Asian Journal of Basic Science & Research
Volume 4, Issue 4, Pages 60-73, October-December 2022

Customer Churn Prediction Using Machine Learning Techniques: the case of Lion Insurance

Edemealem Desalegn Kingawa[1,2,*] & Tulu Tilahun Hailu[1,2]

[1]*Artificial Intelligence and Robotics Center of Excellence, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia.* [2]*Department of Software Engineering, College of Electrical and Mechanical Engineering, Addis Ababa Science and Technology University, Addis Ababa, Ethiopia.*
*Corresponding Author – (Edemealem Desalegn Kingawa): desuking1717@gmail.com**

DOI: http://doi.org/10.38177/AJBSR.2022.4407

Article Received: 15 November 2022 *Article Accepted: 18 December 2022* *Article Published: 26 December 2022*

ABSTRACT

The growth of an insurance company is measured by the number of policies purchased by customers. To keep the company growing and having more customers, the customer churn prediction model is crucial to maintain its competitiveness. Even if the company has good service delivery, it is important to identify the customer's behavior and be able to predict the future churners. The main contribution to our work is the development of a predictive model that can proactively predict the customer who will leave the insurance company. The model developed in this study uses machine learning techniques on lion insurance data. Another main contribution of this study is the labeling of the data using an unsupervised algorithm on 12007 rows with 9 features from which 2 clusters were generated using the K-means++ algorithm. As the cluster results found are imbalanced, the synthetic minority oversampling technique was applied to the training dataset. The Deep Neural Network algorithm turns out to be a very effective model for predicting customer churn, reaching an accuracy of 98.81%. The two years of customer data were obtained from lion insurance and used to train test, and evaluate the model. The Randomized optimization technique was selected for each algorithm. However, the best results were obtained by a deep neural network with a structure of (9-55-55-55-55-55-1). This algorithm was selected for classification in this churn prediction study.

Keywords: *Churn prediction; Clustering; Supervised machine learning; Deep neural learning; Sampling; Encoding.*

1. INTRODUCTION

Insurance is a crucial part of financial planning. The insurance industry is a large institution that protects the assets of organizations or individuals. Therefore, the insurance company is set up to provide important services to its customers. This depends on the premium paid by the customers [1]. So insurance is a way to manage risk. The insurance sector has become one of the main industries in both developed and developing countries.

Modern Insurance was started in Great London when a Fire accident happened in 1666, after destroying more than 30,000 homes; a man named Nicholas Barban began building insurance for the first time. He later introduced the city's first fire insurance corporation. Accident insurance was introduced in the late 19th century and was very similar to modern disability coverage at that time [2].

After 355 years, this type of service has become mandatory by the government. This is because of the enormous benefits that insurance offers to human beings. Except for Ethiopia and Liberia, most of the African countries have been colonized by other countries, especially Europe. It was a frustrating time for Africa, but it also brought new ideas to Africa. As a result, these colonial powers expanded infrastructures like railways, ports roads, and airports to facilitate their power. This new idea enables them to build and expand infrastructures to maximize their benefits. These highly invested infrastructures have brought to Africa a variety of policies and strategies that can be used by insurance companies to prevent risk [3].

Insurance has been a major competitor sector in Ethiopia and around the world. It was established in our country in 1984 G.C its name is known as Ethiopian Insurance Corporation (EIC). Currently, our country has grown to 14 insurance service providers in 37 years [4]. The increasing number of insurance providers raised the level of

competition; one of them is Lion Insurance which was founded in 2007 by 300 stakeholders. The company is one of the highest-earning companies in the private sector in Ethiopia which has initial paid-up capital of Birr 16 million and a subscribed capital of Birr 66.4 million. Currently, Lion insurance has more than 100,000 customers,41 branches, and 600 professionals [5]. The major vision of companies is to sustain for a long time. To achieve competitiveness, companies today strive to understand customer needs and strive to provide services effectively and efficiently that satisfy these needs [6]. Therefore, it is basic to support the insurance sector with different techniques. This paper will attempt to implement a customer's churning predictive model. Predicting customer churn helps the insurance sector by gaining a better understanding of future expected revenue and target to an attempt to prevent customers from discontinuing their subscription with an insurance company. And, since the cost of acquiring a new customer is 5x higher than keeping an existing one, there's plenty of revenue-based reason to do everything in its power to keep those existing customers [7].

2. STATEMENT OF THE PROBLEM

The insurance industry is a large financial service provider, with a billion birr income [8]. For it to grow, it is necessary to consider all the factors that lead to success and failure. In the business domain, machine learning increases the speed at which revenue can grow by using various models for various problems [9]. For this reason, the business model for machine learning that we have chosen is Customer Churn prediction in the case of Lion Insurance.

Here, customers are switching from one insurance company to another for a variety of reasons, and as a result, the company is being adversely affected. While this goes on for years, it is important to identify loyal customers to manage the business and increase the number of customers and make a profit. Insurance companies must retain existing customers or attract new customers. But the chances of that happening are difficult to achieve. According to many researchers, various financial industries are operating in a highly competitive environment [10].

