



Evaluating the Efficacy of Machine Learning Algorithms for Customer Churn Prediction in The Insurance Industry

K.M.S.S.B.Kulasekara ^{a*}, L.S. Nawarathna^b

^a Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya,
20400, Sri Lanka

^b Department of Statistics and Computer Science, Faculty of Science, University of Peradeniya, Peradeniya,
20400, Sri Lanka

^aEmail: sumedhauop@gmail.com

^bEmail: lakshikas@sci.pdn.ac.lk

Abstract

In the insurance industry, large-scale data from customer transactions is generated daily, making it more cost-effective to retain existing clients than acquire new ones due to customer churn. This study aims to identify the best classification algorithm for predicting customer churn, determine the characteristics of churning customers, and model the time until churn. After data preprocessing and addressing dataset imbalance using the Synthetic Minority Oversampling Technique (SMOTE), various predictive models, including logistic regression, naive Bayes, decision trees, random forests, neural networks, extra trees, boosting algorithms, and heterogeneous ensemble learning methods, were applied. K-fold cross-validation was utilized for hyperparameter tuning and preventing overfitting. The results, evaluated using the confusion matrix and Area Under Curve (AUC), showed that Extreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LGBM) classifiers achieved the highest accuracy at 88.57% and 88.53%, respectively, with precision, recall, and F-score also used for assessment. Survival analysis revealed that approximately 80% of policyholders survived the first 200 months, with “having children” being the most significant predictor of customer churn. Unlike standard classification, survival analysis enables modeling of both churn likelihood and expected time-to-churn, accounting for censored observations. This provides actionable insights for intervention timing, a critical advantage in customer retention planning

Keywords: Confusion Matrix; Customer Churn Prediction; Insurance; Machine Learning; SMOTE; Survival Analysis

* Corresponding author. Email address: sumedhauop@gmail.com

1. Introduction

Insurance is the most used risk mitigation strategy in the world. Insurance companies provide unique financial services to the growth and development of every economy [1]. Insurance's importance has increased daily in both developed and developing countries. Insurance is a legal contract between a person and an insurance company, where the insurer promises to provide financial protection (sum insured) against unexpected events for a specific price (payment). Life insurance and General Insurance are the two main types of insurance plans. Life insurance is insurance against life risk only, and general insurance covers other aspects and assets in a person's life, such as motor, health, property, etc. An insurance policy is a written contract between the policyholder (the person or company that gets the policy) and the insurer (the insurance company).

This research focuses on customer retention in the life insurance field, which involves retaining existing customers. Customer retention explains the issue of customer churn, where churn describes the turnover of customers, and churn management describes the company's efforts to identify and control the problem of customer churn [2]. The motivation for churn prediction in insurance comes from the point made in customer relationship management: companies hold valuable customer information in their databases [3]. The data can be used to assess whether a customer could be leaving and the reasons for that. Since it is more profitable to keep existing customers than attract new ones [4], it makes sense for companies to try to predict leaving customers and prevent them from leaving or decreasing purchases.

An insurance company's responsibility is to provide financial compensation for a loss. The choice of insurance provider could be based on various fields, including claim process, customer service, policy features, policy flexibility, and premium cost, among others. Identifying churn customers is a crucial issue that insurance companies face. Customer churn can negatively impact insurance firms because recruiting new customers is sometimes more expensive than retaining existing ones.

The insurance domain defines churn as one who discontinues all his/her policies and stops renewing the policy. Customer churn has been a significant problem that most insurance companies have faced for several years. In order to manage competition, it is important in the insurance field to keep existing clients and win new ones. A loss of a customer may signify a loss in revenue for several years. Customers who are happy with an insurance provider typically stick with them and keep a variety of plans. There are many different reasons to leave an insurance provider. Customers may be offered the same coverage at a lower cost, or a policyholder may have claims that they are not covered by the policy or may not be satisfied with the service they're receiving from the existing provider [5].

According to [6], Machine Learning (ML) can be used to create models from large and complex data that can solve challenging problems, opening doors to big business ideas and opportunities. With the help of an ML algorithm, the movement of the customer can be reduced. Various researchers studied the problem of unbalanced data sets where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction problems. Most of the studies showed that the SMOTE outperformed the others compared to oversampling algorithms. The majority of related work focused on applying only one or two methods of ML to extract knowledge, and the others focused on comparing several strategies to predict churn.

