert categorical agent identifiers into numerical format suitable for modeling.

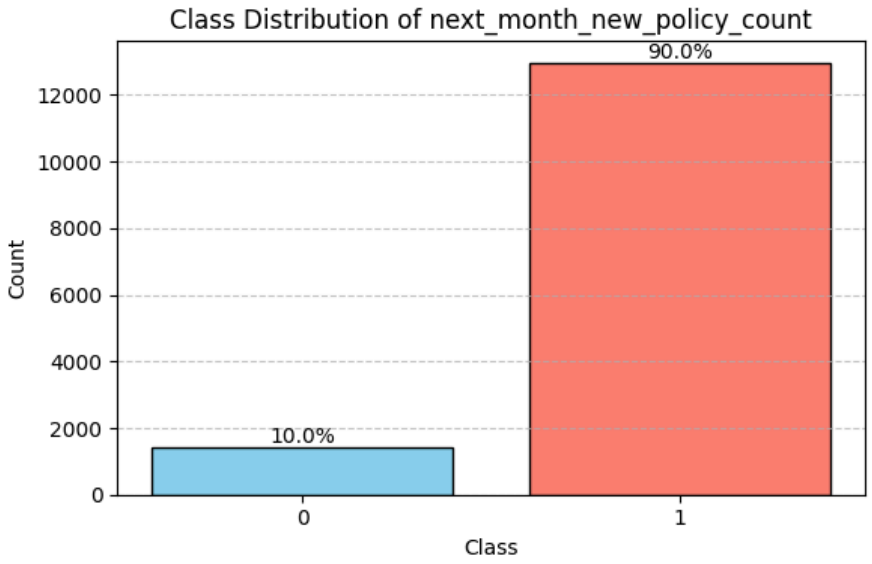
Since our task is to predict NILL agents in the upcoming month, it was important to determine whether the future *new_policy_count* is correlated with past months' data. To address this, we engineered a target variable by shifting the next month's *new_policy_count* to the current month's row. This new target variable was labeled as *next_month_new_policy_count*.

To align with the classification objective (NILL or not), we further transformed the *next_month_new_policy_count* variable into a binary classification target:

