Journal of Theoretical and Applied Information Technology
30th November 2020. Vol.98. No 22

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

A PROPOSED MODEL TO PREDICT AUTO INSURANCE CLAIMS USING MACHINE LEARNING TECHNIQUES

[1*] SHADY ABDELHADI, [2] KHALED ELBAHNASY, [3] MOHAMED ABDELSALAM

[1]Department of information system, Faculty of Commerce & Business Administration, Helwan University, Cairo, Egypt

[2]Professor of Artificial intelligence, Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

[3]Department of information system, Faculty of Commerce & Business Administration, Helwan University, Cairo, Egypt

E-mail: shadyabdelhadi99@gmail.com

ABSTRACT

One of the main challenges facing the insurance companies is to determine the proper insurance premium for each risk represented by customers. Risk differs widely from clients to another, and a Careful understanding of various risk factors assists predict the likelihood of insurance claims based on historical data, Real-world datasets often have missing values, can cause bias in results. the most widely adopted methods for dealing with missing data is to remove observations having missing values, perform a complete case analysis (CCA) and single imputation such as average. these approaches have the disadvantages represent in loss of precision and biased. The main objective of the paper is to build a precise model to predict car insurance claims through machine learning techniques. with a focus on advanced statistical methods and machine learning algorithms that are the most suitable method for handling missing values. we Used available datasets through Kaggle which consists of 12 variables and 30240 cases, the research was carried out by using Artificial Neural Network (ANN), Decision Tree (DT), Naïve Bayes classifiers, and XGBoost to develop the prediction model. The experimental results showed that the model obtained acceptable results The XGBoost model and Resolution Tree achieved the best accuracy among the four models, with an accuracy of 92.53% and 92.22%, respectively.

Keywords: *Machine Learning; Prediction Model; Missing Data; Auto Insurance Claims*

1. INTRODUCTION

Claim prediction is an important operation in the field of the insurance industry because the insurance companies can recommend the right type of insurance policy for each potential policyholder. Risk varies widely from clients to another, Inaccuracies in the prediction of car insurance claims raise the price of the insurance policy for the good driver and decrease the price of the policy for the driver who is not good [1]. by that low-risk customers pay for the damage and loss caused by high-risk customers, so there is no difference between these two groups of customers.

The rate of the insurance premium in many insurance companies are calculated with consideration to different demographic factors, car specifications, and the record of damage caused by the car owner [2].

Mike Kreidler, a member of the board of directors of one of the insurance companies in Washington State pointed out that the rate of insurance premium depended on factors such as the policyholder's age, marital status, and gender, vehicle type, the location of the car owner's residence, the driving pattern, and the claim history [3]. While in Egypt the rate of car insurance premium is dependent on two factors first one price of the car and the second rate of loss. lack of such factors in determining the risk in car insurance leads to computing unfair rates of Insurance premiums because in these cases instead of the customer, the vehicle is insured. That's why most insurance companies experience great loss as far as car insurance as shown in Figure 1, one of the main challenges face the insurance companies nowadays, is to define a proper premium for each risk represented by those customers [4], the majority of insurance companies keep the data on the history of its operations in a data warehouse

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

These huge quantities of data are hiding very important knowledge, which could contribute to increasing profitability, these historical data provide the greatest source of information on claim exposure and is the starting place for insurance claim modeling, in this context, machine learning mainly contributes to creating more accurate predictive models to solve such problems.

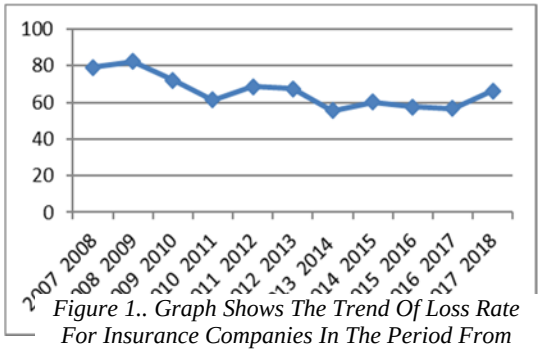

Figure 1.. Graph Shows The Trend Of Loss Rate For Insurance Companies In The Period From 2007 To 2018 In Egypt [5].

