Insurance Renewal Prediction

PGCP AIML Cohort 12 - AIML_Oct_2024 - Group 12

Team

- Akash Mohanty (2024AIML133)
- Ishika Gupta (2024AIML088)
- Namratha S Hegde (2024AIML118)
- Thrilok Attota (2023AIML120)
- Yaswanth Reddy Dasari (2024AIML010)

Mentor: Tapas Chakraborty

Industry Evidence

InsuranceNewsNet
https://insurancenewsnet.com › innarticles

P/C insurers faced with accelerating churn and switching ...

27 Mar 2023 — Property/casualty insurers had one of their worst years in 2022 as runaway inflation, shortages and higher prices drove up claim costs.

Insurance Times
https://www.insurancetimes.co.uk › 1444637.article

Home insurance market seeing considerable customer churn

23 May 2023 — Exactly 44.4% of **insurance customers** switched their policy providers in 2022, reveals

Insurance Asia
https://insuranceasia.com › insurance › exclusive › insu...

Insurers face customer churn as prices rise in select products

20 Jun 2024 — However, **customer** switching **rates** went **up** in the property and casualty (P&C) markets due to **increased** premiums and reduced coverage, according ...

Reuters

India's LIC posts quarterly profit rise on higher premiums from renewed policies

Aug 7, 2025

What We’re Solving

The Need
For insurance companies, renewals are the foundation of consistent revenue and long-term relationships. Securing renewals ensures both financial stability and stronger customer loyalty.

The Challenge
Renewal decisions are often uncertain, influenced by many subtle factors that are hard to anticipate. This unpredictability leads to lost opportunities, declining retention, and weakened trust.

The Approach
By applying machine learning to historical patterns, insurers can move from reactive guessing to proactive prediction. This data-driven perspective provides a clearer understanding of renewal behavior.

The Result
With reliable predictions, insurers can act early, strengthen engagement, and improve overall retention. This leads to healthier customer relationships and sustainable business growth.

Insurance Renewal – Overview

The Foundation

- Renewals are the backbone of insurer revenue and growth
- Retaining customers is more cost-effective than acquiring new ones

The Challenge

- Renewal decisions are uncertain and influenced by many factors
- Missed renewals cause revenue loss and weaken customer trust

Why It Matters

- Predicting renewals helps engage at-risk customers proactively
- Strong renewal rates build loyalty and ensure business sustainability

The Aim

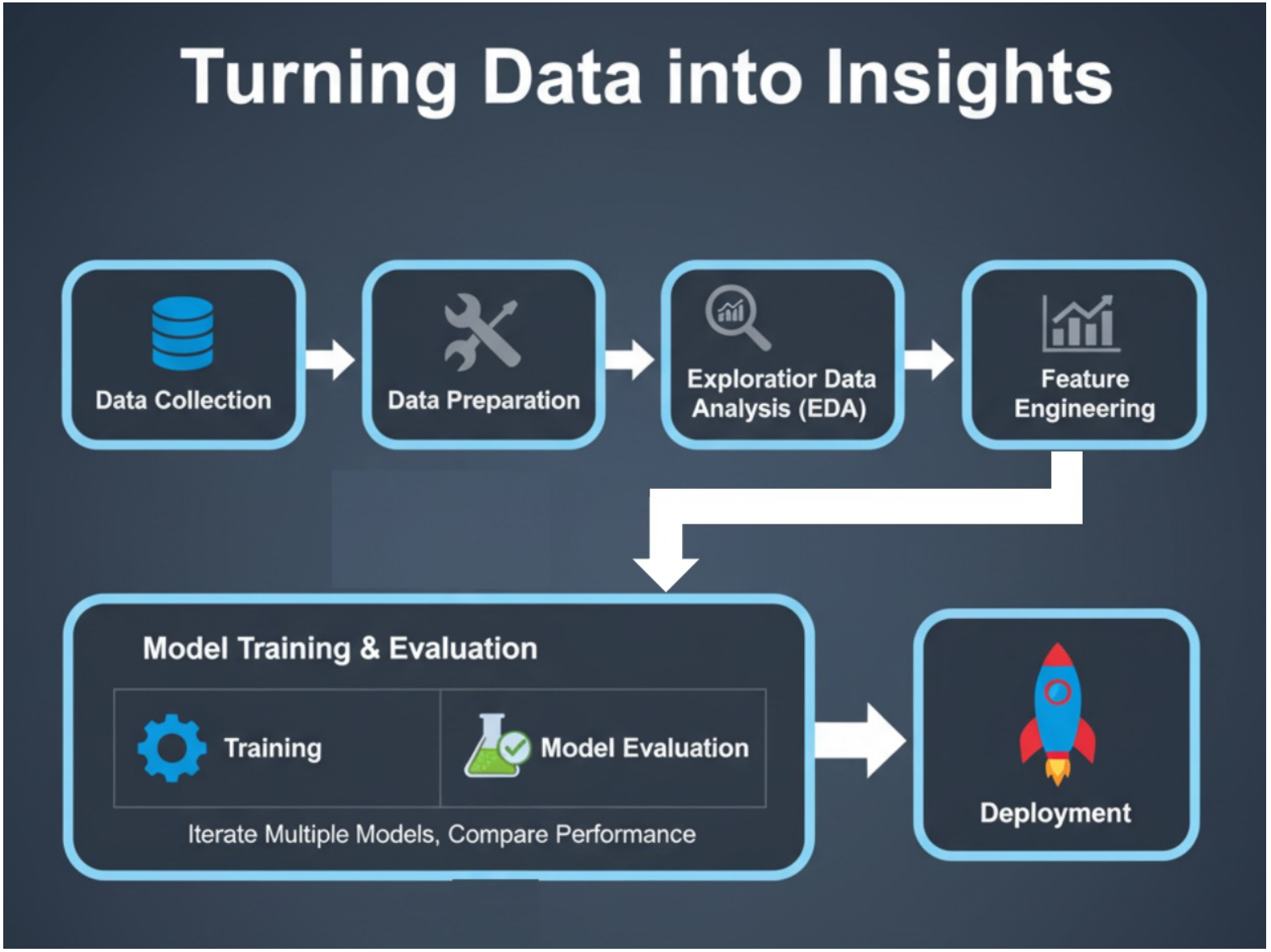
To build a machine learning–based system that predicts the likelihood of customers renewing their insurance policies, enabling proactive engagement and improved retention.