Various researchers studied the problem of unbalanced data sets where the churned customer classes are smaller than the active customer classes, as it is a major issue in churn prediction problems. Most of the studies showed that the SMOTE outperformed the others compared to oversampling algorithms. Churn prediction is generally treated as a classification problem in ML. [5] compared various supervised learning algorithms as well as determining the most efficient classification algorithm based on the data set, the number of instances, and variables (features). Random forest, naive Bayes, decision tree, Support Vector Machine (SVM) radial, logistic regression, and SVM Linear models were used to predict the customer churn, and the random forest algorithm performed the highest accuracy. Random forest and unsupervised techniques were used to identify the customer churn. [7] provided analysis of logistic regression and naive Bayes model for customer churn prediction, and during the analysis, Area Under Curve (AUC) and accuracy matrices were used.

To predict the customer churners in Maltese insurance, [5] proposed six classification algorithms: decision trees, logistic regression, Naive Bayes, random forests, SVM, and survival analysis techniques. The majority of related work focused on applying only one or two methods of ML to extract knowledge, and the others focused on comparing several strategies to predict churn. Therefore, the primary aim of this study is to significantly address customer churn by identifying the churn customers and understanding why these customers switch. To identify the best model for churn prediction, understand which techniques can be used to generate the best churn prediction models and which model generates the best results, comparing accuracy (Accuracy matrices) and interpretability. By ML algorithms, we can predict whether existing customers will churn or not. After that, Survival analysis focuses on the analysis of the time remaining until the churn event occurs and explores its relationships with different features.

2. Materials and Methods

2.1. Data Description

In this work, we used auto insurance customer data sets shared by Merishna S. (Senior Data Scientist & Tech Lead at Collins Aerospace, RTX) for academic research purposes, which includes unique customer details of individuals who initiated their insurance policies between 2005 to 2022 in Texas State USA. In compliance with GDPR and ethical research practices, all customer identifiers and indirect identifiers were removed or pseudonymized prior to analysis. This ensured that the dataset contained only anonymized records, eliminating any possibility of re-identification. All procedures adhered to institutional ethics protocols and the IEEE Code of Ethics.

In those data sets, multiple customers may be assigned to the same address ID (an anonymized code, not PII), and not all customers have demographic information available. After merging datasets, the final data set consists of the training sample 1126209 observations, and the testing sample 554700 observations. The variable Churn was taken as the response variable which is a binary variable, and individual ID, address ID, current annual amount, income, days tenure, customer origin date age in years, date of birth, latitude, longitude, city, state, county, has children, length of residence, marital status, home market value, homeowner, college degree, good credit, and account suspended date were the other variables.

2.2. Methods

First, the preliminary analysis was carried out to understand the patterns and trends of the data. Graphical visualization was used to identify the distribution of the churn event among each variable. This process also involved handling missing values, and duplicate values. The correlations among continuous features were obtained by using the Pearson Correlation test. All categorical features were encoded using label encoding then the relationship between the churn event and each categorical variable was assessed using the Chi-square test and the relationship between the continuous variables was assessed using the Mann-Whitney U test. After data pre-processing and handling the imbalance problem using the SMOTE method, then applied the data normalization technique. Next, the information was divided into two sections, train, and test set in the proportion of 2/3 and 1/3 separately. This study proposes a 10 customer-churn prediction model consisting of Logistic regression, Naive Bayes Model, K-Neighborhood classifier, Neural network, Extreme Gradient Boosting (XGBoost), Lightweight Gradient Boosting (LGBost), Decision Tree, Random Forest, Extra tree classifier, and heterogeneous learning models to determine which models give the better prediction accuracy. The first heterogeneous ensemble learning model consists of logistic regression and an XGBoost algorithm. The second model consists of a combination of logistic regression, random forest, and XGBoost algorithms. Figure 1 shows a high-level overview of the approach proposed for the prediction of customer churn.

