Integrating actuarial models with neural networks for building a fraud detection model for automobile insurance

[1]Rohan Yashraj Gupta, [2]Satya Sai Mudigonda, [3]Pallav Kumar Baruah

[1]Doctoral Research Scholar, [2]Hon.professor, [3]Associate professor,
[1]Department of Mathematics and Computer Science,
[1]Sri Sathya Sai Institute of Higher Learning, Puttaparthi, India.

Abstract : Fraud poses a major risk in any organizations due to its financial implications. To mitigate this, most companies have fraud detection models in place. This work, aim to study the use of Combined Actuarial Neural Networks in building a fraud detection model. In this, the classical generalized linear model (GLM) is embedded in a neural network model. This approach captures the data patterns which was not captured by using a classical generalized model. The dataset used for this study is publically available dataset "carclaims.txt'. It was observed that the combination of a classical generalized linear model with neural network models helped in increasing the sensitivity of the model. The highest sensitivity of 0.88 was observed when GLM was combined with a neural network containing embedding layers. Also, it was observed that the performance of all the models significantly improved once the oversampling of the dataset was done.

Code along with the data used is available in the following link: https://github.com/RohanYashraj/CANN-for-Fraud-Detection

IndexTerms **- Automobile insurance, Data imbalance, Fraud detection, Machine learning, Neural networks, Actuarial models, Classification models.**

I. INTRODUCTION

Fraud detection is an important problem and is taking priority in an insurance organization. According to the FBI, the total estimated fraud in non-health insurance is about $40 Billion every year. The Coalition Against Insurance Fraud (CAIF) report states that about $6 Billion is lost every year towards workers' compensation insurance fraud. According to the study done by the Insurance Research Council (IRC), 15% to 17% of the total claims paid for automobile insurance (third-party) is fraudulent. The study also estimated that around $7.7 Billion was added to auto insurance claims payment in the year 2012 compared to $5.8 Billion in 2002 [1]. These figures indicate that this is a major problem for any insurance business. They act as a major deterrent towards growth. This is not a problem which is overlooked by any insurer. However, despite whatever measure they take to tackle fraud, fraudsters find a way.

There are various antifraud measures that insurers' have. Most of them have SIUs (Special Investigation Units) which helps them to identify and investigate suspicious claims. According, to CAIF about 80% of the insurers had SUIs by 2001. They comprise a small team that is trained to look for a routine type of fraud cases. This is a time taking process and a more complex fraud would go undetected in the majority of the cases.

However, in recent times with the advent of the latest data science tools and techniques, insurers are finding ways to incorporate them into their existing processes. The traditional approach of detecting fraud using business rules are being augmented by data science tools and techniques.

One major problem that is faced in building a fraud detection model for insurance is the data imbalance [2]. Lesser fraudulent data inhibits the model from getting trained better therefore the predictions are skewed towards the non-fraudulent claims. The model becomes very bad at predicting fraudulent claims, which is undesirable. This work uses tabular generative adversarial network-based oversampling with machine learning models. A comparative study with various other machine learning models is also performed.

The paper is organized as follows. Section 2 presents the literature review in this area. Section 3 describes the data and various descriptive statistics about the data being used. Section 4 contains the description of the methodology used in this work. Section 5 contains the result of the study conducted in this work. Finally, section 6 shows the conclusion and future work followed by acknowledgement and references.

II. RELATED WORK

This section contains various works that are done in this area. Subelj et. al. suggested a fraud detection model that uses concepts of social network analysis [3]. This approach allows the relationship between claims to be taken into account which is lost when using machine learning techniques. In most analytics the textual information is often missed out, Wang et. al. have leveraged deep learning application in text analytics to detect automobile fraud [4]. There are various other deep learning approaches found in the literature. The bayesian learning-based neural network proposed by Viaene et. al. is a variation that is performed on closed claims records [5]. There are also works done which model various types of automobile insurance fraud [6]. Rohan et al. have proposed a framework of insurance fraud detection which is a consolidation of various data science and actuarial techniques that are used for building a fraud detection model for any line of business [7]. Nian et. al. have used spectral ranking for detecting an anomaly in the auto insurance claims to detect any fraudulent behaviour in the claims patterns [8]. Tao et. al. have used fuzzy SVMs which uses dual membership. This helps in associating probability to each of the claims being fraudulent or not [9]. Srinivas et al. have used predictive analytics to detect fraudulent claims. In this, they have modelled the time to failure of a specific component of automobile to determine status of the claim [10]. Some works use missing information to identify fraudulent behaviours. Caudil et. al. have used a multinomial logit model to analyse missing information. The aim was to show how misclassified claims can be classified correctly [11].