0 - if the next month is a NILL month

1 - if there is at least one new policy in the next month

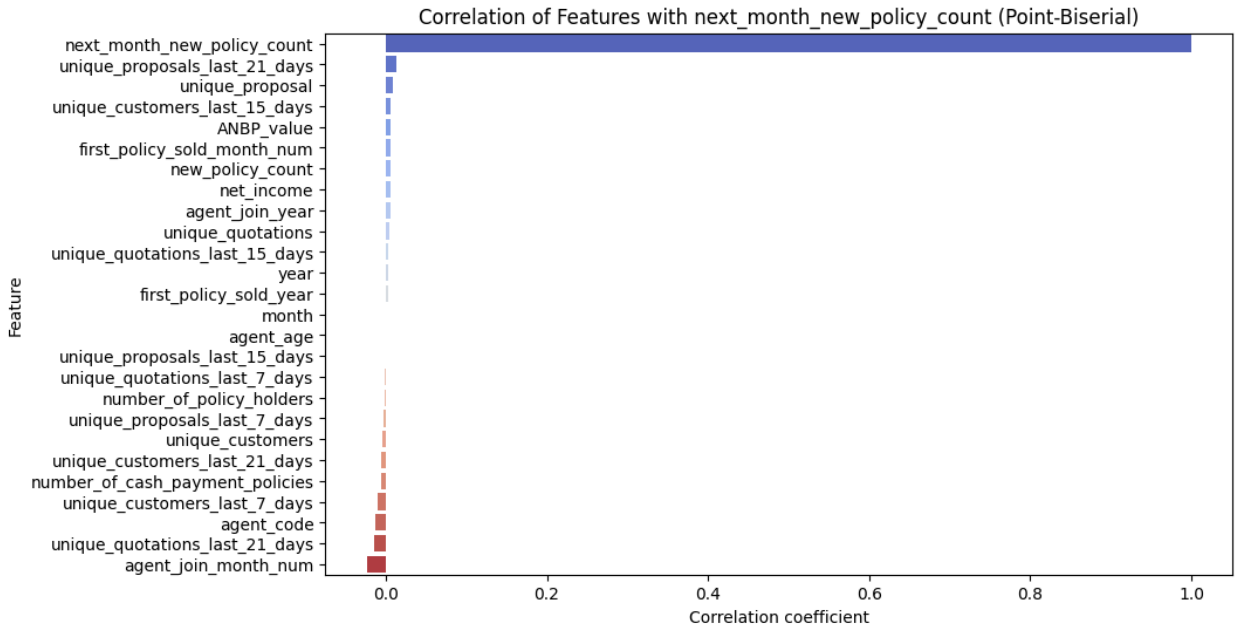
Exploratory Data Analysis



The target variable in this study is **"next_month_new_policy_count"**. The objective of this analysis is to predict this label based on various historical features.

An important characteristic of the dataset is the **class imbalance**, where the majority of the records belong to the **1** class.

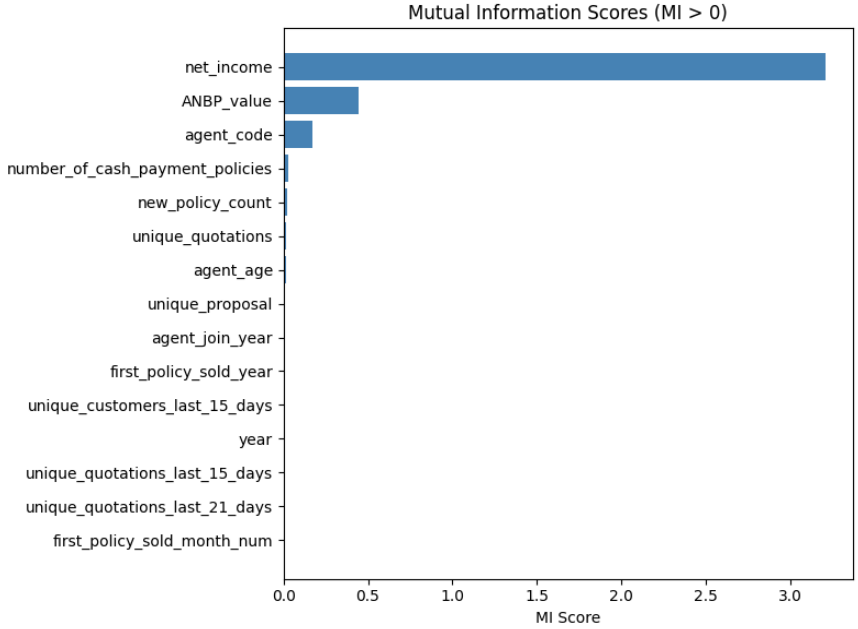
Correlation Analysis



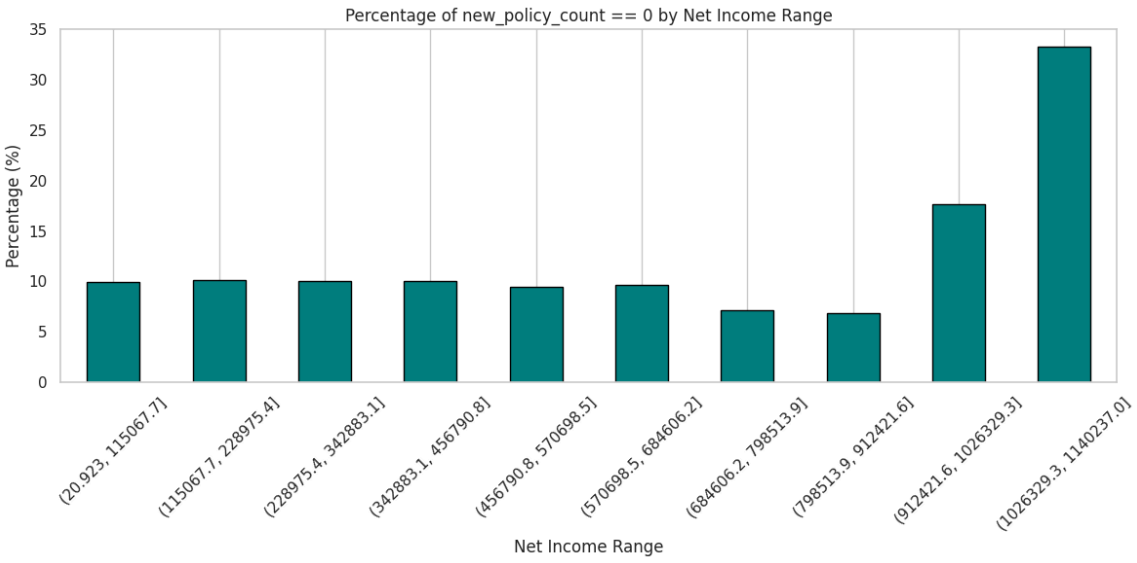
Since our problem involves numerical features and a binary target, we applied Point-biserial correlation (a special case of Pearson correlation) to examine which features have the strongest linear relationship with the target variable. Although we did not observe any notably strong correlations, some features showed a moderate association with the target.