Missing data are a frequent complication of any real-world study. Missing data treatment is an important data quality issue in machine learning, missing data result in a loss of precision and are also a source of bias if observations are not missing completely at random (MCAR). The most widely adopted strategy for dealing with missing data is to delete observations having missing values and perform a complete case analysis (CCA) or use single imputation such as mean, most frequency to impute the missing values, and single stochastic regression [6], result in highly biased estimates when more than 10% of subject's data missing data, and standard errors are underestimated because missing data uncertainty is not incorporated [7].

The multiple imputations are an alternative approach dealing with missing values are imputed depended on the observed data. This technique is an advanced statistical method to deal with missing values as it has the potential to correct the bias in the complete case, alternative analyses, and dealing with the attendant uncertainty of the imputations themselves [8] [9] [10].

The main objective of the research is to build a precise model to predict car insurance claims through machine learning techniques to help insurance companies to improve their pricing decision. with a focus on a proposed approach to Handle missing data using advanced statistical methods and machine learning algorithms that are the most suitable method for handling missing values. We also compare the accuracy of the Decision tree with XGboost, Neural Network, and Naïve Bayes classifiers.

This research consists of five sections. The introduction is followed by Section two, the review of the literature. The methodology is discussed in Section three that separately describes the phases of the proposed model. which in itself comprises three sub-sections presenting data description, data preprocessing, and model. The experimental result is discussed in Section four and finally, the conclusion for further research are offered.

2. LITERATURE REVIEW

In this section, sets of research efforts from knowledge discovery process and machine learning techniques are reviewed. there have been several papers that tackled the problem of claim prediction using data mining, at variance using machine learning techniques.

"Predicting motor insurance claims using telematics data" by Jessica Pesantez- Narvaez was proposed in 2019. This research compared the performances of logistic regression and XGBoost techniques to predict the existence of accident claims with a little number of data training, their results showed that logistic regression is an appropriate model given its interpretability and good predictive ability. But, XGBoost needs effort as regards the interpretation and numerous model-setting procedures to compare with the logistic regression model [11].

In the work by Ranjodh Singh, and et al.in 2019, proposed a system, this system takes images of the damaged car as inputs and produces relevant information such as the cost of repair which would be used in deciding insurance claim amount and, damage localization. Therefore, the predicted car insurance claim they were not considered in this study, but their focus on estimates the cost of repair [12].

Oskar Sucki 2019, The goal of this research is to study the churn prediction. it was found that random forests the best performing model (74% accuracies). The dataset had missing values in multiple fields. After looking at the distributions, it was decided to replace the missing variables with extra attributes that would indicate not having that information [13]. This is only allowed in the event of completely random data loss, so it was necessary to first determine the missing data mechanism by which the appropriate approach to data processing is determined [8][9].

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

Muhammad Arief Fauzan et al in 2018. In this paper, they apply the accuracy of XGBoost for predicting claims. also, compare the performance a set of techniques i.e., AdaBoost, Random Forest, Neural Network, with the performance of XGBoost.

XGBoost gives better accuracies in terms of normalized Gini. Use publicly available datasets from Porto Seguro through Kaggle. There are massive amounts of NaN values in the dataset, in spite of that, this paper handle missing values using mean substitution and median. However, these unprincipled, simple approaches have also been shown to lead to biased results [9]. Therefore, they focus on examining the machine learning techniques that are the most suitable method, such as XGboost for the problem of many missing values [14].

G. Kowshalya, M. Nandhini.in 2018. in this study, data mining techniques are used to predict fraudulent claims and to calculate insurance premium amount for different customers based on their personal and financial details, three classifiers were built to predict fraudulent claims and percentage of the premium amount. The algorithms Random forest, J48, and Naïve Bayes are selected for classification. Depended on the synthetic dataset, the results show Random forest outperforms the remaining techniques. Therefore, this paper does not cover predicting insurance claims, but they focus on fraudulent [15].