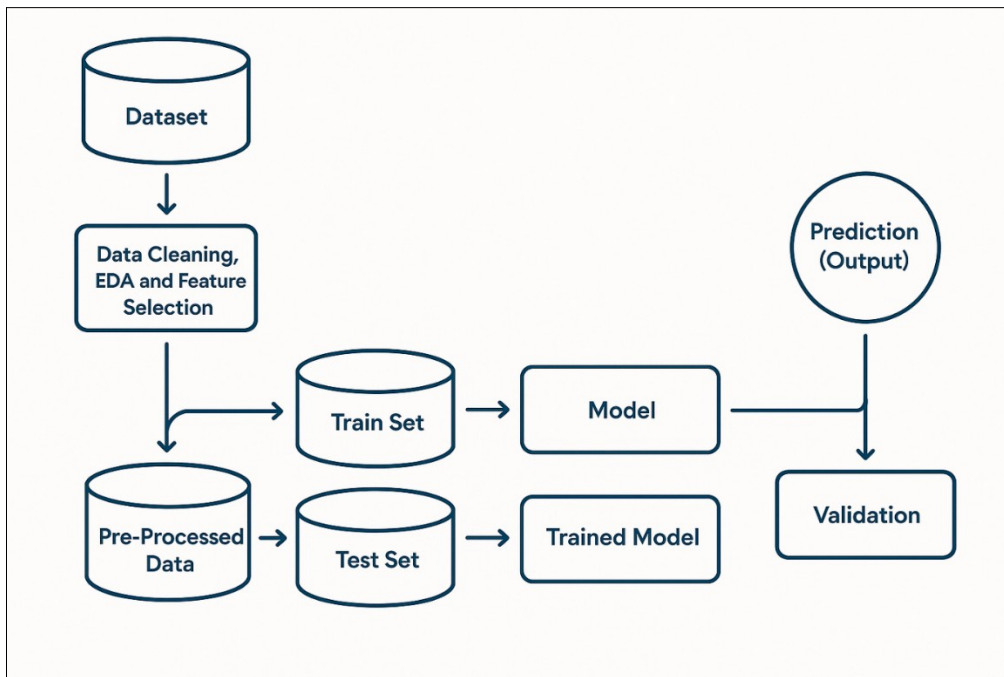


Figure 1. Proposed approach for the prediction of customer churn

2.3. Synthetic minority Oversampling Technique (SMOTE)

Imbalanced datasets often lead to poor model performance in the minority class. To address this, oversampling techniques are used. While simple duplication adds no new information, SMOTE (Synthetic Minority Oversampling Technique) generates synthetic samples to improve learning and balance the dataset [8].

2.4. Machine Learning Models

The following ML classification models were used to develop and compare predictive approaches for customer churn.

Logistic Regression: Logistic Regression is a classification method that estimates the probability of a binary outcome using the sigmoid function. It maps a linear combination of input variables to values between 0 and 1. Coefficients are learned through maximum likelihood estimation [9].

Naive Bayes: This probabilistic classifier applies Bayes' Theorem under the assumption of conditional independence between predictors. Gaussian Naive Bayes is used when features follow a normal distribution, offering fast performance in classification tasks [10].

K-Nearest Neighbors (KNN): KNN is a non-parametric algorithm that classifies a data point based on the majority class among its k nearest neighbors, using distance measures like Euclidean distance. The value of k influences bias-variance trade-offs [11].

Artificial Neural Networks (ANN): ANN is a popular ML technique inspired by the biological neural network in the human brain. An important class of feed-forward neural networks is Multilayer Perceptron (MLP). A MLP with ReLU activation, dropout regularization, and sigmoid output layer was trained using Adam optimizer. It captures complex patterns in data [12] .

Ensemble Learning: Ensemble methods combine multiple models to improve prediction accuracy and robustness. Bagging trains models in parallel while boosting trains sequentially. These techniques reduce errors from bias and variance [13] .

XGBoost: XGBoost is a boosting algorithm that sequentially builds decision trees and emphasizes misclassified instances by assigning higher weights. It is efficient and widely used in structured data tasks [14].

LightGBM: LightGBM is a gradient boosting framework optimized for speed and efficiency. It uses leaf-wise tree growth and focuses on harder-to-classify samples to improve accuracy [14]

Decision Tree: A decision tree splits data based on feature values to classify outcomes [15]. It creates branches and leaf nodes representing decision rules and final predictions, respectively ([16].

Random Forest: Random Forest builds multiple decision trees using random subsets of data and features (bagging). It combines predictions from all trees for improved stability and performance [5].

Extra Trees Classifier: This method builds an ensemble of randomised decision trees by selecting split points and features randomly, increasing model diversity and reducing variance [17].

Heterogeneous Ensemble Learning: This approach combines different types of classifiers to improve diversity and predictive performance, making use of varied learning algorithms [18]. This design strategically combines the interpretability of logistic regression with the non-linear feature interaction capabilities of random forest and the gradient-boosted performance of XGBoost. Such integration is rare in insurance churn prediction research. The ensemble benefits from complementary strengths, enhancing both accuracy and generalizability.

2.5. Model Evaluation

This study evaluates the model using accuracy, precision, recall, F1 score, and AUC-ROC. Accuracy reflects overall correctness, precision measures correct positive predictions, and recall indicates how well actual positives are identified. The F1 score balances precision and recall. The ROC curve shows performance across thresholds, and AUC summarizes this performance, with higher values indicating better prediction [19] .

2.6. Survival Analysis

Survival analysis examines the time until an event (e.g., customer churn) occurs and accounts for censored data cases where the event hasn't yet happened. Unlike traditional regression, it predicts the probability of event occurrence over time. The Cox Proportional Hazards (CPH) model estimates the hazard rate using predictors, where the Hazard Ratio (HR) shows the impact of each variable on churn risk. A HR greater than 1 indicates an increased risk of churn, less than 1 indicates a reduced risk, and equal to 1 indicates no effect on churn. Model performance is measured by the Concordance Index (C-index), which ranges from 0.5 (random) to 1 (perfect discrimination) [20], [21] .