However, because correlation analysis captures only linear relationships, we further computed Mutual Information (MI) scores to identify features that have non-linear or more complex relationships with the target.

Mutual Information Scores

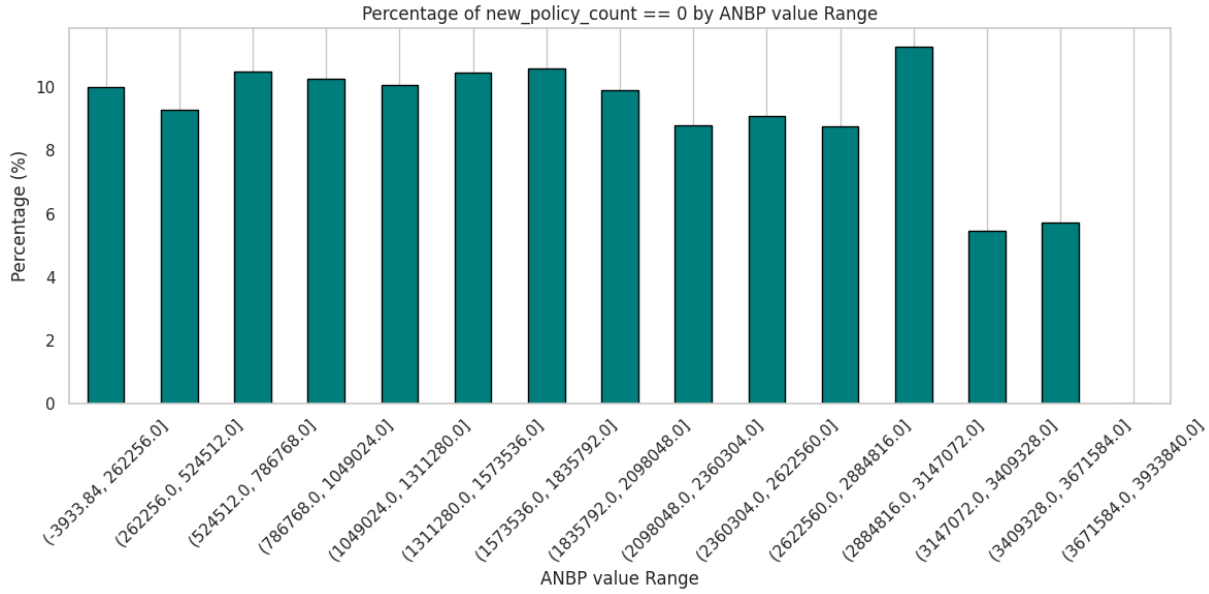


From this analysis, we identified that *net_income*, *ANBP_Value*, *agent_code*, and *number_of_cash_payment_policies* are the features most strongly related to the target variable, next_month_new_policy_count. To gain a deeper understanding of how these features relate to the target, we conducted further visualizations.