Predict whether a customer filed an insurance claim has been proposed forth by the author Matthew Millican et al in 2017 [16]. they use least squares ridge regression, least-squares lasso regression, logistic regression, Naive Bayes, random forests, gradient boosting, and another. Also, there are amounts of NaN values in the data. Replace these missing data with the mean of the feature over all examples in the dataset., mean imputation is convenient because it produces a complete data set. However, convenience is not a compelling advantage because this approach severely distorts the resulting parameter estimates, even when the data are MCAR [17]. Furthermore, Mean imputation methods result in highly biased estimates in all missing data situations when more than 10% of subject's data missing data [7].

Tim Pijl,2017. Applied knowledge discovery technique to construct a framework a step-by-step guide to forecast insurance claims. The results show that dimensionality reduction is not necessarily needed for this problem and those simple techniques,

such as a decision tree or random forest, outperform the more statistically advanced techniques, such as a support vector machine, also use small data set. However, most of the aspersed methods need large amounts of labeled data. Therefore, a much bigger dataset is needed since their dataset is too sparse and this would lead to very volatile results. [18].

Dan Huangfu,2015. the aim of this project was to compare the performances of various statistical models and methods on predicting the bodily injury liability insurance claim payments based on the characteristics of the insured's vehicles in a particular data set, also the data set includes a large number of missing values for the categorical variables. However, the paper didn't handle missing values. Therefore, they focus on examining which machine learning methods to overcome missing data [4].

The above previous works did not consider both big volumes of data and missing values issues in their works but, they depended on common and traditional methods Which studies have proven incorrect. consequently, we focus on advanced statistical methods and machine learning algorithms that are the most suitable method for the problem of claim prediction with many missing values.

3. MATERIAL AND PROPOSED MODEL

3.1. DATA DESCRIPTION

To build the claim predictor, we obtained the data set through the Kaggle site [19]. The training data is used to build a model as a predictor of probabilities a person will file a claim next year. the dataset consists of 12 variables,30240 cases. There some missing values in Years' Experience.

Table 1.Description of the Dataset

Name	Description
Target [class label]	whether or not a claim was filed for that policy holder last year
Age	Age of the client
Gender	Male / Female
Engine power	Engine power or horsepower
Credit history	It's a three-digit figure that represents your history of borrowing
Years' Experience	the more driving experience
Annual claims	average annual claims in past
Marital Status	marred / single
Vehicle type	type of car owned by the client [car , Van , Truck , Utility]
Miles driven	total miles driven by client
Size of family	family size of Clint
State	Countries of the client

