PREDICTION OF THE INSURANCE CONTRACT RENEWAL FOR VEHICLE

FATMA MANLAIKHAF[1]

[1]EMI, FST BENI MELLAL, Sultan Moulay Slimane University, Beni-Mellal, Morocco

E-mail: [1]manlaikhaffatma@gmail.com

ABSTRACT

Convincing customers to renew their insurance contract is an important task for every insurance company. To do so, prior to the expiry date, insurance companies call and send out a series of SMS to their customers in order to persuade them. However, in Morocco, we have an average of 40% nonrenewal of vehicle insurance for contracts lasting from three to six months, and these type of contracts represent 75% of the customers. In our study, we want to focus on the reasons for non-renewal in order to predict whether a customer will be able to renew their contract or not, as well as optimize sales productivity to focus on customers with a higher chance of renewal, then persuade the customer to ensure a longer duration and to provide a better proposal. Our goal in this work is to use Machine Learning (ML) in the field of vehicle insurance to extract meaningful information from data. We used approximately 4000 customer portfolios to investigate the various non-renewal issues. In this scenario, we predict contract renewal using ML methods such as logistic regression, decision tree, random forest, K-NN, and SVM. We also examine and compare the performance of various models. In terms of accuracy, kappa, and AUC, SVM beat other approaches, with values of 0.96, 0.81, and 0.90, respectively.

Keywords: *Machine Learning, Insurance Contract, Artificial Intelligence, ML Models, Classification Approach*.

1. INTRODUCTION

The concerns in IT knowledge enhancement are in balancing new and profound information with incremental knowledge. Staying updated with cutting-edge tech is crucial for competitiveness and innovation, yet neglecting proven methods and best practices can be detrimental. A holistic approach is essential, integrating various sources and prior knowledge to build upon a strong foundation while fostering a culture of continuous learning.

Lately the trend for every industry is to use machine learning to optimize and increase the gains [2, 3, 5, 9–13]. In our work, we will focus on the insurance industry and especially for its non-life insurance category. Due to the high percentage of vehicle insurance non-renewals, Al Horia wants to focus on the reasons for non-renewal in order to predict whether a customer can renew their contract or not, as well as optimize sales productivity to focus on customers with a higher chance of renewal, then convince the customer to ensure a longer duration and to give them a better proposal. This is generally done with supervised machine learning models.

Prediction accuracy enables the insurance sector to better tailor its services and lowers the cost of auto insurance for more drivers [8]. ML techniques are increasingly being used by insurance companies instead of traditional and conventional methods since they provide a more thorough means of generating a more accurate and representative result.

McKinsey & Company (Columbus 2017) did a study on artificial intelligence and company profit margins [1]. They showed that the businesses that fully embraced artificial intelligence projects have generated a profit margin from 3% to 15%. However, selecting a suitable ML predictive model has yet to be fully addressed. In this study, we investigate more powerful ML techniques to make an accurate prediction for renewals occurrence by analyzing the big dataset given by Saham Assurance (Agency Al Horia insurance), a large automotive company based in Morocco, and applying the ML methods using the dataset, such as logistic regression, decision trees. We also evaluate and compare the performance of these models.

This primary goal of this paper is to accurately predict, using ML algorithms, whether a customer can renew their contract or not. As a result, the model needs to take into account certain factors that vary across clients, such as the type of car or the price of the contract.

2. RELATED WORK

As machine learning algorithms are used for driver performance monitoring and insurance market analytics, automotive insurance companies have a lot of motivation to implement them in their business. In a number of papers, including Smith et al [14]., ML models have been used to predict the insurance sector's outcomes, using machine learning models. According to [14], several machine learning models, such as neural networks and decision trees, were tested in order to determine if a policyholder submitted a claim or not, and the insurance company's impact on the case study was discussed as well.

For predicting claims severity, three machine learning methods were compared [19]. The neural networks were found to be the best predictor by the authors. Similarly, the thesis ”Research on Probability based Learning Application on Car Insurance Data” is a satisfactory solution to the same problem [15]. For determining whether a claim is valid or not, they only used a Bayesian network. According to client information, Kowshalya et al. [17] used three classifiers: random forests, J48, and naive Bayes algorithms to create predictions about fraudulent claims. Based on the findings, random forests perform better than the other techniques. Another topic is fraudulent claims, not insurance claims forecasting. Furthermore, a model can also predict the severity of vehicle damage virtually, as well as the amount of funding needed to fix it [16]. As shown above, insurance providers analyze their customers' data in many different ways by using machine learning. So, this paper [20] does not predict insurance claims, but instead estimates repair costs. In this study, insurance companies will be able to predict whether or not their customer relationships will be renewed after the first period of obtaining new insurance, whether it be automobile, life, or property insurance. This study is based on five classifiers, and it forecasts the customer's potential turnover. These classifiers are: LR, RF, KNN, AB, and ANN. Among the models tested, random forests had the best performance.

To predict the frequency of motor insurance claims, two competing methods are used, XGBoost and logistic regression in [21]. This study showed that the XGBoost model is slightly better than logistic regression.

Based on the above studies, we can conclude that the XGBoost model is the best model for classification in the insurance industry in recent studies which used some machine learning models in the insurance industry [21, 22]. Each study contained 2767 and 30,240 observations, respectively, while in [15], authors show that the nave Bayes model can correctly predict occurrences of claims.

In our paper, we use dataset that contained almost 2700 observations with 20 variables, and our results showed that although Random Forest is a useful model, SVM is significantly better. Hence, SVM model can be used to solve our problem. The literature review findings are summarized in Table 1 **(See** Appendix A, table 1)**.**.