However, it is difficult to identify between the churner and the non-churners using traditional and manual work due to a large amount of data. Therefore, the process of model building is a complex task. The predictive models provide a way for the insurance company to attract new customers and retain the existing loyal customers. Customer management and decision-making can play a significant role in improving your organization's image and attracting new customers. This gives an important strategy to keep customers and different strategic ways that must be considered by the insurance company.

Therefore, to solve this problem and address the gaps, it is necessary to know churners before they churn, so it seems very important to develop a model that predicts the future churners by applying machine learning techniques. This research also raises the following research questions:-

- Which customer characteristics are key to predicting customer churn behavior?
- How can unsupervised algorithms and classification algorithms be applied to customer churn prediction?
- Which churn prediction model generates the most accurate churn prediction results for the Lion Insurance Industry?

3. RESEARCH OBJECTIVES

The main objective of this research is to apply machine learning techniques to an existing motor insurance customer's data to obtain the best model which can predict churning customers. Specifically, this research aims to:-

- Reviewing concepts from literature and works related to customer churn prediction.
- Identify variables that are important to predict churn and non-churn.
- Perform data preparation tasks on the collected data for the correctness of the model building process.
- Develop the predictive models using different machine learning models.
- Evaluate the performance of the trained models to select the best model.
- Develop a prototype for the selected model.
- Provide concluding remarks and recommendations for further research works in this area.
- Identify limitations and future work on the proposed areas.

4. LITERATURE REVIEW

Machine learning is a mathematical algorithms concept intended to act like human intelligence by learning from the environment which was introduced by a man called Arthur Samuel, in 1952, in the new era of big data, machine learning algorithms are considered as a working horse [11]. Which have been broadly practical in a variety of areas such as recognizing patterns (PR), natural language processing (NLP), and computational learning (CL). Computers often do their job by comparing decisions with pre-written programs. Machine-learning models, once learned, can be used without any programming to do their jobs. Because ML models can make decisions based on their previous experience [12]. Developing models that can be learned from data or experience and making decisions or predictions is another feature of the machine learning model. In previous decades, machine learning (ML) has got an enormous inspiration in our everyday life, tasks including Traffic prediction, Fraud Detection, remote assistance, malware detection, exploit development and Product recommendations, Online Customer Support, Predictions while Commuting, and Videos Surveillance.

According to Quantilus [13], machine learning can be used as a gateway to big business ideas and opportunities by developing models from vast and complex data that can solve complicated problems. Enable business organizations to more quickly identify profitable opportunities and potential risks. This author explains in his article that machine learning techniques can be used for different purposes. For example, in transportation, online marketing, and financial services, Machine learning can help to adjust financial securities or assess risk for credit and insurance.

Based on a business perspective lion insurance company was selected for this study. The Lion Insurance company is private insurance in Ethiopia. Customers are the main source of profit for any industry since they are considered an important asset. Nowadays, companies need to focus not only on convincing the new customer but also on how to keep the customer loyal to themselves. Churners are people who move from one company to another for a variety

of reasons. To reduce customer churn, Companies need to address this problem by building machine learning models that accurately predict the behavior of their customers and making precautionary decisions. Customer churn is a binary classification problem that has two classes called churner and non-churner.

Customer churn is defined as the movement of people from one service provider to another. Identifying the cause of particular churn behavior and providing what the customer wants is also important for the intended purpose. Currently, it's a major challenging problem for business companies. Churn is an intention of customers to discontinue business with the company and move to another company within a certain period [14]. The customers who stop using the company's products are usually called churners [15]. Churn is also called attrition and is often used to indicate a customer leaving the service of one company in favor of another company [16]. The insurance domain defines churn as one who discontinues all his/her policy and stops renewing the policy [17]. Churners are mostly divided into two groups, Voluntary, and non-voluntary churners. Non-voluntary churn is the type of churn in which the service is purposely terminated by the company because of abuse of service and non-payment of services this type of churn can easily be identified.

Voluntary churn is even more difficult to identify. This type of churn occurs when a customer decides to terminate a contract with a service provider voluntarily. This can be categorized into two groups, incidental churn, and deliberate churn. Incidental churn occurs when circumstances prevent the customer from seeking further service. Examples of accidental churn include changes in the customer's financial situation or moving to another geographical location where the company does not operate. In deliberate churn, a customer purposely chooses a competitor. This kind of churn happens when a customer decides to change to another competitive company because of poor quality service and dissatisfaction, economic causes, and Technology causes.

Focusing on the insurance industry, Mr. Bereket Tilahun and other experts forwarded their opinion as to why a customer un-renewed his/her policy. Among the major reasons they stated for customer churn, some of these are the length of time taken to process claims, Lack of trust in the insurance company due to political and personal problems, poor performance and Lack of good image, and High premium fees value.