The Roadmap

- Gather and structure historical customer and policy data
- Engineer meaningful signals to capture behavioral and transactional patterns
- Explore and analyze renewal trends through data visualization and EDA
- Build and evaluate predictive models (Logistic Regression, Random Forest, XGBoost, Neural Networks)
- Compare model performance using metrics like **Recall, F1-score, Precision, and PR-AUC**
- Identify key factors influencing renewal decisions through feature importance and explainability methods
- Provide actionable insights for retention campaigns and customer communication strategies
- Validate predictions through pilot testing with real customer segments to assess business impact

The Outcome

- Reliable predictions of renewal likelihood
- Early identification of at-risk customers
- Data-driven strategies for retention and engagement
- Strengthened customer trust and sustainable business growth


Turning Data into Insights
Data Collection
Data Preparation
Exploratior Data Analysis (EDA)
Feature Engineering
Model Training & Evaluation
Training
Model Evaluation
Iterate Multiple Models, Compare Performance
Deployment

Scope & Status

Work Completed

Data Loading and Preparation

- Imported customer and policy datasets
- Performed sampling and balancing

Data Preprocessing

- Corrected data types
- Handled missing values
- Detected and treated outliers
- Encoded categorical variables
- Addressed class imbalance

Exploratory Data Analysis (EDA)

- Visualized renewal trends
- Performed correlation analysis
- Identified key influencing variables

Feature Engineering

- Derive new variables
- Transform behavioral data into model-ready features

Next Steps

Data Readiness for Modeling

- Finalize train-validation-test splits
- Ensure scaling/normalization if required

Model Development & Tuning

- Train and evaluate models (LogReg, Random Forest, XGBoost, Neural Nets)
- Cross-validation and recursive feature elimination
- Feature importance and explainability (e.g., SHAP)
- Hyperparameter tuning and optimization

Model Comparison & Reporting

- Compare models using Recall, F1, Precision, and PR-AUC
- Select best-performing model
- Document findings and business insights

Technical Setup

- **Operating System –** Windows / Mac
- **Development Environment –** Jupyter Notebook (Python)
- **Libraries Used –** scikit-learn, pandas, matplotlib, numpy, seaborn, imblearn
- **Dataset Source –** Kaggle: Insurance Renewal Prediction (LazyPredict)
- **Dataset Type –** CSV flat file (~80k records, mix of numeric & categorical features)
- **Version Control –** GitHub repository: PredictMyPolicy

Influencing Factors

Category	Features	Dtype
Customer Demographics	age_in_days, Income	Numeric
Payment Behavior	perc_premium_paid_by_cash_credit, Count_3-6_months_late, Count_6-12_months_late, Count_more_than_12_months_late, no_of_premiums_paid	Numeric
Risk & Underwriting Factors	application_underwriting_score	Numeric
Policy Attributes	premium	Numeric
Distribution & Geography	sourcing_channel, residence_area_type	Categorical
Target Variable	renewal	Binary （Binary classification problem)

Missing but Important Renewal Factors

Factor	How it Influences Renewal
Claim History	Frequent or large claims may reduce renewal likelihood due to higher perceived risk.
Type of Insurance	Renewal behavior differs across product lines (e.g., health vs. motor vs. life).
Gender & Marital Status	Demographics can shape risk appetite and renewal decisions.
Policy Add-ons & Riders	Extra benefits improve stickiness and increase renewal probability.
Customer Satisfaction	Poor service or unresolved complaints drive non-renewals.
Competitor Switching	Availability of cheaper/better alternatives can cause churn.

Loading and Reading Dataset

- Dataset from Kaggle (~80k records)
- Identified target variable: Renewal (binary 0/1)
- Reviewed feature composition: numeric (age, income, premium, late counts) & categorical (sourcing channel, residence type)
- Performed initial analysis:
  - Dataset size and structure (~80k records)
  - Target variable validation (renewal = 0/1)
  - Feature type verification (numeric & categorical)
  - Missing values check
  - Duplicate records check
  - Invalid or inconsistent values check (e.g., “Urban” vs “Urbn”)

Problems with Dataset

Minimal Feature Set

- Dataset has only **basic info**: income, age, premiums, late payments…
- Missing **key churn drivers**, e.g.:
  - Policy type
  - Claim history / claim delays
  - Complaints / service issues
  - Cross-sell / loyalty (multiple products)

Severe Class Imbalance

- Renewal rate ≈ **93%** (non-renewals ≈ 7%)
- Model learns to **predict renewal always** → looks accurate but misses churn
- Imbalance **weakens predictive power** on the segment we care about most (churners)

Business Implication

- Current data = **demographics + payments only**
- Lacks **core insurance behavior signals** and is **imbalanced**
- Together, these issues **limit model's ability to predict churn effectively**

Data Preprocessing & Transformation

1. Dropped irrelevant column (`id`)

The `id` column is only a unique identifier.
It does not carry any predictive power for renewal.
Keeping it could mislead the model into finding false patterns.

2. Converted age_in_days → age_in_years

Days are too granular and less interpretable.
Years are more meaningful for business understanding (young vs mid-age vs senior).
Helps in creating age groups/buckets if needed (e.g., 20–30, 30–40).

3. One-hot encoding for categorical features (sourcing_channel, residence_area_type)

ML models (like Logistic Regression, XGBoost) need numeric inputs.
One-hot encoding converts categorical variables into binary indicators (0/1).
Preserves information without introducing ordinal bias (e.g., channel A ≠ channel B, but both get their own dummy variable).

Missing Value Analysis

Late Payment Features (Count_3–6m, Count_6–12m, >12m)

Always missing **together** → suggests *systematic missingness*, not random.
Likely indicates **"no late payments"** → imputed with **0**.
Since these features are highly predictive, care taken not to remove variance with blanket imputation.

Application Underwriting Score

Continuous risk score influenced by income, premium, age, and number of premiums paid.
Mean/median imputation would distort distribution.
Used **KNN imputation** to preserve underlying relationships.

Outlier Analysis

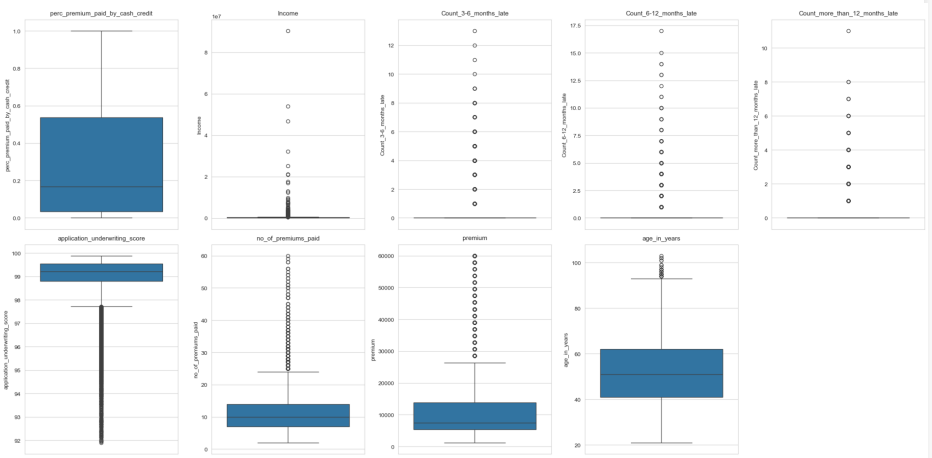
- **Method Used** – Interquartile Range (IQR) method applied to all numeric features.