3. Results

3.1. Preliminary Analysis

This insurance dataset comprises 1,680,909 (1.6 million) unique customer details who initiated their insurance policies between 2005 to 2022 in Texas State USA. It includes information on whether their accounts were suspended or not in 2022.

Data understanding began with identifying the type of data using either quantitative or qualitative and determining the data measurement level used (nominal, ordinal, interval, and ratio). The dataset used in this research included some duplicated data and missing values. To handle those, duplicated data values were removed in preprocessing by keeping the first duplicate row. The count of missing values was 253,719, found in the variables Latitude, Longitude, City, and County. These variables were not used in model training or evaluation and were retained only for reference purposes. Missing entries in these variables were replaced with zeros to maintain dataset structure and avoid processing errors, following the approach described by [22], which is considered acceptable for non-predictive features.

Besides that, the variables latitude, and longitude, which are not relevant to churn prediction, are discarded. As it was introduced in the methodology, the data set had 22 variables, and it can be divided into 14 numerical and 8 categorical variables with ID numbers (individual and address), dates, longitude, and latitude values. Categorical and numerical data can be statistically analyzed to get a high level of understanding.

As presented in Table 1, the average age for policyholders involved in this dataset is 55.8 and the average income involved in this dataset is 81498 dollars. According to the range of the continuous factors, income value spread in ranged between 5000 and 25000. Most of the policyholders are homeowners (82.010%), and most of them are married (62.37%). More than half of the households have children (52.21%).

Table 1. Descriptive Statistics of numerical variables

	Annual Amount (in dollars)	Age (in years)	Income (in dollars)
Count	1680909.0	168.00	1680909.00
Mean	940.31	55.82	81497.73

	Annual Amount (in dollars)	Age (in years)	Income (in dollars)
Std	246.54	14.55	54381.00
Min	-194.31	23.00	5000.00
25%	772.10	45.00	47500.00
50%	935.00	55.00	70000.00
75%	1102.99	64.00	87500.00
Max	2269.37	114.00	250000.00

Days-tenure, and Age in Years variables were plotted to find the outliers present in the data. As the data was highly imbalanced, an analysis was carried out to identify if the outliers fell into the minority class. Most outliers presented in the majority class can be removed using the Inter Quartile Range (IQR) method (removed all points that lie outside the range defined by the quartiles $\pm (1.5 * IQR)$), as it could not lead to important information loss.

3.2. Exploratory Data Analysis of Visualization

The dataset was visualized to identify the features that influence customer churn in the insurance industry. Mostly, bar plots and pie charts are used for categorical variables to identify the counts of categories, while histograms are used to plot numeric variables to determine the distribution. More than 80% of customers have good credit. 18% of customers are not homeowners. Half of the policyholders have children and more than 60% of customers are married and have college degrees.

The majority of customers have been using the insurance service for just the first 25 months and almost more than 200 months. This could be potential because different customers apply different types of insurance policies. Higher density in the age range 50 and 60 for the “Churned” category, which may indicate that middle-aged customers are more likely to churn. Conversely, if the low density is in the 20 and 30 age range for the “Churned” category, it suggests that young customers are less likely to churn.

There might be a higher churn rate among single customers compared to married customers. Customers with a college degree may have a lower churn rate compared to those without a college degree. This implies that having a college degree is associated with higher customer retention. A higher proportion of customers with good credit churned compared to those with poor credit. All five graphs show that the churn rate is higher in married customers compared to single customers. Also, the churn rate is higher among homeowners and customers who have children.

3.3. Association Between Variables

The chi-square test is used to evaluate the categorical variables that are associated with the decision of churn, which can be used in predictive model building. The test produced significant results ($p\text{-value} \leq 0.05$) to indicate that some of the variables have an association with the decision to churn.

Table 2. Chi-Squared Test: p-values for variable comparison

Attribute	Chi-Square Value	Degrees of Freedom	p- value
Has children	765.622	1	< 0.0001
Marital status	34.778	1	< 0.0001
Homeowner	324.951	1	< 0.0001
College degree	508.581	1	< 0.0001
Good credit	89.101	1	< 0.0001

According to Table 2, all variables had an influence on whether a customer churns or not. Therefore, the Null hypothesis is rejected and concluded that there are significant differences in the categorical predictors and characteristics of customers who churn and those who do not.

After performing the Mann-Whitney U test, the results showed statistically significant differences between customers who churn and those who do not in terms of days tenure ($p < 0.0001$), income ($p < 0.0001$), age in years ($p < 0.0001$), and current annual amount ($p < 0.0001$). These low p-values, all below the significance level of 0.05, indicate that these numerical features have a significant impact on customer churn.