Acquiring a new customer is five to six times more than keeping existing ones is considered a difficult situation in many sectors [18]. Being able to predict the customer leaving the company is like creating a huge additional source of revenue. It also enhances the image of the company. This will greatly contribute to the growth of the company. The main purpose of customer churn prediction is to identify the behavior of customers who are highly inclined to leave a company. Therefore, for an insurance company to retain its customers it is important to know why they are leaving the company; this is obtained by extracting knowledge from the data gathered.

With the help of a machine learning algorithm, the movement of the customer can be reduced. Machine learning is an automated data analysis method or system for building an analytical model. It is the area of artificial intelligence that depends on the idea that systems learn from data, identify patterns, and make decisions with little human intervention. There are three machine learning categories: Unsupervised, semi-supervised, and supervised learning. Supervised Learning is the function of machine learning to discover the hidden Patterns from labeled data sets. Unsupervised learning is a machine learning task of discovering hidden patterns from unlabeled datasets; it is the

direct inverse of supervised approaches. The semi-supervised learning approach is one of the machine learning approaches which is practically learned faster, better, and cheaper. The semi-supervised learning approach falls between the supervised and unsupervised learning approach because it uses both small sets of labeled and large sets of unlabeled data [19].

5. RELATED WORKS

Kamala Kannan and Mayilvahanan [20] utilized data mining techniques to build Churn Prediction Model Using a Support Vector Machine algorithm. The experiment was conducted using a normalized k means algorithm for dataset preprocessing to eliminate redundant data and ensure quality clusters. The researcher also used the Min-Max normalization technique because the dataset is limited. The study used the SVM (support vector machine) algorithm by comparing processing time using SVM alone SVM with PSO. Accordingly, The SVM with PSO prediction model has much greater accuracy results than SVM alone. In terms of accuracy SVM with PSO performs well which is better than SVM alone. The dataset is IBM Watson Analytics Telco Customer Churn data which provides information on behavior to retain customers.

Spiteri and Azzopardi [21] proposed six classification algorithms which are decision trees, logistic regression, Naive Bayes, random forests, support vector machine (SVM), and survival analysis techniques to predict the customer churners in a motor policyholder dataset provided by Maltese motor insurance. The main aim of the study was to identify the risk factor associated with churn and build a model and survival analysis based on the data to indicate the time until churn. To do this, the researchers used the following techniques to extract only the important features. Brouta algorithm to remove irrelevant variables, then information gain, gain ratio, chi-square test of independence, recursive feature elimination, and the random forest algorithm was also used for extracting feature importance and ranking, then after the Boruta algorithm was applied to the data, and association rules were used to search for multiple independent characteristics that appear frequently in the dataset and create rules. The Boruta algorithm automatically discarded nationality and town description variables. The data used for this study was gathered from different sources so that data integration was done to reduce redundancy and data inconsistency. For the customer churn problem, the datasets are commonly Imbalanced as the majority of the policyholders usually decide to stay with their company.

Therefore the researchers applied both under-sampling and over-sampling techniques to balance the dataset. The total dataset used for this study was 72,445 for training and 4,186 for testing. The researchers put all of the above ideas together and then implemented classification algorithms on the dataset provided above. The 10-fold cross-validation was used during the implementation of the classification algorithm. Before that for the Support vector machine and Logistic regression categorical data was changed to numerical data. The experiment showed that the random forest model very effective technique to predict customer churn for a motor insurance company, reaching the best accuracy. The survival analysis was done on the data to model time until churn is approximately 90%. This means that most customers want to stay with the company and renew the contract.

Bellani [22] experimented on two insurance policies; third-party liability and comprehensive (kasko) policy. The specific dataset used for this study is from an insurance company in Portugal, as well as commercial, vehicle, policy

details, and external information from the census. On the dataset provided data cleaning, transformation, and reduction (especially, for redundancy) were implemented. The data was divided as 70:30 (training: testing). The models used to build the predictive models are generalized linear models, random forests, and artificial neural networks. The parameter tuning method was used to maximize model performance. The researcher builds two predictive models for a compulsory and comprehensive policy. An artificial neural network is recommended for a compulsory motor insurance policy. The first layer ANN model is 15 neurons and the second layer consists of 4 neurons, with an AUC of 68.72%, a sensitivity of 33.14%, and a precision of 27%. For the comprehensive motor insurance policy, the Random forest model was also suggested with 325 Decision trees AUC 72.58%, sensitivity 36.85%, and precision 31.70%.

Mohammad et al. [23] Conducted research to predict customer churn for telecommunication companies. The researcher proposed Multilayer Perceptron (MLP) neural network approach to predict customer churn in Malaysian telecommunication companies. The various variables are required to be extracted to build the model based on Customer Demographics, Customer Relationship Data, Billing Data, and Usage Data. The proposed Multilayer Perceptron (MLP) results were compared with the most popular churn prediction models such as Multiple Regression Analysis and Logistic Regression Analysis. Finally, neural networks have gained best results than the other statistical methods. The performance of the model was just based on accuracy, sensitivity, and specificity. The data was arranged in four sets. For the training phase, 78 churn data and 58 non-churn data were used. For the testing stage, 13 churner data and 10 non-churner data were used. MLP neural networks used nine algorithms.

