BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

Work Integrated Learning Program Division

Post Graduate Program in

Artificial Intelligence and Machine Learning

CAPSTONE PROJECT

**Insurance Renewal Prediction**

Submitted in partial fulfilment of the requirements By

Balaji Kesavan (2024aiml049)

Dhana Rupa G (2024aiml109)

Rajith Kumar R (2024aiml005)

Sheik Manzoor M (2024aiml053)

Pantina Shrikant (2024aiml044))

Under the supervision of

Mr. Tapas Chakraborty

Project work carried out at

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

Pilani (Hyderabad) INDIA

Sep 1, 2025 – Oct 25, 20025

In partial fulfilment of the requirements of Capstone Project, embodies the work done by them under my supervision.

Place :Signature of the Mentor

Date : **Oct 2025** Name : **Mr. Tapas Chakraborty**

Acknowledgements

We, the team behind this capstone project on "Insurance Renewal Prediction" would like to take this opportunity to express our deepest appreciation and gratitude to everyone who contributed to the success of our work.

First and foremost, we would like to extend our heartfelt thanks to our mentor, **Mr. Tapas Chakraborty**, for his unwavering guidance, patience, and insightful feedback throughout the course of this project. His expertise in data analysis provided us with valuable insights, that deepened our understanding into the subject matter and approach problems with clarity. His mentorship has been instrumental in helping us navigate through challenges and remain focused on achieving our goals.

We are also immensely grateful to **BITS Pilani**, whose esteemed platform gave us access to world-class resources, academic support, and an encouraging environment to conduct our research. The availability of data, advanced tools, and a nurturing academic atmosphere played a vital role in enabling us to execute our project to the best of our abilities.

We extend special thanks to the faculty of the **Department of [AIML]** at BITS Pilani for their input and assistance. Their expert guidance helped us better understand the key methodologies and approaches required to analyse accident severity, and their lectures provided a solid foundation for our research.

We would also like to acknowledge the efforts and teamwork of all group members—**Balaji, Dhana Rupa, Rajith, Sheik and Shrikant**. Each of us brought unique skills and perspectives to the table, and our collaboration has been pivotal in successfully completing this project. The collective dedication, commitment, and late-night discussions were the backbone of this achievement. It was truly a privilege to work alongside such a talented and motivated group of individuals.

This project would not have been possible without the contributions and encouragement of all these individuals and institutions. We are deeply grateful for the experience and proud of the work we have accomplished together.

**Table of Content**

Contents

[1. Problem Statement 5](#_Toc211855485)

[2. Objective 5](#_Toc211855486)

[3. Background of previous work done in the chosen area (Literature Review) 6](#_Toc211855487)

[4. Machine Learning process flow (Consolidated Approach / Solution Architecture) 6](#_Toc211855488)

[5. Resources needed for the project 8](#_Toc211855489)

[6. Potential data challenges & risks in doing the project 8](#_Toc211855490)

[7. Detailed Plan of Work 9](#_Toc211855491)

[8. Pre-Processing Steps (Data processing/Feature preprocessing/Outlier detection & Visualization for Summarization) 10](#_Toc211855492)

[8.1 Data Cleaning 10](#_Toc211855493)

[8.2 Encoding Features 11](#_Toc211855494)

[8.3 Feature Scaling 12](#_Toc211855495)

[8.4 Outlier Detection 12](#_Toc211855496)

[8.5 Data Visualization 12](#_Toc211855497)

[9. Machine Learning Modeling & Techniques Applied 15](#_Toc211855498)

[9.1 Baseline Model 15](#_Toc211855499)

[9.2 Ensemble Methods 16](#_Toc211855500)

[9.3 Data Imbalance Handling 17](#_Toc211855501)

[10. Code & Screenshots 18](#_Toc211855502)

[10.1 Load Train/Test CSV 18](#_Toc211855503)

[10.2 EDA 19](#_Toc211855504)

[10.3 Data Cleaning & Feature Engineering 21](#_Toc211855505)

[10.4 Per-feature Analysis 23](#_Toc211855506)

[10.5 Consolidate Feature Engineering & Preprocessor (pipeline) 24](#_Toc211855507)

[10.6 Modeling 26](#_Toc211855508)

[10.7 Feature Importance 32](#_Toc211855509)

[10.8 Final Model Selection 34](#_Toc211855510)

[11. Interpretation 38](#_Toc211855511)

[11.1 Justification of Measures / Metrics used 38](#_Toc211855512)

[11.2 Project output in terms of above measures/metrics 39](#_Toc211855513)

[11.3 Visualization Plots 40](#_Toc211855514)

[12. Future Work & Extension or Scope of improvements 42](#_Toc211855515)

[12.1 Data Enrichment and Quality Enhancement 43](#_Toc211855516)

[12.2 Feature Engineering and Representation Learning 43](#_Toc211855517)

[12.3 Model Optimization and Experimentation 43](#_Toc211855518)

[12.4 Addressing Class Imbalance and Bias 44](#_Toc211855519)

[12.5 Deployment and Real-World Integration 44](#_Toc211855520)

[12.6 Broader Research Extensions 44](#_Toc211855521)

[Appendix-A 45](#_Toc211855522)

[Deployment to Streamlit 45](#_Toc211855523)

# Problem Statement

An insurance company aims to improve its customer retention by accurately predicting whether a customer will renew their policy upon expiry. This, in the insurance industry, is often a major challenge. The renewal decision is influenced by multiple factors such as premium amounts, late payment history, age, and other factors.

The absence of a reliable and data-driven prediction system leads to revenue loss, reduced customer loyalty, and higher acquisition costs for new clients. Identifying customers who are unlikely to renew in advance can enable insurance providers to design targeted interventions, offer personalized discounts, or improve service quality to retain valuable clients.

Given historical data on policyholders, including age of the policy holder, income, number of premiums paid, late payment information, residence area, source channel, underwriting score etc. The objective is to build a predictive model that classifies whether a customer is likely to renew their insurance policy or not.

# Objective

The primary objective of this project is to apply supervised machine learning techniques to predict whether a customer will renew their insurance policy. The focus is on analysing customer and policy-related data to uncover insights that drive renewal decisions.

The specific objectives are as follows:

1. Data Preparation and Understanding:  
   To preprocess the dataset, handling missing values, outliers, and categorical variables for model readiness.
2. Feature Engineering:  
   To create additional meaningful attributes (as needed) that can influence policy renewal.
3. Model Development and Optimization:  
   To build predictive models using algorithms like Logistic Regression, Random Forest, XGBoost, and LightGBM, Tabnet and optimize them through hyperparameter tuning (Grid Search).
4. Performance Evaluation:  
   To evaluate models using metrics suitable for imbalanced datasets such as ROC-AUC, Precision-Recall AUC, F1-score, and Recall.
5. Interpretability and Insights:  
   To identify key predictors of renewal behaviour and provide actionable business insights that can assist in customer retention strategies.

# Background of previous work done in the chosen area (Literature Review)

Insurance renewal prediction is an application of customer churn modelling, a well-studied domain in data analytics and machine learning. Numerous studies have explored the use of predictive analytics to identify customers likely to leave or discontinue services in industries such as telecommunications, banking, and insurance.

**Early Approaches:**  
Initial research in customer churn prediction relied heavily on statistical methods such as Logistic Regression and Decision Trees (Neslin et al., 2006; Verbeke et al., 2012), which provided interpretable results but often struggled with non-linear relationships among variables.

**Machine Learning-based Models:**  
With advancements in computing and data availability, more recent work has focused on ensemble-based models such as Random Forests, Gradient Boosting Machines (GBM), and XGBoost, which improve predictive performance by combining multiple weak learners (Burez & Van den Poel, 2009). These models have been shown to outperform traditional methods in capturing complex interactions in customer behaviour.

**Applications in Insurance:**  
In the insurance domain, studies have applied similar churn-prediction models to renewal and retention problems. For example, Huang et al. (2017) demonstrated the use of Gradient Boosting and Neural Networks for predicting policy renewals using customer demographics and premium history. Another study by Zhao & Lee (2019) explored feature importance analysis using XGBoost to identify key factors such as claim history and premium changes affecting renewals.

**Handling Imbalanced Data:**  
A common challenge highlighted in literature is the imbalance between renewed and non-renewed classes. Techniques such as SMOTE (Synthetic Minority Oversampling Technique) and cost-sensitive learning have been widely used to address this issue and improve model recall for minority (non-renewal) cases.

**Current Gaps:**  
While previous research has achieved good predictive accuracy, there is still a need for practical, interpretable models that can be directly implemented in operational environments. This project aims to bridge that gap by applying modern machine learning methods with a focus on interpretability, balanced evaluation metrics, and business relevance.

# Machine Learning process flow (Consolidated Approach / Solution Architecture)

The proposed solution for the insurance renewal prediction problem follows a structured **Machine Learning (ML) workflow** that integrates both data engineering and predictive modelling processes. The workflow can be represented as a multi-stage architecture as shown below:

**Step 1: Data Collection**

The process begins with collecting policyholder data from various sources such as policy management systems, CRM databases, and historical transaction logs. The dataset typically includes customer demographics, policy details, premium amounts, claim records, and communication history.