Table 1 is a chronological table that shows different studies for using ML models in the insurance industry, where LR is a logistic regression model, GAMS is a generalized additive model, RF is a random forest model, KNN is a K-nearest neighbor model, NN is a neural network model, DT is a decision tree, AB is an AdaBoost model, MLP is a multi-layer perceptron model, SGB is a stochastic gradient boosting model, SVM is a support vector machine model, MSE is the mean square error, and RMSE is the root mean square error.

3. BACKGROUND

To understand the problem, it is essential to understand the insurance renewal process, ML, and classification. We explore the following terms. Before the expiry date insurance agencies launch a series of SMS and calls to remind and convince customers to renew their contracts. Thus, there is a need for an effective approach and a more reliable ML model to assess the contract renewal, a model that can read and interpret vast databases containing hundreds of consumer details provided by the agency Al Horia insurance subsidiary of the insurance company Saham Insurance. Saham Insurance is the leader in the field of Non-Life insurance, N°1 in Automobile and Health. With more than 481 general agents, SAHAM Assurance has the most extensive exclusive network in Morocco, allowing it to ensure a very strong regional presence and to develop a policy of proximity with all of its customers. The group develops complete and personalized financial solutions for the benefit of individual and institutional clients in all market segments, through all of its business lines: Sanlam Personal Finance, Sanlam Emerging Markets, Sanlam Investments, Santam and the newly created Sanlam Corporate. The group's areas of expertise operates in insurance, financial engineering, retirement, trusts, wills, damages, asset management, risk management, capital market activities and wealth management. Thus, they provided a dataset containing 20 variables with almost 2700 observations. These observations include customer information that the agency collected over several years. **(See** Appendix A, table 1**).**

3.1 Machine Learning

In the last decade, machine learning has gained high demand in data analysis, and it is gradually being adopted in computer science research [23]. Several sources of data have been generated rapidly, resulting in its rapid dissemination. It enables individuals and organizations to understand their datasets in more detail. Forbes's research has also indicated that one in ten companies now uses ten or more AI applications: 21% of the applications are based on fraud detection, 26% on process optimization, and 12% on opinion mining [24]. With the minimum amount of human involvement, machine learning can help machines develop their expertise by learning from the data and defining models, and, using learning algorithms, a prediction can be made through logic and conditions [5]. There are a significant number of applications for machine learning in industries available, including predictive models for online shopping, fraud detection in banks, or even spam filtering in email inboxes [25]. To generate reliable forecasts, the model must generalize well [26], which is the underlying principle behind these implementations. A major goal of machine learning is to discover patterns and models in data that can be used to predict future outcomes [27]. Data analytics and artificial intelligence (AI) research aim to learn complex trends, features, and relationships from large volumes of data. The basis for the application of AI in information discovery applications is machine learning and data mining based approaches. The essence of the intelligent machine learning method is the comparison between the objective to be achieved and the result derived by the machine. In many research fields, this method has proved to be successful [28], including in analysis of the insurance industry [15, 16, 18, 22]. The generic machine learning algorithm will receive input data and split this data into two sets, the first called the training and the second called the test dataset. Using training data, a model is trained to make predictions and determine whether the model is capable of generalizing. Testing the model on actual data allows it to determine if the predictions are accurate.

3.2 Machine Learning Approach to Predict a renewal contract

The problem of predicting insurance contract renewals for vehicles poses significant challenges due to its complex and dynamic nature. The validity of the study depends on various factors, such as the size and representativeness of the dataset, the quality of predictive models used, and the accuracy of the results obtained. This study incorporates a diverse and extensive dataset, employs important five machine learning algorithms, and exhibits strong predictive performance metrics so its validity is necessarily valid and meaningful for all dataset. The original study likely contributed to the general body of knowledge by shedding light on the factors influencing insurance contract renewal in the context of vehicles. It might have identified key variables, patterns, or customer behaviors that significantly impact renewal decisions. This additional knowledge can indeed be an improvement, as it could help insurance companies optimize their strategies, tailor policies, and offer incentives to increase contract retention rates. Moreover, it may pave the way for further research, fostering a deeper understanding of customer retention dynamics in the insurance industry and leading to more effective predictive models in the future.

Predicting whether a contract will be renewed is the purpose of the prediction. The output is Y = [0, 1] with 0,1 denoting either the client has decided not to renew the contract or that he will renew it. This machine learning model aims to predict the probability of renewing a contract by a customer. Consequently, the problem can be categorized as a binary classification [29], where renewing is a 1 and not renewing is a 0. There are a variety of classification algorithms available. Depending on the data state, some perform better and others worse.

$$\Pr(Y = 0|X = x_i) \quad (1)$$

$$\Pr(Y = 1|X = x_i) \quad (2)$$

Where X is a collection of instances x_i that represents all of the known information of the i-th policyholder.

3.3 Classifiers

2.2.2 Regression analysis

In linear regression, a variable of response (target) is estimated from a set of predictor variables. When a binary variable is the target, however, linear regression is not appropriate [30]. Regression evaluation with logistic regression (LR) is appropriate for binary-dependent variables. In LR, different independent features are analyzed in relation to a binary target variable. Several similarities exist between this method and linear regression. For dichotomous problems, LR is a multivariable learning algorithm. Models with two outputs, such as yes/no decision making, perform best using this classification method [31]. The model is therefore suitable for predicting either renewal or non-renewal of a vehicle insurance policy. There are many similarities between LR and linear regression in terms of its functionality. Unlike the binary target variable, linear regression provides a continuous output rather than a categorical one. LR performs with a single output variable,y_i , where i = 1, ...n, and each yi can hold one of two values, 0 or 1 (but not both). This follows the Bernoulli probability density function, as in the following equation:

$$p(y_i) = (\pi_i)^{y_i} (1 - \pi_i)^{1-y_i} \qquad (3)$$

This takes the value 1 when the probability is π_i , and 0 when the probability is $1 - \pi_i$; the interest in this is when $y_i = 1$ with an interesting probability π_i. The classifier will then produce the output of the predicted label; π_i i is equal to 1 if it is greater than or equal to its threshold (by default, 0.5).