This study transforms categorical values for instance, the categorical variable Marital Status contains two distinct classes (Married and Single), and the converted values are (0, 1) s using label encoding into a numeric value between 0 and the number of classes minus 1. The final dataset was obtained after handling missing values, duplicate values, data transformation, outlier handling, and normalizing the dataset. In this research, the final dataset was split into 2/3 training and 1/3 test datasets based on previous research [5].

3.4. Class-Imbalanced Dataset Handling

In this auto insurance dataset, the data are seriously imbalanced, with 88.5% being non-churn and 11.5% churn. SMOTE generates synthetic examples of the minority class to balance the class distribution. For this, it creates new instances of minority class samples based on their nearest neighbors in the feature space.

3.5. Machine Learning Methods

In this phase, the pre-processed data obtained was used to build the ML model to predict customer churn. The evaluation results of each model for churn prediction are summarized in Table 3 and accuracy matrices of all the models are presented. We notice that the XGBoost classifier and LGBBoost classifier received the best values in every evaluation metric. Precision, recall, and F1 scores are threshold-dependent metrics calculated at a fixed classification threshold (usually 0.5), where the model is optimized to balance false positives and false negatives effectively [23]. A relatively lower AUC despite high precision and recall suggests that the model performs well at the selected threshold but has moderate discriminatory power when considering all thresholds. This may also indicate some degree of overfitting, potentially caused by the synthetic samples introduced through SMOTE [24].

Table 3 . Performance Metrics of Machine Learning Models

Model	Accuracy	Precision	Recall	F1 Score	AUC
Naive Bayes	0.6542	0.91	0.67	0.77	0.63
Logistic Regression	0.6520	0.92	0.66	0.77	0.65
KNN	0.8099	0.91	0.87	0.89	0.65
Decision Tree	0.8192	0.90	0.89	0.90	0.58
Neural Network	0.8709	0.93	0.93	0.93	0.70
Extra Tree	0.8740	0.91	0.95	0.93	0.68
Random Forest	0.8809	0.91	0.96	0.93	0.69
XGBoost	0.8857	0.92	0.96	0.94	0.70

LGBoost	0.8853	0.92	0.95	0.94	0.70
Logistic Regression + XGBoost	0.8838	0.92	0.95	0.94	0.70
Logistic Regression + LGBoost	0.8843	0.92	0.95	0.94	0.70
Logistic Regression + XGBoost + RandomForest	0.8847	0.92	0.95	0.94	0.70

All the Heterogeneous models (Last three models in Table 3) presented a precision of 0.92, a recall of 0.95, and an F1 score of 0.94. The area under the ROC curve came out at 0.70 of all heterogeneous models. The ensemble performed well compared to single models and offered the benefit of combining different learning methods. The results were compared with previous churn prediction studies. [5] and [15] found similar success with random forests but did not explore ensemble methods. Our heterogenous ensemble, combining Logistic Regression, Random Forest, and XGBoost, improved accuracy.

The XGBoost algorithm yielded the highest accuracy with the following parameters: a subsample ratio of 0.8, 200 boosting rounds, a minimum child weight of 3, a maximum depth of 7, a learning rate of 0.2, and a column subsample ratio of 0.8. These choices were made to fine-tune the model for better predictive accuracy.

Figure 2 shows the feature importance in the highest accuracy gain models which is the XGBoost algorithm. when analyzing the importance of each feature through the model. Some of the features contribute a lot to the model while some are less. The “has children” is the most important feature for XGboost in determining customer churn.

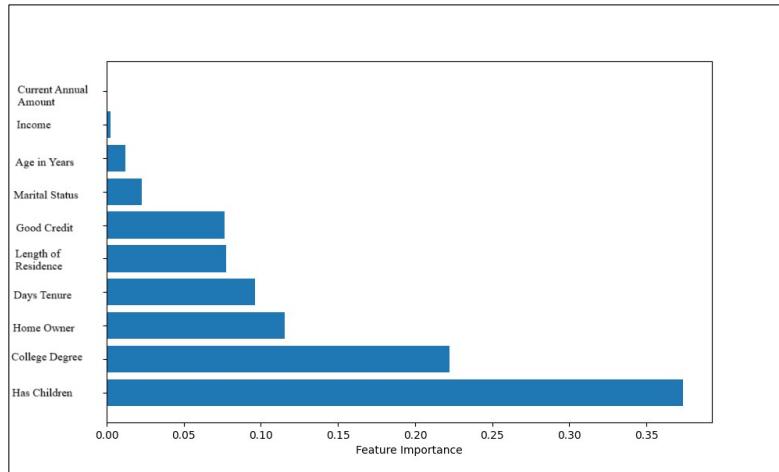


Figure 2. Feature Importance Analysis with XGBoost Model

3.6. Survival Analysis

Survival analysis was used to model the time till churn. As indicated in Figure 3, 95% of the policyholders survived during the first 15 months after subscribing to the policy. The company was able to retain more than 80% of its customers after 210 months. There is a constant decrease in the survival probability between 20 and 200 months. After 200 months, the survival probability has drastically dropped.