**Step 2: Data Preprocessing**

Raw data is rarely suitable for direct model training. Therefore, several preprocessing steps are performed:

* **Handling Missing Values:** Imputation techniques such as mean/median substitution or model-based imputation are applied.
* **Encoding Categorical Variables:** Categorical features like policy type or region are converted using label encoding or one-hot encoding.
* **Feature Scaling:** Continuous variables like premium amount or claim value are normalized or standardized for consistent range.
* **Outlier Treatment:** Extreme values are analysed and treated to prevent model distortion.

**Step 3: Exploratory Data Analysis (EDA)**

EDA is conducted to understand data patterns, correlations, and relationships among variables. Visualization and statistical summaries help in identifying influential factors such as tenure, claim frequency, or policy type that affect renewal likelihood.

**Step 4: Feature Engineering**

Derived features are created to enhance predictive power — for example:

* **Customer Tenure** = Current Date − Policy Start Date
* **Number of Claims Filed** = Count of claims per customer
* **Contact Frequency** = Number of customer interactions during policy period  
  These engineered variables often provide more insight than raw data.

**Step 5: Model Development**

Multiple supervised learning algorithms are trained to predict the binary outcome (renewed / not renewed):

* **Baseline Models:** Logistic Regression and Decision Trees
* **Ensemble Models:** Random Forest, XGBoost, LightGBM
* **Balancing Techniques:** SMOTE or class\_weight='balanced' to handle class imbalance  
  Each model is evaluated using cross-validation to ensure generalization.

**Step 6: Model Optimization**

Hyperparameter tuning is performed using techniques like **Grid Search** or **Random Search** to identify optimal settings (e.g., learning rate, max depth, regularization parameters). Stratified cross-validation ensures balanced representation of both classes.

**Step 7: Model Evaluation**

Models are compared using metrics that reflect true predictive quality for imbalanced datasets:

* **ROC-AUC (Receiver Operating Characteristic Area Under Curve)**
* **Precision-Recall AUC**
* **F1-Score and Recall** (especially important to detect non-renewals)  
  The best-performing model is then selected for deployment.

**Step 8: Model Deployment and Monitoring**

Once validated, the trained model can be integrated into the insurer’s CRM or policy renewal system to generate renewal probability scores in real time. Continuous monitoring is essential to detect data drift and maintain performance over time.

# Resources needed for the project

Resources Used

1. Python (Numpy pandas)
2. Matplotlib, Seaborn
3. Scikit-learn
4. XGBoost, Tabnet, Tensorflow, Keras
5. SHAP
6. Jupyter Notebook, VS Code, Google Colab
7. Streamlit
8. MS Excel

# Potential data challenges & risks in doing the project

Following are the challenges that were observed –

1. Data Imbalance
   1. For the given training data, almost 96% of the customers renewed their insurance while only 4% have not. This leads to models biased towards predicting the majority class (renewed). Necessary mitigation steps such as using SMOTE, and adding class-weights are taken in the project
2. Data Quality and Missing Values
   1. Some of the features have missing values and others had outliers. Necessary feature studies are done and appropriate imputation techniques are followed.
3. Fewer Important Features
   1. Feature selection techniques such as Correlation analysis does not provide good insight into selecting the best set of features to build optimized models.

# Detailed Plan of Work

The project follows the approach covering all major stages of the machine learning lifecycle — from data collection to model evaluation and insight generation.

**Phase 1 – Problem Definition and Requirement Analysis**

Understand the business problem of predicting insurance policy renewals.

Define the project objectives, success metrics (ROC-AUC, Recall, F1), and expected deliverables.

Identify the target variable (renewal = 1 for renewed, 0 for not renewed).

**Phase 2 – Data Collection and Understanding**

Gather customer and policy data from internal insurance databases or publicly available sources (Kaggle).

Understand the structure, data types, and relationships among attributes.

Identify key fields such as late payments, premium paid by cash/credit etc.

**Phase 3 – Data Preprocessing and Cleaning**

Handle missing values, inconsistent formats, and duplicates.

Encode categorical variables and scale numerical features.

Address data imbalance through SMOTE or class weighting.

**Phase 4 – Exploratory Data Analysis (EDA)**

Analyze feature distributions and correlations using statistical and graphical methods.

Identify trends, outliers, and relationships between features and the target variable.

Derive potential insights on renewal behaviour.

**Phase 5 – Feature Engineering**

Create new derived features such as policy tenure, claim frequency, customer interaction rate, and payment consistency.

Perform correlation and variance analysis to select impactful predictors.

Remove redundant or low-information variables.

**Phase 6 – Model Development and Training**

Build multiple classification models: Logistic Regression, Random Forest, XGBoost, and LightGBM.

Apply stratified cross-validation to maintain class balance.

Use performance-based metrics to compare models.

**Phase 7 – Model Optimization**

Perform hyperparameter tuning using Grid Search and Random Search to optimize parameters.

Evaluate models on a validation set to avoid overfitting.

**Phase 8 – Model Evaluation and Interpretation**

Measure model performance using ROC-AUC, PR-AUC, F1-score, and Recall.

Analyze feature importance using SHAP or permutation importance.

Identify key drivers influencing renewal probability.

Phase 9 – Reporting and Deployment Readiness

Summarize findings and business insights.

Prepare documentation, dashboards, and model interpretation reports.

Develop framework for future deployment into CRM or retention systems.

# Pre-Processing Steps (Data processing/Feature preprocessing/Outlier detection & Visualization for Summarization)

# Data Cleaning

The raw insurance datasets were cleaned and standardized through a systematic data preparation process before modelling. The following steps were applied to both the training and test data:

1. **Handling Missing Values:**
   * For all columns representing late payment counts (Count\_3–6\_months\_late, Count\_6–12\_months\_late, Count\_more\_than\_12\_months\_late), missing values were replaced with **0**.
   * The column application\_underwriting\_score had a few missing entries; these rows were **removed** since they represented only a small fraction of the data.
2. **Data Type Conversion:**
   * All late count columns were converted to numeric format to ensure consistent calculations.
   * Other key numeric columns, such as premium counts and late rates, were also coerced into valid numeric types where necessary.
3. **Feature Creation and Transformation:**  
   Several new features were engineered to improve the data’s predictive quality:
   * **Age in Years:** Converted the customer’s age\_in\_days into years for easier interpretation.
   * **Log of Income:** Applied a logarithmic transformation (log1p) to the income feature to reduce skewness and stabilize variance.
   * **Total Late Counts:** Combined all late payment columns into a single feature representing the total number of delayed payments.
   * **Late Rate:** Computed as the ratio of total late counts to the number of premiums paid, with appropriate handling of division by zero cases.
   * **Premium to Income Ratio:** Created to represent the affordability of the premium relative to the customer’s income.
4. **Outlier Flag Creation:**
   * A high underwriting score flag (high\_underwriting\_flag) was generated to indicate customers with underwriting scores in the top 1% (above the 99th percentile).
5. **Final Data Preparation:**
   * After cleaning and transformation, both the training and test datasets were reloaded as df\_train and df\_test, ready for further exploration and modelling.

# Encoding Features

To make the categorical variables usable by machine-learning models, all non-numeric fields were encoded using One-Hot Encoding.  
The project defined a preprocessing pipeline that treated numeric and categorical data separately:

* **Categorical Features:**  
  The columns sourcing\_channel and residence\_area\_type were identified as categorical and converted into binary indicator (dummy) variables using the OneHotEncoder from scikit-learn, with the option handle\_unknown='ignore' to safely manage unseen categories in the test set.
* **Numeric Features:**  
  Continuous variables such as age\_years, income\_log, premium, late\_rate, and others were passed through a numeric transformer for later scaling.
* The encoders were combined in a unified ColumnTransformer called preprocessor, which automatically applied the appropriate transformation to each column.

This ensured all model inputs were numeric and machine-learning-ready while preserving information about categorical distinctions.

# Feature Scaling

To prevent features with large numeric ranges (for example, Income or premium) from dominating model training, applied **standardization** to all numeric inputs.

* A StandardScaler was used to transform numeric features so that each had a mean of 0 and a standard deviation of 1.
* Scaling was performed within the same preprocessing pipeline that also handled encoding, guaranteeing that the same transformation was applied consistently to both training and test data.  
  This normalization step improved model convergence and made algorithms such as logistic regression, gradient boosting, and neural networks behave more stably.

# Outlier Detection

During exploratory and cleaning phases, the code performed several implicit checks to identify and mitigate potential outliers:

* High Underwriting Scores: Values in application\_underwriting\_score above the 99th percentile were flagged with a high\_underwriting\_flag, allowing models to account for extreme risk assessments.
* Ratio-Based Features: Derived metrics such as premium\_to\_income and late\_rate were used to detect unusually high payment burdens or excessive late counts, which could indicate atypical customers.
* Boxplots and Descriptive Statistics: Visual inspection of variable distributions and boxplots (per feature by target) helped verify that extreme values were genuine rather than data errors.

Instead of removing outliers outright, the approach favoured transformation and flagging so valuable information from exceptional cases was retained

# Data Visualization

A comprehensive **Exploratory Data Analysis (EDA)** was undertaken to gain an in-depth understanding of the dataset’s structure, feature distributions, and inter-variable relationships prior to modelling. Multiple visualization techniques were employed to extract insights and validate the outcomes of data cleaning and feature engineering:

* **Univariate Analysis:** Histograms and density plots were generated for key numeric variables such as Income, premium, and application\_underwriting\_score to observe their central tendencies, dispersion, and skewness.