One major challenge that is dominant in automobile fraud detection is data imbalance [12]. In general, there is less than 10% of the total claims records that are fraudulent. Therefore, handling this is a essential step to improve the performance of the model [2].

There are various innovative methods to handle this issue. Broadly these are classified into undersampling and oversampling techniques [13], [14]. The most widely used oversampling techniques are the SMOTE proposed by Chawla et. al. [15]. There are various variations to this like Borderline-SMOTE [16], GAN-SMOTE [17]. Some other popular oversampling methods are the ADASYN for imbalanced learning [18], MWMOTE [19]. Nikhil et. al. have used supervised learning methods like SVM, Decision Trees and Random Forests wherein they have handled the data using MWMOTE [20].

This work combines classical actuarial models in neural networks to improve the efficacy of actuarial models. These models are also known as Combined Actuarial Neural Networks (CANN). Mario et. al. have proposed this model and implemented it for building effective pricing models for insurance business [21]–[23]. This model is studied in this work and its effectiveness is explored in the field of automobile fraud detection.

III. Data and Descriptive Statistics

The dataset used for this work is the publically available "carclaims" dataset. This dataset is used for fraud detection modelling purpose. The data points here represent the automobile claim record containing various information about the driver, car, policy type, policy information, etc. There are a total of 32 features in this dataset with 25 of these features being categorical, 6 features being nominal and 1 class variable. There are a total of 15420 claim records with only 6% of them being fraudulent records. This makes the classification process bit challenging as there is too little fraudulent data for the model to learn.

Data Pre-Processing

The dataset had to be pre-processed before it could be used for the study. Feature with various categorical variables was clubbed together such that variables representing similar information were replaced with a single variable. This was done so that in the final dataset the number of variables was lesser compared to the original dataset. This facilitates the learning of the machine learning model.

Table 1 - Aggregating the categorical variables

Feature	Variables	After pre-processing
Make	Accura	Make1
	BMW	
	Ferrari	
	Jaguar	
	Lexus	
	Mercedes	
	Nisson	
	Toyota	
	Mazda	
	Chevrolet	Make2
	Dodge	
	Pontiac	Make3
	Saturn	
	Porche	Make4
	Saab	
Vehicle Price	20000 to 29000	20000 to 39000
	30000 to 39000	
	40000 to 59000	40000 to 69000
	60000 to 69000	
Days: Policy Accident	1 to 7	1 to 15
	8 to 15	
	none	more than 30
Age Of Vehicle	new	less than 4 years
	2 years	
	3 years	
	4 years	
	5 years	4 to 6 years
	6 years	
	7 years	
Address Change Claim	under 6 months	0 to 3 years
	1 year	
	2 to 3 years	
Number Of Cars	2 vehicles	more than 1
	3 to 4	
	5 to 8	
	more than 8	

Feature Engineering

Apart from this, all the month features in the dataset were encoded serially from 1 to 12. This was done to engineer a new feature in the dataset "days_diff" which is known as reporting delay in business terms. This allowed us to remove all the month and days related to the dataset (7 features). Finally, three other features 'PolicyNumber', 'PolicyType' and 'RepNumber' were removed this resulted in the final dataset having 23 features (32+1-7-3 = 23). This dataset was used for all the study which is presented in this work. All the dataset, before and after the pre-processing is available in the GitHub repository (link provided in the abstract).

Apart from this, there were other model specific data pre-processing performed on the dataset.