- **Rationale** – Detect values that lie beyond the expected distribution range.

- **Approach**

  Calculated Q1, Q3, and IQR for each feature

  Flagged observations outside $Q1–1.5\times IQR, Q3 + 1.5\times IQR$ as outliers



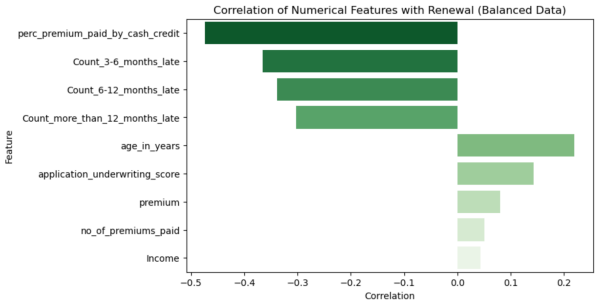
Late Payment Features – Outlier

Distribution

Heavily zero-inflated; most customers have no late payments, smaller subset has positive counts.

IQR Result

Both bounds = **0** → technically flags every positive count as an outlier.



Interpretation

Positive counts are **not errors**, but valid signals of delinquency.

Strong **negative correlation with renewal** → critical predictors.

Key Insight

If all missing values were imputed with 0, these columns would collapse to all zeros

effectively **erasing the feature**.

Correlation analysis shows these features are among the **most important**.

Decision

Preserve in raw form; retain positive counts; avoid blanket zero-imputation.

Application Underwriting Score Outliers

IQR Bounds → 97.72 to 100.64

Outliers Detected

3,381 rows (454 distinct values), including values such as 91.9, 91.96, 92.03, etc.

Renewal Rates

Outliers: **87.84%** vs Non-outliers: **94.00%**

Outliers were **6.16% more likely not to renew**

Interpretation

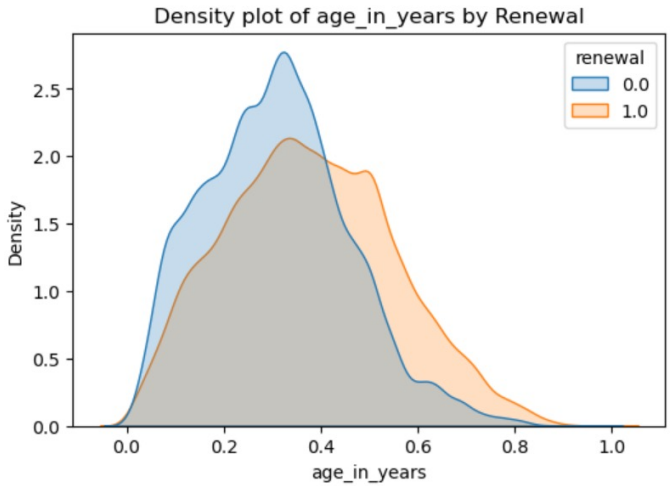
Low underwriting scores reflect higher risk customers

Predictive of non-renewal behavior

Decision

Retained outliers, as they carry strong predictive signal

Age Outliers



IQR Upper Bound → 93.5 years

Outliers Detected → 44 records (ages 94–103)

Renewal Rates → Outliers: **95.45%** vs Non-outliers: **93.74%**

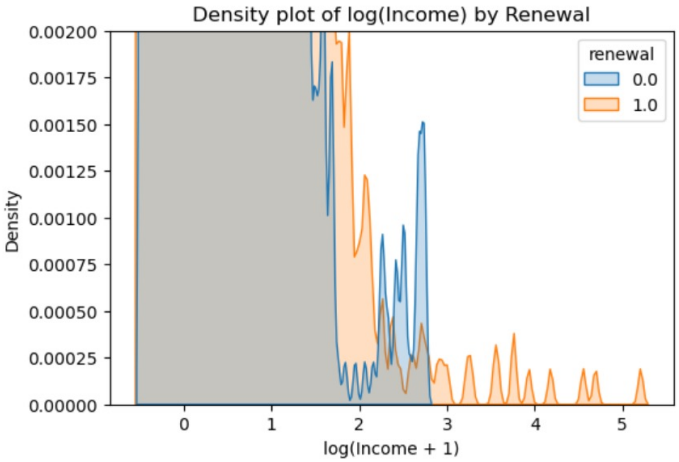
Interpretation →

These are realistic elderly customers, not data errors
Density plot of renewal by age shows older customers are **more likely to renew**
Renewal rates show **no adverse effect** → outliers retained

Note →

In real-world insurance, certain policy types may not be offered to very elderly customers
Since **policy type is not part of our dataset**, we cannot assess this interaction fully

Income Outliers



IQR Bounds → –108,110 to 468,210

Outliers Detected → 3,428 rows (2,233 distinct values) up to 470,0

Renewal Rates → Outliers: **95.22%** vs Non-outliers: **93.67%**

Interpretation →

Outliers correspond to **very high-income customers**

High-income customers (right tail) show a stronger **renewal tendency** compared to lower-income customers.

These customers show **higher renewal loyalty**

Outliers retained as they are **business-relevant segments**

Premium Outliers

IQR Upper Bound → 26,400

Outliers Detected → 4,523 rows (16 distinct values)

Pattern Observed → Structured increments (28,500 → 30,600 → … → 60,000)

Renewal Rates → Outliers: **95.49%** vs Non-outliers: **93.64%**

Conclusion →

- These are valid premium brackets for **high-value policies**
- No adverse impact on renewal → **outliers retained**

Number of Premiums Paid – Outliers

IQR Upper Bound → 24.5 premiums

Outliers Detected → 1,426 rows (34 distinct values), up to 60 premiums

Renewal Rates → Outliers: **93.13%** vs Non-outliers: **93.75%**

Interpretation →

Outliers represent **long-tenured, loyal policyholders**
Renewal behavior is consistent with rest of population
Outliers retained as they provide valuable information

Logical Consistency Checks

Late Payments vs Premiums Paid

- 242 rows (~0.3%) recorded **more late payments than premiums paid**
- Example: Customer with 2 premiums but 6 late payments
- Indicates data recording issues or unrealistic behavior

Premiums Paid vs Age

- Some records showed **premiums paid greater than customer's age**
- Unrealistic under the assumption of **yearly premium cycle**
- Suggests potential anomalies in reporting or data entry