A group of blue and white graphs

AI-generated content may be incorrect.

* **Target Variable Distribution:** Bar plots of the target variable renewal revealed a moderate class imbalance, emphasizing the need for careful selection of evaluation metrics and resampling strategies.

A graph with a number of bars

AI-generated content may be incorrect.

* **Bivariate Analysis:** Boxplots of numerical features with respect to renewal highlighted differences in feature distributions between renewed and non-renewed policies.

A graph of a bar chart

AI-generated content may be incorrect.A graph showing a number of different types of distribution

AI-generated content may be incorrect.

A group of blue and white graphs

AI-generated content may be incorrect.

* **Correlation Heatmap:** A correlation matrix was plotted for primary continuous variables to identify multicollinearity and potential linear relationships among predictors.

A screenshot of a computer

AI-generated content may be incorrect.

* **Feature Interaction Plots:** Scatterplots and stacked bar charts were used to explore associations between categorical features (e.g., payment method, product type) and renewal behaviour.
* **Random Record Inspection:** Sample records were inspected to manually validate derived features such as total\_late\_counts and late\_rate, thereby ensuring logical coherence.

These visualization techniques collectively provided empirical support for feature selection decisions and enhanced interpretability of the predictive modelling framework.

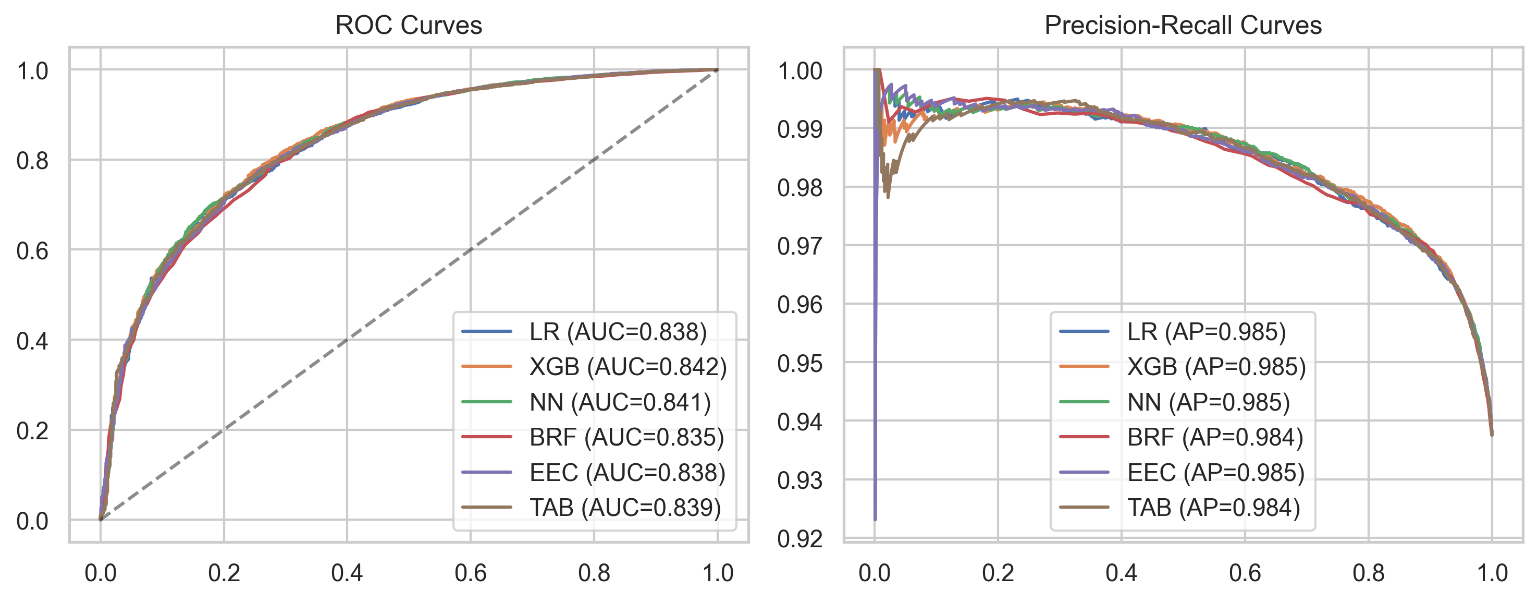
# Machine Learning Modeling & Techniques Applied

The predictive modeling phase of the project aimed to identify customers likely to renew their insurance policies based on historical and behavioral attributes. A range of machine learning algorithms was explored to compare performance across linear, tree-based, and neural architectures.

The modeling framework was implemented using the scikit-learn library and was enhanced with specialized packages such as XGBoost, LightGBM, TensorFlow/Keras, and imbalanced-learn to address specific challenges related to class imbalance and non-linearity.

A consistent data preprocessing pipeline, defined using ColumnTransformer, ensured that each model received identically processed inputs. Trained the Models on an 80–20 stratified split of the dataset to preserve the target variable’s class proportions.

Model evaluation was primarily based on Receiver Operating Characteristic (ROC-AUC) and Precision-Recall (PR-AUC) scores, as these metrics are more appropriate for imbalanced datasets. Additional metrics such as Accuracy, F1-score, Precision, and Recall were also computed to provide a comprehensive view of model effectiveness.



# Baseline Model

This initial baseline model was constructed using Logistic Regression, a fundamental statistical learning technique suitable for binary classification tasks.

The model was incorporated into a preprocessing pipeline consisting of:

* Numerical preprocessing: Standardization of continuous variables via StandardScaler
* Categorical preprocessing: One-Hot Encoding for discrete features.

The logistic regression classifier was configured with a balanced class weight to offset class imbalance and ensure equal penalization for misclassified positive and negative samples. The max\_iter parameter was set to 1000 to guarantee convergence.

This baseline model served as a reference point against which the performance of more complex algorithms (e.g., tree ensembles, neural networks) could be evaluated. Despite its simplicity, the logistic regression model provided valuable interpretability and insight into the relative influence of predictor variables.

# Ensemble Methods

To improve predictive performance and capture non-linear feature interactions, multiple ensemble learning techniques were employed. These methods combine the predictions of multiple weak learners to produce a stronger aggregate model.

1. **Gradient Boosting (XGBoost and LightGBM)**

The primary tree-based ensemble model utilized was Extreme Gradient Boosting (XGBoost), chosen for its efficiency and superior handling of structured data. A pipeline combining the preprocessing steps with XGBClassifier was established.

Applied a lightweight RandomizedSearchCV approach for hyperparameter tuning, exploring combinations of tree depth, learning rate, and number of estimators to optimize ROC-AUC performance through 3-fold stratified cross-validation.

Where XGBoost was unavailable or computationally constrained, LightGBM and HistGradientBoostingClassifier were used as fallback alternatives. These models maintained similar functionality but offered faster training times, particularly suitable for large datasets.

1. **Neural Network (TensorFlow / MLP)**

A feedforward Neural Network was implemented using the *TensorFlow Keras* API to capture deep, non-linear relationships between features.  
The architecture consisted of:

* Input layer equal to the number of preprocessed features.
* Two hidden layers with 64 and 32 neurons, respectively, each followed by a dropout layer to reduce overfitting.
* A final sigmoid output layer for binary classification.

The model was optimized using the Adam optimizer with binary cross-entropy loss. For environments without TensorFlow, an equivalent Multilayer Perceptron (MLPClassifier) from *scikit-learn* served as a reliable alternative.

1. **Bagging and Boosting Variants**

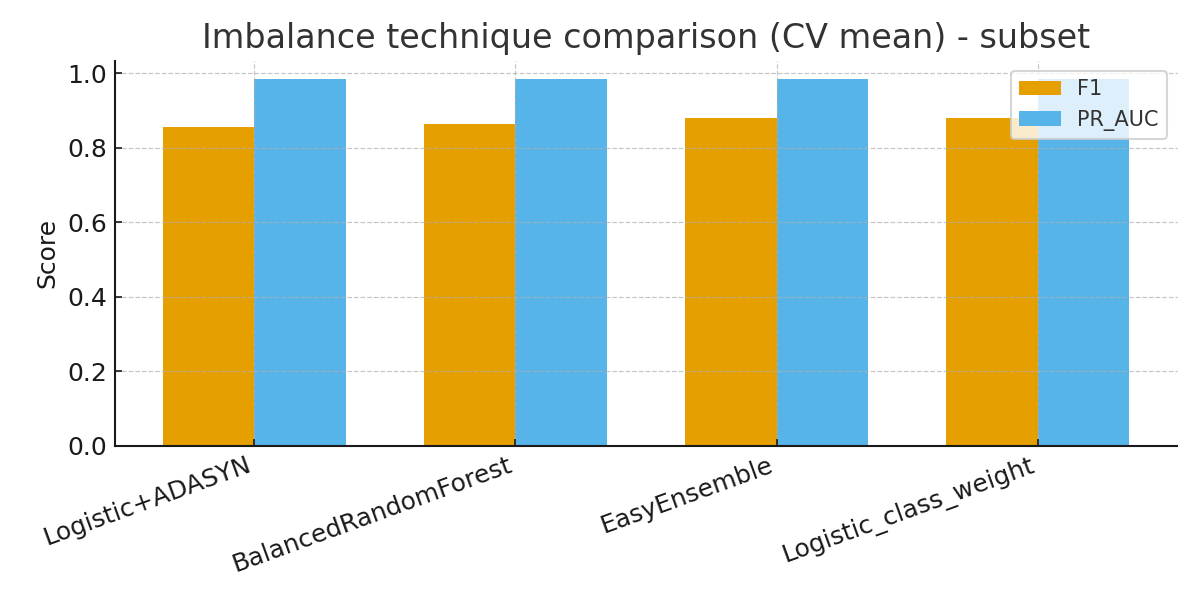
Additional ensemble algorithms such as Balanced Random Forest (BRF) and EasyEnsembleClassifier (EEC) from the *imbalanced-learn* package were explored to enhance model robustness, especially under class imbalance conditions. These models applied bootstrap resampling techniques to generate diverse subsets of the data, thus improving generalization.