IV. METHODOLOGY

This section describes the methodology adopted in this study and also details various models used in this work.

The entire study is conducted in three different setups and for each of the setups, six models are studied. In total 18 models are studied and presented in this work. The three setups are differentiated in the way the data imbalance was handled. In the very first setup baseline (raw data) was used. This means that the data used was highly imbalanced in nature. In the second setup, undersampling of the dataset was done. Undersampling reduces the majority class such that the number of majority and minority class records are almost equal in number. In the final setup, data oversampling was done. The oversampling was done using SMOTE. This was done to increase the number of fraudulent claims in the dataset.

Each of the models used in this work is described below:

Generalized Linear Model (GLM)

Generalized linear models are an extension of linear models [24]. In this study, GLM models are used to classify a claim as fraud or not. This is achieved by considering binomial distribution. The associated link function for binomial distribution is Logit function denoted by:

$$g(u) = log(u/(1-u))$$

This results in the output of the model are in the range of [0,1]. To determine a threshold for classifying a record as fraud or not, optimization functions are used which outputs a threshold that gives the highest accuracy in the considered model.

An additional use of GLM models is that interaction between variables can also be model. In the dataset considered in this study, various interactions were considered [25]. The variable interaction was arrived at by consulting industrial experts.

- Age of the individual & Make of the car
- Driver Rating & Marital Status
- Vehicle Price & Make
- Fault & Age

Apart from these interaction variables, there were other changes in the variable of age made as shown by Mario et. al. [22]. This is listed below

- Age + log(Age) + I(Age ^2) + I(Age ^3) + I(Age ^4)

Simple Neural Net (SimpleNN)

A simple neural net as described in this work has three dense layers with 20 nodes in the first dense layer, 15 nodes in the second, 10 nodes in the third and finally an output layer with one node. For each of the dense layer, the tanh activation function was used and in the output layer sigmoid activation function to ensure that the final output was in the range of [0,1]. Figure 1 shows the diagrammatic representation of the SimpleNN used in this work.

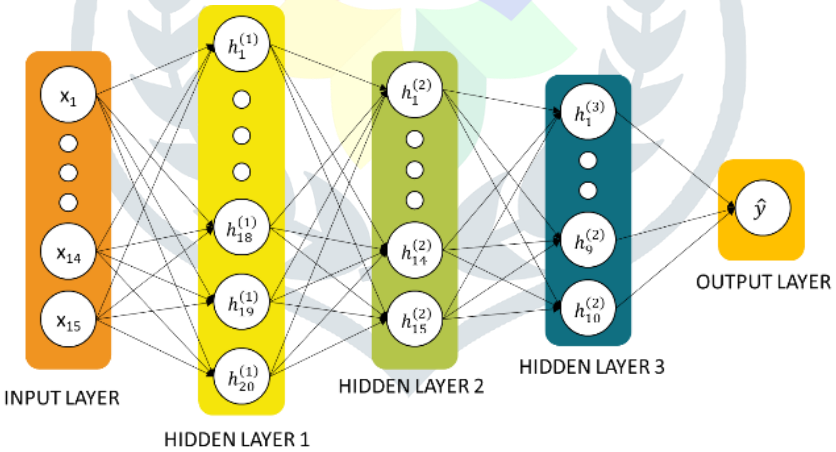

Figure 1 - SimpleNN architecture

The neural nets are trained with 400 epochs and a batch size of 500. Apart from the SimpleNN, there are two other variations used in this study. They are called “Neural Net plus” & “Neural Net with embedding layer” in this work.

Neural Net plus (NNplus)

Neural net plus is a combination of three neural networks. Each of these neural networks has the same architecture as that of SimpleNN, with the difference being in the input layer. The original data is divided into three different parts, which is then fed into each of these networks respectively. The dataset is divided such that in the first set there are all the vehicle-related features. In the second set, there are all claimant related features and finally, in the third set, all other remaining features are placed.

As mentioned each of the networks has the same architecture as that of SimpleNN. The output of these networks is connected to the output layer which has a sigmoid activation function. The visual representation of the NNplus architecture is shown in Figure 2.