$$\text{if}(p(y = 1)) \geq \text{theinstance} \in \text{class}(y = 1) \qquad (4)$$
$$\text{if}(p(y = 1)) < \text{theinstance} \in \text{class}(y = 1) \qquad (5)$$

3.3.2 Decision tree

There are many uses for decision trees, but the most common use is the solution of classification and regression issues. The tree structure represents the classifier, where each node represents the data variable, each branch represents the decision rules, and each node represents the output. The graph consists of two nodes. Among them is a decision node, which contains various branches used for decision-making. The second node represents the results of these decisions, which is a leaf node. While decision trees provide many advantages, they typically don’t outperform more complex algorithms. In contrast, random forests and gradient boosters were developed by combining multiple decision trees to create more efficient algorithms. Additionally, there are several types of DTs, including CART, C4.5, C5.0, and more [32]. The structure of a general decision tree is presented in figure 7 (see Appendix A, figure 1).

3.3.3 Random forest

With random forests, a wide variety of decorrelated trees are created through the use of bagged decision trees. These algorithms are very popular because they provide good predictive performance and require few hyperparameters. In general, the Leo Breiman algorithm is considered to be the most authoritative implementation of random forests [33]. The regression of individual trees creates a predictive value for random forests. It resolves to over-fit [34]. The following is an example of a random forest model:

$$g(x) = f_0(x) + f_1(x) + f_2(x) + ... + f_n(x) \qquad (6)$$

Where g is the final model, i.e., the sum of all models. Each model f(x) is a decision tree.

3.3.4 K-Nearest neighbor

An approach to supervised learning is K-nearest neighbor. It is regarded as one of the simplest ML models. K-NN makes the assumption that the new data will be compared to the existing data, and it will include the new data in the category that is closest to the existing classes. Although it can be applied to classification regression, its primary usage is in classification. Since it takes some time to learn from the training dataset, it is often referred to as a lazy model [35]. When new data is received, K-NN classifies it into the closest available category based on the training data stored in the K-NN model. K-NN modes only store the dataset through training. It might also be computationally ineffective.

3.3.5 Support vector machines

A Support Vector Machine (SVM) is a machine learning algorithm used for classifying and predicting data. With SVM classifiers, new data points can be assigned to a specific category by building a model. As a result, it is considered a non-probabilistic linear classifier. Linear classification can be accomplished with SVMs. The kernel trick can also be used to perform non-linear classification with SVMs. In this way, we are able to map inputs into feature spaces with high dimensions implicitly.

4. EVALUATION MODELS (PREDICTION PERFORMANCE)

Classifier models can be evaluated with several metrics such as how well they fit datasets and how well they perform on unknown samples (Hossin and Sulaiman 2015). For classification problems, accuracy alone cannot always be reliable, as it can provide bias for the majority class, which may result in high accuracy for the majority class and low accuracy for the minority class, which makes it unreliable as a prediction tool, especially when the data are imbalanced (Ganganwar 2012). The majority of policyholders do not renew their car insurance contracts, making them an excellent example of imbalanced data. Therefore, the bias would be towards not renewing classes if accuracy was used. In order to examine the effectiveness of our intervention, we employ other measures, such as F-measure and area under the curve (AUC). Understanding how a confusion matrix works is essential for understanding accuracy.

4.1 Confusion Matrix

Binary classification problems are solved using a confusion matrix. It is useful when determining whether or not class outputs were correctly predicted. A predicted class is represented in the rows of Table 2, whereas an actual class is represented in the columns [36]. Positive and negative instances are represented by TP and TN, respectively, while incorrectly classified positive and negative samples are represented by FP and FN.

Table 2. Confusion matrix.

	Actual Positive	Actual Negative
Predicted positive	True positive (TP)	False negative(FN)
Predicted negative	False positive (FP)	True negative (TN)

In car insurance prediction renewal, true positive would represent not renew, and true negative would represent a renew.

$$accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (7)$$

4.2 Kappa Statistics

Kappa statistics significantly measure besides the accuracy, because it considers the probability of a correct prediction in both classes 0 and 1. Kappa is essential for datasets with extreme class imbalance, such as auto insurance renew, since a classifier can achieve high precision by merely guessing the most common class.

$$K = \frac{pr(a) - pr(e)}{1 - pr(e)} \qquad (8)$$

4.3 Sensitivity and Specificity

When determining the ratio of positive classified examples, the sensitivity (true positive rate) is used to compare the predicted positive class to the actual positive example. By comparing the predicted negative class to the actual negative, the specificity (true negative rate) calculates the ratio of negative classified examples correctly.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (9)$$

$$Specifity = \frac{TN}{FP+TN} \qquad (10)$$

4.4 Precision and Recall

In order to determine whether a class belongs to the right category, the precision metric is used. Recall is another useful metric, which measures how well the model can detect class type on the basis of the fraction of positive classes that are classified correctly [36].

$$Precision = \frac{TP}{TP+FP} \qquad (11)$$

$$Recall = \frac{TP}{TP+FN} \qquad (12)$$

4.5 The F-Measure

It is also known as the F1 score or the F-score and represents a model's performance by combining precision and recall. F-measure can be calculated as follows:

$$F - measure = \frac{2 * precision \times recall}{recall + precision} = \frac{2 \times TP}{2 \times TP + FP + FN} \qquad (13)$$

Due to their weakness of being less resistant to changes in class distribution, accuracy, precision, and recall suffer from weakness as the area under the receiver operator characteristics curve (ROC) is required; AUC metrics, also known as receiver operating characteristics (ROC) or global classifier performance metrics, are often used to compare overall performance among different classification schemes [37]. Previously tested metrics may not perform as well if the test set's distribution of positive and negative instances changes. A change in the proportion of positive to negative instances or the distribution of classes, however, has no effect on the ROC curve.