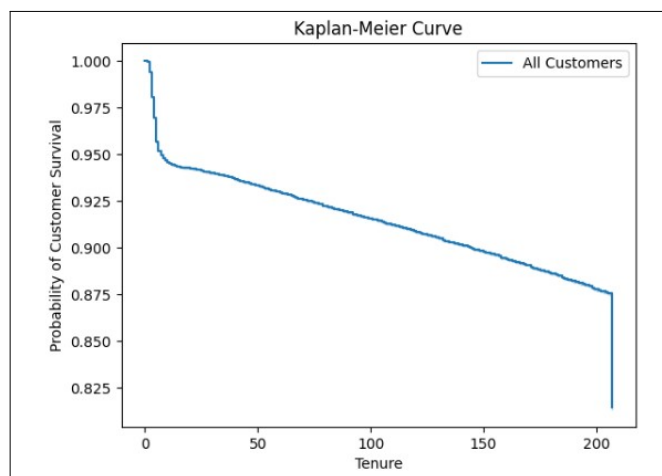


Figure 3. Survival curve showing the activity during the period of 220 months

A log-rank test was carried out to analyze group-wise churning probabilities and to find if there was a statistical significance between groups. Figure 4 shows group-wise survival curves.

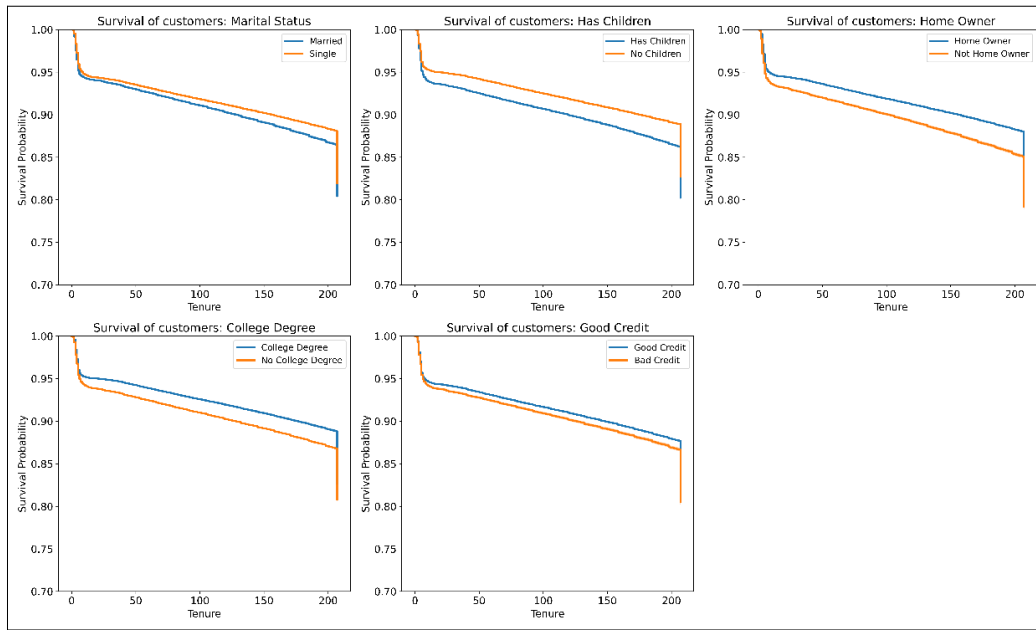


Figure 4. Kaplan–Meier survival curves for has children, college degree, good credit, marital status and homeowner variables

From the above Figure 4, the following can be concluded. The Single customers (orange line) are slightly above the married customers (blue line) for most of the tenure period. Therefore, Single customers have a slightly higher survival probability than married customers over time. Furthermore, across all the graphs, Single customers, those without children, homeowners, customers with a college degree, and customers with good credit each show a slightly higher survival probability over time compared to their counterparts. The survival probabilities of single homeowners with a college degree, and who don't have children, are high. In other words, if the customer is a single homeowner with a college degree, and who doesn't have children, he or she is less likely to churn. This can be explained through various lenses.

Financial Stability and Home Investment: Since single people frequently have stable incomes, they can view their homes as long-term investments. This viewpoint could be a factor in the desire for service subscriptions to be consistent [25]. **Less Lifestyle Change:** Customers frequently re-evaluate services and products because of important life changes, such as the birth of children [26]. Lower churn rates may result from single homeowners without children experiencing these changes less frequently.