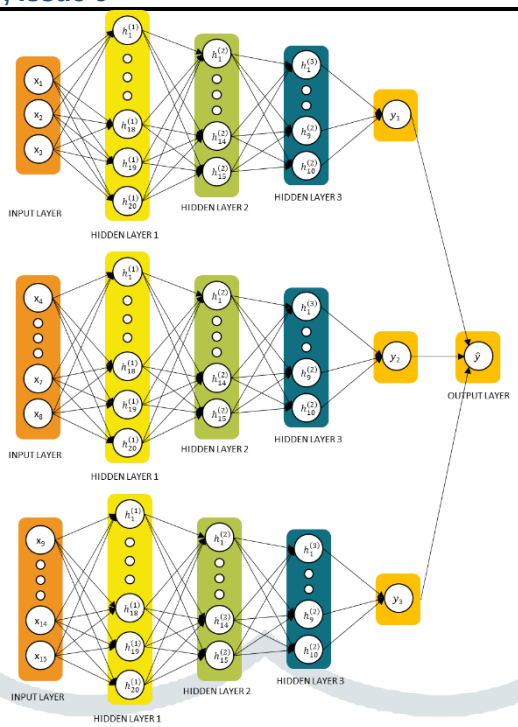

Figure 2 - NNplus architecture

Neural Net with embedding layer (NNembedd)

Another variation of neural networks testes in this study is called neural net with embedding layer (NNembedd). In this, there are fundamentally two different neural networks whose outputs are combined to get the final prediction. However, there is a small addition to this. In the first neural network, there are two embedding layers of dimension two for the features "VehiclePrice" and "Make". This is done because these features are categorical in nature and the embedding performed on this is better compared to a one-hot encoding [22]. The architecture is better understood in Figure 3.

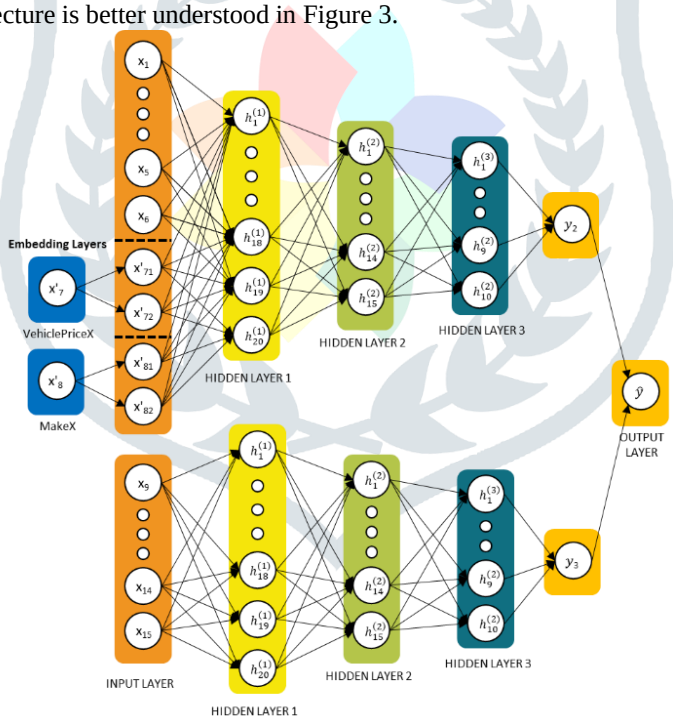

Figure 3 - NNembedd architecture

Combining GLM and Neural Nets (CANN)

The GLM model and various neural nets are combined using a concept of skip connection as proposed by Mario et. al. [22]. This is referred to as CANN (Combined Actuarial Neural Network). The output of the GLM model and neural network models are provided as nodes in the penultimate layer of the neural network which is then finally connected to the output layer. This modification in the architecture captures the data patterns which was not captured by using a classical generalized model. Figure 4 shows the visual representation of the CANN architecture.