Oyeniyi [24] designed a data mining model capable of clustering customer and churn prediction. The experiment was conducted using K-means clustering for customer clustering and the JRip algorithm to generate rule sets. Researchers used 10-fold to avoid training and testing on the same data that could lead to false results. The data was split as 66% for training and the remaining 34% data for testing. The Waikato Environment for Knowledge Analysis called WEKA tool was used both for classification and clustering including preprocessing tasks.

Swaminathan [25] experiments with clustering algorithms for their effectiveness. The patient data was used to experiment. Five cluster algorithms have been successfully analyzed, which are K-means, K-medoids, fuzzy c means, Hierarchical method, and DBSCAN. K-means analytics algorithm is suitable for customer churn analysis, which requires pre-processing steps for incorrect and missing values. When the number of clusters is increased K-medoids take more iteration. It is also more expensive than K-means, it computes all pair-wise distances. Fuzzy c means the cluster we want to define the number of iterations and clusters. The hierarchical method does not cluster all objects in a single step so it consumes more time. KNIME analytics platform was used for visualizations and analysis of those clustering algorithms. Based on the analysis result K-means was best over the rest.

a. Research gap

Detailed study of state-of-the-art approaches in predicting customer churn was performed for this study. Many researchers have tried to build churn prediction models using different machine learning techniques. Still, classifying customer churn as per their character is one of the challenges that need to be studied. The purposes of this study are to support insurance experts, insurance company and other financial sectors to make on-time

decisions for building a good image of their company, to attract new customers, and retain the existing customers. To develop a good prediction model, it is challenging to understand the reasons that make customers churn. Therefore, cluster-based labeling has been proposed before and in recent times. Researchers have extensively used the usual pattern without looking at customer data. Customer data is highly significant, depending on the sharpness of the study. This is because it requires the good analytics and clustering algorithms to accurately group customer data. The proposed algorithms have more valuable than previous techniques. Hence, the deep neural network wills tune parameters for model performance optimization, Modified version of K-means used for clustering customer data, Cross validation to prevent model over-fitting and under-fitting, Missing data will be handled using imputation techniques and SMOTE focused on data balancing which is oversampling techniques.

6. RESEARCH METHODOLOGY

According to Kamiri and Mariga's [26] study, Most of the machine learning research was done by a quantitative research approach with an experimental research design. Researchers use more than one algorithm to solve a problem. The research design for this study is the Experimental method due to using more than one algorithm.

a. Data Collection

Data collection is a process of collecting and organizing data for an intended purpose and using it as an input for a predetermined purpose. It is also used to identify the type of problem by feeding the data. Artificial Intelligence is a process of solving problems, using a vast amount of data. So the data collection process is like the backbone of machine learning and artificial intelligence. This study looked at motor insurance customer data like Type of policy, type of cover, vehicle brand, year of make, number of risks, carrying capacity, purpose, the time needed for processing claim, and Premium. The dataset for this study would be collected from Lion Insurance which is located around the Black Lion Hospital. As listed above the dataset might include a type of policy, Carrying Capacity in quintal, brand, type of cover, premium price, Number of risks, Year of Make, etc. There are two types of data collection methods called primary and secondary. The secondary data collection method was used for this study.

b. Data Pre-processing

The data obtained for this study need to be properly prepared and standardized to ensure a quality prediction model. Data preprocessing plays a big role in building quality models because algorithms are dependent on the quality of the data that they operate on. If the data is insufficient and inappropriate, machine learning algorithms might result in less accurate and less understandable results [27].

Therefore, data preprocessing performs the most important job for us by enabling us to build quality models. It includes missing data handling, Encoding, data transformation (often normalization and standardization), clustering the Data, Resampling, and feature selection. The first two steps are useful for a more accurate and complete dataset, and the third one is typically used to have more uniformly distributed data and to minimize data variability because transformed data can be easily used by both people and computers.

The existing data is unlabeled so it is not suitable for machine Learning. Therefore, it is important to use an unsupervised algorithm to label the dataset. For labeling, the data modified version of K-means which is the

Kmeans++ algorithm was employed to group the data based on their similarities. The Lion insurance data was highly imbalanced. An imbalance Class occurs when the values in one class are a small number of values in another class. Prediction models with imbalanced data are biased toward the majority class; therefore, the model produces an incorrect result. This is because the number of customers who renew the contract is not the same as the number of customers who switch to another company. To overcome this problem this study used Oversampling technique for the dataset. Finally, the fourth step is used to obtain required independent variables which have more relation with the dependent feature and help in building a good model with the ExtraTreeClassfier algorithm, which was used to rank features according to their importance.

c. Architectural Design

In this study, various ways we're involved in the machine learning cycle and major steps being carried out from the transformation of data to model building. Overall architectural design is put below in Figure 1.