5. DATASET

In this study, we analyze a dataset given by ASSURANCES AL HORIA Agency, which is related to direct marketing campaigns (phone calls). The classification goal is to predict whether the customer will renew (1/0) to a term deposit (variable 'renew'). In this article we want to use 2700 customers' portfolios, to study the different factors of renewal / non-renewal of the contracts. This is a test database of 22 columns and 2700 rows, Although it may take more time than needed to choose the best algorithm suited for your model, accuracy is the best way to go forward to make your model efficient.

6. PROPOSED MODEL

In this paper, we developed a model to predict Insurance Contract Renewal For Vehicle by applying ML techniques. The stages of the proposed model shown in Figure 2.

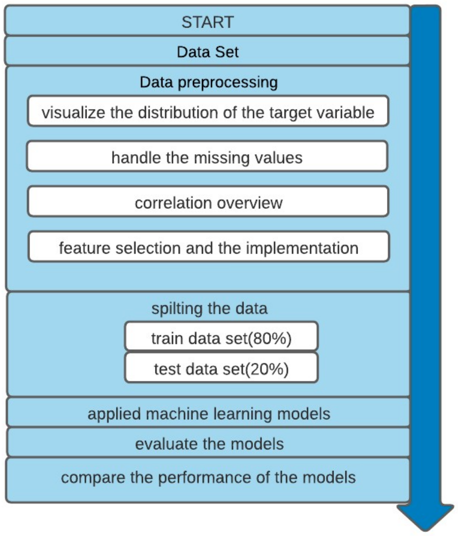

START
Data Set
Data preprocessing
visualize the distribution of the target variable
handle the missing values
correlation overview
feature selection and the implementation
spilting the data
train data set(80%)
test data set(20%)
applied machine learning models
evaluate the models
compare the performance of the models

Figure 2: Overall structure of the proposed model.

6.1 Data Preprocessing

The dataset consists of 20 variables. Each of these attributes has its relation to the Insurance Contract Renewal for Vehicle, which is our dependent target variable. The data is checked and modified to apply the data to the ML algorithms efficiently. We begin by considering the variable answer (dependent), target.

6.1.1 Renew variable

Our target column is a binary variable that contains two classes (1,0), 1 if a customer has renewed and a 0 if a customer has not renewed. Figure 5 shows the distribution of 1 and 0 for the target column.

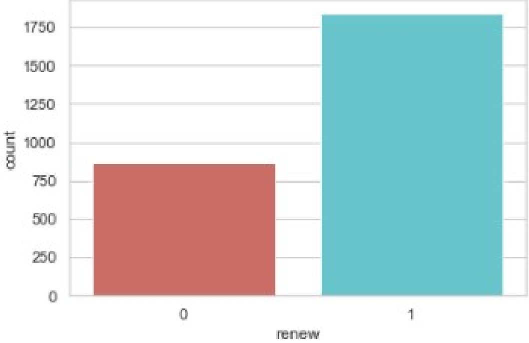


Figure 3: Histogram of the distribution of target values.

The figure shows that the target variable with class 0 having 32% observations and class 1 having 68% observations.

6.1.2 Correlation overview

In the figure of Correlation matrix (see appendix A, figure 2), "Green color" explain relation with another column, for example the relationship between renew and seniority or between job and renew, then between renew and convention with employer of customers.

6.1.3 Hyper-parameter optimization

Grid searches were conducted to find hyper-parameters that would yield optimal performance for the models. A 10-fold cross-validation technique was used based on accuracy as an evaluation metric. Table 3 shows the hyper-parameter tuning on the models used in this paper, where mtry is the number of randomly selected predictors, model is the model type, k is the number of nearest neighbors. C is the penalty parameter of the error term, maxdepth is the max tree depth, the kernel is the main hyperparameter of the SVM. It maps the observations into some feature space, penalty is the fact of shrinking the coefficients of the less contributive variables toward zero. This is also known as regularization. K is the number of neighbors, and criterion is the parameter that determines how the impurity of a split will be measured.

Table 3. The Hyper-Parameter Tuning On The Models Used In This Paper.

Model	Parameters	Range	Optimal Value
Regression Analysis	1.penalty	1.[l1,l2]	1.L2
DT	1.Criterion 2.max depth	1.['gini', 'entropy'] 2.[1,20]	1.Gini 2.8
RF	1.mtry 2.criterion: 3.max depth: 4.estimators:	1.[2,100] 2.[Gini, entropy] 3. [2, 6, 8, 10, 20] 4. [8, 10, 16, 20, 24, 200]	1.2 2. gini 3. maxdepth=2 4.100
KNN	1.k 2.weights	1. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] 2.[uniform, distance]	1
SVM	1.C 2.Kernel	1. [1, 10, 100, 1000] 2.[poly, linear, rbf]	1.1 2.linear

6.1.4 Features selection and implementation

In this study, we used the assurance Al Horia agency dataset. The study aims to predict the rate of renewal using various ML models. The dataset contains 22 variables and 2700 observations. Every observation includes the specific details of different insurance contract, . The dataset is separated into two parts; the first part is called the training data, and the second part is the test data. The training data make up about 80% of the total data used, and the rest is for test data. These models are trained with the training data and evaluated with the test data. For this study, R x64 4.0.2 is used for implementing the models. For classification, we used accuracy, error rate, kappa, sensitivity, specificity, precision, F1 and AUC as measures of evaluation.