Extreme Late Payment Behavior

183 customers (~0.23%) had **≥6 late payments** across any lateness bucket

Out of these:

92 renewed (50.3%)

91 did not renew (49.7%)

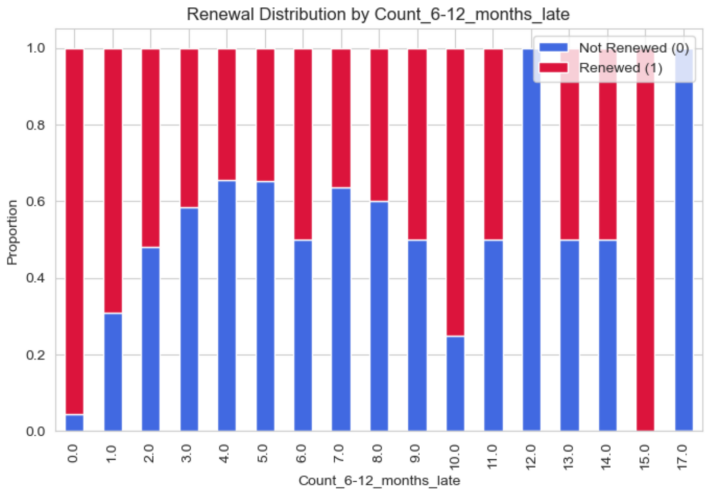
Renewal rate ≈ **50%**, unlike the population average of **~95%**

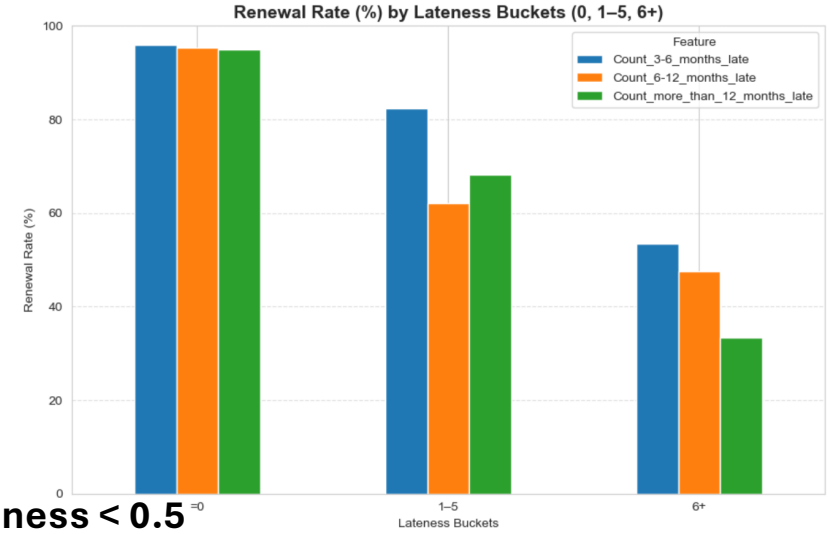
Such cases behave like **noise/random**, not consistent patterns

For churn modeling:

Removing 92 **renewed records (~0.12%)** improves clarity

Treating ≥6 late payments as **high churn risk** is logical and business-aligned





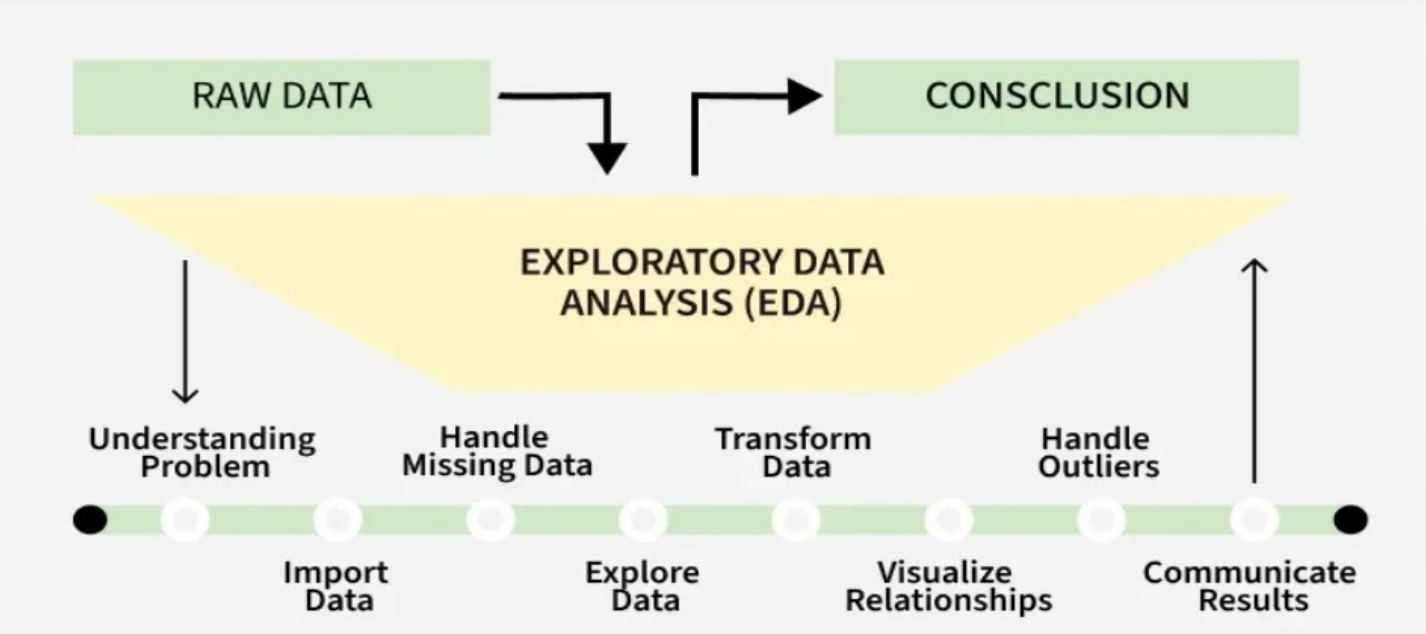
KNN Neighbor Analysis (excluding lateness & target):

For these extreme renewed cases, the **average neighbor lateness < 0.5**

Confirms they are **very different from their peer group**

Strengthens the case that these records are **anomalies**

Exploratory Data Analysis (EDA)



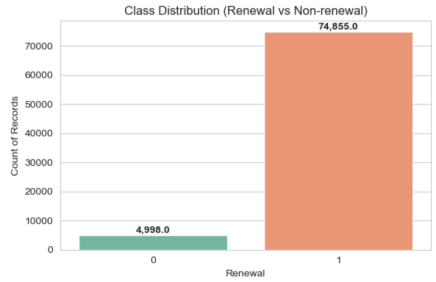
Having addressed **data quality and logical consistency**,