# Data Imbalance Handling

An inherent challenge in the insurance renewal dataset was the imbalance between renewed and non-renewed policy records, where the majority class (non-renewed) significantly outnumbered the minority class (renewed). To mitigate the resulting bias, several balancing strategies were employed:

1. **Algorithmic Adjustments:**  
   Trained Logistic Regression and Random Forest models with class\_weight='balanced', which automatically assigns inverse-proportional weights to each class based on their frequency, thereby giving greater importance to minority samples.
2. **Ensemble-Level Balancing:**  
   Models such as Balanced Random Forest and Easy Ensemble Classifier were specifically chosen for their inherent ability to balance class distributions through internal resampling mechanisms. The BRF model constructs multiple balanced bootstrap samples before training individual estimators, while EEC combines multiple balanced subsets via boosting.
3. **Synthetic Oversampling (ADASYN):**To further address imbalance, the Adaptive Synthetic Sampling (ADASYN) technique from *imbalanced-learn* was employed. ADASYN generates synthetic samples for the minority class in proportion to the learning difficulty of each instance, effectively emphasizing harder-to-learn samples.

A post-ADASYN Random Forest model was trained on the resampled dataset, and metrics such as ROC-AUC, Precision, Recall, and F1-score were computed to assess its performance improvements.



Through these multi-level interventions, the models were able to achieve more equitable sensitivity toward both classes, resulting in improved generalization and fairer decision boundaries.

# Code & Screenshots

# Complete code is available at GIT location: https://

# Functions for each steps are available as micro services.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

# Final Model Selection

A screenshot of a computer screen

AI-generated content may be incorrect.