Table 4. Summary of Hazard Ratio, confidence intervals, and p-values

Variable	Coef.	Exp (Coef.)	CI (Coef.)	CI (Exp.)	P-value
Curr Ann Amt	0.00	1.00	[0.00, 0.00]	[1.00, 1.00]	< 0.0001
Age In Years	-0.02	0.98	[-0.02, -0.02]	[0.98, 0.98]	< 0.0001
Income	-0.00	1.00	[-0.00, -0.00]	[1.00, 1.00]	< 0.0001
Has Children	0.11	1.11	[0.10, 0.12]	[1.10, 1.12]	< 0.0001
Length Of Residence	-0.02	0.98	[-0.02, -0.02]	[0.98, 0.98]	< 0.0001
Marital Status	0.02	1.02	[0.01, 0.03]	[1.01, 1.03]	< 0.0001

Homeo w n e r	-0.01	0.99	[-0.02, 0.00]	[0.98, 1.00]	0.06
College Degree	-0.12	0.88	[-0.13, -0.11]	[0.88, 0.89]	< 0.0001
Good Credit	-0.05	0.95	[-0.06, -0.04]	[0.94, 0.96]	< 0.0001

As presented in Table 4, The hazard ratio (HR) given under $\exp(\text{coef})$ against each predictor shows that the “current annual amount” and “income” features with $\exp(\text{coef}) = 1.0$, that do not affect the hazard. The “college degree” feature with $\exp(\text{coef}) = 0.88$ (i.e., less than 1. Which means a decrease in hazards and hence better survival. So, having a college degree reduces the hazard by a factor of 0.88 and hence if a customer has a college degree, it’s a good sign for better survival.

The p-value for “Has children” is < 0.05 and HR is 1.11, indicating a strong relationship between the “Has children” value and increased risk of churn. Holding the other covariates constant, a value of “Has children = 1” is associated with poor survival. Here person with a “Has children = 1” value has an 11% higher risk of churn. The C-index of this CPH model, approximately 0.61, suggests that the model has a moderate level of discrimination on the test. It indicates that the model has some ability to distinguish between customers who will churn and those who will not.

4. Discussion

In discussion, this research aimed to develop ML models for predicting customer churn and identify the most effective model for the insurance industry. The methodology involved analyzing big data from the insurance field using various ML algorithms, including logistic regression, naive Bayes, k-nearest neighbor, decision tree, random forest, XGBoost, LGBost, extra tree classifier, and heterogeneous ensemble learning. To address the imbalance problem in the dataset, the SMOTE was employed, as it effectively handled the issue of misrepresented majority classes. The results showed that XGBoost and LGBost algorithms achieved the highest accuracy, with 88.57% and 88.53% respectively. The heterogeneous ensemble learning algorithm, combining logistic regression, random forest, and XGBoost, had the third-highest accuracy at 88.47%. XGBoost consistently performed well across all evaluation metrics, including precision, recall, specificity, and ROC score, making it the best classifier for predicting customer churn in the insurance domain. LGBost and the heterogeneous ensemble learning algorithm also demonstrated strong performance.

This study contributes to the existing body of research by evaluating the effectiveness of different ML methods in predicting customer attrition in the insurance industry. The findings can assist insurance companies in developing effective customer retention strategies by highlighting the most accurate predictive algorithms. Survival analysis revealed that approximately 80% of customers remain with the company for up to 210 months, with "having children" identified as the most significant predictor of customer churn in both ML and survival analysis. This underscores the importance of familial status in influencing customer retention decisions.

Researchers can apply these models to real-world scenarios with similar data features to identify churn customers and make predictions. Although this study focused on the life insurance domain, the approach could be adapted to other customer relationship management systems facing similar challenges.

Additional future work may incorporate alternative data sources, such as customer interaction logs or unstructured text from service communications, to enrich feature sets. Advanced modeling techniques, including deep neural networks and online learning frameworks, could be evaluated for improved real-time churn prediction capabilities.