We now focus on exploring **patterns, distributions, correlations, and feature relationships**

Aim → uncover signals that explain **renewal vs non-renewal**

Class Imbalance

Observation → Dataset is highly imbalanced: **~93% renewals vs 7% non-renewals**



Risk → Training directly would bias models toward the majority class (renewals)

Mitigation Strategies

Tree-based models → Used **class weights** to counter imbalance

Statistical models

Applied **SMOTE/ADASYN oversampling** to generate synthetic minority samples

Evaluation Metrics

Accuracy alone is misleading

Use **Balanced Accuracy, F1-score, and PR-AUC** to ensure performance reflects minority class capture

Residence Area Type

Exploratory Analysis

Renewal rates nearly identical for **Urban vs Rural** customers

Suggests very **minimal predictive power**

Multivariate Interaction

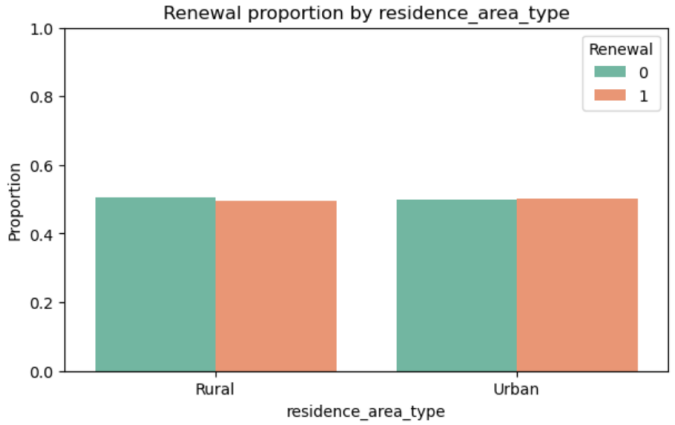
Tested interactions with **premium, age, income, late payments, underwriting score, and sourcing channel**

Differences in renewal rates remained negligible (**<3%**)

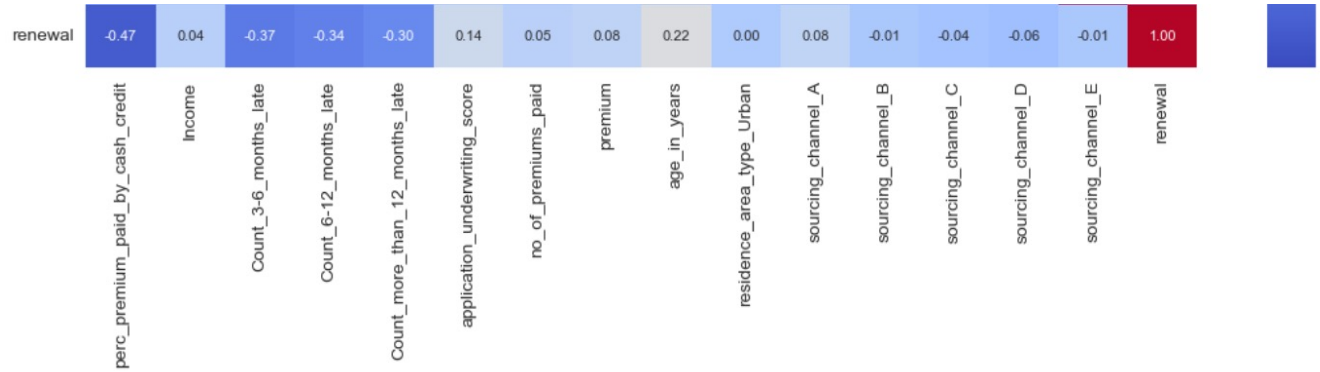
Conclusion

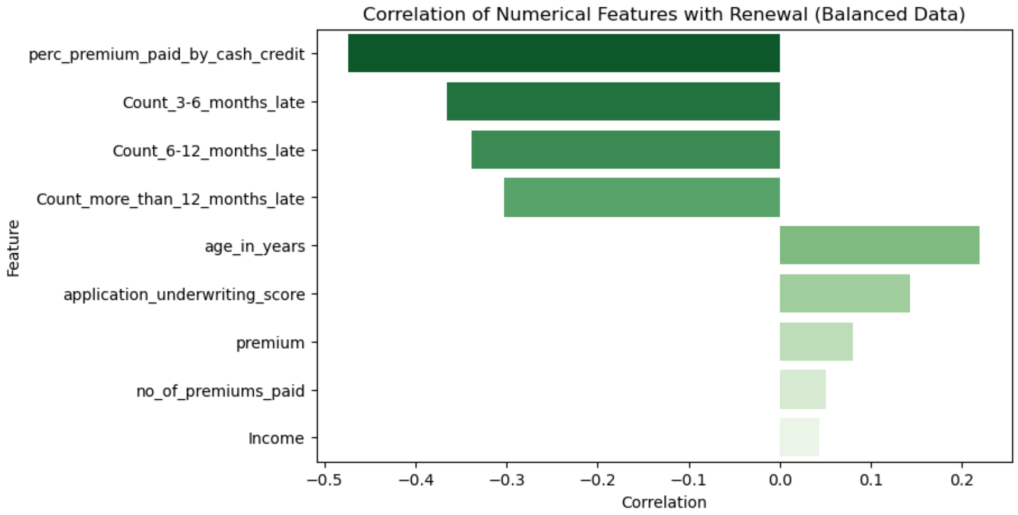
Tree-based models → Feature can be retained (automatically down-weighted if weak)

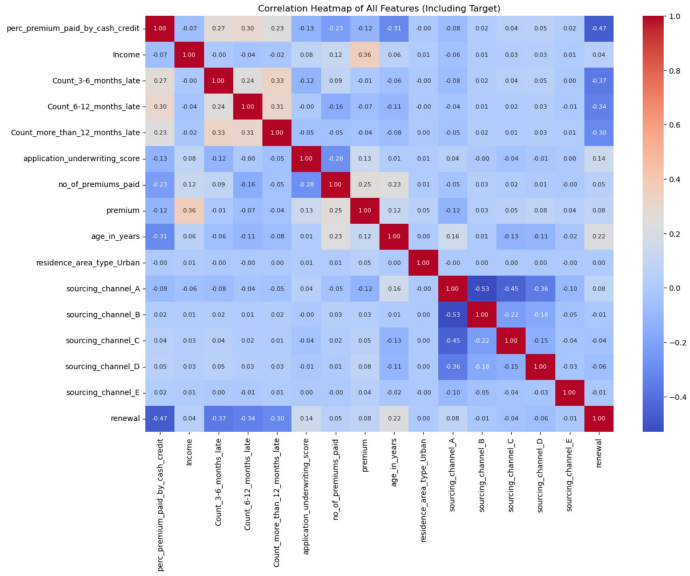
Statistical models → Safe to drop to reduce noise and improve model stability