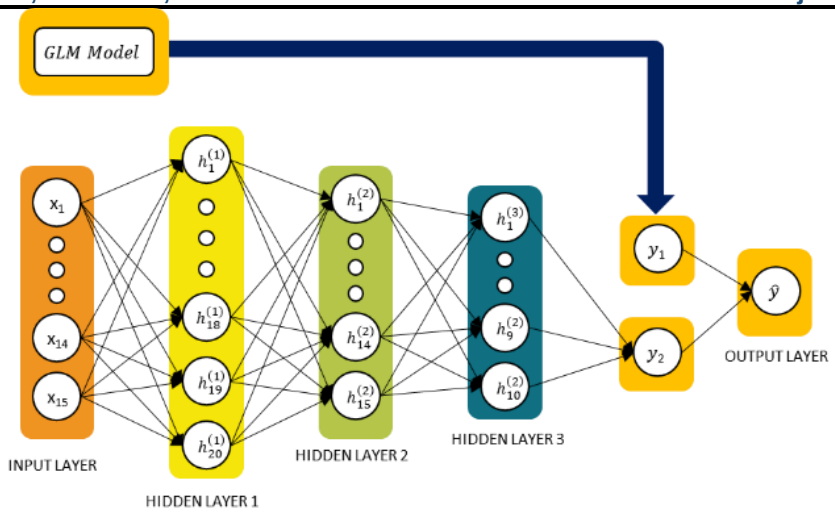

Figure 4 - CANN architecture

V. RESULTS

This section contains the results of various studies carried out in this work. As mentioned in the "methodology" section, three studies are shown in this work.

In the beginning, the dataset used was baseline data. The percentage of fraud cases in this was 6% of the total number of claims. Baseline data were used and classification models were built and tested. Table 2 shows the results of this study. It can be observed that nesting GLM model with various neural network models has improved the performance with respect to various metrics. "GLM+NNplus" has the highest sensitivity. However, the precision of all the models is very low. This indicates that there are too many false positives in the prediction. This is undesirable because investigating costs are very high.

Table 2 - Results using baseline data

Model	Sensitivity	Specificity	Precision	Accuracy	F1 Score	AUC-ROC
GLM	0.7396	0.7348	0.1392	0.7351	0.2343	0.8065
SimpleNN	0.7337	0.7149	0.1298	0.7160	0.2206	0.7858
GLM + SimpleNN	0.7692	0.6714	0.1195	0.6767	0.2068	0.7671
NNplus	0.7456	0.7328	0.1392	0.7335	0.2346	0.7980
GLM + NNplus	0.8107	0.7033	0.1367	0.7091	0.2340	0.8030
GLM + NNembedd	0.7751	0.6309	0.1085	0.6388	0.1904	0.7254

Figure 5 shows the bar chart of performance metrics for various models. It can also be observed that the F1-score of all the models is very low. The major reasons for such poor performance of the model is the high imbalance nature of the dataset.

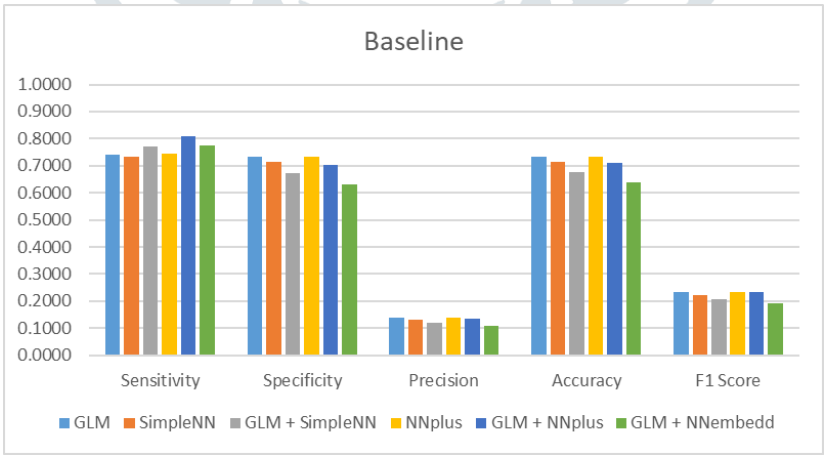

Figure 5 - Comparison of various models using baseline data

In the next two studies, this issue is handled using undersampling and oversampling.