7. RESULTS

This section analyzes the results obtained from each classification and compares the results to determine the best model with a highly accurate prediction. Every model used in this study was evaluated according to a confusion matrix, precision, recall, F1-score, and AUC; then, we compared all classifier models to explain why SVM was selected as the best classifier (see Table 4).

Table 4. Model performance.

Model	Accuracy	Error Rate	Kappa	AUC
RF	0.91	0.09	0.786	0.88
KNN	0.88	0.12	0.70	0.81
SVM	0.96	0.04	0.81	0.90
RA	0.75	0.25	0.49	0.74
DT	0.92	0.08	0.81	0.89

Model	Sensiti-vity	Specifi-City	Preci-sion	Rec-all	F1
RF	0.79	0.96	0.92	0.96	0.98
KNN	0.66	0.96	0.88	0.96	0.92
SVM	0.86	0.94	0.96	0.94	0.98
RA	0.64	0.85	0.72	0.85	0.78
DT	0.84	0.96	0.94	0.96	0.95

Table 4 presents the evaluation of all classifiers of ML used in this study. The range of accuracy values for all ML models was between 75% and 96%. SVM was the best model, with a high accuracy of 96% and a kappa coefficient of 0.81. The results showed that SVM was most likely to solve renewal prediction problems correctly. The DT model achieved good classification, with an accuracy of 92%. Regression Analysis showed the lowest accuracy of 75% and a kappa coefficient of 0.25. From the table, we obtain that SVM had the highest sensitivity, which means that 86% of the samples detected as positive were actually positive. The specificity for the SVM model explains that 94% of the true negative samples were correctly classified. Figure 11 shows the comparison between the techniques on various performance measures. It shows that, according to the accuracy, error rate, kappa, sensitivity, precision, F1 and AUC the best model was SVM and the worst was Regression Analysis. According to specificity, and recall the best model was KNN and the worst was Regression Analysis. Thus, we conclude that SVM showed the best performance.

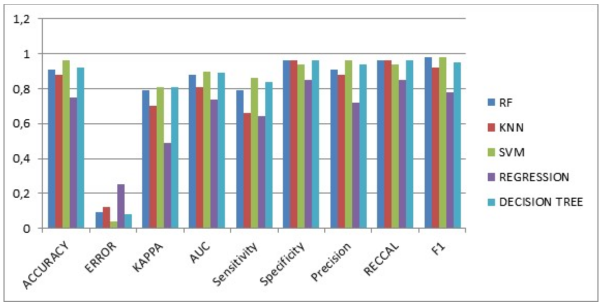


Figure 5: Comparison between the models

The ROC curves are shown in Figure 6. ROC offered the overall performance variable of the classification as its threshold for discrimination (Ariana et al. 2006). The AUC is known as the general quality index of the classifiers. The value of 1 for the AUC means a perfect classifier, while 0.5 means a random classifier. Based on the AUC comparison of the classifiers, the SVM score was 0.965, which was the best, followed by Decision tree (0.89), Random forest (0.88), and KNN (0.85). These findings suggest that SVM, DT, and RF had satisfactory performance (AUC > 0.80), whereas Logistic Regression had a poor performance.
(see appendix A, figure 3).

7.1 Variable Importance

Variable importance is a technique that tests the contribution of each independent variable to the outcome prediction. The variable importance for all data features is shown in Figure 7 It starts with the most influential variables and ends with the variable with the smallest effects (Kuhn and Johnson 2013).

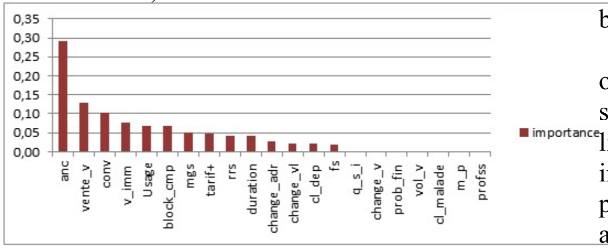


Figure 7: The rank of features by importance based on the random forest algorithm

8. CONCLUSIONS

As big data and data science methods, such as machine learning (ML), become more prevalent, they play an increasingly important role in managing insurance contracts. A data quality check before preparation and cleaning can help deal with a bias in favor of a majority class in an imbalanced dataset. Similar to other industries, insurance companies use ML analytics to optimize marketing strategies, improve their business, increase profits, and reduce costs. Two machine learning techniques are presented in this paper for analyzing insurance renewal predictions and comparing their performance using various metrics. The aim was to make the insurance prices fair to the client by using ML models to predict the renewal of the insurance contract in the next year. As a result, insurance companies can create a model that is accurate enough to predict risk in order to make automotive insurance more accessible to more clients. It is imperative that workers are trained routinely and consistently in order to adapt and use these new techniques effectively. In order to maximize efficiency and understand some of these algorithms' limitations, regulators and policymakers must make fast decisions. We suggested that the performance metrics should not be limited to one criterion, such as the AUC. Thus, we used six performance metrics to increase the modeling process transparency, because regulators will need to ensure the transparency of decision-making algorithms to prevent discrimination and a potentially harmful effect on the business. According to the results of this paper, ML methods can be used to predict contract renewals in insurance companies relatively accurately and with fairly good performance. In addition, SVM and Random Forest appear to perform well based on Table 4, but SVM performs the best due to its better performance metrics. The SVM model meets both functional and non-functional requirements, based on the classifier model results. SVMs were therefore stronger and more generalizable as the best prediction models due to this feature.

The research's contribution to the problem of insurance contract renewal prediction is satisfactory, and it aligns well with previous literature on data-driven approaches to improve insurance operations. Implementing these predictive insights can lead to competitive advantages for insurance companies and improved customer satisfaction. However, continuous research and updates are necessary to adapt to evolving market dynamics and ensure the sustained relevance and effectiveness of the predictive models.