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

3.2. PROPOSED MODEL

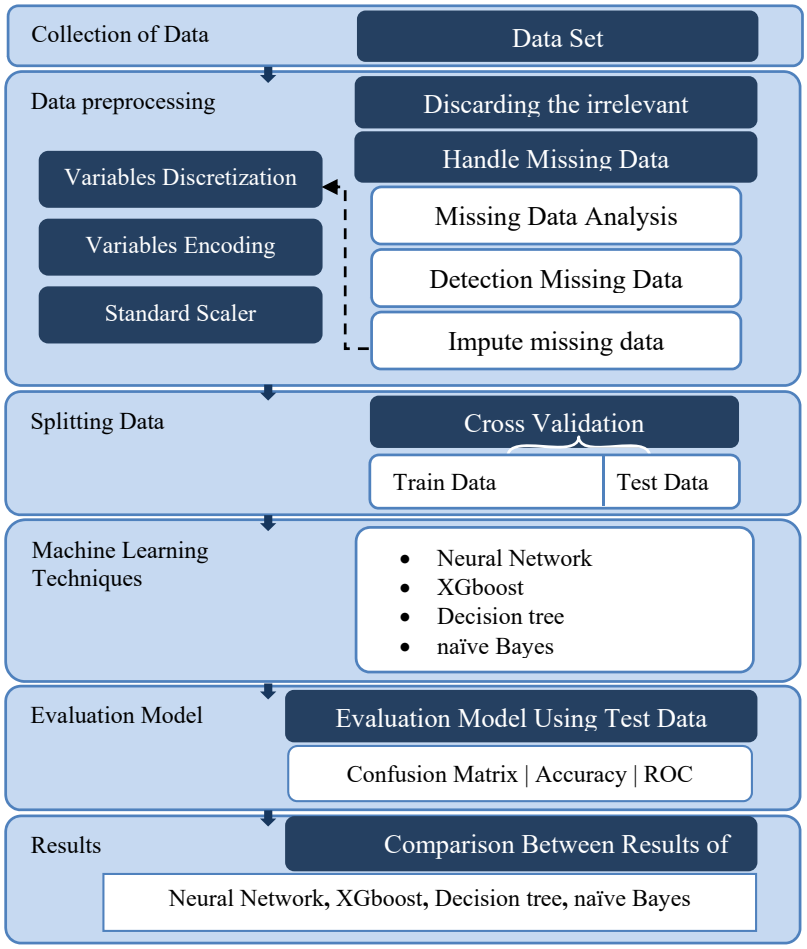

Figure. 2. Overall structure of the proposed model

In this research, we designed a model to predict car insurance claims by applying Machine Learning Techniques (MLT) based on the customer's data show in figure (2). phase to prepare the proposed model consists of (1) Data collection phase, (2) Data preprocessing the data before applying the MLT, (3) data splitting into data training and data testing, (4) Selection of classification models, (5) evaluation phase to evaluate the accuracy of the built model using a machine learning technique.

3.2.1. DATA PREPROCESSING

To improve the predictive effect of our proposed model, the raw data, which are often there missing values, inconsistent, Therefore, it is important to preprocess the data before developing the predictive model. The following steps have been done to achieve enhancing in this section.

3.2.1.1. HANDLE MISSING DATA

There are many missing values existing in the dataset about driver's information, the following is an analysis of missing data using SPSS tool.

1) **MISSING DATA ANALYSIS**

The dataset contains not a few numbers of missing values for the quantitative variables. There are 605 cases missing value in one column called years' experience figure 2 and 3 shows the counts and percentage of missing values.

On the level of the variables, Figure 3 shows that there is Univariate among 12 variables that contain missing values. As for the level of the rows, there are 605 rows that contain missing values from 30,240 rows, and there are 605 cells out of 362,275 cells that have missing values.

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

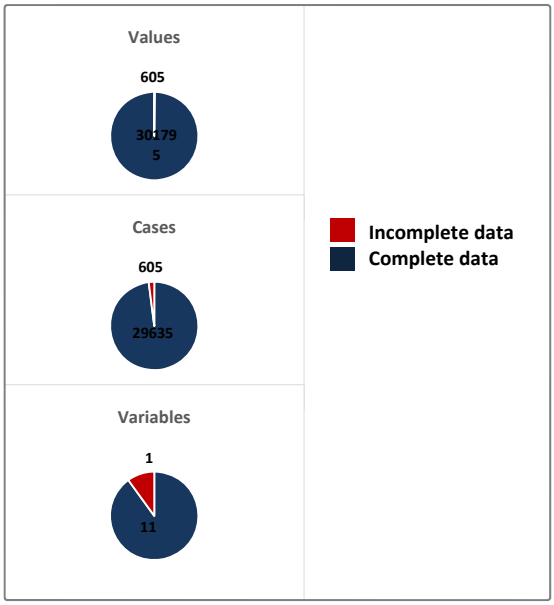

Figure 3. Overall summary of missing data

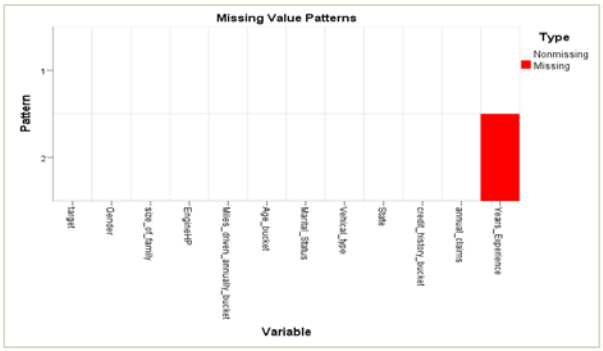

Figure.4. overall missing value patterns

The patterns chart presents missing value patterns for the analysis variables. Each pattern match to a group of cases with the same pattern of incomplete and complete data. Figure 4 shows, Pattern number one represents cases that have no missing values, while Pattern number two represents cases that have missing values on years' experience variable.