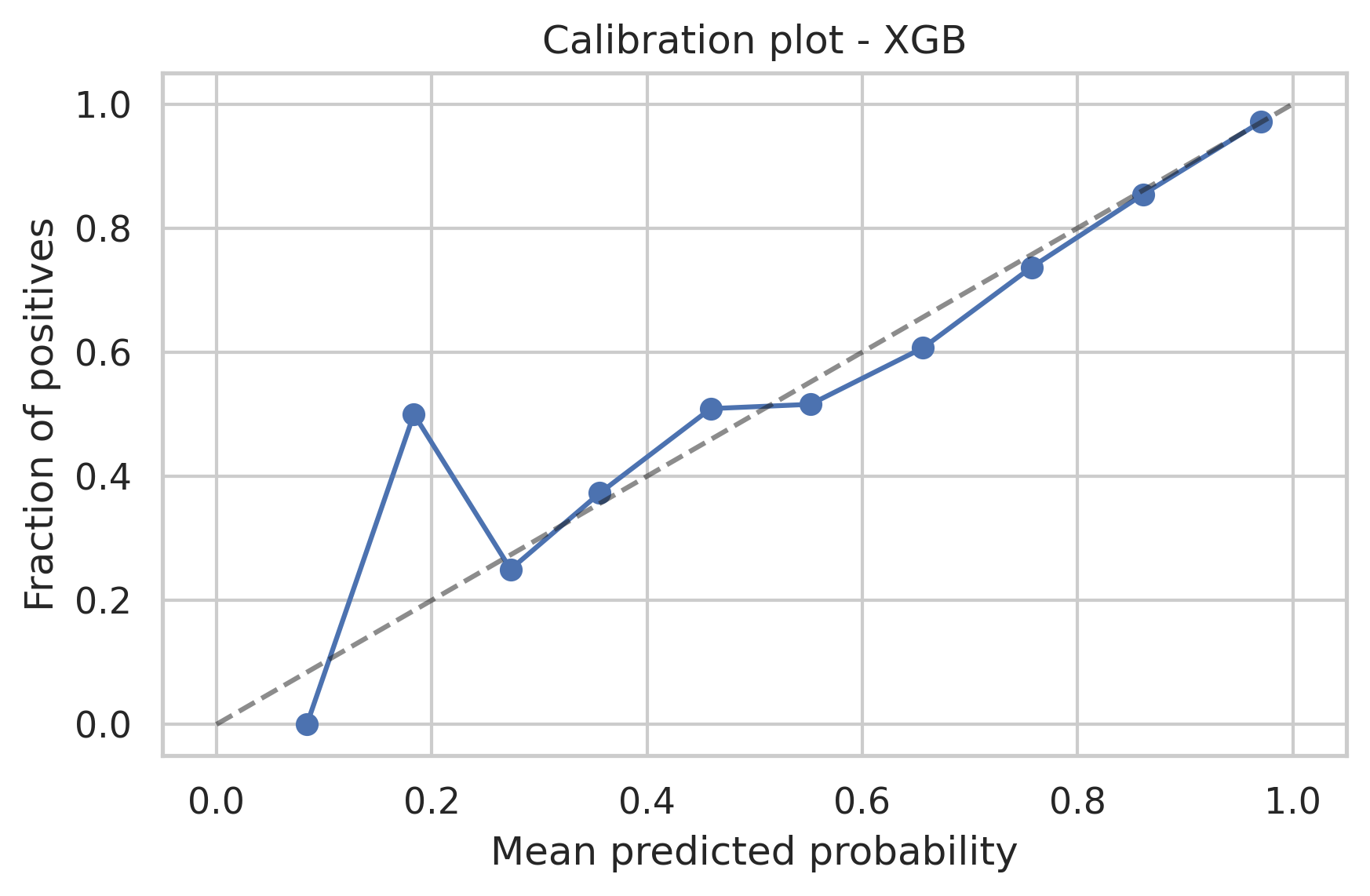
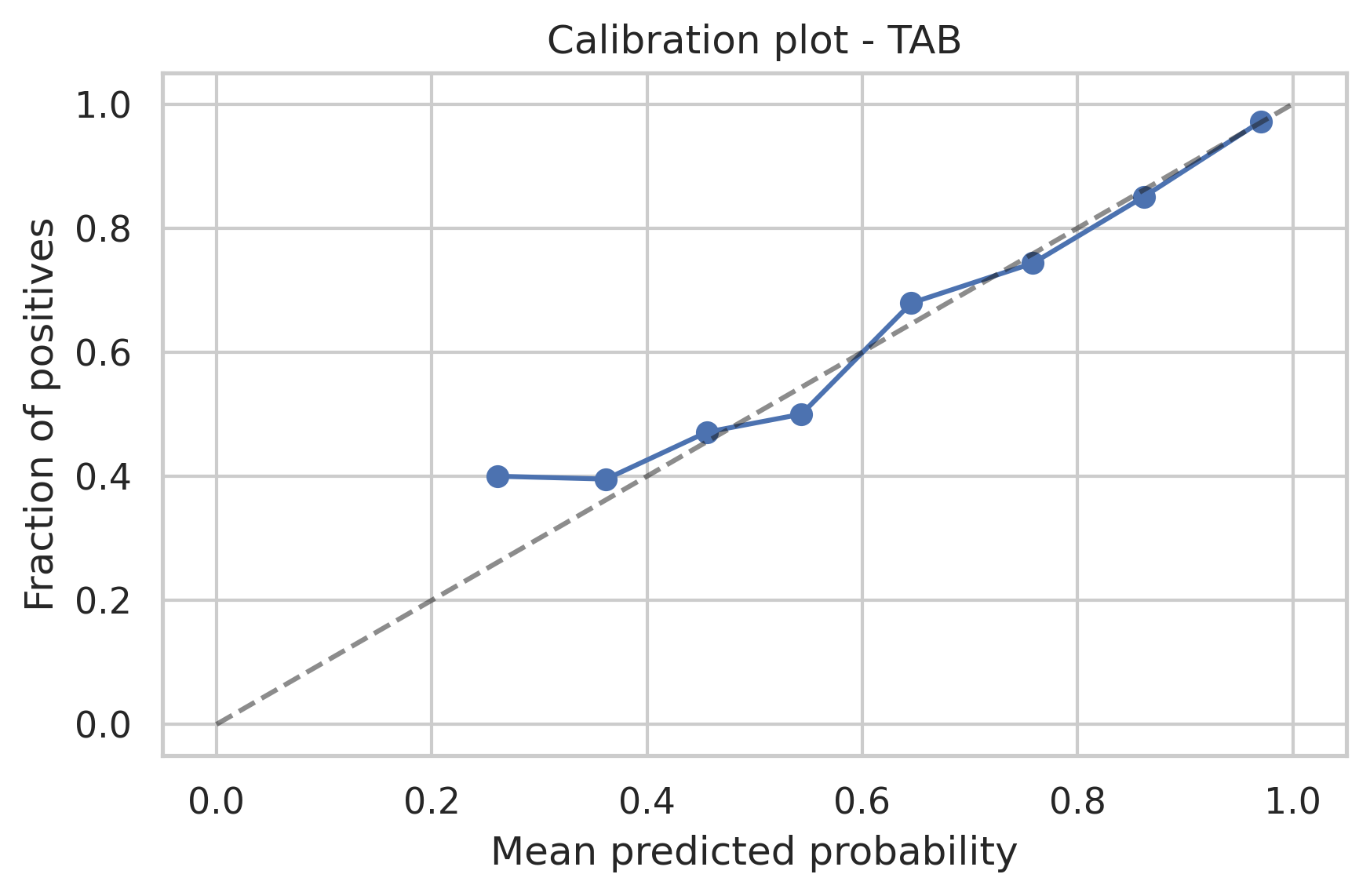
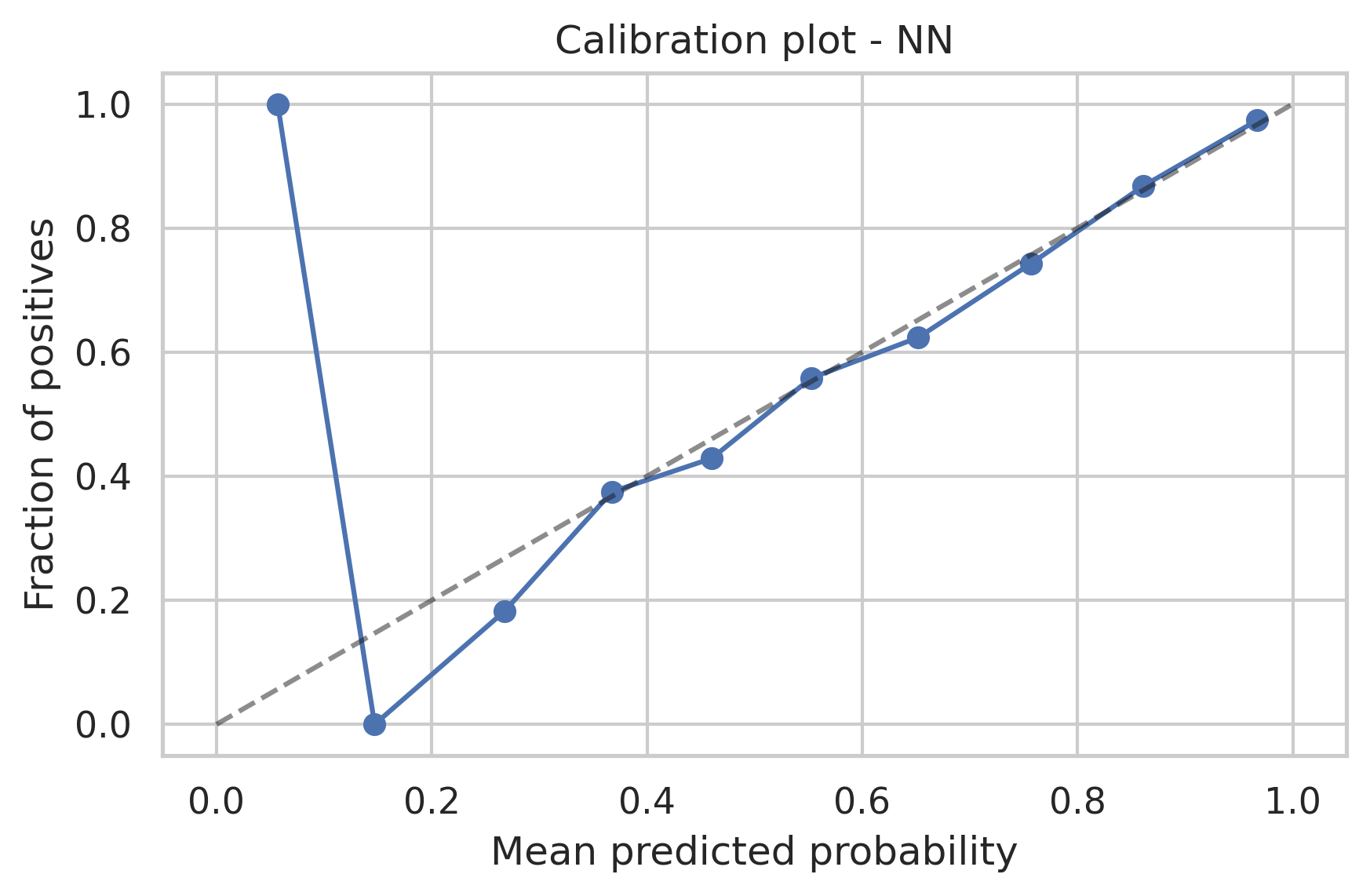
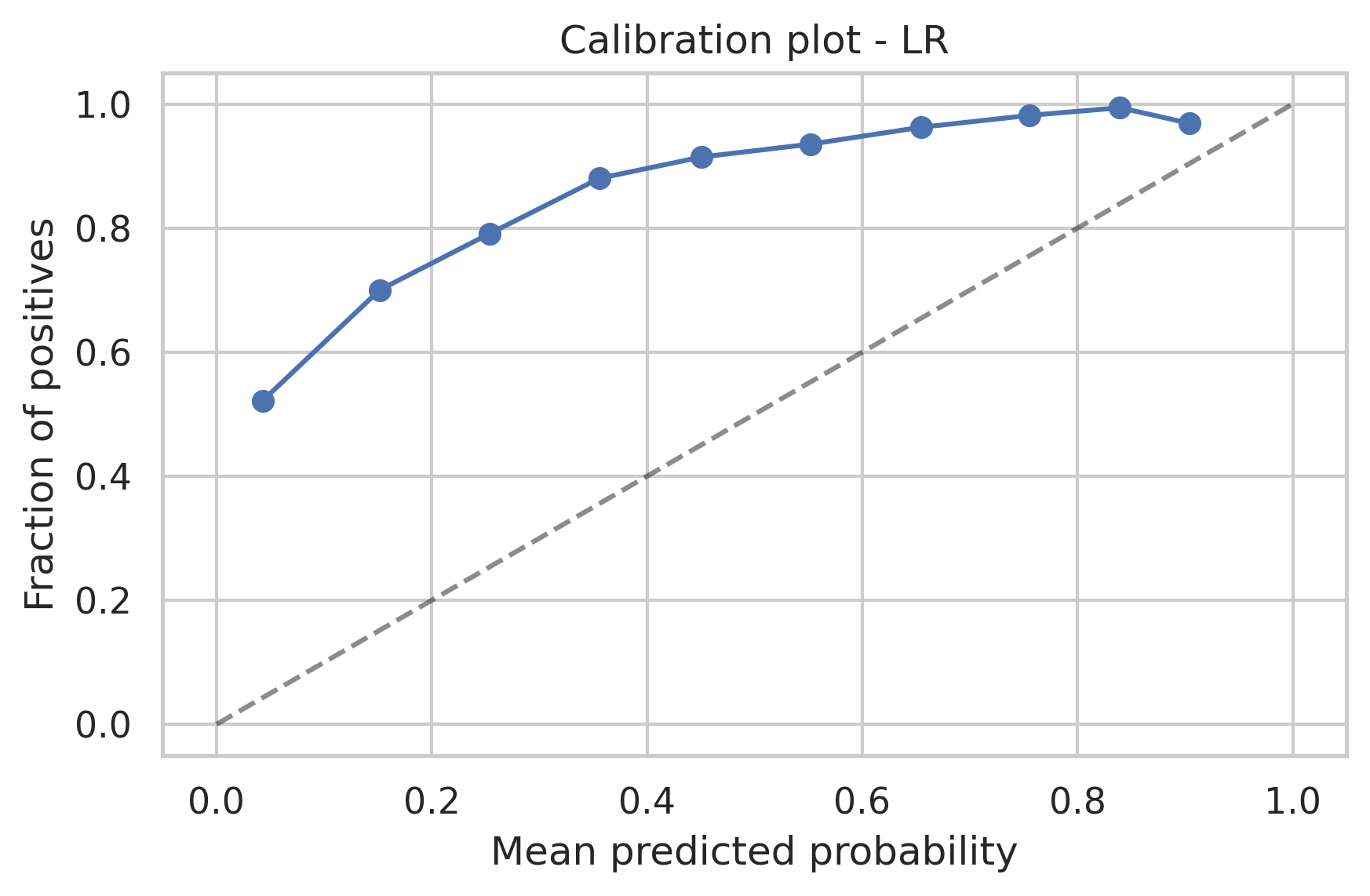
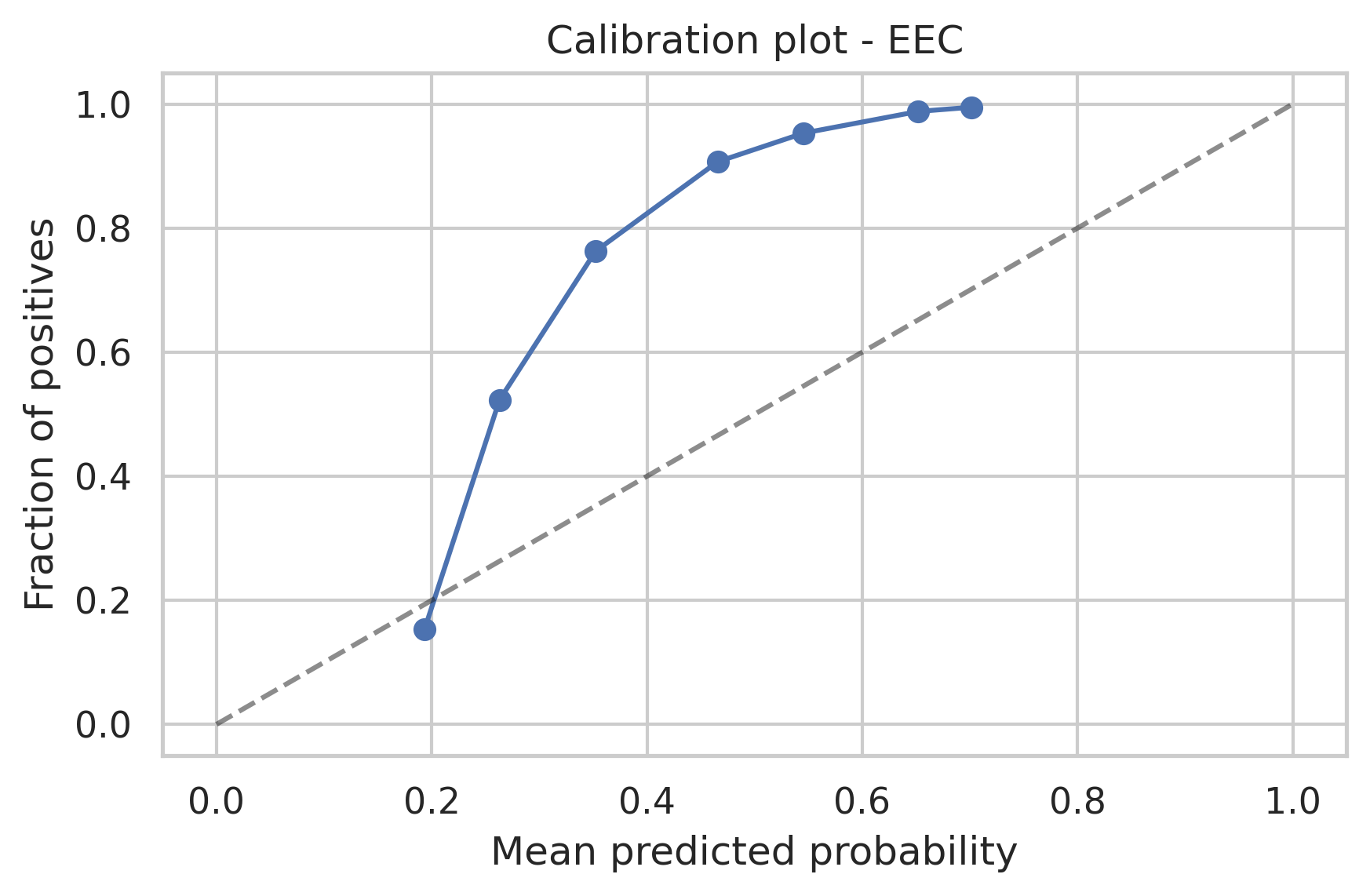
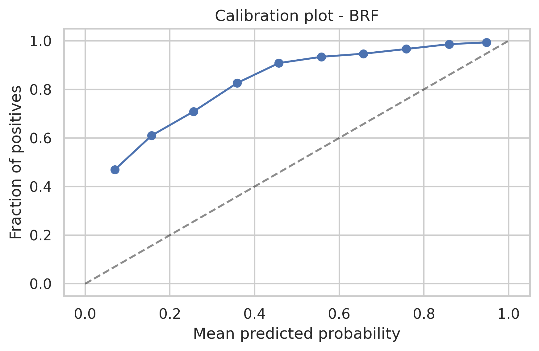
# Interpretation