REFERENCES:

[1] Columbus, Louis. "McKinsey's State of Machine Learning and AI, 2017." Forbes. Available online: https://www.forbes.com/sites/louiscolumbus/2017/07/09/mckinseys-state-of-machinelearning-and-ai-2017 (accessed on 17 December 2020) (2017).

[2] Felipe Cucker and Ding uan Zhou, Learning Theory: An Approximation Theory Viewpoint.

[3] V. Monbet, July 2017, Machine Learning, environmental data.

[4] Tutorials Point (I),2019, Machine Learning with Python.

[5] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. Machine learning basics. Deep Learning 1: 98–164.

[6] Hossin, Mohammad, and M. N. Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process 5: 1.

[7] Tutorials Point (I),2019, Machine Learning with Python.

[8] AN, Su Hyun, Seong Hee YEO, and Minsoo KANG. ”A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree.” Korea Journal of Artificial Intelligence 9.1 (2021): 9-14.

[9] S Lyaqini and M Nachaoui. Identification of genuine from fake banknotes using an enhanced machine learning approach. In International Conference on Numerical Analysis and Optimization Days, pages 59–70. Springer, 2021.

[10] S Lyaqini, M Nachaoui, and A Hadri. An efficient primal-dual method for solving non-smooth machine learning problem. Chaos, Solitons & Fractals, 155:111754, 2022.

[11] S Lyaqini, M Nachaoui, and M Quafafou. Non-smooth classification model based on new smoothing technique. In Journal of Physics: Conference Series, volume 1743, page 012025. IOP Publishing, 2021.

[12] Soufiane Lyaqini and Mourad Nachaoui. Diabetes prediction using an improved machine learning approach. Mathematical Modeling and Computing, 8(4):726–735, 2021.

[13] Soufiane Lyaqini, Mohamed Quafafou, Mourad Nachaoui, and Abdelkrim Chakib. Supervised learning as an inverse problem based on non-smooth loss function. Knowledge and Information Systems, pages 1–20, 2020.

[14] Smith, Kate A., Robert J. Willis, and Malcolm Brooks. ”An analysis of customer retention and insurance claim patterns using data mining: A case study.” Journal of the operational research society 51.5 (2000): 532-541.

[15] JING, Longhao, ZHAO, Wenjing, SHARMA, Karthik, et al. Research on Probability-based Learning Application on Car Insurance Data. In : 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017). Atlantis Press, 2018. p. 59-63.

[16] DEWI, Kartika Chandra, MURFI, Hendri, et ABDULLAH, Sarini. Analysis Accuracy of Random Forest Model for Big Data–A Case Study of Claim Severity Prediction in Car Insurance. In : 2019 5th International Conference on Science in Information Technology (ICSITech). IEEE, 2019. p. 60-65.

[17] KOWSHALYA, G. et NANDHINI, M. Predicting fraudulent claims in automobile insurance. In : 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018. p. 1338-1343.

[18] SINGH, Ranjodh, AYYAR, Meghna P., PAVAN, Tata Venkata Sri, et al. Automating car insurance claims using deep learning techniques. In : 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM). IEEE, 2019. p. 199-207.

[19] WEERASINGHE, KPMLP et WIJEGUNASEKARA, M. C. A comparative study of data mining algorithms in the prediction of auto insurance claims. European International Journal of Science and Technology, 2016, vol. 5, no 1, p. 47-54.

[20] STUCKI, Oskar. Predicting the customer churn with machine learning methods: case: private insurance customer data. 2019.

[21] PESANTEZ-NARVAEZ, Jessica, GUILLEN, Montserrat, et ALCANIZ, Manuela. Predicting motor ˜ insurance claims using telematics data—XGBoost versus logistic regression. Risks, 2019, vol. 7, no 2, p. 70.

[22] ABDELHADI, Shady, ELBAHNASY, Khaled, et ABDELSALAM, Mohamed. A proposed model to predict auto insurance claims using machine learning techniques. Journal of Theoretical and Applied Information Technology, 2020, vol. 98, no 22.

[23] GERON, Aur´elien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. ” O'Reilly ´ Media, Inc.”, 2022.

[24] Columbus, Louis. 2018. Roundup of Machine Learning Forecasts and Market Estimates, 2018. Forbes Contrib.

[25] SCHMIDT, Jonathan, MARQUES, M´ario RG, BOTTI, Silvana, et al. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 2019, vol. 5, no 1, p. 1-36.

[26] GONC¸ ALVES, Ivo, SILVA, Sara, MELO, Joana B., et al. Random sampling technique for overfitting control in genetic programming. In : European Conference on Genetic Programming. Springer, Berlin, Heidelberg, 2012. p. 218-229.

[27] D'ANGELO, Gianni, TIPALDI, Massimo, GLIELMO, Luigi, et al. Spacecraft autonomy modeled via Markov decision process and associative rule-based machine learning. In : 2017 IEEE international workshop on metrology for aerospace (MetroAeroSpace). IEEE, 2017. p. 324-329.

[28] D'ANGELO, Gianni, FICCO, Massimo, et PALMIERI, Francesco. Malware detection in mobile environments based on Autoencoders and API-images. Journal of Parallel and Distributed Computing, 2020, vol. 137, p. 26-33.

[29] KOTSIANTIS, Sotiris B., ZAHARAKIS, Ioannis D., et PINTELAS, Panayiotis E. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 2006, vol. 26, no 3, p. 159-190.

[30] SABBEH, Sahar F. Machine-learning techniques for customer retention: A comparative study. International Journal of Advanced Computer Science and Applications, 2018, vol. 9, no 2.

[31] MUSA, Abdallah Bashir. Comparative study on classification performance between support vector machine and logistic regression. International Journal of Machine Learning and Cybernetics, 2013, vol. 4, no 1, p. 13-24.