2) DETECT MISSING DATA MECHANISM

Rubin and et al, introduced a framework to classify missing data problems. This research has produced three Concepts so-called missing data mechanisms that illustrate how the probability of a missing data relates to the observed data, [20] missing at random (MAR) refers to that a methodical relationship exists between one or more measured attributes and the probability of missing values. missing completely at random (MCAR) it assumes that missing data is completely unrelated to the observed data, finally missing not at random (MNAR) when the probability of missing data on a variable Y is related to the values of Y itself [21].

In this step, we apply Little's missing completely at random test to see if the data was completely randomly lost or not, to determine the mechanism by which the data was lost based on it we will determine the best way to handle the missing data. The following is the formulation of hypotheses:

$$H_0 = MCAR$$
$$H_1 \neq MCAR$$

Table 2..results of Little's MCAR test

EM Covariances[a]

	target	size_of_family	EngineHP	Years Experience	annual claims
target	.248				
size of family	.007	5.228			
Engine	-8.074-	-.507-	17515.718		
Years Experience	-.421-	.066	-429.273-	97.704	
Annual claims	.020	-.001-	36.012	-4.257-	1.173

a. Little's MCAR test: Chi-Square = 294.882, DF = 4, Sig. = .000

which indicates the refusal of the null hypothesis and the acceptance of the alternative hypothesis, meaning that the data is not missing completely at random.

The practical problem with the Missing At Random (MAR) mechanism is that there is no way to confirm that the likelihood of missing values on Y is solely a function of other measured variables [21], but using analyzing data in SPSS (tabulated cases), we found that the missing data in the years' experience, has to do with the observed data in the age variable, In another form, found that most people under the age of 27 in the Age column are the ones who lose their data in the "Years of Experience" column, which indicates that the missing data is

missing at random, i.e. it cannot be deleted because it has a bias towards the default variable, but it must be compensated for the value of its alternative.

3) **IMPUTE OF MISSING VALUES**

Multiple imputation technique is becoming one of the most important methods to deal with missing values as it has the potential to correct the bias in the complete case and alternative analyses [8], and unique in the sense that it presents a mechanism for dealing with the attendant uncertainty of the imputations themselves [9]. Subsequently, we use it for handling missing data in the dataset Higher education institutions should provide methods to support and develop the educational process.

A multiple imputation method consists of three steps: the imputation step, the analysis step and the pooling, Figure 7 shows Scheme of main steps in multiple imputation, the imputation phase produces work, m = 3), each of which include various estimates of the missing values. the goal of the analysis phase is to analyze the filled-in data sets, the analysis phase yields m sets of parameter estimates and standard errors, so the purpose of the pooling phase is to combine everything into a single set of results [17] [21]. Equation 1 indicates the method of multiple imputations algorithm [22] [23] [24].

$$Y = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 X_3} \quad (1)$$

The pooling step is given by:

$$\overline{\boldsymbol{\beta}}_{\boldsymbol{MI}} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \widehat{\boldsymbol{\beta}_{\boldsymbol{i}}} \quad (2)$$

Where $\widehat{\boldsymbol{\beta}_{\boldsymbol{i}}}$ is the estimate of interest from the completed dataset number , $\overline{\boldsymbol{\beta}}_{\boldsymbol{MI}}$ is the estimate obtained from multiple imputation, and $\boldsymbol{m}$ is the number of imputed datasets[25].