References

- [1] J. F. Outreville, "The Relationship between Insurance Growth and Economic Development: 80 Empirical Papers for a Review of the Literature," *SSRN Electronic Journal*, 2012, doi: 10.2139/ssrn.1885401.
- [2] S. Y. Hung, D. C. Yen, and H. Y. Wang, "Applying data mining to telecom churn management," *Expert Syst Appl*, vol. 31, no. 3, 2006, doi: 10.1016/j.eswa.2005.09.080.
- [3] M. Scriney, D. Nie, and M. Roantree, "Predicting customer churn for insurance data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2020. doi: 10.1007/978-3-030-59065-9_21.
- [4] V. Kumar and W. Reinartz, "Creating enduring customer value," *J Mark*, vol. 80, no. 6, 2016, doi: 10.1509/jm.15.0414.
- [5] M. Spiteri and G. Azzopardi, "Customer churn prediction for a motor insurance company," in *2018 13th International Conference on Digital Information Management, ICDIM 2018*, 2018. doi: 10.1109/ICDIM.2018.8847066.
- [6] I. M. Enholm, E. Papagiannidis, P. Mikalef, and J. Krogstie, "Artificial Intelligence and Business Value: a Literature Review," *Information Systems Frontiers*, vol. 24, no. 5, 2022, doi: 10.1007/s10796-021-10186-w.
- [7] I. Bose and X. Chen, "Hybrid models using unsupervised clustering for prediction of customer churn," *Journal of Organizational Computing and Electronic Commerce*, vol. 19, no. 2, 2009, doi: 10.1080/10919390902821291.
- [8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002, doi: 10.1613/jair.953.
- [9] H. Jain, A. Khunteta, and S. Srivastava, "Churn Prediction in Telecommunication using Logistic Regression and Logit Boost," in *Procedia Computer Science*, 2020. doi: 10.1016/j.procs.2020.03.187.
- [10] L. Ali *et al.*, "A Feature-Driven Decision Support System for Heart Failure Prediction Based on χ^2 Statistical Model and Gaussian Naive Bayes," *Comput Math Methods Med*, vol. 2019, 2019, doi: 10.1155/2019/6314328.
- [11] P. Cunningham and S. J. Delany, "K -Nearest Neighbour Classifiers," *Multiple Classifier Systems*, 2007.
- [12] A. Sharma and P. Kumar Panigrahi, "A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services," *Int J Comput Appl*, vol. 27, no. 11, 2011, doi: 10.5120/3344-4605.
- [13] I. N. Nyoman Mahayasa Adiputra and P. Wanchai, "Customer Churn Prediction Using Weight Average Ensemble Machine Learning Model," in *Proceedings of JCSSE 2023 - 20th International Joint Conference on Computer Science and Software Engineering*, 2023. doi: 10.1109/JCSSE58229.2023.10202105.
- [14] N. T. M. Sagala and S. D. Permai, "Enhanced Churn Prediction Model with Boosted Trees Algorithms in the Banking Sector," in *2021 International Conference on Data Science and Its Applications, ICoDSA 2021*, 2021. doi: 10.1109/ICoDSA53588.2021.9617503.
- [15] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," *Expert Syst Appl*, vol. 36, no. 3 PART 1, 2009, doi: 10.1016/j.eswa.2008.06.121.
- [16] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: a machine learning approach," *Computing*, vol. 104, no. 2, 2022, doi: 10.1007/s00607-021-00908-y.

- [17] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, 2006, doi: 10.1007/s10994-006-6226-1.
- [18] M. P. Sesmero, A. I. Ledezma, and A. Sanchis, "Generating ensembles of heterogeneous classifiers using Stacked Generalization," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 5, no. 1, 2015, doi: 10.1002/widm.1143.
- [19] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit*, vol. 30, no. 7, 1997, doi: 10.1016/S0031-3203(96)00142-2.
- [20] P. Sobreiro, J. Garcia-Alonso, D. Martinho, and J. Berrocal, "Hybrid Random Forest Survival Model to Predict Customer Membership Dropout," *Electronics (Switzerland)*, vol. 11, no. 20, 2022, doi: 10.3390/electronics11203328.
- [21] A. Alabdallah, M. Ohlsson, S. Pashami, and T. Rögnerdsson, "The Concordance Index Decomposition - A Measure for a Deeper Understanding of Survival Prediction Models," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4024162.
- [22] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of Machine Learning Research*, vol. 8, 2007.
- [23] D. M. W. POWERS, "Estimation of high affinity estradiol binding sites in human breast cancer EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION," *Journal of Machine Learning Technologies*, vol. 2, no. 1, 2011.
- [24] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans Knowl Data Eng*, vol. 21, no. 9, 2009, doi: 10.1109/TKDE.2008.239.
- [25] D. R. Carroll and R. Cohen-Kristiansen, "Evaluating Homeownership as the Solution to Wealth Inequality," *Economic Commentary (Federal Reserve Bank of Cleveland)*, 2021, doi: 10.26509/frbc-ec-202122.
- [26] N. Koschate-Fischer, W. D. Hoyer, N. E. Stokburger-Sauer, and J. Engling, "Do life events always lead to change in purchase? The mediating role of change in consumer innovativeness, the variety seeking tendency, and price consciousness," *J Acad Mark Sci*, vol. 46, no. 3, 2017, doi: 10.1007/s11747-017-0548-3.