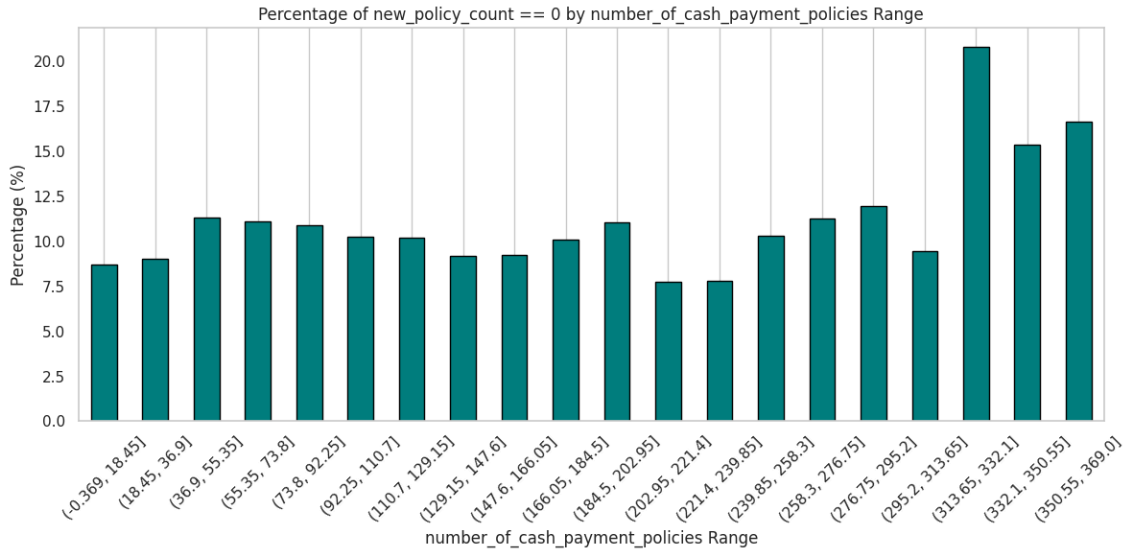


Net Income: We observed that when an agent's net income in the previous month is higher, the probability of being NILL in the next month tends to increase. A possible explanation is that

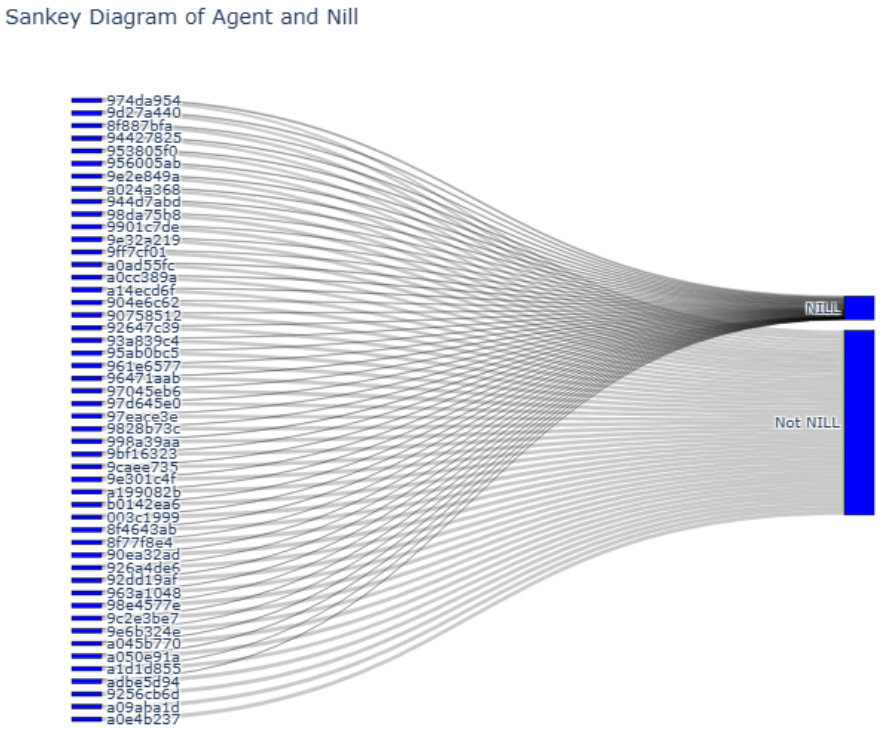
agents who had a very strong month may take a break or become less active in the following month. Alternatively, they might have already met their sales targets and therefore reduce their efforts in the subsequent period.



ANBP Value: In contrast, when the ANBP_Value (Annualized New Business Premium value) in the previous month is higher, the probability of being NILL in the next month decreases. This suggests that agents who sold valuable policies are more likely to maintain momentum and continue selling policies, reducing their likelihood of having zero new policies in the following month.