Higher late payment counts → strongly associated with non-renewal



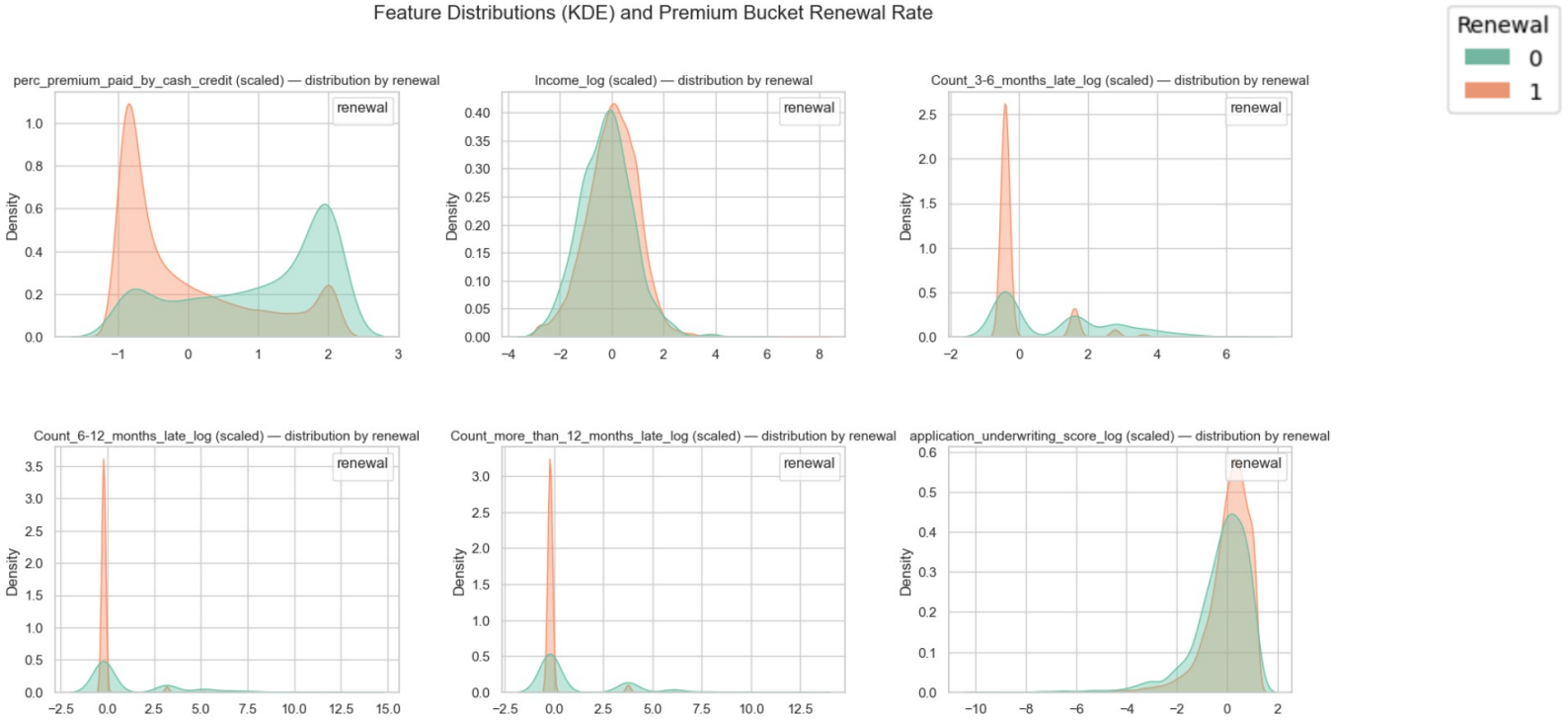




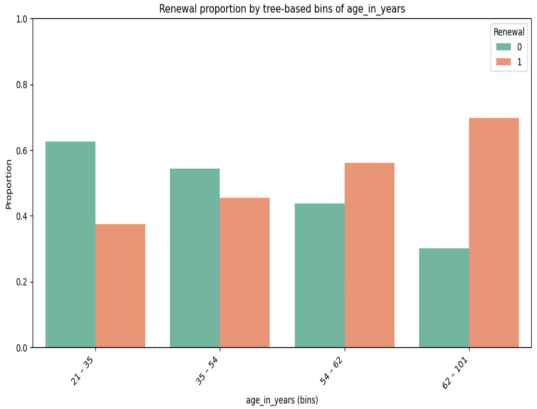
Density Plots

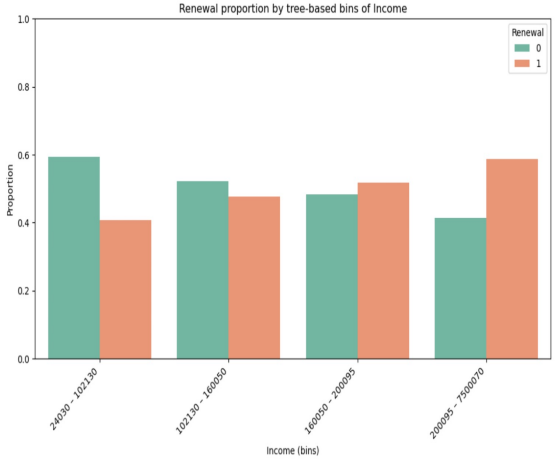
Heavy overlap between renewers and non-renewers across most features

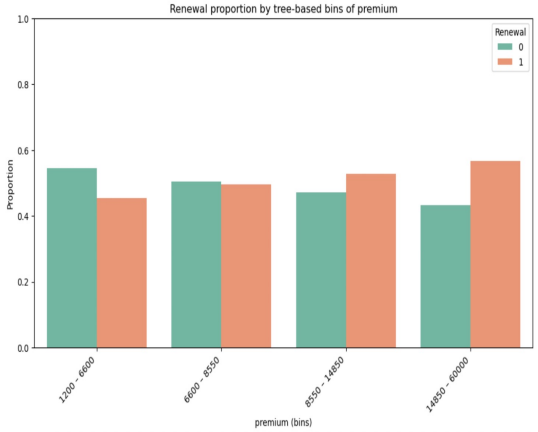
% Premium Paid by Cash/Credit → clearer separation, with non-renewers peaking at higher values