[32] SONG, Yan-Yan et YING, L. U. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 2015, vol. 27, no 2, p. 130.

Appendix A

Table 1. Different studies for using ML models in the insurance industry to get a representative overview.

Article & Year	Purpose	Algorithms	Performance Metrics	The Best Model
(Abdelhadi et al. 2020)	Classification to predict claims Occurrence	J48, NN, XGB, naïve base	Accuracy ROC	XGBoost
Raghavan & El Gayar, 2019	I. Fraud Detection Using Machine Learning And Deep Learning	KNN, RF, SVM, CNN, RBM) and DBN	ROC Curve (AUC), Matthews Correlation Coefficient (MCC) and Cost of failure	SVM
Jessica Pesantez-Narvaez, Montserrat Guillen and Manuela Alcañiz 2019)	Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression	LR , XGB	ROC, AUC , accuracy, Specificity,Sensitivity	XGBoost
(Pesantez-Narvaez et al. 2019)	Classification to predict claims Occurrence	XGB, LR	Sensitivity, Specificity, Accuracy, RMSE, ROC	XGBoost
(Dewi et al. 2019)	Regression to predict claims severity	RF	MSE	**RF**
(Stucki 2019)	Classification to predict churn and Retention	LR, RF, KNN, AB, and NN	Accuracy, F-Score, AUC	RF
(Sabbeh 2018)	Classification to predict churn Problem	RF, AB, MLP, SGB, SVM,KNN, CART, Naïve Bayes, LR, LDA.	Accuracy	AB
(Kowshalya and Nandhini 2018)	Classification to predict insurance fraud and percentage of premium amount	J48, RF, Naïve Bayes	Accuracy Precision Recall	RF
(Jing et al. 2018)	Classification to predict claims Occurrence	Naïve Bayes, Bayesian, Network model	Accuracy	Both have the Same accuracy.
(Mau et al. 2018)	Classification to predict churn, retention, and cross-selling	RF	Accuracy, AUC ROC, F-score	RF
(Subudhi and Panigrahi 2017)	Classification to predict insurance Fraud	DT, SVM, MLP	Sensitivity, Specificity Accuracy	SVM
(Fang et al. 2016)	Regression to forecast insurance customer profitability	RF, LR, DT, SVM, GBM	R-squares RMSE	RF
(Weerasinghe and Wijegunasekara 2016)	Classification to predict the number of claims (low, fair, or high)	LR, DT, NN	Precision Recall Specificity	NN
(Günther et al. 2014)	Classification to predict the risk of Leaving	LR and GAMS	ROC	LR
(Smith et al. 2000)	Classification to predict customer retention patterns	DT, NN	Accuracy ROC	NN

Figure 1

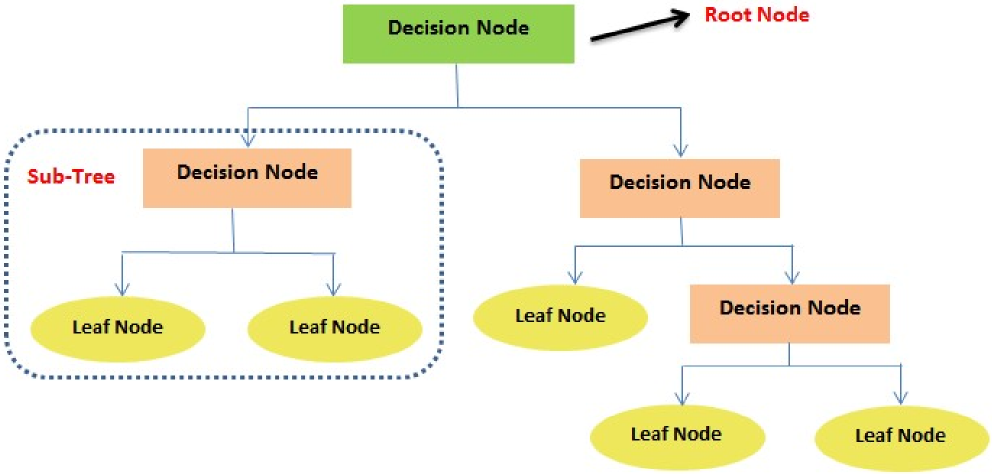


Figure 1: Structure of a general decision tree.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 2