In the second study, the data imbalance was handled using the undersampling of the majority samples. This resulted in randomly removing the majority samples such that the number of fraud and non-fraud claim records in the "undersampled data" is approximately equal. Under this scenario, since there was equal number of both fraud and non-fraud claims the model training was not skewed towards any of the two. Thus, the precision of all the models has increased, indicating that there were lesser false positives in the prediction. This also increased the F1-score. Overall, almost all the performance metrics have increased compared to the baseline data except specificity, accuracy and AUC-ROC value. This is because these metrics are dependent on the majority classes. In the baseline dataset, though the minority cases were less the majority cases were very high. Therefore, the model was trained much better for the majority class hence higher specificity and accuracy.

Table 3 - Results using undersampled data

Model	Sensitivity	Specificity	Precision	Accuracy	F1 Score	AUC-ROC
GLM	0.7449	0.7011	0.5771	0.7166	0.6503	0.7998
SimpleNN	0.8878	0.5950	0.5455	0.6986	0.6757	0.7675
GLM + SimpleNN	0.8010	0.6508	0.5567	0.7040	0.6569	0.7232
NNplus	0.8112	0.6788	0.5803	0.7256	0.6766	0.7945
GLM + NNplus	0.8214	0.6508	0.5629	0.7112	0.6680	0.7808
GLM + NNembedd	0.8571	0.5978	0.5385	0.6895	0.6614	0.7553

However, in the undersampled data, too much of the information is lost about the majority claims. This resulted in the reduced value of specificity, accuracy and AUC-ROC value. Table 3 and Figure 6 shows that sensitivity is highest when a simple neural network is used whereas with respect to other metrics GLM is better.

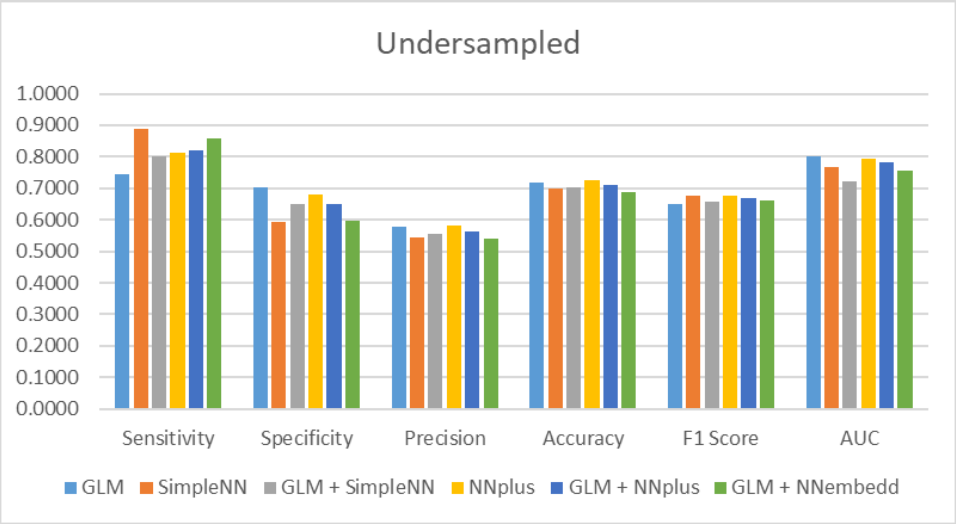

Figure 6 - Comparison of various models using undersampled data

However, the overall performance is not good for any of these. This leads us to the final study where oversampling of the dataset was done. The oversampling technique in this works is SMOTE [15]. Data oversampling helped us to retain the majority samples also the addition of synthetic samples into the dataset increased the number of minority samples in the dataset. The oversampling was done such that the final dataset used had an approximately equal number of fraud and non-fraud cases. Table 4 and Figure 7 shows the result of the model using oversampled data.