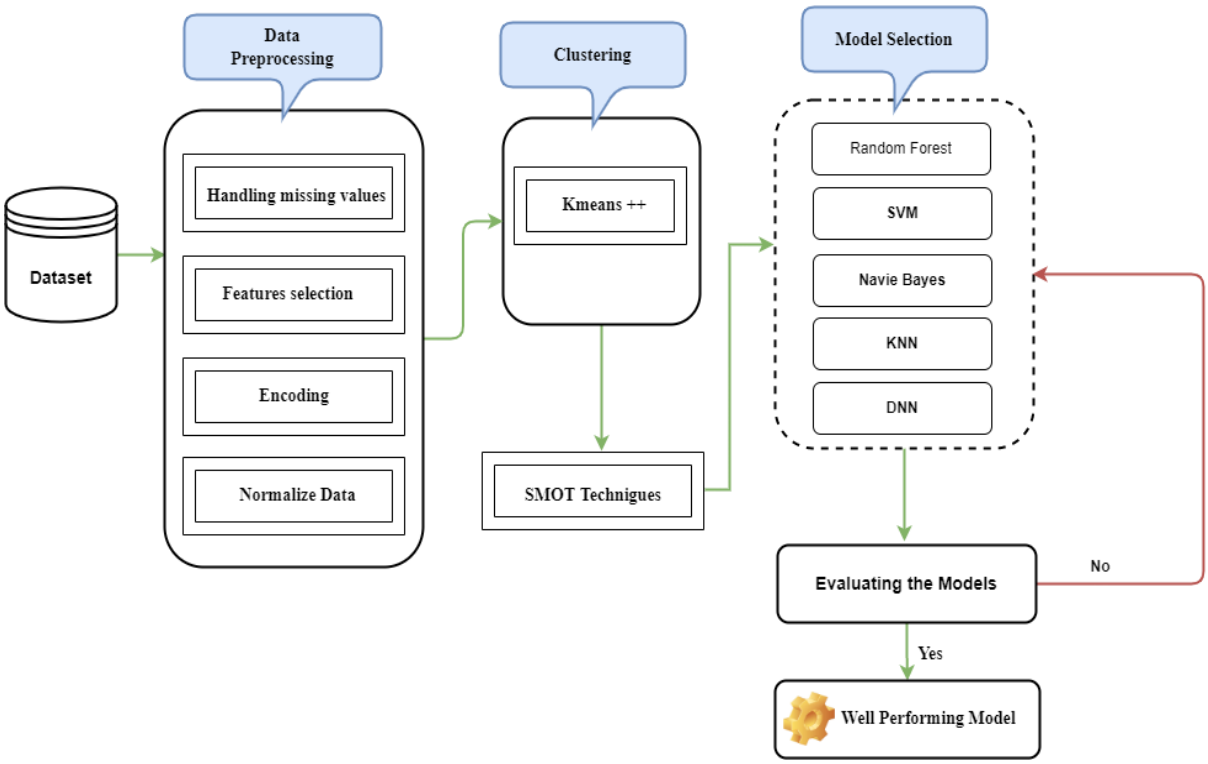

Figure 1. Customer Churn Framework

7. RESULTS AND DISCUSSIONS

A. Data Collection

The data used for this study is collected from the database of a lion insurance company. Motor insurance is the most important type of insurance sold in developing countries to protect the public. The Lion Insurance Customers data contains 25 variables; which contains information of 12207 members of lion insurance who have joined the company from 2013 to 2014 to get motor insurance services. However, most variables are not necessary for this study. Features were selected. It has been observed that there were six columns with the name of Branch code, Serial number, Name of insured, Business Sources, plate Number, and engine number, and so all columns were

dropped out from the dataset using the python drop command. As a result, the variables are reduced. The variables selected for this study are 9 in number. They are the type of policy, type of cover, vehicle brand, type of body, make the year, Number of risks, Carrying capacity of the car, Purpose of the car, and Premium.

B. Data Pre-processing

The Data Preprocessing section included cleaning missing values, Encoding, Normalizing the data (scaling), feature selection, cluster, sampling, and finally data splitting.

In this study, there was one variable - "Number of Risks" missing values over 4%, and these were zero-filled. This is because this missing value indicates customers that are free from risk. Other categorical variables such as type of cover, vehicle brand, type of body, and car purpose missing values were filled by mode. Except for the number of risk variables other numerical variables' missing values were imputed by using the mean.