Number of Cash Payment Policies: We observed that agents who handled a high number of cash payment policies in the previous month have a sharply higher likelihood of having zero new policies in the following month. Possible reasons for this pattern include burnout or exhaustion from a heavy workload, or the possibility that these agents were closing out a large number of existing customers, leading to reduced activity in the next month.



Agent Code: When analyzing agent-wise historical patterns, we found that **200 agents** have never had a NILL month in their entire performance history. The remaining agents have experienced at least one NILL month. Agents without any past NIL months demonstrate consistently good performance.

Feature Engineering

Time-Based Features

These features extract or compute temporal information to capture trends related to time progression and agent activity history:

agent_join_year, ***agent_join_month_num***: Extracted the year and month components from the agent's join date.

first_policy_sold_year, ***first_policy_sold_month_num***: Extracted the year and month components from the date of the agent's first policy sold.

year, ***month***: Extracted the year and month from the main transaction date (year_month) to support time-based grouping and trend analysis.

join_duration: Calculated the duration (in months) between the agent's join date and the current record's date. This measures how long the agent has been active.

time_to_sell: Computed the time (in months) it took for the agent to sell their first policy after joining the company.

To reduce redundancy and prevent data leakage, the original datetime columns, agent_join_month, first_policy_sold_month, and year_month were dropped after extracting these components.

Window-Based Features

These features capture recent trends in agent activity by calculating differences and proportions over different rolling time windows:

proposals_trend: Difference between the unique proposals in the last 7 days and the last 15 days, indicating recent changes in proposal activity.

quotations_trend: Difference between unique quotations in the last 7 days and the last 15 days, highlighting recent shifts in quotation activity.

customers_trend: Difference between unique customers engaged in the last 7 days and the last 15 days, signaling short-term customer acquisition trends.

To understand each agent's performance relative to all other agents within the same month, we engineered proportional features. For a selected set of key activity metrics — such as proposals, quotations, customers engaged, and income

The following features were generated as proportions:

Unique_proposals_last_7_days_proportion
Unique_proposal_proportion
Unique_quotations_last_7_days_proportion
Unique_quotations_proportion
Unique_customers_last_7_days_proportion
Unique_customers_proportion
Number_of_cash_payment_policies_proportion
Number_of_policy_holders_proportion
net_income_proportion

Feature Crossing

past_nill: A binary feature indicating whether an agent has never experienced a NILL month in their historical records. Agents with consistently non-NILL months are marked as 1, others as 0. This helps distinguish agents with strong historical performance.

ML modeling and validation

We tested three machine learning classifiers: XGBClassifier, CatBoostClassifier, and LGBMClassifier. For each model, we performed 10-fold Stratified Cross Validation to ensure robust evaluation given the highly imbalanced nature of our dataset.

Among the models tested, XGBClassifier consistently outperformed the others. Both CatBoostClassifier and LGBMClassifier struggled to generalize well and showed signs of overfitting to the majority class (non-NILL months).

Importantly, we conducted this initial comparison without hyperparameter tuning, to assess each model’s baseline performance.

XGBClassifier — Custom Parameters and Rationale

We retained most default parameters in XGBClassifier but made the following **custom adjustments** to better handle our dataset characteristics:

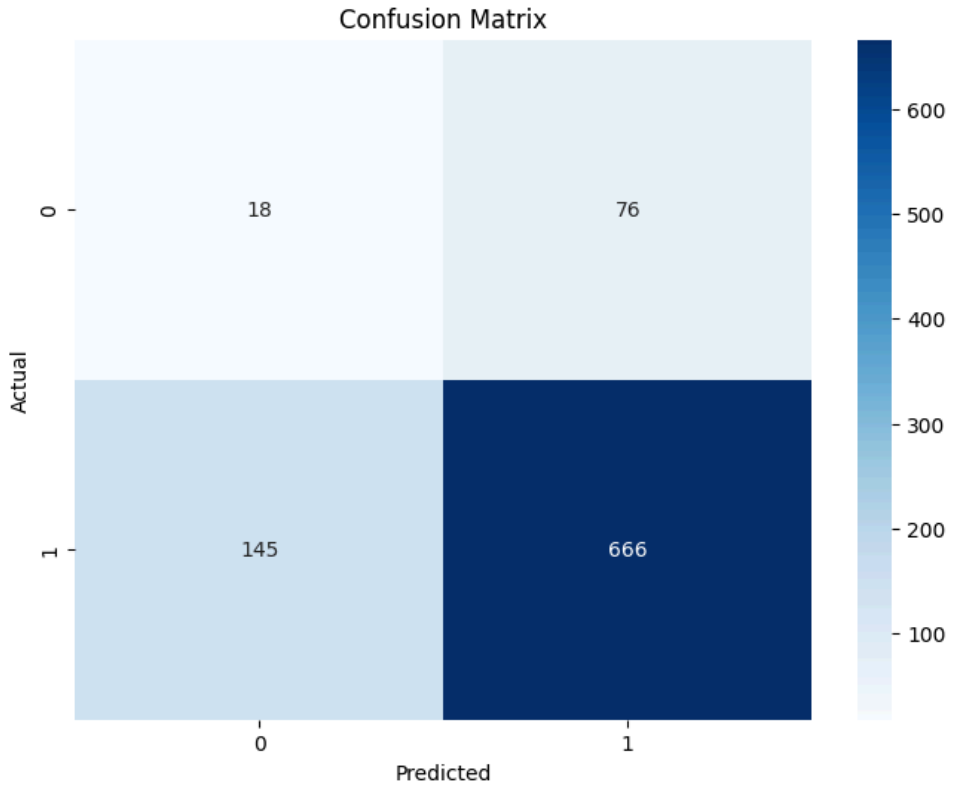
```
xgb_params = {
    'tree_method': 'gpu_hist',  # Accelerates training using GPU
    'device': 'gpu',  # Enables GPU processing
    'enable_categorical': True,  # Efficient handling of categorical variables
    'scale_pos_weight': (y == 0).sum() / (y == 1).sum()  # Balances class weights to counteract imbalance
}
```

Parameter Rationale

- **tree_method** = **'gpu_hist'** and **device** = **'gpu'** were used to **speed up training time**, enabling us to iterate faster on large datasets.
- **enable_categorical** = **True** allows XGBoost to **natively process categorical features** more efficiently, reducing the need for one-hot encoding.
- **scale_pos_weight** is crucial due to the **class imbalance**. By setting it as the ratio of majority to minority class samples, we **penalize misclassification of the minority class** making the model more sensitive to these cases.

Model Performance and Interpretation

After fitting the model with the above parameters, we evaluated it on the test data. The resulting **confusion matrix** is shown below:



Note

In the problem **majority class was labeled as 1** and the **minority class (NILL risk)** as 0. However, this was not ideal for our goal. Since we use F1 score and our priority is to accurately detect NILL agents, it makes more sense to label the minority class (NILL risk) as 1.

Labeling the majority class as 1 **artificially inflates F1 scores** (because true positives are easier to achieve) but does not align with our business objective.

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

As observed, **predicting the minority class** (NILL month agents) remains challenging due to data imbalance. However, adjusting `scale_pos_weight` notably **improved our true negative rate** for identifying at-risk agents.

Without this adjustment, the model **would heavily favor the majority class**, leading to almost all predictions being positive(non-NILL months). Although tuning `scale_pos_weight` slightly reduces overall **F1 score**, it aligns better with **business objectives**:

> The insurance company prioritizes **early detection of potential NILL month agents** over marginal improvements in a general F1 score.

Thus, we **strategically optimized the model for higher recall** of the positive class (NILL risk), ensuring the system provides **actionable insights** even if it trades off slightly on overall accuracy metrics.

List of top factors affecting early performance

Based on our analysis and exploratory data visualizations, we identified several key factors that strongly influence an agent's likelihood of achieving zero new policies (NILL month) in the following month. These factors provide valuable insights into patterns of early performance and activity trends:

Net Income (Previous Month)

Agents with exceptionally high net income in the previous month tend to have a higher chance of experiencing a NILL month in the next period. This pattern may be attributed to agents taking a break after a strong sales month or temporarily reducing activity after meeting performance targets.

ANBP Value (Previous Month)

Higher ANBP value (average net business premium) in the prior month is associated with a lower likelihood of a NILL month next month. This suggests that agents selling high-value policies are more consistent and likely to maintain steady performance.