Table 4 - Results using oversampled data

Model	Sensitivity	Specificity	Precision	Accuracy	F1 Score	AUC-ROC
GLM	0.7945	0.7646	0.7677	0.7794	0.7809	0.8526
SimpleNN	0.8466	0.8808	0.8743	0.8638	0.8602	0.9420
GLM + SimpleNN	0.8500	0.8693	0.8643	0.8598	0.8571	0.9403
NNplus	0.8686	0.9122	0.9064	0.8906	0.8871	0.9532
GLM + NNplus	0.8776	0.8923	0.8886	0.8850	0.8831	0.9507
GLM + NNembedd	0.8821	0.8794	0.8775	0.8807	0.8798	0.9503

The models are better in this scenario. “GLM+NNembedd” is the best in terms of sensitivity with 0.8821 of the model with specificity, precision and accuracy at 0.8794, 0.8775 and 0.8807 respectively. “NNplus” is also a good performing model and is better than “GLM+NNembedd” in all the metrics except sensitivity. However, the management would not be entirely comfortable with this model because explaining a neural network model is difficult. On the other “GLM+NNembedd” is explainable to an extent because GLM is a well-studied model and the working of a GLM is more understood than a purely neural network model.

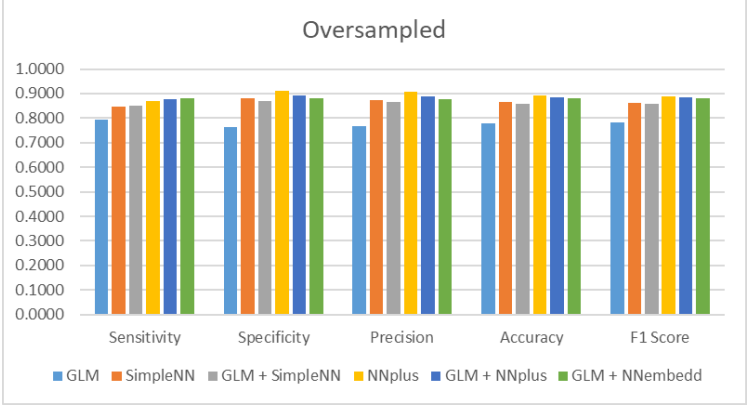

Figure 7 - Comparison of various models using oversampled data

VI. Conclusions And Future Work

In this work, various fraud detection model in automobile insurance was studied by combining actuarial models with neural networks. The work showed that the combination of GLM and neural networks models can be used to improve the GLM model performance. It was also demonstrated that handling data imbalance helps in significantly improving the performance of the models.

In the future, the work can be extended to other lines of insurance business like health. A similar approach can also be tested on other financial institutions like the banking sector.

Acknowledgement

The authors express their gratitude to Bhagawan Sri Sathya Sai Baba, founder chancellor, SSSIHL.

Reference

[1] Corum, "Insurance Research Council Finds That Fraud and Buildup Add Up to $7.7 Billion in Excess Payments for Auto Injury Claims," *Insur. Res. Counc.*, p. 3, 2015.

[2] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study1," *Intell. Data Anal.*, vol. 6, no. 5, pp. 429–449, Nov. 2002, doi: 10.3233/IDA-2002-6504.

[3] L. Šubelj, Š. Furlan, and M. Bajec, "An expert system for detecting automobile insurance fraud using social network analysis," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 1039–1052, Apr. 2011, doi: 10.1016/j.eswa.2010.07.143.

[4] Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decis. Support Syst.*, vol. 105, pp. 87–95, Jan. 2018, doi: 10.1016/j.dss.2017.11.001.

[5] S. VIAENE, G. DEDENE, and R. DERRIG, "Auto claim fraud detection using Bayesian learning neural networks," *Expert Syst. Appl.*, vol. 29, no. 3, pp. 653–666, Oct. 2005, doi: 10.1016/j.eswa.2005.04.030.

[6] M. Artís, M. Ayuso, and M. Guillén, "Modelling different types of automobile insurance fraud behaviour in the Spanish market," *Insur. Math. Econ.*, vol. 24, no. 1–2, pp. 67–81, Mar. 1999, doi: 10.1016/S0167-6687(98)00038-9.