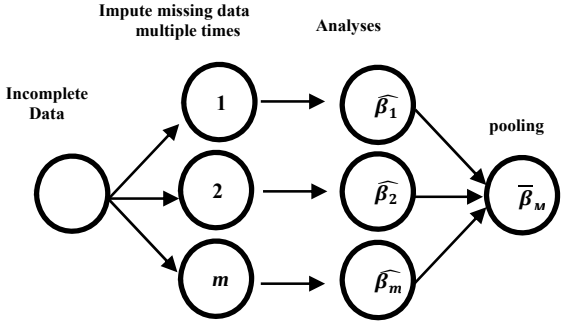

Figure 5 . Graphical depiction of main steps in multiple imputation

To ensure the accuracy of the imputation process, we have performed mean testing to compare the three iterations with each other and also compare them with the average before processing the lost data, and we obtained effective results. The averages were very close to each other before the filling process and after, Table 3 shows the results of the imputation process.

Table.3. Test means for comparison before and after the imputation process.

Years_Experience

Imputation Number	Mean	N	Std. Deviation
Original data	**13.46**	**29635**	**9.877**
1	**13.26**	**30239**	**9.921**
2	**13.26**	**30240**	**9.924**
3	**13.26**	**30240**	**9.917**
Pooled	**13.26**	**30239.7**	

3.2.1.2. DISCRETIZATION

Discretization, as one of the essential data reduction methods. Its main goal is to trans-form a set of continuous variables to discrete variables by dividing its scope into a finite set of disjoint intervals and then relates all intervals with denotation labels [26]. thereafter, we obtain a simple level of knowledge representation, we depend on descriptive statistics (minimum, maximum, quartiles) to discretize the continues variables (engine HP, and years' experience) in datasets, table 4 and 5 show datasets, table 4 and 5 show the results of discretization *process.*

Table.4. Descriptive statistics.

		Engine HP	*Years Experience*
N	*Valid*	***30240***	***30240***
	Missing	***0***	***0***
Range		***925***	***40***
Minimum		***80***	***0***
Maximum		***1005***	***40***
Percentiles	*25*	***111.00***	***5.00***
	50	***141.00***	***10.00***
	75	***238.00***	***20.00***

Table 5. Engine power HP variable after discretization

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	from 80 to 111	**7668**	**25.4**	**25.4**	**25.4**
	from 111.1 to 141	**7612**	**25.2**	**25.2**	**50.5**
	from 141.1 to 238	**7419**	**24.5**	**24.5**	**75.1**
	from 238.1 to 1005	**7541**	**24.9**	**24.9**	**100.0**
	Total	**30240**	**100.0**	**100.0**	

Table 6. Years' Experience variable after discretization

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	from 0 to 5	**8421**	**27.8**	**27.8**	**27.8**
	from5.1 to 10	**6706**	**22.2**	**22.2**	**50.0**
	freom10.1 to 20	**8419**	**27.8**	**27.8**	**77.9**
	from 20 to 40	**6694**	**22.1**	**22.1**	**100.0**
	Total	**30240**	**100.0**	**100.0**	

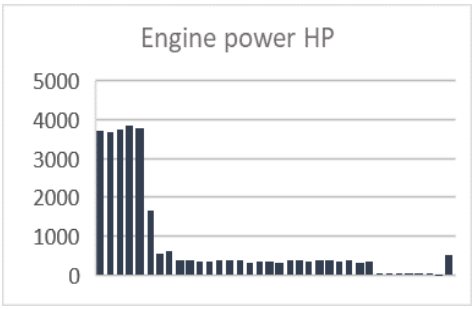

Figure 6. Engine HP before discretization

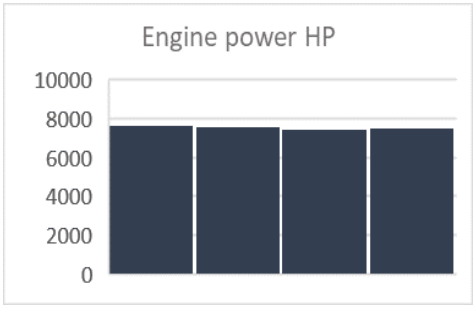

Figure 7. Engine HP after discretization

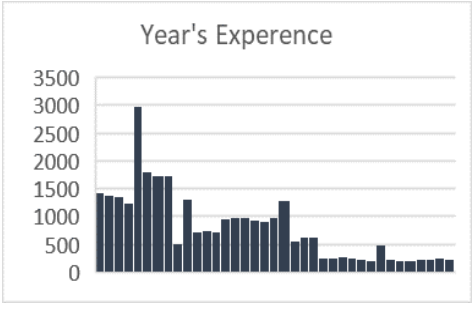

Figure 8. Year's Experence before discretization.