Number of Cash Payment Policies(Previous Month)

Agents who processed a large volume of cash payment policies in the previous month show an increased risk of a NILL month afterward. Possible explanations include agent fatigue, burnout from high workload, or the completion of bulk policy closures leading to a temporary slowdown.

Agent Historical NILL Status (past_nill)

Agents with no history of NILL months across their recorded data exhibit a strong tendency to avoid NILL months in the future as well. We engineered a feature (past_nill) to capture this history, as such consistently performing agents are statistically less likely to experience zero policy months.

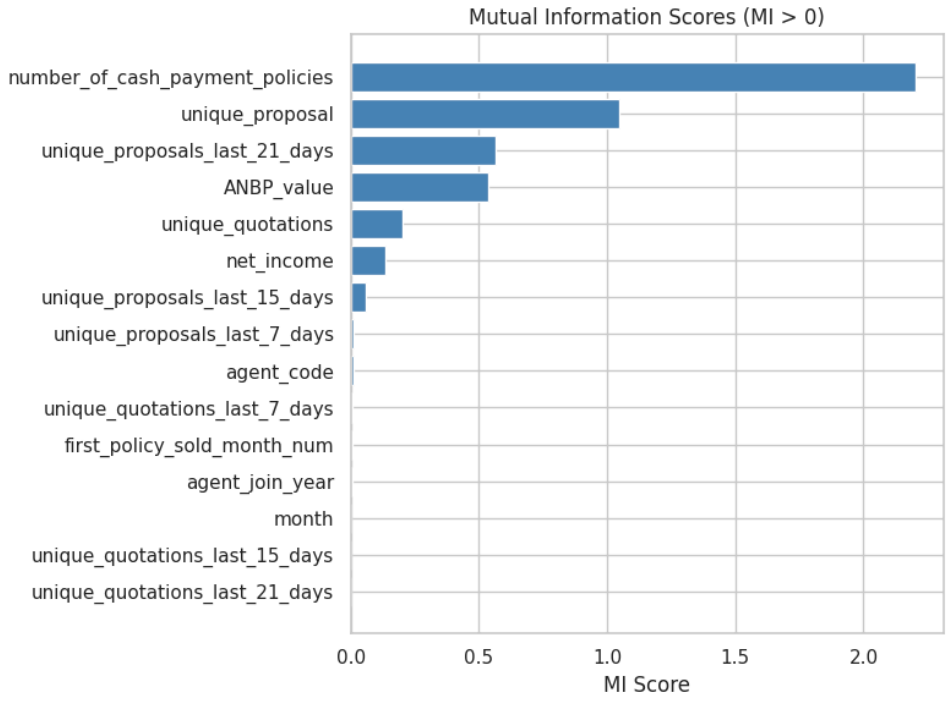
Relative Performance Proportions(Previous Month)

Proportion-based features, such as net_income_proportion, unique_proposals_proportion, and number_of_cash_payment_policies_proportion, reflect how an agent performed relative to peers in a given month. High or low relative performance gives further signals about future activity patterns.

Personalized action plans for at-risk agents

Our Mutual Information (MI) analysis identified two activity-related features for current month's success. The unique_quotations and unique_proposals in that ongoing month hold high relation with that month's performance.

The plot below illustrates the top features ranked by MI scores, with unique_quotations and unique_proposals emerging as highly influential:



These findings imply that encouraging agents to increase their quotations and proposals activity in the current month can substantially reduce their likelihood of experiencing a NILL month.

Recommended Personalized Action Plans

Risk Indicator	Identified Reason	Recommended Action Plan
Low unique_quotations & unique_proposals	Lack of customer engagement	Targeted reminders to boost customer quotations and proposals Assign leads from under-served customer segments
High net_income (Previous Month)	Likely slowdown after strong month (possible break or reduced push)	Gentle prompts to maintain engagement Recognition badges for consistency
High number_of_cash_payment_policies (Previous Month)	Burnout or exhaustion from prior heavy workload	-Recommend balancing workload with digital/online payment policies Encourage re-engagement with lighter tasks (follow-ups, renewals)
Low relative performance proportion (compared to peers)	Lagging behind peers despite steady activity	Personalized coaching sessions)