[7] R. Y. Gupta, S. S. Mudigonda, P. K. Kandala, and P. K. Baruah, "A Framework for Comprehensive Fraud Management using Actuarial Techniques," *Int. J. Sci. Eng. Res.*, vol. 10, no. 3, pp. 780–791, 2019.

[8] K. Nian, H. Zhang, A. Tayal, T. Coleman, and Y. Li, "Auto insurance fraud detection using unsupervised spectral ranking for anomaly," *J. Financ. Data Sci.*, vol. 2, no. 1, pp. 58–75, Mar. 2016, doi: 10.1016/j.jfds.2016.03.001.

[9] Han Tao, Liu Zhixin, and Song Xiaodong, "Insurance fraud identification research based on fuzzy support vector machine with dual membership," in *2012 International Conference on Information Management, Innovation Management and Industrial Engineering*, 2012, vol. 3, pp. 457–460, doi: 10.1109/ICIII.2012.6340016.

[10] R. Srinivasan, S. P. Devi, S. Manivannan, and N. Ethiraj, "Fradulent Claims Detection in Automobile Industry Using Predictive Data Analytics," *Int. J. Mech. Eng. Technol. (IJMET*, vol. 9, no. 11, pp. 1447–1452, 2018.

[11] S. B. Caudill, M. Ayuso, and M. Guillen, "Fraud Detection Using a Multinoamial Logit Model With Missing Information," *J. Risk Insur.*, vol. 72, no. 4, pp. 539–550, Dec. 2005, doi: 10.1111/j.1539-6975.2005.00137.x.

[12] S. Subudhi and S. Panigrahi, "Effect of Class Imbalanceness in Detecting Automobile Insurance Fraud," in *2018 2nd International Conference on Data Science and Business Analytics (ICDSBA)*, 2018, pp. 528–531, doi: 10.1109/ICDSBA.2018.00104.

[13] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory under-sampling for class-imbalance learning," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2006, doi: 10.1109/ICDM.2006.68.

[14] A. Ali, S. M. Shamsuddin, and A. L. Ralescu, "Classification with class imbalance problem: A review," *Int. J. Adv. Soft Comput. its Appl.*, 2015.

[15] N. V Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[16] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Lecture Notes in Computer Science*, vol. 3644, no. PART I, Springer Verlag, 2005, pp. 878–887.

[17] M. Scott and J. Plested, "GAN-SMOTE: A Generative Adversarial Network approach to Synthetic Minority Oversampling for One-Hot Encoded Data."

[18] Haibo He, Yang Bai, E. A. Garcia, and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, pp. 1322–1328, doi: 10.1109/IJCNN.2008.4633969.

[19] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, Feb. 2014, doi: 10.1109/TKDE.2012.232.

[20] N. Rai, P. K. Baruah, S. S. Mudigonda, and P. K. Kandala, "Fraud Detection Supervised Machine Learning Models for an Automobile Insurance," *Int. J. Sci. Eng. Res.*, vol. 9, no. 11, pp. 473–479, 2018.

[21] M. V. Wüthrich and M. Merz, "EDITORIAL: YES, WE CANN!," *ASTIN Bull.*, vol. 49, no. 1, pp. 1–3, Jan. 2019, doi: 10.1017/asb.2018.42.

[22] J. Schelldorfer and M. V. Wuthrich, "Nesting Classical Actuarial Models into Neural Networks," *SSRN Electron. J.*, pp. 1–27, 2019, doi: 10.2139/ssrn.3320525.

[23] A. Ferrario, A. Noll, and M. V. Wuthrich, "Insights from Inside Neural Networks," *SSRN Electron. J.*, pp. 1–64, 2018, doi: 10.2139/ssrn.3226852.

[24] G. Hutcheson, "Generalized Linear Models," in *The SAGE Dictionary of Quantitative Management Research*, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2001, pp. 133–134.

[25] H. Chen and J. K. Lindsey, "Applying Generalized Linear Models," *Technometrics*, vol. 40, no. 2, p. 156, May 1998, doi: 10.2307/1270654.