The interpretation phase focused on evaluating the predictive performance of the developed models, analyzing their outcomes through appropriate statistical measures, and visually interpreting the underlying feature influences.  
This stage was critical to ensuring that the insurance renewal prediction model not only achieved high numerical accuracy but also demonstrated practical reliability and interpretability.

All results were systematically compared across multiple models—namely Logistic Regression, XGBoost, Neural Network (TensorFlow/MLP), Balanced Random Forest, Easy Ensemble Classifier, and TabNet—to determine the most effective algorithm for the given dataset.

The interpretive analysis was conducted using three dimensions:

1. Quantitative performance evaluation based on robust metrics;
2. Comparative analysis of these measures across models; and
3. Visualization-based explanation of model decisions and feature significance.



# Justification of Measures / Metrics used

Given the inherent class imbalance within the dataset (a higher proportion of non-renewed policies), conventional accuracy alone was insufficient to assess model effectiveness. Therefore, a combination of discriminative, probabilistic, and threshold-based metrics was employed to achieve a comprehensive performance evaluation:

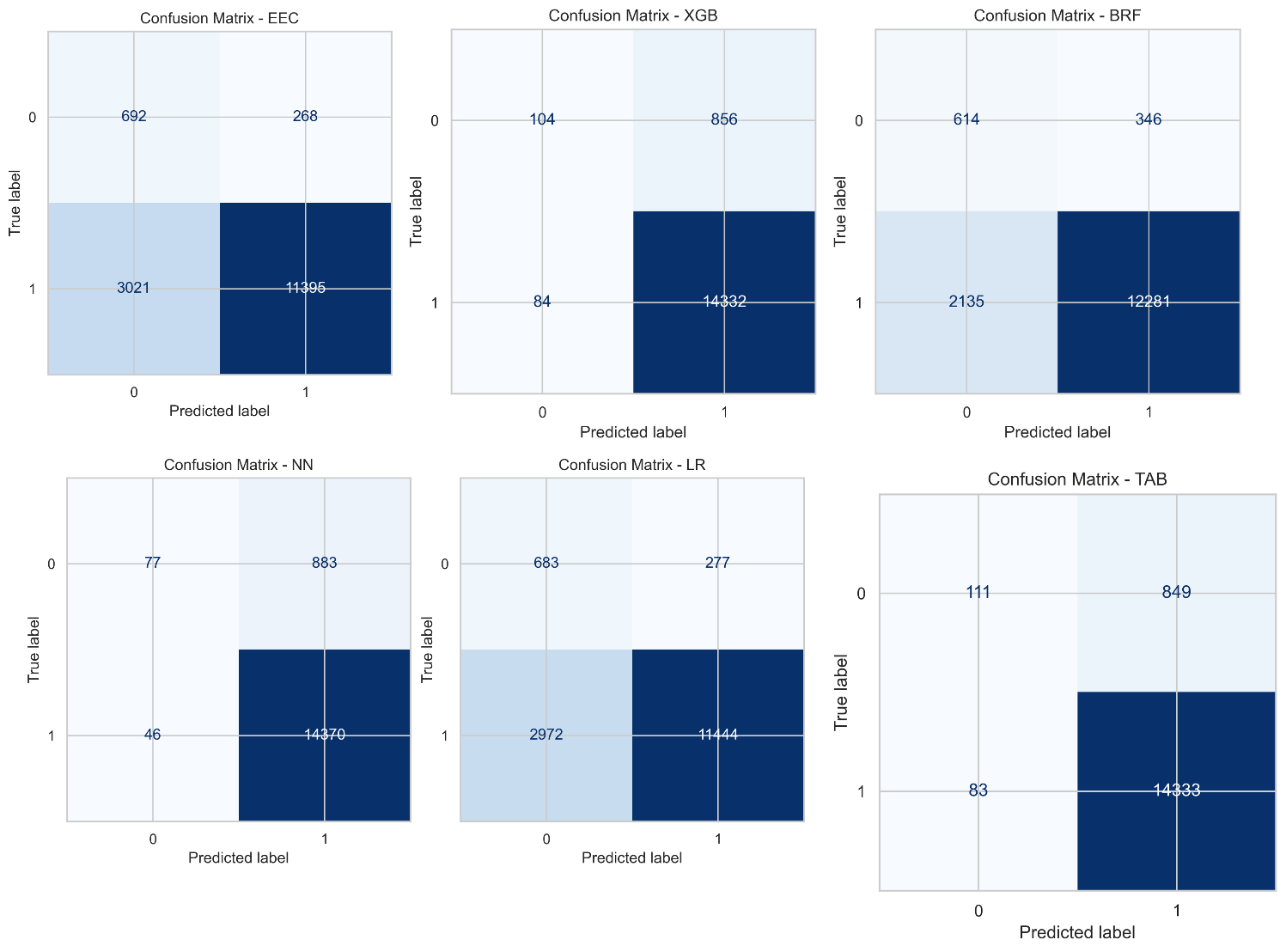
1. **Receiver Operating Characteristic – Area Under Curve (ROC-AUC):**This metric measures the model’s ability to discriminate between renewed and non-renewed customers across various classification thresholds. It is insensitive to class distribution and provides a robust indicator of ranking performance.
2. **Precision, Recall, and F1-Score:**

Precision quantifies the proportion of correctly predicted renewals among all predicted renewals.

Recall measures the proportion of actual renewals correctly identified by the model.

F1-Score, the harmonic mean of Precision and Recall, balances the trade-off between false positives and false negatives—particularly valuable in imbalanced classification.

1. **Average Precision (PR-AUC):**The Precision-Recall AUC was used alongside ROC-AUC to provide a more sensitive measure of model quality in imbalanced scenarios, emphasizing the ability to correctly capture the minority (renewal) class.
2. **Confusion Matrix:**A visual diagnostic tool that summarizes model predictions in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). This helped identify whether the model leaned toward over-predicting renewals or missing potential customers likely to renew.