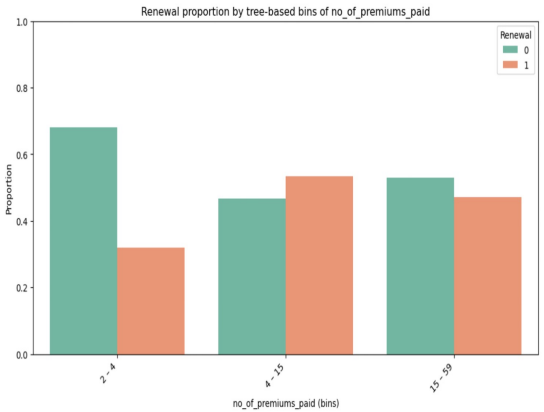


Higher income, **more premiums paid**, and **older age** → positively associated with renewal







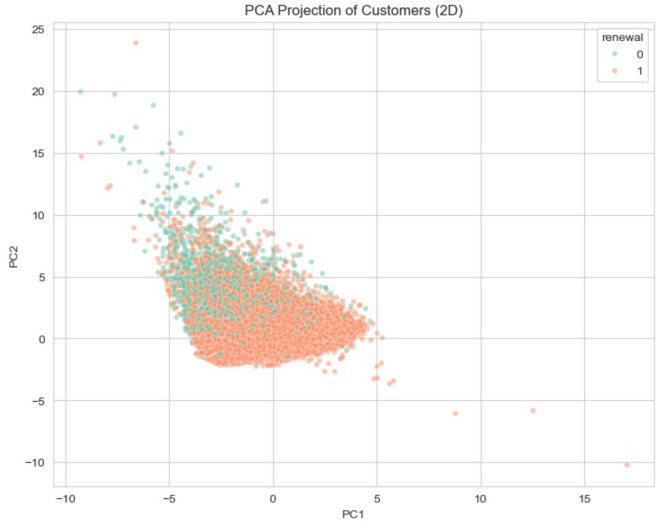


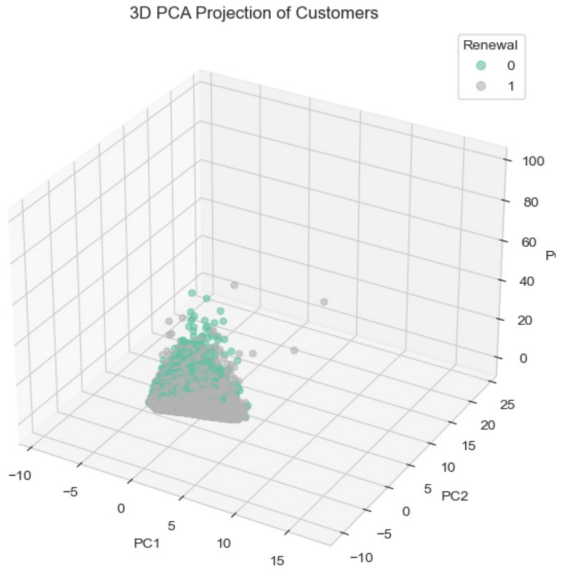
PCA Projection – Customer Renewal

2D & 3D PCA plots show **substantial overlap** between renewers (1) and non-renewers (0).
Indicates that **dominant variance directions** in the data do not cleanly separate the classes.
Renewal behavior is driven by **complex, non-linear feature interactions**, not simple variance.

Implication

Linear models relying only on variance (e.g., plain Logistic Regression without interactions) may underperform.
Tree-based ensembles (Random Forest, XGBoost, Gradient Boosting)
and other non-linear algorithms are better suited to capture these interactions.





Overall EDA Conclusions

Class Imbalance → 93% renewals vs 7% non-renewals; requires resampling/weighting

Feature Correlations →

Negative: % Premium Paid by Cash/Credit, Late Payments
Positive: Age, Underwriting Score, Premium
No single dominant predictor → predictive strength from **combined weak signals**

Categorical Feature (Residence Area Type) → Minimal impact on renewal, low predictive power

Density Plots → Higher income, more premiums, older customers renew more; late payments strongly linked to non-renewal

PCA Visualization → Strong overlap between classes; renewal driven by **non-linear feature interactions**

Modeling Implication → Tree-based ensembles best suited to capture these relationships

Feature Engineering (Applied)

Transformations Applied

- Converted age_in_days → age_in_years for interpretability
- One-hot encoding for categorical features (sourcing_channel, residence_area_type)
- Normalized continuous features (Income, Premium, Age, Underwriting Score) for scale-sensitive models

Conclusion

- These steps made raw data more interpretable and model-friendly
- Ensured compatibility with both statistical and tree-based algorithms

Feature Engineering (New derived features)

New Features to Explore

Total Late Payments = Sum of 3 late payment features → captures overall delinquency
Late Payment Ratio = Total Late Payments ÷ Premiums Paid → delinquency adjusted for tenure
Premium-to-Income Ratio = Premium ÷ Income → affordability signal

Conclusion

These engineered ratios may strengthen predictive signal
Especially useful for statistical models that don't capture interactions automatically

Modeling Summary

Objective
The goal is to build a robust classification model to **predict the likelihood of non-renewal** (customer churn), enabling insurers to proactively engage at-risk customers and improve retention.

Modeling Approach
We are following a consistent pipeline across all models:
Preprocessing and feature engineering (age conversion, one-hot encoding, normalization)
Addressing class imbalance (Class Weights, SMOTE/ADASYN)
Splitting dataset using **train_test_split() with stratification** to preserve class ratios
Training multiple classifiers: **Logistic Regression, Random Forest, XGBoost, SVM, Neural Networks**
Evaluating models with metrics suited for imbalance and churn capture

Evaluation Metrics
Accuracy is avoided due to imbalance (~93% renewals vs 7% non-renewals)
Focus is on:

- **Recall (priority)** → to maximize capture of churners (non-renewals)
- **F1-score** → to balance recall & precision
- **PR-AUC** → to ensure robustness under imbalance