For the sense of a common distribution, the dataset was scaled to a common scale by applying the MinMax normalization technique. The scaling technique was employed specifically on an individual column. Due to the value of the columns varied in min and max value. A correlation matrix has been used to find the relationship between independent and dependent variables and the issue of multicollinearity was detected and the most important feature has been identified in predicting the outcome. In addition to the correlation matrix, another feature selection method was used based on a previous study [28] to find the important predicting features. This method was also identified using sklearn.ensemble ExtraTreeClassifier function to identify the important predicting features. A graph of feature importance was plotted as shown below:

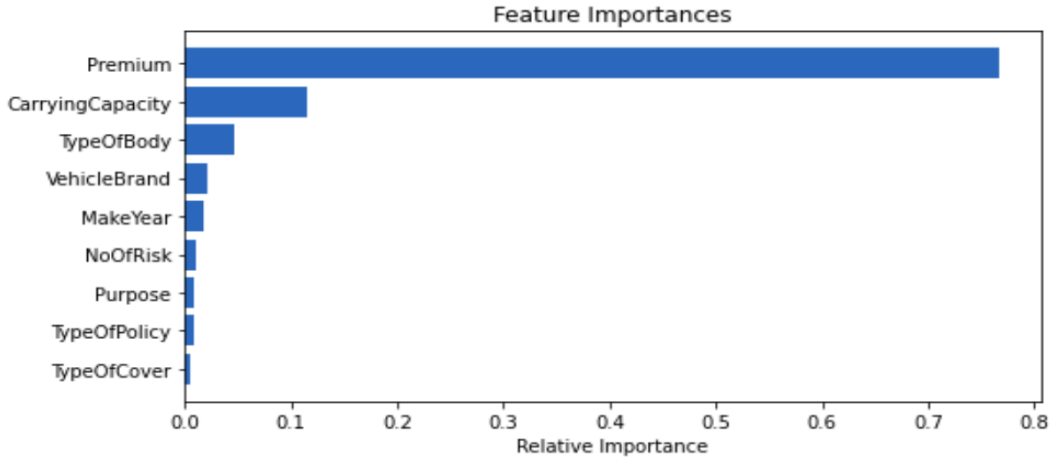

Figure 2. Feature importance graph

As seen from the above graph also the most important feature in predicting the target was ‘Premium’ followed by ‘carrying capacity, ‘TypeOfBody’, and so on. K-means is one of the most popular and significant clustering methods exhibited by Mc. Queen in 1967. The data obtained from Lion Insurance is not labeled. To label, the data an unsupervised Kmeans algorithm was applied. When this algorithm is used, two classes were created. So to decide which data point is most likely to leave the company and which data points indicate the retaining one, the advice of an insurance professional was asked and labeled the dataset accordingly based on their comment and suggestions. For Experimentation modified Kmeans++ was used. From sklearn.cluster import kmeans_plusplus

library was imported and call kmeans = Kmeans (n_clusters=2) function.

C. Machine Learning

Experiments were conducted to find a suitable machine learning model. The most common data problem in machine learning is a class imbalance, The Synthetic Minority Over-Sampling Technique (SMOTE) was used on the training dataset [29]. By using these techniques, the data was balanced. The 10-fold cross-validation technique was used for data splitting. For evaluation, accuracy, precision, recall, and F1 scores were computed from the confusion matrix. For the development of the model, the Randomized Search optimization technique was used for hyper-parameter tuning for Deep Neural Network and other machine learning algorithms.

The class distribution of the cluster is shown in the figure below. The target churn class consists of 907 churners and 11300 for non-churn.

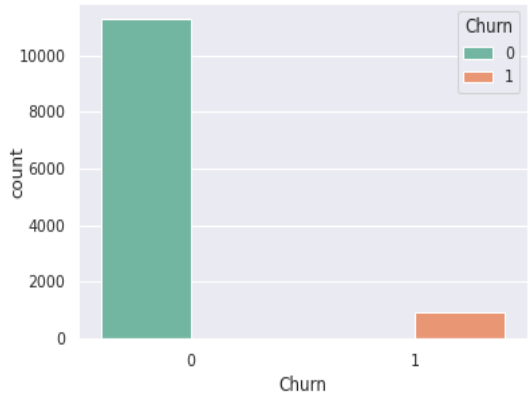

Figure 3. Before SMOTE

Figure 4. After SMOTE

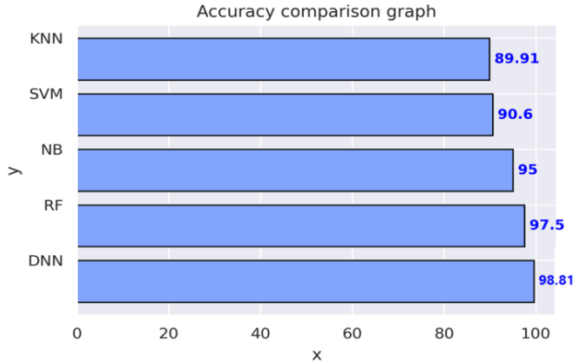

Figure 5. Accuracy Comparison graph

In the experiments, performed in terms of accuracy the deep neural network algorithm has outperformed the Random Forest, SVM, Naive Bayes, and KNN. The graph above was a comparison graph based on the accuracy of the supervised machine learning algorithms used in this study - Deep Neural Network, Random Forest, SVM, Naive Bayes, and KNN.

Table 1. Results of Supervised Machine Learning Models

Evaluation Metric	RF	**SVM**	**Naive Bayes**	**KNN**	**DNN**
Accuracy	97.04%	90.6%	95%	89.91%	98.81%
Precision	100%	93.72%	59.60%	36.80%	86%
Recall	60.0%	90.6%	100%	51.11%	100%
F1 Score	74.99%	91.77%	74.06%	42.79%	92.30%

As indicated in Table 1: The deep neural Network has generated better results when compared with Random Forest, Support vector machine, KNN, and Naive Bayes algorithms.