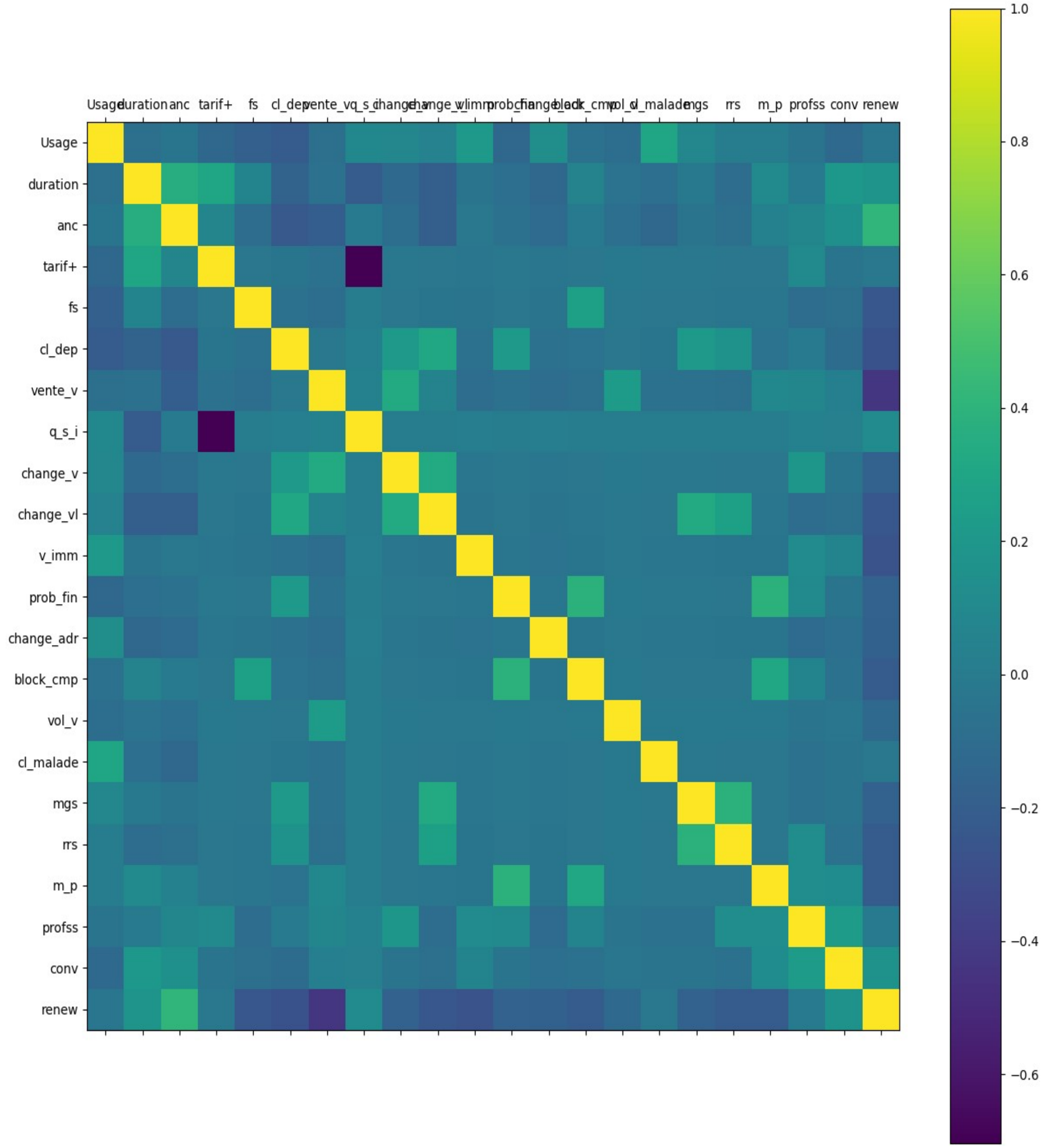
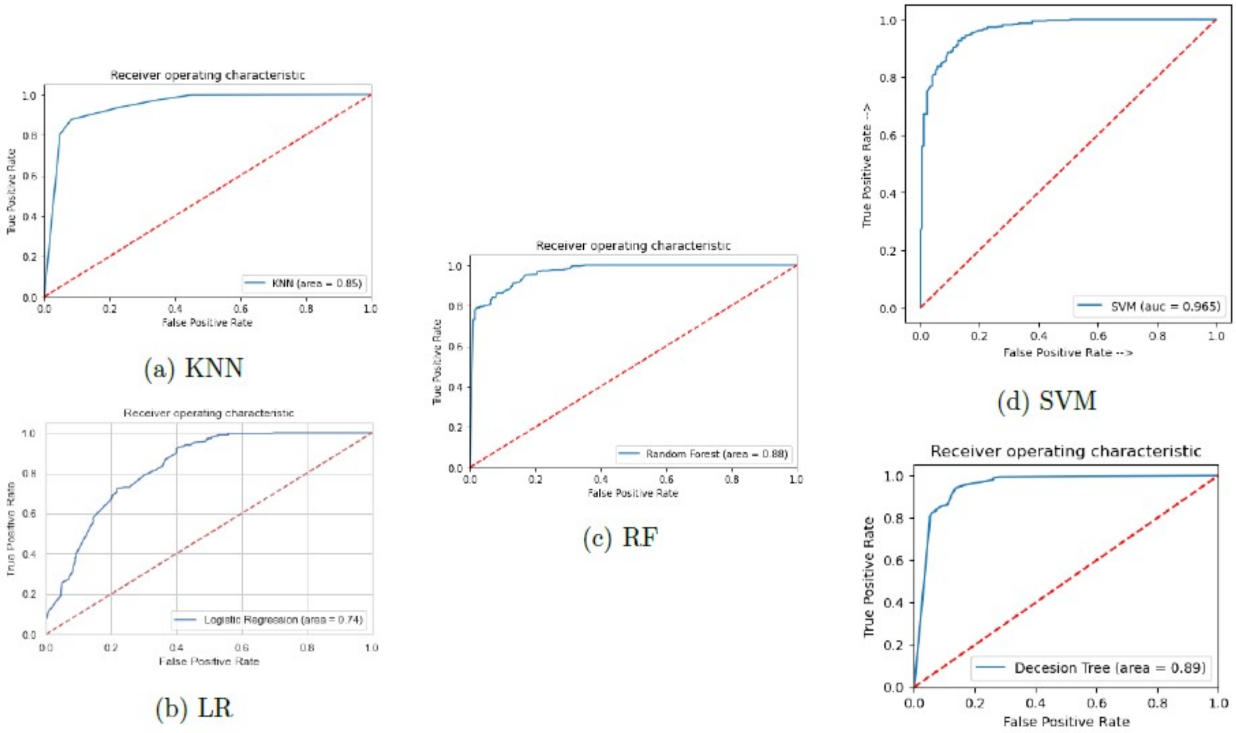


Figure 2: Correlation matrix of all data features

Figure 3



(a) KNN

(b) LR

(c) RF

(d) SVM

(e) DT

Figure 3: ROC and calculated AUC obtained for the classifier models of (a) KNN, (b) LR, (c) RF, (d) SVM, (e) DT