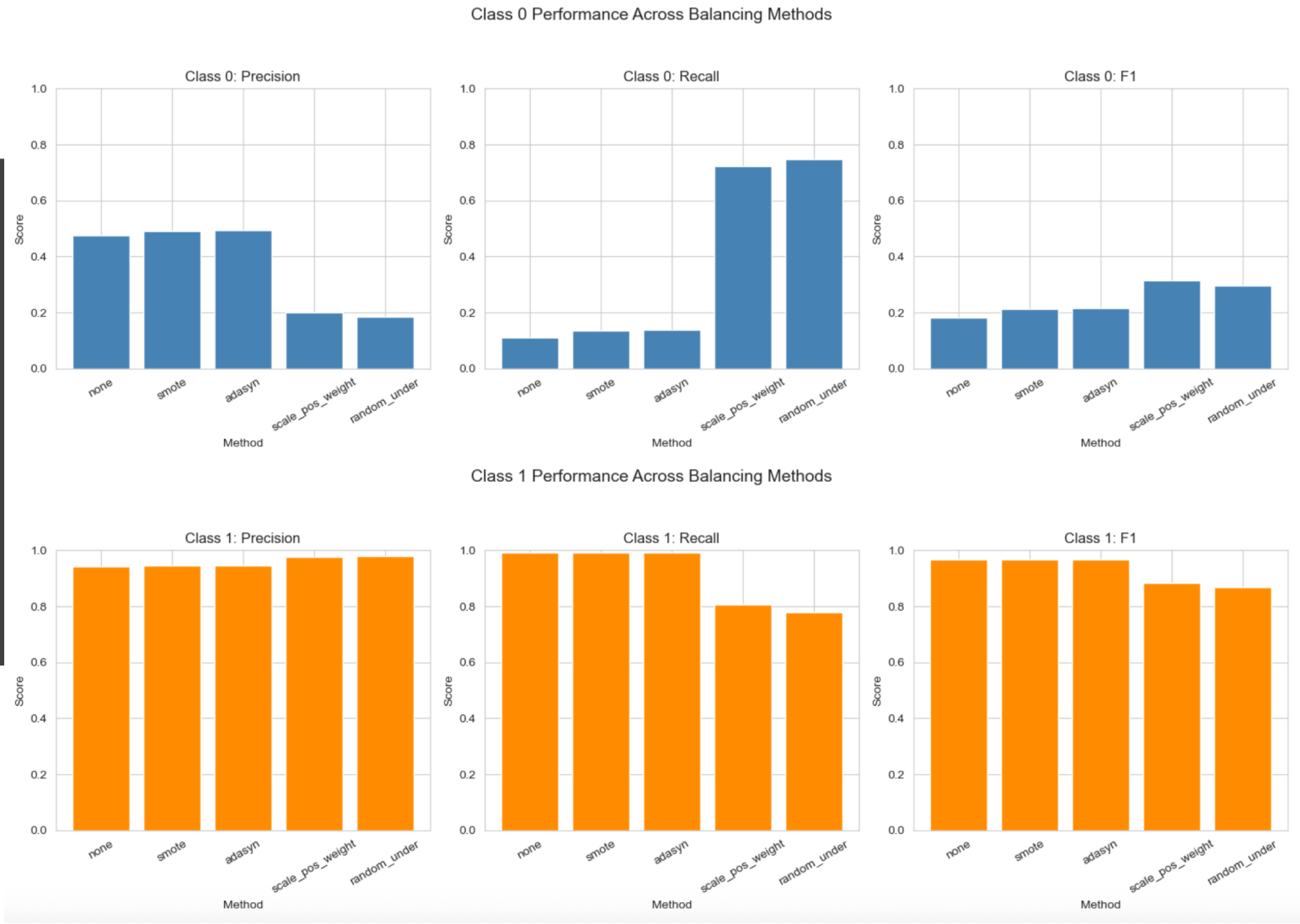
Models Trained & Evaluated(Untouched Data)

Model	Accuracy	Recall (Class 0)	Recall (Class 1)	F1 (Class 0)	F1 (Class 1)	ROC-AUC	Verdict / Notes
Logistic Regression	0.9381	0.123	0.9926	0.1994	0.9678	0.8319	Interpretable & fast baseline. Moderate recall for Class 0.
XGBoost	0.9354	0.145	0.9882	0.2195	0.9663	0.8205	Best F1 on Class 0, fast training, well-balanced.
Gradient Boosting	0.9376	0.124	0.9920	0.1994	0.9676	0.8452	Best ROC-AUC, great balance across metrics.
Random Forest	0.9376	0.111	0.9928	0.1821	0.9675	0.8260	Stable, good fallback, weaker on Class 0 recall.
SVM	0.9374	0.000	1.0000	0.0000	0.9677	0.5066	Ignores Class 0 entirely — not usable for churn detection.

Models Trained & Evaluated(over sampled data)

Model	Accuracy	Precision (Class 0)	Recall (Class 0)	F1 (Class 0)	Recall (Class 1)	ROC-AUC	Notes
Logistic Regression	0.9378	0.5146	0.123	0.1985	0.9923	0.829	Interpretable, very high recall on Class 1, weak on minority (Class 0)
XGBoost	0.9354	0.4517	**0.145**	**0.2195**	0.9882	0.821	Best balance between recall & F1, good generalization
Gradient Boosting	0.9376	0.5082	0.124	0.1994	0.992	0.845	Strong ROC-AUC, still weak Class 0 recall
Random Forest	0.9376	0.5068	0.111	0.1821	0.9928	0.826	Robust baseline, strong Class 1 performance
SVM	0.9374	0	0	0	1	0.507	Completely ignored Class 0, not suitable for imbalanced data

Experimenting Different Balancing methods on XGBoost



	perc_premium_paid_by_cash_credit	age_in_days	Income	Count_3-6_months_late	Count_6-12_months_late
count	79853.000000	79853.000000	7.985300e+04	79853.000000	79853.000000
mean	0.314288	18846.696906	2.088472e+05	0.248369	0.078093
std	0.334915	5208.719136	4.965826e+05	0.691102	0.436251
min	0.000000	7670.000000	2.403000e+04	0.000000	0.000000
25%	0.034000	14974.000000	1.080100e+05	0.000000	0.000000
50%	0.167000	18625.000000	1.665600e+05	0.000000	0.000000
75%	0.538000	22636.000000	2.520900e+05	0.000000	0.000000
max	1.000000	37602.000000	9.026260e+07	13.000000	17.000000

	Count_more_than_12_months_late	application_underwriting_score	no_of_premiums_paid	premium
count	79853.000000	79853.000000	79853.000000	79853.000000
mean	0.059935	99.080903	10.863887	10924.507533
std	0.311840	0.732573	5.170687	9401.676542
min	0.000000	91.900000	2.000000	1200.000000
25%	0.000000	98.830000	7.000000	5400.000000
50%	0.000000	99.220000	10.000000	7500.000000
75%	0.000000	99.550000	14.000000	13800.000000
max	11.000000	99.890000	60.000000	60000.000000

Modeling Considerations

Outliers in features (Premium, Income, Age, Tenure, Underwriting Score) were found to be **valid business cases** → retained

Statistical models (e.g., Logistic Regression) may underperform due to sensitivity to outliers and linear assumptions

Tree-based ensemble models (Random Forest, XGBoost, Gradient Boosting) are better suited → can handle outliers, class imbalance, and non-linear feature interactions

We still implemented **multiple models** for benchmarking and fairness, but tree-based approaches are **expected to perform best**

Q&A