D. Discussion

As seen in the above Results the Deep Neural Network (DNN) model has the highest accuracy of 98.81%. So this model was chosen as the best model for predicting customer churn in this Lion Insurance dataset. In general, the number of raw data used for this model development process is 12007. This means that the more data there is, the more Deep Neural Learning algorithm outperform because the more data there is, the machine learning algorithms will fail. Thus, for this study deep neural network models were selected for the lion insurance customer dataset. The parameter selection for Deep Neural Network was done using Randomized Search CV() searching algorithm, based on the best-tuned parameters like number of Neurons, activation function, optimizer, batch size, and Epoch. For this study, to determine the number of hidden layer rules thump was used which is 2/3 * the sum of the input layer and output layer was used.

Premium, type of the body, vehicle brand, make the year, type of policy, type of cover, number of risks, Purpose, carrying capacity, and Churn are the basic variables to build a customer churn prediction model.

Some of the literature reviews for this study, in a few of the literature, indicate that the support Vector machine has been a good result. In one of the research conducted on online Auto insurance services by Hur and Lim (2005), the Support Vector Machine has achieved more promising results than the logit model and Artificial Neural Network.

In another research focus Spiteri and Azzopardi (2018) on motor insurance customer datasets, the Random Forest has outperformed Naive Bayes, Decision tree, logistic regression, and Support Vector Machine. In this research, the experimentation was focused on both imbalanced and balanced datasets. In both cases, the random forest achieved the highest accuracy.

In another research by Seymen, Dogan, and Hiziroglu (2021) conducted on retail industry datasets, deep learning

has achieved better classification and prediction results than logistic regression and artificial neural network models.

Thus, a Deep Neural Network model built using the lion insurance customer data, will achieve high accuracy (98.81%) than the other supervised machine learning algorithms like Random Forest, Support Vector Machine, K-nearest neighbor, and Naive Bayes, to predict the customer churn.

8. CONCLUSION AND RECOMMENDATION

A. Conclusion

The study of the machine learning approach for customer churn prediction is presented, and a churn prediction model based on machine learning is proposed for Lion Insurance Company. The dataset for the experiment was collected from Lion Insurance Branch which is located around black lion hospital. The dataset includes customer pieces of information such as type of policy, type of cover, premium, vehicle brand, type of body, make the year, number of risks, start date, end date, plate number, and carrying capacity from 2013 to 2014. Before conducting several experiments on the dataset Kmeans ++ algorithm was applied on it to label the dataset. Before conducting several experiments on the dataset Kmeans ++ algorithm was applied on it to label the data. After the data was ready for the classification algorithms, the experiments were conducted with selected machine learning algorithms, and comparisons were performed to choose the best model based on its performance. The experimental results show that the Deep Neural Network algorithm produces a promising result. As a result, the Deep Neural Network model is selected for predicting the churning customers.

B. Recommendation and Future work

Some future work was identified throughout this study, which may be carried out. Here in this research data was collected only from one branch of lion insurance. In the future, another branch of the lion insurance data can be explored and analyzed. In this regard, information from other insurance companies is not included in this study. Inclusion was good but was not possible due to time and resource constraints. But in the future, it could be added and expanded. More study is needed to handle DateTime data type variables. The time needed to process claim requests and the reason customers switch the company should also be included in future work.

The Five machine learning algorithms were used in this study on the lion insurance data. Further other deep learning algorithms can be explored as well. For the future, we are interested in train and testing our model on all lion insurance branches and other insurance institutions to predict churners. The model is also recommended to train and test on a large number of data with complex configurations.

In addition to this, different clustering algorithms can be applied and evaluated based on their evaluation metrics, the best should be selected. To apply this research inside the lion insurance company, we recommend developing a web-based application that takes excel and CSV files and gives predicted results of churners and gives helpful expert advice to the company.

Due to resources and time constraints, the findings of this study did not test the data of other insurance companies. It could also be interesting to including features, such as premium, of the other competitors.

Declarations

Source of Funding

This research did not receive any grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing Interests Statement

The authors declare no competing financial, professional, or personal interests.

Consent for publication

The authors declare that they consented to the publication of this research work.

References

[1] Insurance Information Institute, Insurance Handbook- An insurance guide: what it does and how it works. 2010.

[2] S. R. Corporate, A History of Insurance in Japan, p. 4, 2017, [Online]. Available: www.swissre.com.

[3] D. J. Kuss, M. D. Griffiths, J. F. Binder, and B. Street, Metadata, citation and similar papers at core.ac.uk, pp. 1–19, 2013.

[4] I. Institutions and O. In, Insurance Companies Name & Address. 2012.

[5] Yitagesu kebede, establishment of lion insurance, 2007. https://www.anbessainsurance.com/profile.php.