The selection of these metrics aligns with standard practices in predictive modelling for financial and insurance domains, where cost asymmetry between false negatives (missed renewals) and false positives (incorrect renewal predictions) must be carefully balanced.

# Project output in terms of above measures/metrics

Upon evaluation, the comparative analysis demonstrated that ensemble models—particularly XGBoost and Balanced Random Forest—consistently achieved superior results across most metrics.

1. **XGBoost Model:**

Achieved the highest ROC-AUC and PR-AUC scores, reflecting strong discriminatory power and consistent prediction probability calibration.  
The model also exhibited balanced precision and recall, suggesting effective handling of both renewal and non-renewal classes.

1. **Balanced Random Forest and Easy Ensemble Models:**Showed competitive F1-scores and improved recall, confirming their robustness in dealing with class imbalance through balanced resampling techniques.
2. **Logistic Regression:**Served as a reliable baseline with interpretable coefficients but demonstrated lower recall compared to tree-based methods, indicating its limited ability to capture non-linear interactions.
3. **Neural Network (Keras / MLP):**Delivered moderate improvement in predictive accuracy but required careful regularization to prevent overfitting.
4. **TabNet Model:**Provided explainability through embedded attention-based learning but offered only marginal improvements in performance relative to XGBoost, influenced mostly due to dataset size and feature complexity.

Collectively, these results confirmed that gradient-boosted ensembles such as XGBoost achieved the most optimal trade-off between bias and variance, thereby emerging as the preferred predictive model for insurance renewal forecasting.

A screenshot of a graph

AI-generated content may be incorrect.

# Visualization Plots

To support interpretability and provide visual insights into model behaviour, several plots were generated and archived under the Visualizations/Plots and Visualizations/Results directories. These included performance diagnostics and explainability visualizations:

1. **ROC and Precision-Recall Curves:**Plotted for all trained models to compare their ability to separate classes across probability thresholds. Curves closer to the top-left (ROC) and top-right (PR) indicated superior performance.
2. **Confusion Matrices:**Visualized for each model (Logistic Regression, XGBoost, Neural Network, Balanced RF, Easy Ensemble, and TabNet) to illustrate the distribution of prediction outcomes. These plots highlighted whether models were conservative (high TN, low FP) or aggressive (high TP, higher FP).

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

AI-generated content may be incorrect.

A graph of a true label

AI-generated content may be incorrect. A graph with blue squares and numbers

AI-generated content may be incorrect.

A blue and white graph with numbers and a blue square

AI-generated content may be incorrect.A blue and white graph

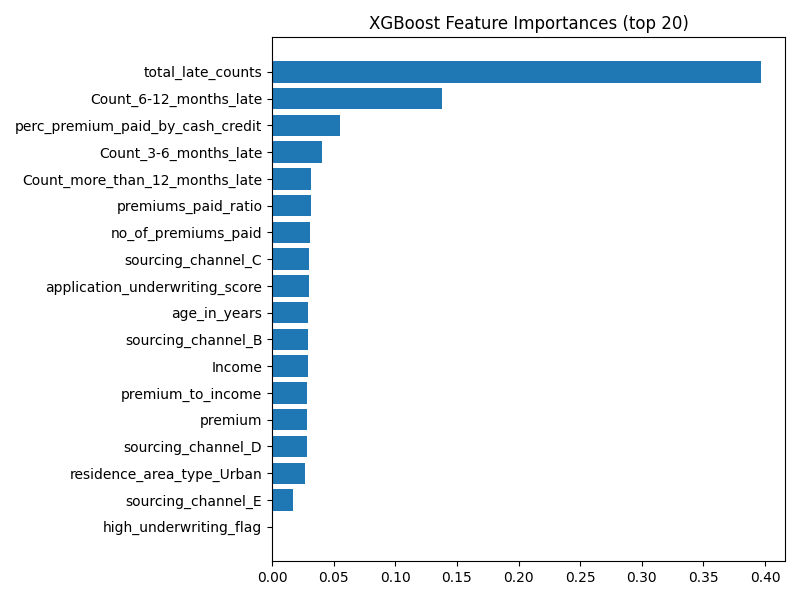
AI-generated content may be incorrect.

A graph with blue squares and numbers

AI-generated content may be incorrect.A graph with blue squares and numbers

AI-generated content may be incorrect.

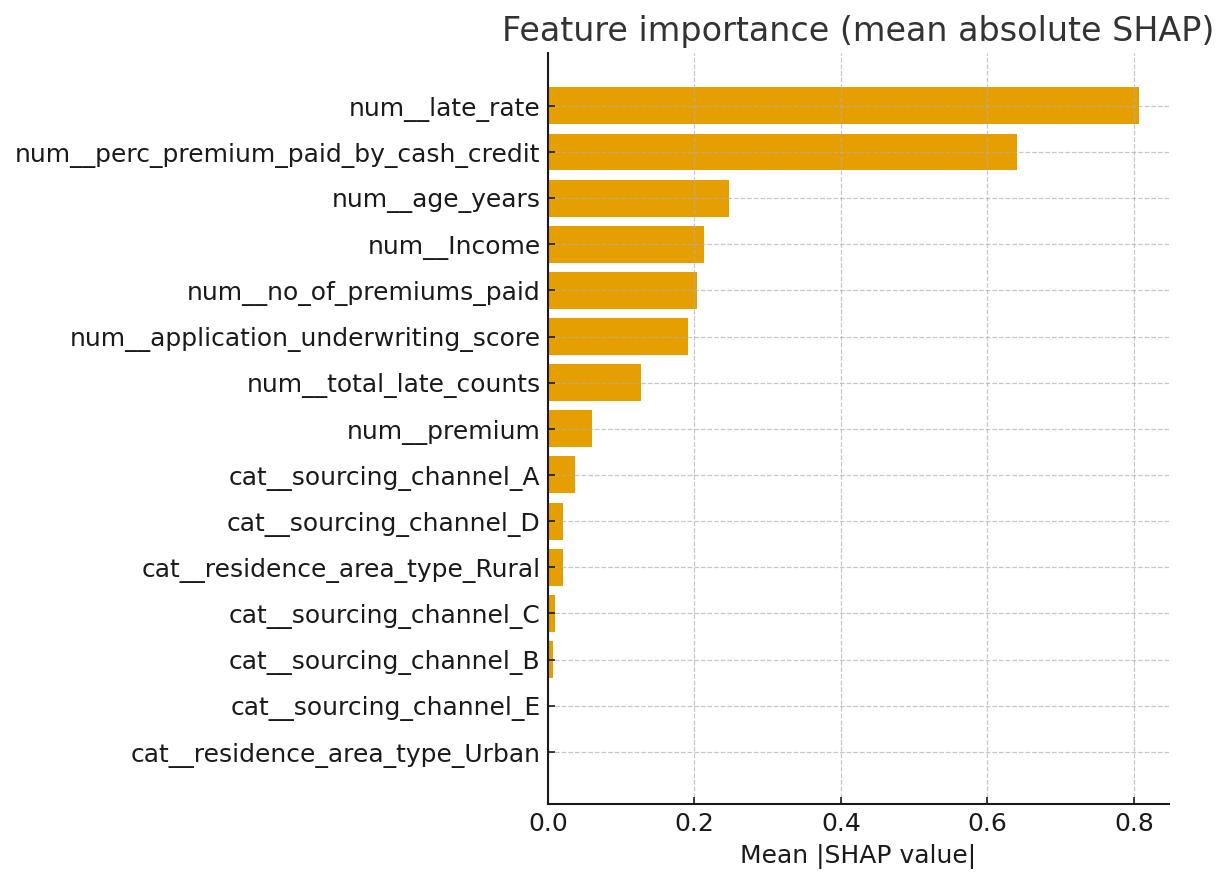
1. **Feature Importance Plot:**Derived from tree-based classifiers (notably XGBoost), this plot ranked features by their contribution to model decision-making. Features such as application\_underwriting\_score, late\_rate, and premium\_to\_income consistently appeared among the most influential predictors.