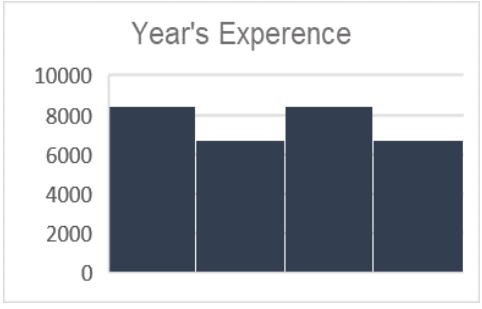

Figure 9. Year's Experence after discretization.

3.2.1.3. VARIABLES ENCODING

there some algorithms that do not work with categorical data, such as gender, marital status, Vehicle type ...etc., so must be converted to numerical variables for representing categorical variables.

3.2.1.4. STANDARDIZATION

because of the different qualities of the features, standard scaler processing generally plays a crucial role in transforming raw data into a dimensionless index, we applied on all attributes in the data set.to make each index value is at the same scale level. to standardize training set, Z-standard was employed which is a common method in statistics. Equation 2 indicates standard scaler Z-standard.

$$Z_i = \frac{x_i - \overline{x}}{s} \quad (3)$$

where x_i is the input variable ($X_1 X_2 \ldots . X_n$), (x) is the average of input variables, S is the sample standard deviation.

3.2.2. MODELING

car insurance data set is divided into two parts, 70 percent of which is the training set and 30 percent of which is testing set. The training data is used to model a fitted and logical model. As for testing data, it is utilized to calculate the accuracy of the prediction model. In this paper, four widely-used classification models are used, such as Decision Tree, Neural Network, Naïve Bayes, and XGBoost.

3.3. SCOPE AND LIMITATION

a) The paper focused on predicting the likelihood that the driver will file an insurance claim and did not focus on predicting the insurance premium or the level of risk, and accordingly the problem is classified as supervisory learning**.**

b) The research dealt with processing missing quantitative data and was not exposed to the qualitative data.

c) The missing data pattern was exclusive on Monotone Pattern just

4. MODELS EVALUATION AND EXPERIMENTAL RESULTS

In this work, we used confusion matrix and accuracy to evaluate the performance of the Classifier as shown in Table 7.

The confusion matrix known as the contingency table is a specific table layout that displays the performance of a model. For binary classification, it contains two rows and two columns that report the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). Figure 8, illustration the confusion matrix shape.

	Model Prediction Positive	Model Prediction Negative
Truth: Positive	TP	FN
Truth: Negative	FP	TN

Figure. 20.Confusion matrix

	Model Prediction Positive	Model Prediction Negative
Truth: Positive	**4043**	**142**
Truth: Negative	**535**	**4352**

Figure. 11. Confusion matrix of XGBOOST

	Model Prediction Positive	Model Prediction Negative
Truth: Positive	**4048**	**58**
Truth: Negative	**647**	**4319**

Figure. 12. Confusion matrix of decision tree

	Model Prediction Positive	Model Prediction Negative
Truth: Positive	***3922***	***184***
Truth: Negative	***548***	***4418***

Figure. 13. Confusion matrix of neural network

	Model Prediction Positive	Model Prediction Negative
Truth: Positive	**3524**	**582**
Truth: Negative	**871**	**4095**

Figure. 14. Confusion matrix of naïve base

Table 7. Experimental results of classification models.