[6] M. Harris and H. J. Harrington, Service Quality in the Knowledge Age, Measuring Business Excellence, vol. 4, no. 4, pp. 31–36, Jan. 2000, doi: 10.1108/13683040010362562.

[7] P. E. Pfeifer, The optimal ratio of acquisition, Journal of Targeting, Measurement & Analysis for Marketing, vol. 13, no. 2, pp. 179–188, 2005.

[8] M. Negash, K. Venugopal, and S. Asmare, Identifying and Analyzing of Factors Contributing for Growth of Non–Life Insurance Gross Premium a Developing Country Perspective: Case of Insurance Industry in Ethiopia, SSRN Electronic Journal, Jan. 2018, doi: 10.2139/ssrn.3430324.

[9] A. Punoo, M. Otta, M. Fayaz, and S. Abdul Khader, Machine Learning in Insurance. 2022.

[10] D. R. Biggar, Competition and Related Regulation Issues in the Insurance Industry, SSRN Electronic Journal, vol. I, no. June, 2005, DOI: 10.2139/ssrn.185068.

[11] I. El Naqa and M. Murphy, What Is Machine Learning?, 2015, pp. 3–11.

[12] Thomas H. Davenport, What are some popular machine learning methods?, 2019. https://www.sas.com/en_us/insights/analytics/machine-learning.html.

[13] Quantilus, Machine Learning Important and How will it Impact Business?, 2020. https://quantilus.com/why-is-machine-learning-important-and-how-will-it-impact-business/.

[14] V. Bhambri and I. Research Scholar, Singhania University, Pacheri Bari, Jhunjhunu, Rajasthan, Data Mining as a Tool to Predict Churn Behavior of Customers, vol. 2, pp. 85–89, 2012.

[15] G. Nie, W. Rowe, L. Zhang, Y. Tian, and Y. Shi, Credit card churn forecasting by logistic regression and decision tree, Expert Systems with Applications, vol. 38, no. 12, pp. 15273–15285, 2011.

[16] A. Sharma and P. Kumar Panigrahi, A Neural Network based Approach for Predicting Customer Churn in Cellular Network Services, International Journal of Computer Applications, vol. 27, no. 11, pp. 26–31, 2011.

[17] C. Huigevoort, Customer churn prediction for an insurance company, no. April, p. 99, 2015.

[18] A. Ayache, M. Calciu, M. Fradon, and F. Salerno, Analytic solution to find optimal balance between customer acquisition and retention spending. Solution analytique pour trouver le meilleur équilibre entre les dépenses d'acquisition et rétention de clientèle.

[19] S. Vluymans, Multi-label Learning, Studies in Computational Intelligence, vol. 807, pp. 189–218, 2019, DOI: 10.1007/978-3-030-04663-7_7.

[20] T.Kamalakannan, P.Mayilvahanan Efficient Customer Churn Prediction Model Using Support Vector Machine with Particle Swarm Optimization, International Journal of Pure and Applied Mathematics, vol. 119, no. 10, pp. 247–254, 2018.

[21] M. Spiteri and G. Azzopardi, Customer churn prediction for a motor insurance company, 2018 13th International Conference on Digital Information Management, ICDIM 2018, 2018, pp. 173–178, 2018.

[22] C. Bellani, Predictive Churn Models in Vehicle Insurance, 2019, [Online]. Available: http://hdl.handle.net/10362/90767.

[23] M. R. Ismail, M. K. Awang, M. N. A. Rahman, and M. Makhtar, A multi-layer perceptron approach for customer churn prediction, International Journal of Multimedia and Ubiquitous Engineering, vol. 10, no. 7, pp. 213–222, 2015, DOI: 10.14257/ijmue.2015.10.7.22.

[24] O. A and A. A. B, Customer Churn Analysis In Banking Sector Using Data Mining Techniques, African Journal of Computing & ICT, vol. 8, no. 3, pp. 165–174, 2015.

[25] I. Franciska and B. Swaminathan, Churn prediction analysis using various clustering algorithms in KNIME analytics platform, Proceedings of 2017 3rd IEEE International Conference on Sensing, Signal Processing and Security, ICSSS 2017, pp. 166–170, 2017, DOI: 10.1109/SSPS.2017.8071585.

[26] J. Kamiri and G. Mariga, Research Methods in Machine Learning: A Content Analysis, International Journal of Computer and Information Technology(2279-0764), vol. 10, pp. 764–2279, Mar. 2021.

[27] S. B. Kotsiantis and D. Kanellopoulos, Data preprocessing for supervised leaning, vol. 1, no. 2, pp. 1–7, 2006.

[28] M. R. Khan, J. Manoj, A. Singh, and J. E. Blumenstock, Behavioral Modeling for Churn Prediction: Early Indicators and Accurate Predictors of Custom Defection and Loyalty, 2015 IEEE International Congress on Big Data, pp. 677–680, 2015.

[29] D. Elreedy and A. F. Atiya, A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance, Information Sciences, vol. 505, pp. 32–64, 2019.