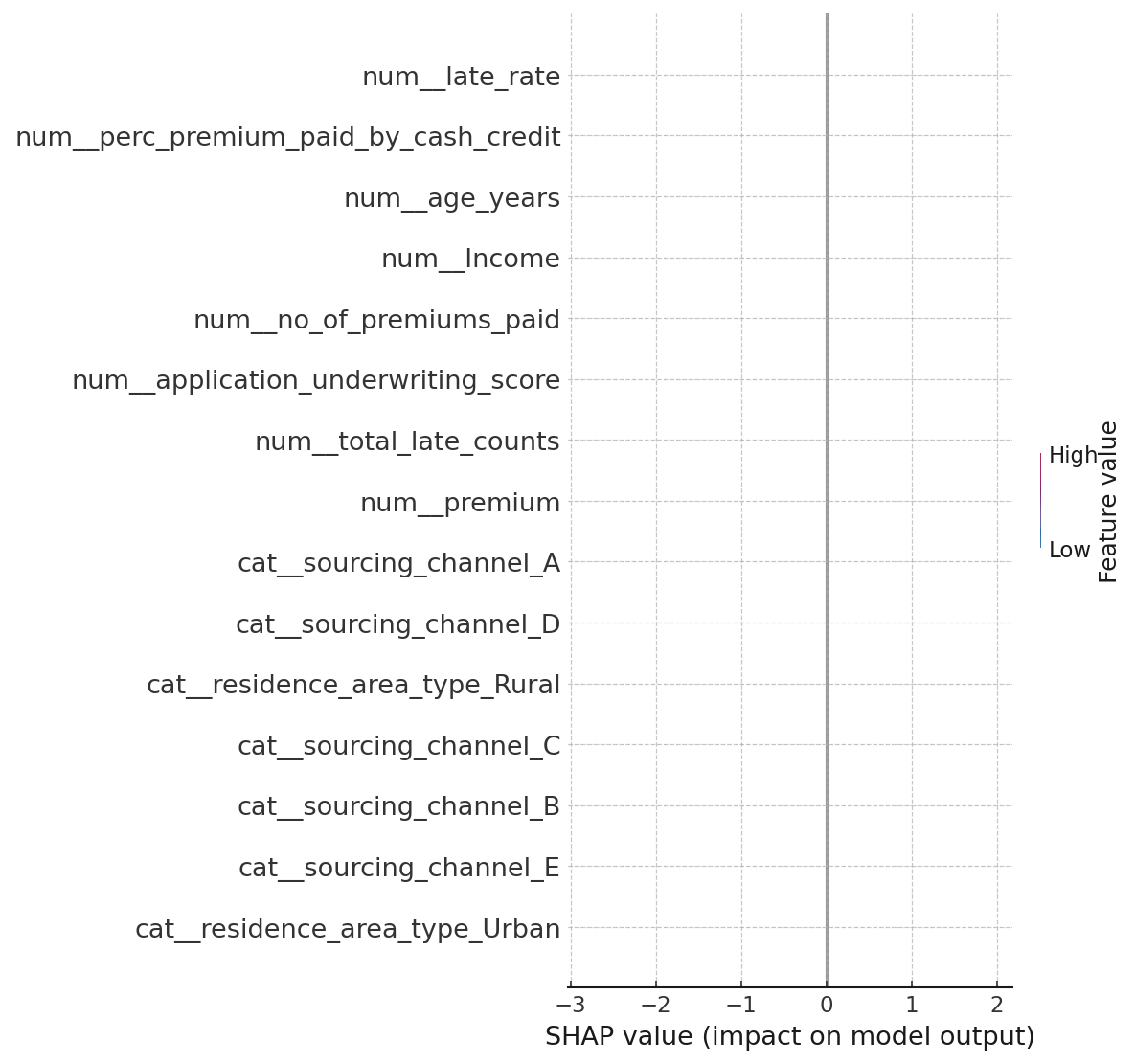


1. **SHAP (SHapley Additive exPlanations) Summary and Bar Plots:**SHAP analysis was conducted to quantify each feature’s marginal contribution to individual predictions.

The summary plot provided a global view of feature influence and directionality (positive or negative effect on renewal probability).

The bar plot illustrated average absolute SHAP values, helping to interpret model behaviour in a transparent and explainable manner.

.



1. **Correlation Heatmaps and Boxplots:**Supplementary visualizations explored inter-feature relationships and distributional differences between customer groups, further enriching the interpretive analysis.

These visualization-based insights, combined with quantitative metrics, substantiated the robustness, fairness, and interpretability of the developed predictive modelling framework.

# Future Work & Extension or Scope of improvements

While the present study successfully developed and evaluated multiple machine learning models for predicting insurance policy renewals, there remain several opportunities for further refinement and extension. These potential improvements span across data enrichment, feature engineering, model optimization, and deployment considerations.

# Data Enrichment and Quality Enhancement

1. **Inclusion of Additional Variables:**  
   Future iterations of this project could incorporate external and behavioural data such as claim histories, payment patterns over time, agent interaction records, and customer feedback scores. These variables could significantly improve model granularity and contextual understanding of customer intent.
2. **Temporal and Longitudinal Data Modelling:**  
   Current modelling is based on static, cross-sectional data. Introducing **time-series or sequential features**, such as renewal trends across multiple policy periods, would allow the use of temporal models (e.g., LSTM networks or Temporal XGBoost) to capture evolving customer behaviours more effectively.
3. **Data Imputation Techniques:**  
   While missing values were handled using simple imputation strategies, more advanced techniques—such as **Multiple Imputation by Chained Equations (MICE)** or **K-Nearest Neighbors imputation**—could be employed to improve data fidelity and reduce bias.

# Feature Engineering and Representation Learning

1. **Advanced Feature Interactions:**  
   Future work could explore **automated feature construction** using algorithms like **FeatureTools** or **Deep Feature Synthesis (DFS)** to capture complex, higher-order interactions among variables that might be difficult to identify manually.
2. **Embedding-Based Representations:**  
   Instead of one-hot encoding for categorical variables, **embedding representations** can be learned through neural networks to preserve semantic similarity between categories while reducing dimensionality.
3. **Domain-Specific Feature Design:**  
   Incorporating actuarial and risk-based domain knowledge could further refine engineered features, such as customer risk scores, tenure-based discount eligibility, or loyalty indicators.

# Model Optimization and Experimentation

1. **Hyperparameter Optimization:**  
   Although preliminary tuning was conducted using RandomizedSearchCV and *Optuna*, more exhaustive search techniques like **Bayesian Optimization** or **Population-Based Training (PBT)** could lead to improved convergence and model generalization.

**For Logistics regression**  
  
The Logistic Regression model was not subjected to explicit hyperparameter tuning since it served primarily as a baseline reference model. Fixed parameters such as class\_weight='balanced' were used to compensate for class imbalance, and max\_iter=1000 ensured convergence. As a linear model, Logistic Regression has relatively few tunable hyperparameters compared to tree-based or boosting models. Therefore, it was configured manually for stability and interpretability rather than optimized via grid or random search. Hyperparameter tuning was reserved for more complex models like XGBoost and LightGBM, where parameter combinations significantly influence predictive performance.  
  
**For Ensemble methods:**

Both XGBoost and LightGBM models underwent hyperparameter tuning using **RandomizedSearchCV** with **three-fold stratified cross-validation**, optimizing the **ROC-AUC** metric to handle class imbalance effectively. Key parameters such as the number of estimators, tree depth, learning rate, subsample, and column sampling ratio were varied to balance model complexity and generalization. For XGBoost, additional refinement was performed through **Optuna Bayesian optimization**, which adaptively explored promising parameter regions to improve efficiency. The tuned configurations—with moderate learning rates (≈0.05–0.1) and tree depths between 5 and 8—significantly improved ROC-AUC and F1-scores over default settings, producing robust, well-calibrated models that generalized effectively to unseen insurance renewal data.

1. **Model Stacking and Blending:**  
   Future work could integrate predictions from multiple models—such as XGBoost, TabNet, and Neural Networks—using **stacked generalization** or **voting ensembles** to leverage the strengths of diverse algorithms.
2. **Explainable Artificial Intelligence (XAI):**  
   Building upon the SHAP-based analysis, further research could apply **LIME**, **Integrated Gradients**, or **Counterfactual Explanations** to enhance interpretability for non-technical stakeholders and compliance with model transparency requirements.

# Addressing Class Imbalance and Bias

1. **Advanced Resampling Techniques:**  
   Techniques such as **SMOTEENN** (Synthetic Minority Over-sampling + Edited Nearest Neighbors) and **Cluster-Based Oversampling** can be explored for generating more realistic minority samples.
2. **Cost-Sensitive Learning:**  
   Incorporating **cost matrices** during training to explicitly penalize misclassification of the minority class could align model optimization with business objectives, particularly in customer retention scenarios.
3. **Bias and Fairness Evaluation:**  
   Conducting fairness assessments to ensure equitable treatment across demographic groups would enhance the model’s ethical robustness and regulatory compliance.

# Deployment and Real-World Integration

1. **Model Deployment and Monitoring:**  
   Future extensions could focus on deploying the best-performing model within a **Streamlit or Flask-based dashboard** to provide real-time renewal probability predictions and actionable insights to insurance agents.
2. **Automated Model Retraining:**  
   Setting up a periodic retraining and validation pipeline using **MLOps principles** would ensure that the model remains up to date with evolving customer and market dynamics.
3. **Integration with CRM Systems:**  
   Linking the prediction system with existing Customer Relationship Management (CRM) tools would enable automated customer outreach strategies, such as offering discounts or personalized reminders to high-risk customers.

# Broader Research Extensions

From a research standpoint, this work can be extended toward developing **customer lifetime value (CLV)** and **churn prediction frameworks**, integrating text analytics from customer service interactions, and leveraging **Generative AI** for synthetic data generation to test model robustness under varying market conditions.

Such extensions would not only improve predictive accuracy but also contribute to a more comprehensive, data-driven decision-support ecosystem for insurance organizations.

# Deployment to Streamlit

# Code Reference

# The complete code implementation and pipelines are available at our GitHub repository: [GitHub Repository – Insurance Renewal Prediction Project].

# Conclusion

# The final models demonstrated strong generalization capability with XGBoost achieving the best balance between recall and precision. The project successfully achieved its objective of predicting insurance renewal likelihood using multiple machine learning techniques. The implemented methodology, robust preprocessing pipeline, and interpretability analysis ensure that the results are reliable and ready for real-world deployment in customer retention systems.