Techniques	Accuracy	ROC
J48	**92.22 %**	**0.981**
NN	***91.93***	***0.841***
XGboost	**92.53**	**0.986**
Naïve base	**83.84 %**	**0.680**

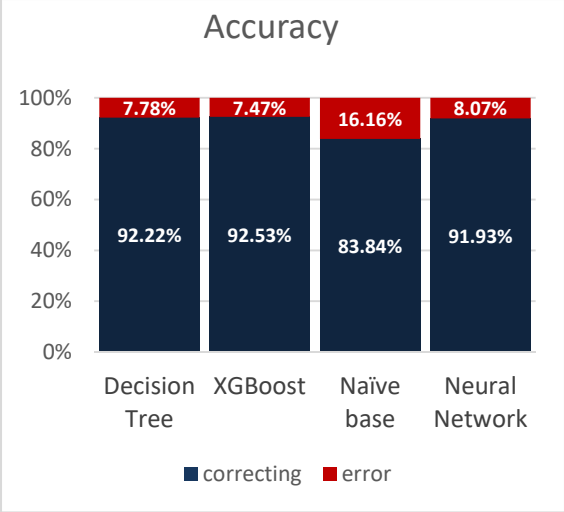

Figure.11. Comparison Of Prediction Accuracy Among Four Algorithms.

Table 7 shows that the XGBoost technique outperforms other methods in testing datasets. From Table 7, we also see that the decision tree (j48) gives better accuracies than neural networks and Naïve base. The results in Table 3 also show that multiple imputations give precision estimations for missing values, we find that the mean before is close to the mean after the imputing process, and not bias.

5. CONCLUSION

Claim prediction is One of the main challenges in the field of the insurance industry. Through it, car insurance companies can prepare the right type of insurance policy for each policyholder. Numerous insurance companies use traditional methods to analyze client details, moreover, the volume of the historical claim data is usually big data. also, there are numerous missing data for many attributes of the data. Therefore, we applied advanced statistical methods and machine learning algorithms that can handle these problems. This work constructed a model to predict insurance claims, four classifiers were built to predict the claims. The algorithms XGBoost, J48, ANN, and Naïve Bayes are selected for classification, The XGBoost, j48 model performed the best among the four models.

REFERENCES

[1] Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).

[2] Hanafizadeh, P., & Paydar, N. R. (2013). A data mining model for risk assessment and customer segmentation in the insurance industry. International Journal of Strategic Decision Sciences (IJSDS), 4(1), 52-78.

[3] Kreidler, M. (2008). Guide to auto insurance. Washington State Office of the Insurance Commissioner. Retrieved from www.insurance.wa.gov

[4] Huangfu, D. (2015). Data Mining for Car Insurance Claims Prediction.

[5] General Authority for Supervision and Control,2018, Statistical yearbook, Egypt

[6] Eekhout I, de Boer R, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology 2012; 23:729e32.

[7] Eekhout, Iris, et al. "Missing data in a multi-item instrument were best handled by multiple imputation at the item score level." *Journal of clinical epidemiology* 67.3 (2014): 335-342.

[8] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

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

[9] Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

[10] Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science*, *33*(2), 142-159.

[11] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70.

[12] Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.

[13] Stucki, Oskar. "Predicting the customer churn with machine learning methods: case: private insurance customer data." (2019).

[14] Fauzan, M. A., & Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction. Int. J. Adv. Soft Comput. Appl, 10(2).

[15] Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.

[16] Millican, M., Zhang, L., & Kimball, D. (2017). CS 229 Final Report: Predicting Insurance Claims in Brazil.

[17] Enders, C. K. (2010). Applied missing data analysis. Guilford press.

[18] Pijl, T., van de Velden, M., & Groenen, P. (2017). A Framework to Forecast Insurance.

[19] https://www.kaggle.com/datasets

[20] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

[21] White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine, 30(4), 377-399.

[22] Kasza, J., & Wolfe, R. (2014). Interpretation of commonly used statistical regression

[23] Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research, 20(1), 40-49.

[24] Yuan, Y. C. (2000, April). Multiple imputation for missing data: Concepts and new development. In Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference (Vol. 267).

[25] Lee, K. J., & Simpson, J. A. (2014). Introduction to multiple imputation for dealing with missing data. Respirology, 19(2), 162-167.

[26] Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., & Herrera, F. (2012). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4), 734-750.